

31th International Conference on Information Modelling and Knowledge Bases

EJC 2021

Marina Tropmann-Frick
Bernhard Thalheim
Hannu Jaakkola
Yasushi Kiyoki

September 7-9, 2021, Hamburg Germany

Impressum

Editors

Bernhard Thalheim
Information Systems Engineering
Dept. of Computer Science
Christian-Albrechts-University Kiel
24118 Kiel, Germany

Marina Tropmann-Frick
Fakultät Technik und Informatik
Department Informatik
University of Applied Sciences Hamburg
Berliner Tor 5 20099 Hamburg, Germany

Hannu Jaakkola
Information Technology and Communication
Dept. of Computer Science
Tampere University, Pori
P.O.Box 300, FIN-28101 Pori, Finland

Yasushi Kiyoki
Graduate School of Media and Governance
Keio University
P.O.Box 300, 5322 Endoh
Fujisawa, Kanagawa Japan, 252-0882

Kiel Computer Science Series

Kiel Computer Science Series (KCSS)
ISSN 2193-6781 (print version)
10.21941/kcss/2021/6

2021/6 dated 2021-09-03
ISSN 2194-6639 (electronic version)

Electronic version, updates, errata available via <https://www.informatik.uni-kiel.de/kcss>

The editor can be contacted via <http://www.mip.informatik.uni-kiel.de>

Published by the Department of Computer Science, Kiel University
Electronic version and errata available via <https://www.informatik.uni-kiel.de/kcss>

Please cite as:

M. Tropmann-Frick, B.Thalheim, H. Jaakkola, Y. Kiyoki. Proceedings of the International Conference on Information Modelling and Knowledge Bases (EJC 2021)

Number 2021/6 in Kiel Computer Science Series. Department of Computer Science, 2021.
Dissertation, Faculty of Engineering, Kiel University.

```
@book{EJC2021ConferenceProceedings,
```

```
  TITLE = {Proceedings of the International Conference on Information  
          Modelling and Knowledge Bases (EJC 2021)},
```

```
  YEAR = {2021}
```

```
  editor = {M. Tropmann-Frick, B.Thalheim, H. Jaakkola, Y. Kiyoki},
```

```
  publisher = {Kiel University},
```

```
  volume = {2021/6},
```

```
  series = {KCSS},
```

```
  organization = {Department of Computer Science, Faculty of Engineering}
```

```
  isbn = {ISSN 2194-6639},
```

```
  source = {https://www.informatik.uni-kiel.de/kcss}
```

```
  doi = {10.21941/kcss/2021/6}
```

```
}
```

© Marina Tropmann-Frick, Bernhard Thalheim, Hannu Jaakkola, Yasushi Kiyoki

Preface

Information Modeling and Knowledge Bases has become an important technology contributor for the 21st century's academic and industry research that addresses the complexities of modeling in digital transformation and digital innovation, reaching beyond the traditional borders of information systems and computer science academic research.

The amount and complexity of information itself, the number of abstraction levels of information, and the size of databases and knowledge bases are continuously growing. Conceptual modelling is one of the sub-areas of information modelling. The aim of this conference is to bring together experts from different areas of computer science and other disciplines, who have a common interest in understanding and solving problems on information modelling and knowledge bases, as well as applying the results of research to practice. We also aim to recognize and study new areas on modelling and knowledge bases to which more attention should be paid. Therefore, philosophy and logic, cognitive science, knowledge management, linguistics and management science as well as machine learning and AI are relevant areas, too.

In the conference, there will be three categories of presentations, i.e., full papers, short papers and position papers. The international conference on information modelling and knowledge bases originated from the cooperation between Japan and Finland in 1982 as the European Japanese conference (EJC). Then professor Ohsuga in Japan and Professors Hannu Kangassalo and Hannu Jaakkola from Finland (Nordic countries) did the pioneering work for this long tradition of academic collaboration. Over the years, the organization extended to include European countries as well as many other countries. In 2014, with this expanded geographical scope, the European Japanese part in the title was replaced by International. The conference characteristics include opening with a keynote session followed by presentation sessions with enough time for discussions. The limited number of participants is typical for this conference.

The 31st International conference on Information Modeling and Knowledge Bases (EJC 2021) held at Hamburg, Germany constitutes a research forum exchanging of scientific results and experiences drawing academics and practitioners dealing with information and knowledge. The main

topics of EJC 2021 cover a wide range of themes extending the knowledge discovery through Conceptual Modelling, Knowledge and Information Modelling and Discovery, Linguistic Modelling, CrossCultural Communication and Social Computing, Environmental Modeling and Engineering, and Multimedia Data Modelling and Systems extending into complex scientific problem solving. The themes of the conference presentation sessions; Learning and Linguistics, Systems and Processes, Data and Knowledge Representation, Models and Interfaces, Formalizations and reasoning, Models and Modelling, Machine Learning, Models and Programing, Environment and Predictions, Emotion Modeling and Social Networks reflected the coverage of those main themes of the conference.

The contributions of this proceeding of the 31st International Conference of Information Modeling and Knowledge Bases feature twenty-one reviewed, selected, and upgraded contributions that are the result of presentations, comments, and discussions during the conference. Suggested topics of the call for papers include, but are not limited to:

Conceptual modelling: Modelling and specification languages; Domain-specific conceptual modelling; Concepts, concept theories and ontologies; Conceptual modelling of large and heterogeneous systems; Conceptual modelling of spatial, temporal and biological data; Methods for developing, validating and communicating conceptual models.

Knowledge and information modelling and discovery: Knowledge discovery, knowledge representation and knowledge management; Advanced data mining and analysis methods; Conceptions of knowledge and information; Modelling information requirements; Intelligent information systems; Information recognition and information modelling.

Linguistic modelling: Models of HCI; Information delivery to users; Intelligent informal querying; Linguistic foundation of information and knowledge; Fuzzy linguistic models; Philosophical and linguistic foundations of conceptual models. Cross-cultural communication and social computing: Cross-cultural support systems; Integration, evolution and migration of systems; Collaborative societies; Multicultural web-based software systems; Intercultural collaboration and support systems; Social computing, behavioral modeling and prediction.

Environmental modelling and engineering: Environmental information systems (architecture); Spatial, temporal and observational information

systems; Large-scale environmental systems; Collaborative knowledge base systems; Agent concepts and conceptualization; Hazard prediction, prevention and steering systems.

Multimedia data modelling and systems: Modelling multimedia information and knowledge; Contentbased multimedia data management; Content-based multimedia retrieval; Privacy and context enhancing technologies; Semantics and pragmatics of multimedia data; Metadata for multimedia information systems.

The EJC 2021 will be held online and hosted by the Department Informatik of University of Applied Sciences Hamburg, Germany. Due to regulations caused by Corona virus this year conference is transformed into a virtual event, held on September 7-9th. Presentations, prepared according to the instructions, have 10 minutes time slot. Virtual conference is organized with Zoom platform. We thank all colleagues for their support in making this conference successful, especially the program committee, organization committee, and the program coordination team, especially Naofumi Yoshida who maintains the paper submission and reviewing systems and compiles the files for this book.

The Editors

Marina Tropmann-Frick

Bernhard Thalheim

Hannu Jaakkola

Yasushi Kiyoki

Model-Based Reasoning

Bernhard THALHEIM¹

Christian-Albrechts-University Kiel, Christian-Albrechts-Platz 4, 24118 Kiel, Germany

Abstract. Models are used everywhere, in daily life, sciences, engineering, and thoughts. They represent, support, and enable our thinking, acting, reflecting, communication, and understanding. They are universal instruments. Model-based reasoning is, however, different from those that we use in 'exact' sciences and is far less understood. The notion of model is becoming nowadays well-accepted. Model-based reasoning is far less understood and a long lacuna. This keynote aims at closing this gap.

Keywords. model-based reasoning, plausible reasoning, approximative reasoning, abduction, induction, explanation, hypotheses, empiric

1. Introduction

Humans intensively use models everywhere, at any time, for any reason, by everybody, for everybody, and at any sphere of human activity. Pupils learn natural sciences through models. They transform complex, abstract, or partial ideas, systems, and theories into more easily to understand and simpler to use things, i.e. humanise them in dependence on human abilities. They are already used to deploy models with their first thoughts. The very first intellectual instrument we use is a model. It is not surprising that babies quickly develop their own models or at least concepts of the 'mother' and 'father'. They cannot yet use a natural language but they know already models of their surroundings [12]. Later they realise that their models are completely different from those of their contemporaries.

We are using the word 'model' widely in our daily life as well as in sciences and engineering. Models are also widely used in the social sphere, in religion, in communication, in interaction, and collaboration. They must not be correct but should be useful as an instrument (*'model for'*). Models can be understood as a collection of competing interpretations, perception, prehension, ideas, comprehension, imaginations, or conceptions about the world a human observes and understands, each with a utility core, which nevertheless must prove to be progressive over time. This wide usage of models direct us to consider models as the fourth sphere of our life beside sensing and reflecting the *world of the being*, acting and mastering the *world we create*, and the *intelligible world* of science, knowledge, and concept(ion)s. Models can be considered as our 'third reflection

¹bernhard.thalheim@email.uni-kiel.de

eye' (*'model.of'*) we use for comprehension, acceptance, understanding, finding our way around, socialising, communication, planing, and actuation.

The theory and practice of models and modelling is already fairly rich (see, for instance, [33]) and resulted in a large body of knowledge for almost all disciplines of science and engineering. Let us start with five observations:

Models have not to be called 'model' since we use them anyway as some kind of explicit means or instruments. They are far older than the explicit naming of things as models. The oldest explicit model we definitely know as model is older than 4.000 years. Models often appear as conceptions, notions, visions, images, view(point)s, concepts, pictures, basic orientations, performances, comprehensions, representations etc.

Mental models are used by everybody in an explicit but often implicit form. They can be less rigid than notions we use in science and engineering but they are useful. As such we consider ideas, imaginations, perceptions, prospects as a mental picture, beliefs, conceivability, visions, and imaginations as models.

Models are also used in manufacturing for instance as a template, pattern, reference, presentation, prototype, origin for production, master copy, and sample.

Sciences and technology might claim that they are model-free but are typically 'downright contaminated' with models. Look, for instance, to claims made in Computer Science. There is neither a systematic study of models nor a modelling sub-discipline. However, almost anything is done by use of models.

Models can but don't have to be explicitly designed for usage. Engineering as the approach to create the artificial [27] is based on models as a starting point for construction, as a documentation of the construction, as a means for developing variations, as a thought and imagination instrument, and as an artifact within the creation process. There are, however, objects that became models at a far later stage of their existence.

These few observations allow us to expect that models are one of our main instruments similar to the broad usage of languages. Moreover, modelling is an activity that comes before we learn a language and accompanies us our entire life. So, we should ask ourselves first: What is a model? Next, we have to ask: what is model-based reasoning? The first question got already hundreds of answers. The third one almost none. Therefore, the paper aims at answering the third question after gaining an understanding of the first and second answer.

2. Modellkunde – Towards a Study of Models and Modelling

2.1. The Notion of Model

The notion we use since [29] generalises almost all notions or pre-notions used and known so far in general model theory [13,20,21,28,33]²:

²Its advantage is that all notion we have seen so far can be understood as a parametric specialisation. More specific notions can be declined by parameter refinement and hardening from this notion.

“A **model** is a well-formed, adequate, and dependable instrument that represents origins³ and that functions in utilisation scenarios.

Its criteria of well-formedness⁴, adequacy⁵, and dependability⁶ must be commonly accepted by its community of practice (CoP) within some context and correspond to the functions that a model fulfills in utilisation scenarios.” [30]

This notion also allows consideration of the *model-being of any instrument*⁷. Anything – any thought and any thing – can be a model as long as it is used as such. The model-being is, therefore, an assignment for an instrument that is used in scenarios.

2.2. Functions of Models in Scenarios

Models function in application scenarios, i.e. they have in those scenarios a function^{8,9}. Typical functions in science and engineering scenarios are reflection, illustration, visualisation, being a theory surrogate, guiding thoughts and activities, aiding for theory construction, mediating, and substituting theories.

Models are used instruments. The instrument-being is, thus, a pre-requisite for the model-being. The means that models have to be optimised on the function that the model has in the given application scenario. Instead of considering holistic models, model suites with a sophisticated and explicit association schema among models in the model suite are far better accommodated to model-based reasoning and deployment in scenarios. A scenario consists of a task space and an envisioned delivery space. Instruments may functions in a variety of ways. Therefore, a model may serve in several functions. Also, a scenario may consist of a collection of scenarios. The upper part in Figure 1 depicts this ‘landscape’ of the model-being.

³The ‘origin’ is different from ‘original’. ‘Origin’ means the source of something’s existence or from which it derives or is derived. It points to the place, event, the point of origination, the initial stage of a developmental process, etc. where something begins, where it springs into being.

⁴*Well-formedness* is often considered as a specific modelling language requirement.

⁵The criteria for *adequacy* are analogy (as a generalisation of the mapping property that forms a rather tight kind of analogy), being focused (as a generalisation of truncation or abstraction), and satisfying the purpose (as a generalisation of classical pragmatics properties).

⁶The model has another constituents that are often taken for granted. The model is based on a background, represents origins, is accepted by a community of practice, and follows the accepted context. The model, thus, becomes *dependable*, i.e. it is justified or viable and has a sufficient quality. Most notions assume dependability either as a-priori given or neglect it completely.

⁷We note that the instrument-being is based on the function that a model plays in some scenario.

⁸The word ‘function’ has seven word fields for the noun and three for the verb. We use here the meaning of a function that is associated with purpose, role, use, utility, usefulness, i.e. what something is used for.

⁹The word ‘function’ is often considered a synonym of ‘goal’ or ‘purpose’. We distinguish the three word and use a layered approach: *Goal* is definable as a ternary relation between initial state, desired states and community of practice who may assess the states and follow their beliefs, desires, and intentions. *Purposes* extend goals by means, e.g. methods, techniques, and operations. *Functions* embed the model into practices in applications and, thus, relate the purpose to the application, i.e. as a role and play of the model in an application scenario.

2.3. The Model-Being of Things and Thoughts

The model-being is determined by the function of an instrument in an application scenario. Nothing is a-priori a model. Things and thought have not to be models forever. Models have their journey in the model-being. They can be used in one function, remain to be useful or pass away as model. They can be used in a different function at a later stage. Criteria for the model-being seem to be necessary for some demarcation, i.e. a discrimination between things and thoughts as different and distinct on the basis of their characteristics or attributes. The demarcation can be derived from the model-being of an instrument and from the instrument-being of something:

1. A model functions in scenarios. It may functioning well, optimally, flawlessly, properly, satisfactorily, or primarily. Or barely and poorly.
2. A model may have several functions. The function might change during model's existence.
3. Functions can be characterised. This characterisation is an essential element of the mission, determination, meaning and identity of something¹⁰.
4. Functioning may be matured. The maturity level depends on the model objectives.
5. Functioning can only be defined for specified scenarios. There is no universal function of a model.
6. Model functions determine the adequacy and dependability.

Typical engineering functions are blueprint for realisation, starting point, prescription, mould, guide, companion, modernisation, integration, replacement, deploy, informative, recording, and assess. These functions, the usefulness, the utility, and the quality in and of use determines whether an instrument is a model. The instrument-being is based on the actual, practised, skilled, ideal, and desired play of a role in an application scenario. There are two main roles: the reflection of those origins the model have to be represented and the achievable result through use of the instrument. The instrument-being depends on the temporal, spatial, and disciplinary context of the community of practice.

The model-being is based on three viewpoints that determine the model utility as a mediator (see Figure 1): (1) the model-being as a 'model_of'¹¹ (2) the ends for the model-being as a 'model_for'¹² (3) the model-being based on the mediator function of the instrument¹³. Mediation includes transfer of main properties of origins that are essential in the given scenario to the result by means of the model, i.e. the model 'transports' those properties to the results of model application as *invariants*. The explication of the mediation can be directly given as

¹⁰B. Mahr [19,20] introduced the notion of *cargo* as a carrier of main properties and objectives of origins to important issues for the result. It describes the instrument, the main functions, the forbidden usages, the specific values of the instrument, and the context for the usage model.

¹¹Representing a collection of origins that are really of interest and relevant, i.e. the whereof as the source for the models existence or from which it derives or is derived.

¹²Essentially, a plan that is intended to achieve and that (when achieved) terminates behavior intended to achieve it, i.e. the for what as the cause or intention underlying model usage.

¹³Determined by a function in a given context and scenario, i.e. whereby as the helper that offers benefits and supporting means for achieving a result.

an informative model of the model (as an instrument). Informative models [32] are, thus, essentially the ‘product insert’ of the model. This model of the model is used as some kind of a leaflet or model suite insert that represents the essentials of the model suite. That means, we use already a model suite consisting of at least two models.

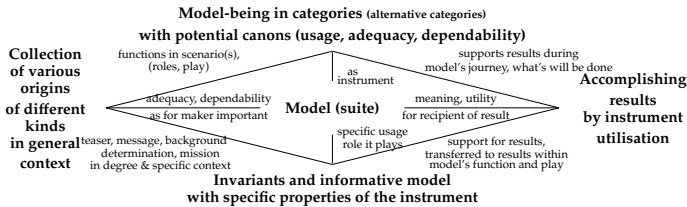


Figure 1. Characterising the model-being of instruments according to reflected origins and results to be accomplished by means of the model

The large variety of models in life, science, engineering, and thought seems to prevent development of a general ‘Modellkunde’ as a systematic study of models and modelling. There are (a) perception or mental, (b) representation or reflection, (c) communication or collaboration, (d) actuation and activity, (e) guidance or steering, (f) thought or reasoning, (g) substitution or sense-making, (h) socialisation or interaction, and (i) orientation models. These nine kinds of model follow however their specific adequacy and dependability canon and provide means for their usage. We, thus, realise that we have essentially nine different model categories of models.

As any instrument, a model has its own additional properties that are neither of importance for origins nor for results, its own authority, its obstinacy¹⁴, its profile (goals, purposes, functions) and anti-profile, its modus agendi and mode action, and its materiality. A model may, thus, also be misleading, disorienting, and of lower quality than other models.

3. Model-Based Reasoning Beyond Classical Logics

Sciences are oriented on true statements and consistency of theories, at least to certain extent. This explains the omnipresence of deductive systems as the main reasoning mechanism. Models must not be true. They can be contradictory or even paraconsistent. Models should be useful at least in some application scenario, for some time, for some community of practice, in some context, for some origins, for some results, within some background, on the basis of some supporting and enabling mechanisms, and within human restrictions. Model-

¹⁴M.W. Wartofsky already states in [36]: “There is an additional trivial truth, which may strike some people as shocking: anything can be a model! ... And although it is the case that anything can be a model of anything else, it is taken as a model which makes an actual out of a potential model; and every case of being taken as a model involves a restriction with respect to relevant properties.”

based reasoning does not have to be entirely based on classical logics. A similar observation can be made for engineering¹⁵. The study of models and model-based reasoning has also to be based on other kinds of reasoning.

The study of models is also concerned with obstacles, mismatches, limitations, and restrictions of model-based reasoning. Despite the common belief in most books and research on a theory of models (e.g. [28]), model-based reasoning is, however, rather seldom based on deduction and deductive-nomological reasoning.

3.1. The Obstinacy of Classical Logics

Classical mathematical logics mainly considers deductive systems and various mechanisms of deduction. Already C.S. Peirce [22] distinguished three reasoning mechanisms: deduction, induction, and abduction. Their difference is illustrated in Table 1 for a set of premises, supporting means, and results: These three reason-

	Deduction	Induction	Abduction
Reasoning style	Rule-Data-Result	Data-Result-Rule	Rule-Result-Data
First	general rules	observed primary phenomena	general rules
Second	specific observations	dependent secondary phenomena	dependent secondary phenomena
Finally	conclusion for observations and new rules	rule supposition and questions	potential (causal) explanations

Table 1. Subduction: deduction, induction, abduction

ing styles are well-known. Deduction is considered to be the main mechanism. It is the basis for Mathematical Logics. Deductive reasoning is based on three postulates that are too restrictive: (A) completeness of the specification, (B) agreement on the background and the matrix, and (C) context-independence. The Peirce triangle is more general, however: From its *abductive* suggestion, *deduction* can draw a prediction which can be tested by *induction*. Abduction is well-known. Induction is far less accepted since rule suppositions are only hypothetical results and have to be revised whenever primary and secondary phenomena are not matching anymore. Many researchers, e.g. K.R. Popper¹⁶, however, strictly deny usefulness and utility of induction and avoid usage of induction.

¹⁵See, for instance, [31] on the problematic side of first-order predicate logics for database engineering.

¹⁶[23]: "Induction simply does not exist, and the opposite view is a straight-forward mistake. ... I hold that neither animals nor men use any procedure like induction, or any argument based on repetition of instances. The belief that we use induction is simply a mistake."

3.2. Inductive Model-Based Reasoning

Induction is the most prominent and important reasoning mechanism in daily life and for model building based on evidences or observations. Inductive conclusions are uncertain due to the incompleteness of observations. Worlds that are potentially infinite are and will be, however, never completely observable. The inherent incompleteness of the world of phenomena shows that induction is the best logical mechanism for human reasoning.

We distinguish between:

- Induction in broad sense as explanatory inferences, as well as analogical and ‘more-of-the-same’ inferences in the style: ‘*All observed Xs have property P*’ to ‘*The next X observed will have property P*’. It includes explanatory inferences, as well as analogical and ‘more-of-the-same’ inferences.
- Induction in narrow sense is based on a random sample (with test/validation set) and results in simple enumerative induction (or the straight rule).

Induction degrees are either strong inductive argument based on authority, on evidence, or stronger inductive argument based on better evidence.

The inductive reasoning schema is based on given knowledge \mathcal{K} , beliefs \mathcal{B} , models \mathcal{M} known so far, and data \mathcal{D} observed. Within the setting of some reasoning systems Γ , we assume $\mathcal{K} \cup \mathcal{M} \not\models_{\Gamma} \mathcal{D}$.

The induction task aims at discovery of a formula α (not uniquely defined) such that

- it is coherent with \mathcal{K} , \mathcal{M} , and \mathcal{D} and
- that allows to explain \mathcal{D} , i.e.
 - * $\mathcal{K} \cup \mathcal{M} \cup \mathcal{D} \not\models_{\Gamma} \neg\alpha$ and
 - * $\mathcal{K} \cup \mathcal{M} \cup \{\alpha\} \models_{\Gamma} \mathcal{D}$

Inductive reasoning schemata can be extended to Solomonoff induction bound by Kolmogorov complexity. In this case, sophisticated inductive reasoning generates most relevant, most simple, and preferred generalisations from facts and/or observations. The conclusions can be revised whenever the fact or observation set is extended. Inductive reasoning inherits the obstinacy of the representation language of facts and observations.

Induction is a kind of compilation-so-far reasoning with uncertainty, preferences for conclusions, and complexity reduction for result presentation.

Induction is transfer of likely truth from a number of observations to a general principle. It is based on conjecture spaces and specific approaches, experience, (tacit) knowledge, parsimony, economy, clever sampling, and wise experimentation. Induction is a very strong modelling principle. We appreciate statistical, probabilistic, possibilistic, eliminative, and mathematical induction. Induction can be treated as ‘blind’ search (depth-first, breath-first). It can also be clever search for suppositions as humans like to do.

3.3. Abductive Model-Based Reasoning

Abductive reasoning (e.g. [2,18,24]) is a kind of concise reasoning that infers particular cases from general observations and rules. It is a weak kind of inference because we cannot say that the explanation is true, but that it can be true.

Premises are given in the form:

- D is a collection of data, facts, observations
- M explains D within a given reasoning mechanism.
- No other model can explain D as well as M does.

Conclusion: Therefore, the model M is probably acceptable.

Abductive model-based reasoning is a process that tries to form plausible models for some situations. It covers also abnormal situations. The inference result is a model, which is somehow acceptable within the given reasoning mechanism and, thus, could explain the occurrence of the given facts. This approach can be used for detection of good explanations and especially good causal explanations.

A typical abductive hypothetical reasoning schema is the following:

1. Searching somehow anomalous, surprising, or disturbing phenomena and observations.
2. Observing details, little clues, and tones.
3. Continuous search for hypotheses and noting their hypothetical status.
4. Aiming at finding what kind or type of explanations or hypotheses might be viable to constraint the search in a preliminary way.
5. Aiming at finding explanations (or ideas) which themselves can be explained (or be shown to be possible).
6. Searching for "patterns" or connections that fit together to make a reasonable unity.
7. Paying attention to the process of discovery and its different elements and phases.

Abductive reasoning also allows to consider negative information by modus tollens

$$\frac{H \rightarrow I, \neg I}{\neg H}$$

The Mathematical Model of Meaning [15] is a third kind of abductive reasoning schema that is used for categorisation of observations:

- Empirical observations can be represented by data representing the importance of some feature for the observation.
- Importance data should be normalised, e.g. 0, ..., 10.
- Data can be represented as a table (or matrix) with some features/indicators as attributes.
- Attributes (in the universal world approach) can be related to categories. The database is then a universal relation with tuples where those values that are $\neq 0$ show belongingness to a category.
- Multiplication of tuples from the observations with the feature-category matrix results in a tuple that characterises the belongingness of an observation to a category.

Feature $F_j (1 \leq j \leq f)$ are relevant $d_{i,j} (1 \leq i \leq m, 1 \leq j \leq f)$ for observations $o_i (1 \leq i \leq m)$. These features belong to categories $C_k (1 \leq k \leq R)$ by a knowledge or abduction matrix $c_{jk} (1 \leq j \leq f, 1 \leq k \leq R)$.

$$\begin{pmatrix} F_1 & \dots & F_f \\ o_1 & d_{1,1} & \dots & d_{1,f} \\ \dots & \dots & \dots & \dots \\ o_m & d_{m,1} & \dots & d_{m,f} \end{pmatrix} \times \begin{pmatrix} C_1 & \dots & C_R \\ F_1 & c_{1,1} & \dots & c_{1,R} \\ \dots & \dots & \dots & \dots \\ F_f & c_{f,1} & \dots & c_{f,R} \end{pmatrix} = \begin{pmatrix} C_1 & \dots & C_R \\ o_1 & r_{1,1} & \dots & r_{1,R} \\ \dots & \dots & \dots & \dots \\ o_m & r_{m,1} & \dots & r_{m,R} \end{pmatrix}$$

The result of multiplication of the observation matrix with the abduction matrix is a matrix of relevance of a category C_1, \dots, C_R for an observation o_i .

3.4. Principles and Assumptions of Model-Based Reasoning

Models are instruments that properly function in utilisation scenarios. The utility is given by the quality of appropriateness in use. Therefore, we have to understand which objects, artifacts, and thoughts can really be used as models whenever we base reasoning on models. Appropriateness of models is a specific variant of the design principle 'form-follows-function'.

Goals, purposes, and function must be well-defined, well-thought and achievable: In most cases, utilisation scenarios are neither an ad-hoc, nor chaotic, or nor trial-and-error flows of work. They must not be fully defined. We have to understand to a greater or lesser extent what should be done and, especially, which instruments might be useful in which way on which grounds. The profile of an instrument is given by the goal we follow, by the means we could use for our goal (i.e. purposes of the instrument), and by the way how the instrument is going to be used according to the purpose (i.e. function of the instrument). Instruments shall be effective. From the other side, appropriateness of an instrument is also determined whether the goal is accomplishable.

Models are mental compilations of observed worlds: Models are a product of our thoughts. As a referent, we observe some situation in our world. Following the consideration by [16] on the three analogies by Platon (analogy of the cave, of the sun, and of the divided line), the referent recognises shadows in the observable world, builds some comprehension based on the thoughts and his/her intellect, and uses some language (not necessarily natural one; potentially some visual one) for reflection by terms, e.g. signs and images.

The background of models strikes through or is limiting reasoning: As already noted, models are often only given as the normal model while the deep model is implicit and the matrix of model application is commonsense in a discipline. As long as the deep model and the matrix are unconditionally acceptable and have not to be changed, the results of model functioning are reliable. There are, however, reasons to reconsider this background and these application frames. The potential and capacity of a model is restricted to these assumptions. Models are, however, not really context-free. They have their anti-profile also due to restrictions and their focus.

Evidential reasoning as initial point for model development: Evidential reasoning starts with evidences or clues and compiles guesses or conclusions, thus, providing hints about possible or likely conclusions with an explicit representation of uncertainty. Evidences, thus, support or refute hypotheses about the current status of the existing and observable situation. Unobservable propositions can be then determined on the basis of observable evidence, e.g. the observable data are used to reason on the almost unobservable real and micro-data. Evidential reasoning should be distinguished from causal reasoning which orients on explaining observable evidences by a hypothesised cause. It has its limits which should be integrated into this reasoning style.

Living in a world without necessity for a universal world formula — Almost plausibility and inherent incompleteness: Models have to be incomplete whenever they are based on the principles of reduction, decontextualisation, vagueness, and ignorance. Models used in some domain have not to be consistent. A property that is acceptable in this case is coherence what means that sub-models of models which express the same set of properties are compatible and partially homogeneous to certain degree (so-called non-adjunctive model suites) [14]. Inconsistency is handled in a controlled way by many-facetted coherence without integration. A classical example is Bohr's theory of atom and the system of Maxwell's equations. Paraconsistency treats a collection of models as consistently as possible without requiring full consistency. Model suites represent then some kind of 'knowledge islands' with partial bridge axioms.

It is surprising that neither form-follows-function, form-restricts-function, function-and-form-determine-techniques, nor inherent incompleteness and almost plausibility have been explicitly discussed in model theory and practice. A model property that is commonly accepted is well-formedness (some times called 'beauty', stronger well-defined) of models. They allow proper application of methods that support functioning of models.

4. Model-Based Reasoning Mechanisms

Model-based reasoning is completely different from classical logical reasoning techniques. It might use deduction. Models don't have to be true, consistent, fully integrateable with other models, based on a homogen understanding, at the most recent state-of-art, or acceptable by everybody. They can be certain to some limited extent, somehow coherent or even paraconsistent, heterogenous, representing islands without homogeneity, combine various generations of knowledge, or personal opinions. The two lists are not complete but demonstrate the difficulty to develop a sophisticated theory of model-based reasoning. Instead, let us consider some of the most essential reasoning procedures for models.

4.1. Plausible Reasoning

Models must not be complete and are considered within the given but changeable context. Models don't have to be true. They have to be useful and functioning

as instruments in the given scenario. Models focus and scope on certain parts while neglecting others, i.e. they are using approaches of ignorance. Therefore, they are typically incomplete. The premise set does not strictly allow for every conclusion with certainty. Model development is often based on inductive and evidential reasoning which is another source for incompleteness. Model-based reasoning has to additionally use more appropriate reasoning mechanisms. We especially use approaches based on plausibility and approximation.

Plausible reasoning stand for reasoning with uncertain conclusions for both certain or uncertain premisses. Typical forms are abductive, analogical, autoepistemic, counterfactual, default, defeasible, endorsement-backed, presumptive, and non-monotonic reasoning techniques.

The classical approach to plausible reasoning is given by the following schema:

The lack of soundness makes the conclusion plausible with a certainty below 1.0 based on evidence $CertF(\alpha|e)$ and reasonable (called believable) $ReasonF(\alpha|e)$ with reasonability below or equally certainty.

Certainty factors and reasonability factors may follow empirical rules to aggregate pieces of evidence, e.g.

$$CertF(\alpha|e_1 \wedge e_2) = \min(CertF(\alpha|e_1), CertF(\alpha|e_2)) \text{ and}$$

$$CertF(\alpha|e_1 \vee e_2) = \max(CertF(\alpha|e_1), CertF(\alpha|e_2))$$

in the Dempster/Shافر reasoning style or in the possibility theory style.

These rules shall be applied in dependence on the context since they may lead to unpredictable, problematic, and counterintuitive results. Negation can be handled as negation-by-failure or in a multi-valued or paraconsistent form [25].

4.2. Approximative Reasoning

Models must not be precise although precision is necessary whenever models are used for automatic generation of solution from a given model [4,7,10]. Instead, model can be reduced, abstracted, truncated, imprecise, and raw. We follow the principle of parsimony and economy. Models must support efficient and effective thinking and actuation. The final and optimal solution might not exist at all or is infeasible both in time for its generation and in space for its presentation. Although, tools might not exist.

Approximation supports aggregative, generative, imprecise and robust reasoning. Approximate reasoning based on models is a common form to avoid complexity throwback due to over-detailing. Typical kinds are reasoning systems supporting aggregation and cumulation, generalisation and categorisation, imprecision, heuristics, robust thinking, and shallow consideration. These reasoning styles are used in daily life and especially for models, e.g. best characterised by the Austrian saying 'paßt schon' (fits somehow, fits already, close enough, suits, somehow convenient). In Computer Science, approximate algorithms provide a reasonable solution to problems at polynomial time instead of optimal solutions computable at (hyper-)exponential time in dependence on the problem complexity measured by Kolmogorov complexity [17]. The principle of Occam's razor orients on models as 'simple' as possible. We, thus, do not miss simple models. We may also use approximative rules with preference and simplicity in the Solomonoff style [8], e.g. for model-based explanation.

4.3. Hypothetical Model-Based Reasoning

Hypothetical model-based reasoning is based on the following schema:

1. Given a hypothesis model M that implies a statement E which describes observable phenomena.
2. The statement E has been observed as true.
3. The conclusion is that M is true.

The method of the hypothesis is not deductively valid because wrong hypotheses can also have real consequences.

Different assumptions are considered in order to see what follows from them, i.e. reasoning about alternative possible models, regardless of their resemblance to the actual world. Potential assumptions with their possible world conclusions assertions are supported by a number of hypotheses (allowing to derive them). Inductive model-based reasoning can be combined with abductive reasoning.

Hypothetical model-based reasoning restricts inductive reasoning by specific forms of inductive conclusions:

1. *statistical inductive generalizations*, in which the premise that x percent of observed A's have also been B's, so that the conclusion is, x percent of all A's are B's;
2. *predictive conclusions*, in which the premises are that x percent of the observed A's have also been B's, and a is an A and where the conclusion is that a is a B;
3. *direct conclusions*, in to which the premises are that x percent of all A's are also B's, and that a is an A and where the conclusion is that a is a B, and
4. *conclusion by analogy* in which the premises are that certain individual objects have the properties F_1, \dots, F_n and a have the properties F_1, \dots, F_{n-1} , and where the conclusion is that a also has the property F_n .

4.4. Model-Based Explanation

The reasoning schema used for inductive reasoning can be extended to model-based explanation that describes, explains, illustrates, clarifies and characterises in a guiding way in a mediation scenario essential, central and in the given scenario important elements of complex origins in a comprehensible, concrete and coherent form for the recipient.

The reasoning schema for model-based explanation, elaboration, and comprehension can be defined as follows:

Given

some theoretical background \mathcal{T} based on the context, background, knowledge, concepts, etc.,

a model class \mathfrak{M} with orders for preference \leq and simplicity \lesssim ,

a deducability operator $\#$ as an advanced operator for deductive, abductive, inductive, non-monotonic, approximative, and plausible derivation of conclusions, and

data under consideration O (observations) from the data space derived from input model suite data and prepared for analysis while being $\mathcal{T} \# O$ (non-trivial for \mathcal{T}) and $\mathcal{T} \# \neg O$ (not in conflict with \mathcal{T}).

The model M_E from \mathfrak{M} is an **explanation model** for O within \mathcal{T} if it explains O within \mathcal{T} , i.e. $\mathcal{T} \sqcup M_E \Vdash O$ while being non-trivial (or parsimonious) for O , i.e. $M_E \not\vdash O$ and coherent with O and \mathcal{T} , i.e. $M_E \not\vdash \neg O$ and $\mathcal{T} \sqcup M_E \not\vdash \perp$.

Based on the orders in \mathfrak{M} we may be interested in the weakest (best) explanation model that is additionally parsimonious, i.e. $M_E \not\vdash O$.

Model-based explanation is not consolidative modelling that uses the model as a surrogate for the system, for instance, by consolidating known facts about the system for purposes of analysis whether the model adequately represents the system. Model-based explanation explores how the world would behave if various models were correct. Many details and mechanisms of a system are uncertain. The model has not to be a reliable image of the world. Relevant ‘ground truth’ data for evaluating model may not be obtainable. We, thus, identify an ensemble of plausible models and modelling assumptions, identify the range of outputs predicted by plausible models under plausible assumptions, and identify the relationship between modelling assumptions and model outputs. A trick is to find assumptions that have a large impact on model outputs. Another trick is to identify predictions that are robust across different sets of modelling assumptions.

There is no methodological approach for derivation of a good explanation model. It seems that the embracement method is the strongest one. This method considers at the same time the generation of models from one side and the generation of partially explaining models from the other side. These partially explaining models can be seen as hypotheses which would form a good explanation model together with the generated model. A simple embracement method are Mill’s methods of agreement, difference, joint methods of agreement and difference, residues, and concomitant variations.

4.5. The Model as Mediator in Empiric Model-Based Reasoning

Empiric investigation and reasoning is based on data spaces. Data are structured according to properties of parameters. It is often not possible to observe all parameters. We may distinguish outer parameters that can be observed and inner parameters that cannot be accessed or are not yet observed or are not yet observable. This separation is similar to genotype and phenotype observations. The problematic accessibility for inner parameters has already been discussed by Platon in his analogy of the cave (see, for instance, [16]). Development of an understanding based on outer parameters is a real challenge that is difficult to overcome. The data space should give an insight into potential quantitative observation concepts or conceptions. We need an insight into the data space for the inner parameters in order to reason on the reality situation.

Empirical reasoning starts with an investigation of data sources for the outer parameters and develops some quantitative observation concepts that might be embedded into a theory offer. A *theory offer* is a scientific, explicit and systematic discussion of foundations and methods, with critical reflection, and a system of

assured conceptions providing a holistic understanding. A theory offer is understood as the underpinning of technology and science similar to architecture theory [26] and approaches by Vitruvius [35] and L.B. Alberti [1]. Theory offers do not constitute a theory on their own, rather are some kind of collection consisting of pieces from different and partially incompatible theories, e.g. sociology theories such as the reference group theory, network theories, economic theories such as the agent theories, Darwinian evolution theories, subjective rationality theories, and ideology theories.

The main target is however, to form a theory that is based on the data, that is based on concept or conceptions, and that allows to draw conclusions on this theory. Concepts or conceptions to be developed should be qualitative and theory-forming. If qualitative concepts cannot be drawn then we need quantitative concepts that allow reasoning. Figure 2 displays this challenge. The challenge is solved if a number of functions exist.

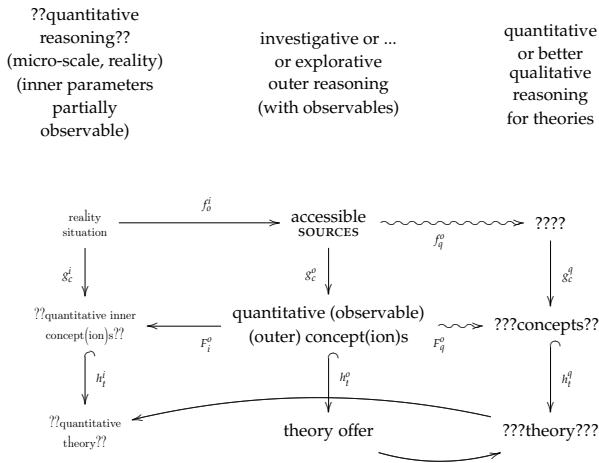


Figure 2. The challenge of empirical reasoning targetting on qualitative concepts and theory development with only partially known data for the inner parameters

The best solution for this challenge would be if we can map the inner parameters to the outer ones by a function f_o^i and use some kind of abstraction function g_c^o for association of data to concept(ion)s. In this case, we might succeed in constructing functions g_c^i resp. F_i^o from the reality situation resp. outer concept(ion)s to quantitative inner concept(ion)s. Then we could use the theory embedding of outer quantitative concept(ion)s h_i^o for construction of such inner functions h_i^i . This would also result in a coherence condition and a commuting diagram $F_i^o(g_c^o(f_o^i(situation))) = g_c^i(situation)$. We will be able to use the em-

bedding function h_i^q for construction of a corresponding supposition h_i^q for inner parameter theories.

The next step is a construction of a reasoning system. We use some aggregation function f_q^q for compiling sources in support of concepts by a function g_c^q and for embedding these concepts into a theory by h_i^q . If we succeed then we can use the theory offer for the construction of a theory for reasoning and as the next step for mapping this theory back the to inner quantitative theory.

We arrive therefore with the big challenge of empiric research: *How we can close the gap between quantitative theory offers and qualitative theories?*

This program for empirical reasoning is not really feasible. The construction of the functions is a higher-order challenge. Instead we can use model-based reasoning as displayed in Figure 3. The model is then used as a mediating means between qualitative and quantitative reasoning. The model is at the same time (1) a means, (2) a mediator, and (3) a facilitator [3], i.e. (1) an instrumentality for accomplishing some end, (2) a negotiator who acts as a link between quantitative and qualitative issues, and (3) an instrument that makes reasoning easier. Since models are more focused, we do not have to have fully-fledged functions. Instead, we can concentrate on the main issues. At the same time, we properly support qualitative reasoning based on our data spaces. This would also allow to formulate proper hypotheses from the model world to the quantitative world. The validity power of the model would then support qualitative reasoning.

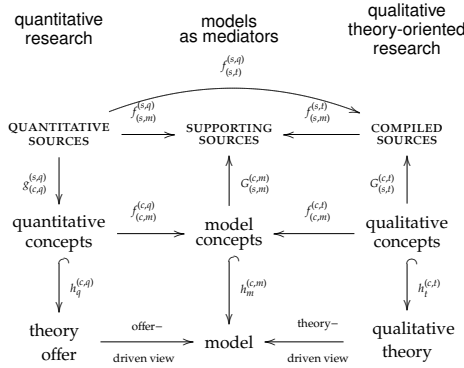


Figure 3. Models as integrating and mediating instrument in empiric research

In this case, we may succeed in constructing insights that go far beyond data-backed reasoning, e.g. in data science. We could then also construct massives of supporting sources for our models. The model accommodates the quantitative theory, the theory offers, and the qualitative theory.

Theories can be built on the basis of theoretical concepts which are supported by sources. Quantitative concepts should be associated with qualitative concepts. The association can only be developed in the case when the association among the data has been clarified. So far, the explanations that can be generated are mainly developed for explaining the observations made on the basis of outer data.

4.6. Meta-Model Reasoning Used for Model-Based Steering

Thinking in models should be supported by a systematic methodology. Model-based steering use meta-models (i.e. *steering models*) as a guiding or motivating model that directs the direction of model-based reasoning. This kind of meta-reasoning enables us to explore potential opportunities in the opportunity and possibility spaces. We arrive at some proposal as a result of model-based reasoning, i.e. putting forward or stating something for consideration by making or offering a formal plan or suggestion for acceptance, adoption, or performance. The *proposal model* is used in a second step as an additional origin. The trick we use is then based on second-order modelling. We know already *governing models*. Such meta-models make and administer the selection of opportunities based on possibilities, regulate, and control model-based reasoning while keeping exploration under control. They exercise a deciding or determining influence on selection and thoughts.

As a result of steering and governing, we obtain an *advice model* as a recommendation regarding a decision or course of conduct. It could be considered to state an opinion about what could or should be done about a situation or problem, what is going to be recommended offered as worthy to be followed. Advice models are used as a counsel and denote an opinion as to a decision or course of action.

We may use meta-techniques such as specific question-answer forms (or, more specifically, query-answer or input-output forms [9]). These question-answer forms have their inner meta-structure and inner meta-flow that could be used in investigative research, e.g. what-if analysis, what-would-be-if, 5-why-drill-down, rolling-up distancing, context-enhancement, assumption-slicing with attention restriction, why-it-must-be, why-this-question-and-not-other, why-not-rephrasing, observation-in-context dependencies, immersion-into-context, why-finish, how-we-can-know, why-this-question, question-reformulation by opening or closing the parameter space, and parameter-space-reduction by dicing with tolerance of errors, e.g. by principal component analysis. Essentially, these meta-techniques are steering models.

We use the steering model for driving into a problem space and detecting opportunities and possibilities. Sciences, engineering and daily life are full of such ‘wisdom’ techniques.

This approach can be generalised to meta-models for research, i.e. moulds¹⁷. Methodologies are simple moulds. They provide a guidance for a flow of work. Frameworks are complex moulds that can be adapted to the given situation¹⁸. Civil engineering uses moulds as frame on which something can be constructed.

Steering models are used to control or to direct or to guide the course of actuation. They set, follow, pursue, and hold to a course of action and reasoning and especially a hint as to procedure. Steering models are used as a piece of advice or information concerning the development of a situation. They allow to control a situation so that it goes in the direction that you want. They enable to take a particular line of action. Such meta-models are models of the models, of the modelling activities, and of the model association within a model suite. Their goal is to improve the quality of model outcomes by spending some effort to decide what and how much reasoning to do as opposed to what activities to do. It balances resources between the data-level actions and the reasoning actions. A typical case is design of activities in data mining or analysis [11] where agents are preparation agents, exploration agents, descriptive agents, and predictive agents. Meta-models for a model suite contain decisions points that require macro-model control according to performance and resource considerations. This understanding supports introspective monitoring about performance for the data mining process, coordinated control of the entire mining process, and coordinated refinement of the models. Meta-level control is already necessary due to the problem space, the limitations of resources, and the amount of uncertainty in knowledge, concepts, data, and the environment.

Steering models extend the origin' collection (see Figure 1) by meta-reasoning origins. These origins enable us to use second-order cybernetics [34], i.e. to continuously reason on insight we got in previous steps and to change our mind whenever new insight has been obtained. The methodology follows a mould of continuous changing method application.

This approach is based on control by meta-reasoning [6] as displayed in Figure 4. We distinguish the activity layer that is based on methods ground level, the meta-modelling level that contains the modelling methods level and rules the selection of actions, and the meta-meta-modelling level that contains the abstraction to meta-modelling methods and controls the middle level section. This approach is similar to government and binding [5] where the utterance payout is based on a second layer payout selection for the utterance that is again ruled by controller for settling the kind of utterance and its general form.

¹⁷A *mould* is a distinctive form in which a model is made, constructed, shaped, and designed for a specific function a model has in a scenario. It is similar to mechanical engineering where a mould is a container into which liquid is poured to create a given shape when it hardens. In Mathematics, it is the general and well-defined, experienced framework how a problem is going to be solved and faithfully mapped back to the problem area.

¹⁸For instance, analytical solution of differential equations use a set of solution methods formulated as *ansatz*. Database development can be guided by specific frameworks. Artists and investigative researchers are guided by moulds, i.e. by steering or governing models. Software engineering is overfull of meta-models for design, development, and quality management.

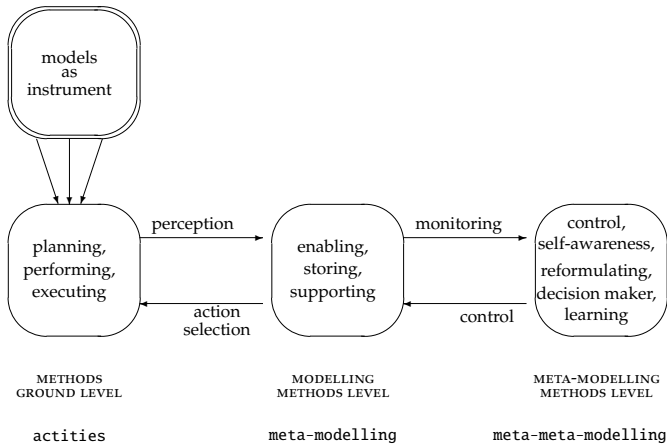


Figure 4. The meta-control mould for meta-reasoning

5. Conclusions for 'Modellkunde' – The Study of Models and Modelling

Models must not be true or consistent. They should be useful as instruments in application scenarios. Usefulness presupposes the existence of techniques for model utilisation. Reasoning is one kind of technique. Reasoning is a daily life practice that is rarely based on deductive systems. Instead, induction and abduction are used for model-based reasoning. These reasoning techniques are especially useful for models since we do not require consistency and truth maintenance.

We demonstrated the power of such techniques by the generalisation to plausible, approximative, hypothetical, and explanatory reasoning. These mechanisms are really sophisticated. For instance, pattern recognition and Modellkunde for pattern can be based on explanatory model-based reasoning in combination with mediator approaches. One of the best achievements is mediator-based reasoning that allows to overcome pitfalls of middle-range theories and badly associated theory offers in empiric research. Instead we use models as a mediating device between empiric and qualitative reasoning. The utilisation of models may be governed by other models. Steering meta-models guide application and usage of models.

This paper can be extended by application of other model-based reasoning techniques such as separation and concentration of concern, playing with ignorance and de-contextualisation, qualitative techniques used for data analysis, and probabilistic calculi.

This paper has been centred around reasoning techniques and model-based reasoning. There are many other techniques beside reasoning that can be applied

to models. Typical techniques are enhancements similar to conceptualisation, model inheritance from generic or reference models, parameter hardening used for inverse modelling in physics, and simulation of behaviour for some of the parameters. Cognitive modelling is another technique that has been left out for this paper. Shallow and deep reasoning techniques are another lacuna for the study of models.

The study of models has to consider also other techniques for model utilisation. Models form a landscape. Some models are partially isolated. These isolated models should be supported by bridging techniques. Models are focused and have, thus, their abstraction level. Model-based problem solving use, therefore, also techniques for generalisation and governed specialisation.

Models can also be composed in vertical or horizontal layering. Models can be also origins for other models. The composition should be supported by techniques similar to nested data warehousing, i.e. roll-up, drill-down, dice, slice, rotate, algebraic construction, peaceful renovation and updating, unnesting, and nesting. Models may consist of a well-associated collection of models, i.e. of a model suite. Association techniques allow also management of coherence of models in a model suite. Some models in a model suite may play the role of a master (or order) model while others are slaves. The model suite should, thus, be enhanced by control models for adaptation of the master models to a given situation, e.g. in inverse modelling.

References

- [1] L.B. Alberti. *On the art of building in ten books*. MIT Press, Cambridge. Promulgated in 1475, published in 1485, 1988.
- [2] A. Aliseda-Llera. *Seeking explanations: Abduction in logic, philosophy of science and artificial intelligence*. PhD thesis, Universiteit van Amsterdam, 1997.
- [3] C. Blättler. *Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung*, chapter Das Modell als Medium. Wissenschaftsphilosophische Überlegungen, pages 107–137. De Gruyter, Boston, 2015.
- [4] E. Börger and A. Raschke. *Modeling Companion for Software Practitioners*. Springer, 2018.
- [5] N. Chomsky. *Some concepts and consequences of the theory of government and binding*. MIT Press, 1982.
- [6] M.T. Cox and A. Raja, editors. *Metareasoning - Thinking about Thinking*. MIT Press, Cambridge, 2011.
- [7] A. Dahanayake, O. Pastor, and B. Thalheim. *Modelling to Program: Second International Workshop, M2P 2020, Lappeenranta, Finland, March 10–12, 2020, Revised Selected Papers*, volume 1401 of CCIIS. Springer Nature, 2021.
- [8] David L. Dowe, editor. *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, volume 7070 of LNCS. Springer, 2013.
- [9] H. Jaakkola and B. Thalheim. Supporting culture-aware information search. In *Information Modelling and Knowledge Bases XXVIII*, Frontiers in Artificial Intelligence and Applications, 280, pages 161–181. IOS Press, 2017.
- [10] H. Jaakkola and B. Thalheim. Models as programs: The envisioned and principal key to true fifth generation programming. In *Proc. 29th EJC*, pages 170–189, Lappeenranta, Finland, 2019. LUT, Finland.
- [11] K. Jannaschk. *Infrastruktur für ein Data Mining Design Framework*. PhD thesis, Christian-Albrechts University, Kiel, 2017.

- [12] M. Kangassalo and E. Tuominen. Inquiry based learning environment for children. In *Information Modelling and Knowledge Bases XIX*, volume 166 of *Frontiers in Artificial Intelligence and Applications*, pages 237–256. IOS Press, 2007.
- [13] R. Kaschek. *Konzeptionelle Modellierung*. PhD thesis, University Klagenfurt, 2003. Habilitationsschrift.
- [14] R. Kauppi. Einführung in die Theorie der Begriffssysteme. *Acta Universitatis Tamperensis, Ser. A, Vol. 15, Tampereen yliopisto, Tampere*, 1967.
- [15] Y. Kiyoki, T. Kitagawa, and T. Hayama. A metadatabase system for semantic image search by a mathematical model of meaning. *SIGMOD Rec.*, 23(4):34–41, 1994.
- [16] C. Lattmann. *Vom Dreieck zu Pyramiden - Mathematische Modellierung bei Platon zwischen Thales und Euklid*. Habilitation thesis, Kiel University, Kiel, 2017.
- [17] M. Li and P.M.B. Vitanyi. *An introduction to Kolmogorov complexity*. Springer, New York, 2008.
- [18] L. Magnani, W. Carnielli, and C. Pizzi, editors. *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*. Springer, 2010.
- [19] B. Mahr. *Visuelle Modelle*, chapter Cargo. Zum Verhältnis von Bild und Modell, pages 17–40. Wilhelm Fink Verlag, München, 2008.
- [20] B. Mahr. Information science and the logic of models. *Software and System Modeling*, 8(3):365–383, 2009.
- [21] B. Mahr. Modelle und ihre Befragbarkeit - Grundlagen einer allgemeinen Modelltheorie. *Erwägen-Wissen-Ethik (EWE)*, Vol. 26, Issue 3:329–342, 2015.
- [22] C.S. Peirce. *The Collected Papers of Charles S. Peirce*, (C. Hartshorne, P. Weiss, and A. W. Burks (eds.)). Cambridge: Harvard University Press, 1931-1966.
- [23] K. R. Popper. *Logik der Forschung*. J.C.B. Mohr (Paul Siebeck), Tübingen, 10th. edition, 1994.
- [24] M. V. Rodrigues and C. Emmeche. Abduction and styles of scientific thinking. *Synth.*, 198(2):1397–1425, 2021.
- [25] K.-D. Schewe and B. Thalheim. NULL value algebras and logics. In *Information Modelling and Knowledge Bases*, volume XXII, pages 354–367. IOS Press, 2011.
- [26] G. Semper. *Die vier Elemente der Baukunst*. Braunschweig, 1851.
- [27] H. Simon. *The Sciences of the Artificial*. MIT Press, 1981.
- [28] H. Stachowiak. *Allgemeine Modelltheorie*. Springer, 1973.
- [29] B. Thalheim. The conceptual model \equiv an adequate and dependable artifact enhanced by concepts. In *Information Modelling and Knowledge Bases*, volume XXV of *Frontiers in Artificial Intelligence and Applications*, 260, pages 241–254. IOS Press, 2014.
- [30] B. Thalheim. Conceptual models and their foundations. In *Proc. MEDI2019, LNCS 11815*, pages 123–139. Springer, 2019.
- [31] B. Thalheim. Semiotics in databases, keynote paper. In *Proc. MEDI2019, LNCS 11815*, pages 3–19. Springer, 2019.
- [32] B. Thalheim and A. Dahanayake. Comprehending a service by informative models. *T. Large-Scale Data- and Knowledge-Centered Systems*, 30:87–108, 2016.
- [33] B. Thalheim and I. Nissen, editors. *Wissenschaft und Kunst der Modellierung: Modelle, Modellieren, Modellierung*. De Gruyter, Boston, 2015.
- [34] S. A. Umpleby. Second-order cybernetics as a fundamental revolution in science. *Constructivist Foundations*, 11(3):455–465, 2016.
- [35] Vitruvius. *The ten books on architecture (De re aedificatoria)*. Oxford University Press, London, 1914.
- [36] M.W. Wartofsky. *Conceptual foundations of scientific thought*. MacMillian, New York, 1968.

Remark: More detailed information on our research papers can be found on research gate in [collections](https://www.researchgate.net/profile/Bernhard_Thalheim) at https://www.researchgate.net/profile/Bernhard_Thalheim .

See also the youtube channel "Bernhard Thalheim", videos at <https://vk.com/id349869409> or <https://vk.com/id463894395> (in German).

Polymorphism of Intelligence – A Look at Human and Artificial Intelligence

Hannu JAAKKOLA^{a1}, Bernhard THALHEIM^b and Jaak HENNO^c

^aTampere University, Finland

^bChristian Albrechts University at Kiel, Germany

^cTallinn University of Technology, Estonia

Abstract: Computers were originally developed for executing complex calculations fast and effectively. The intelligence of the computer systems was based on arithmetic capabilities. This has been the mainstream in the development of computers until now. In the middle of the 1950s a new application area, Artificial Intelligence (AI), was introduced by researchers. They were interested in using computers to solve problems in the way intelligent beings do. The architecture, which supported calculations, was harnessed to perform tasks associated with intelligent beings, to execute inference operations and to simulate human sense. Artificial intelligence has had several reincarnation cycles; it has reappeared in different manifestations since this area became of interest to researchers. All the time a lot of discussion about the intelligence of these systems has been ongoing – are AI based systems and robots intelligent? What is the difference between human and machine intelligence? Abilities related to intelligence include the ability to acquire and apply knowledge and skills, as well as the ability to learn. AI provides different manifestations of the term “intelligence”: human intelligence incorporates a wide variety of different types of intelligence, and the meaning of artificial intelligence has varied over time as well. In our paper we will look at this term, especially to provide a means for comparing human and artificial intelligence, and have a look at the learning capability related to it.

1. Introduction

Artificial intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings². *Intelligence* (synthesis of several sources) is defined as the ability to acquire and apply knowledge and skills; the ability to learn, understand, and think in a logical way about things; the skilled use of reason; the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria.³ Key aspects related to intelligence are the ability to apply knowledge, *reasoning*, *learning capability*, *the ability for abstract thinking*, and *the aim of using intelligence to affect something*. Artificial intelligence instead “*simulates*” *human intelligence*. What are the differences and similarities of these two kinds of intelligence? This is the starting point of our paper.

¹ Corresponding Author: Hannu Jaakkola, Hannu.jaakkola@iki.fi.

² Britannica. <https://www.britannica.com/technology/artificial-intelligence>

³ Oxford: <https://www.lexico.com/definition/intelligence>; <https://www.oxfordlearnersdictionaries.com/>;

Merriam-Webster <https://www.merriam-webster.com/dictionary/intelligence>

Human Intelligence (HI) is manifold, as is Artificial Intelligence (AI). AI has a certain importance in ICT because it is one of the current emerging technologies. It is also an example of recurring technologies, which has reappeared in waves time after time. The characteristics of intelligence have changed over time, as well as the driving force behind it and the opportunities provided by it. Artificial intelligence (AI) is intelligence demonstrated by machines Human Intelligence is displayed by humans and animals that involves *consciousness and emotionality*". This points out something that AI is not able to handle (yet).

The term "Artificial Intelligence" was coined by John McCarthy in 1955. Late in the 1950s, he introduced the programming language Lisp, which provided a means for developing computer programs having the ability for self-modifying code dynamically in run time. Computers were developed to conduct complex calculations; this architecture had to be harnessed to support tasks related to intelligent systems with software level support; Lisp was the first effort in this area. The dynamic modification of the code implemented a primitive *learning capability* in a time (1950s) when the processing power of the computers was low, the availability of data to process was limited and access to it was slow.

AI systems create knowledge from a variety of data elements, based on built-in "intelligence". The DIKW (Data, Information, Knowledge, Wisdom) pyramid in Figure 1 illustrates the cultivation process from data to wisdom (original source [1], several applications in the literature)

- *Data* is conceived of as symbols or signs. To simplify, data is a representation of something, having itself no exact meaning or interpretation.
- Information associates semantics or meaning to the data.
- Knowledge is processed or organized from information based on the rules or algorithms and makes information utilizable in the use context.
- Wisdom is applied knowledge in its use context.

The role of *intelligence* is connected to the cultivation process introduced above. We need intelligence to create meaning for data (-> information) and further to handle the information to create useful knowledge from it. Wisdom provides guidelines for utilizing the knowledge and transfers it to actions and behavioural patterns.

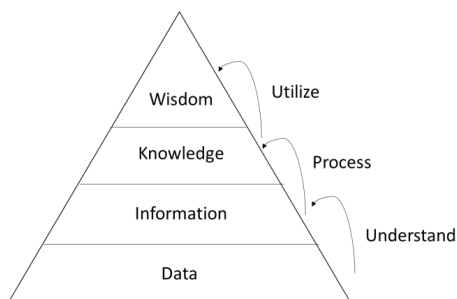


Figure 1. DIKW Pyramid.

The DIKW pyramid can be connected to both human and artificial intelligence. In both cases we need *data* to handle and the (*intelligent*) *process* to refine it into an applicable form in the use context. However, the human way and machine way are different. How different will be discussed in the following sections. Intelligence, in addition to data processing capability, also needs the capability to *learn* – i.e., the ability to change future behaviour based on the experiences or external knowledge available.

The aim of this paper is to find (at least partially) the answer to the following questions:

- What is Intelligence?
- What are the differences between HI and AI?
- What is the intelligence of AI?
- What are the key elements of (Artificial) Intelligence?

The paper is structured in the following way. *Section 2* deals with the characteristics of HI, compared to AI. The focus in *Section 3* is on the intelligence of AI – what are the functionalities for implementing intelligence in these systems? *Section 4* handles two key aspects of intelligence: communication and learning. *Section 5* concludes the paper.

2. Human Intelligence

2.1. The Potential of AI and HI

Human intelligence is far broader than artificial intelligence. According to [14], we can distinguish three kinds of human reasoning systems: brain-based central nervous system with reasoning, the partially autonomous vegetative nerve system with body system control, and the governing survival nerve system for reproduction. AI research is centred around the first system and only concentrates on one type of intelligence: creative or problem-solving intelligence. There are, however, four other kinds of intelligence that cannot (yet) be supported: emotional (or social) intelligence, self-reflection (or spiritual or existential) intelligence, body intelligence as the second human reasoning system, and survival intelligence. These four types are supported and partially governed by the central nerve systems while interacting with it.

Several specific types of creative intelligence can be distinguished that are so far only partially covered and not yet well-supported by AI research:

- linguistic, narrative, or verbal intelligence including metaphorical intelligence,
- musical intelligence,
- abstract intelligence including analytical intelligence, logical-mathematical intelligence, and numerical intelligence,
- visual intelligence,
- practical intelligence including application intelligence, and practical wisdom,
- imaginative intelligence,
- physical-kinaesthetic intelligence, and
- spatial intelligence.

Creative intelligence also covers intuitive intelligence, including crystallized intelligence. This specific type cannot so far be supported at all.

Thesis 1a: *HI can only be properly supported by AI if the specific kind and type of intelligence is well-understood.*

Intelligent human behaviour requires above all a great deal of knowledge about details. Knowledge is to be distinguished from intelligence. Intelligence operators (such as the very large variety of inferential reasoning) allow us to derive new knowledge or at least insight from knowledge, experience, observations, models and intuition. We often hypothesise in 'unknown territory' and develop hypotheses, models or even theories. Detailed knowledge is helpful to 'falsify' hypotheses, i.e. to eliminate them as certainly false, if they contradict facts or experience.

Consciousness, learning, creativity, freedom, communication behaviour are not understood in an algorithmic fashion as yet, e.g. decision making under restrictions, competing simultaneous objectives and uncertainties about the future.

Currently, the AI hype aims at the development of deep learning mechanisms. However, this is only the first step towards emulation of HI. Consciousness, feeling and other topics, such as creativity or will, must be understood before they can be supported by AI.

Thesis 1b: *AI and IT may handle a regular and typical case far better. Anything else is beyond the horizon.*

We have discovered that Turing-based computation is mainly based on an algorithmic treatment by deductive systems. It is an incarnation of the digital. Turing-based computation is limited by the second Rice's theorem [25] that has been extended by many non-computability and undecidability results. It shows that with current technology it is, thus, impossible to build safe and secure software systems. Advanced reasoning mechanisms such as induction and abduction are not yet well-supported. We need a more sophisticated support for reasoning. For instance, logics research has shown that there are true properties in arithmetic that cannot be proven by deduction. The Turing machine model is not at all the only kind of computation. Analogical, plausible and approximative computation is badly covered by Turing computation. Neural networks are so far very simplistic networks. The web of neurons in living systems is far more sophisticated and provides advanced computations that cannot yet be emulated by our machines. There exist many kinds of computation (deterministic, non-deterministic, randomized and quantum), each of which is characterized by a class of computationally equivalent mechanisms. This is also the case of cognitive systems, which are simply specialized non-uniform evolutionary computational systems supplied by the information delivered, thanks to their own sensors and effectors, from their environment. There is evidence that the computational power in human intelligence might be bounded by the Σ_2 level of the arithmetic hierarchy [35], which is far beyond computability.

Thesis 1c: *AI needs an IT that goes far beyond Turing-based computation and logical reasoning by deduction and is not only based on digitalisation.*

2.2. Some Limitations of Artificial Systems

Modern IT systems are wasteful/extravagant in their energy requirements. One may wonder why a complete operating system must be loaded before using the computer as a typewriter. Human systems, on the other hand, are energy-efficient. AI systems are oriented on a luxurious 'featuritis' with many tools that cannot be handled by a single human. Additionally, human reasoning is energy-minimalised, while AI computation, as well as any kind of current computation, is energy-extravagant and recklessly wasteful.

Humans also use other reasoning systems. These systems are parsimonious and energy-optimised, which is not the case for current AI systems. They form some kind of super-organism that are 'living' structures whose ability to survive depends on appropriate coordination of the interaction of individual systems, which are themselves viable, just as a human being is made up of billions of living cells. Self-organisation and self-optimisation usually go through an (evolutionary) process and the process must be 'attractive' enough for the individual elements involved to join together to form a whole.

Thesis 2a: *HI follows the utility value paradigm in a parsimonious manner, whereas AI is mainly based on the marketable value paradigm in a lavish manner.*

We are used nowadays to the 'goodies' of modern AI solutions such as smartphones. At the same time, we do not realise that we are sacrificing the cultural achievements necessary for HI. A good example is the loss of education depth due to the dominance of technical solutions such as Wikipedia or other web services which reduce knowledge to what the monopoly player considers as opportune. Anything else vanishes.

Thesis 2b: *The monopoly game played by the big internet players is oriented to the brainsickness of simple, direct consumers. The AI monopoly is not interested in intelligent consumers. HI is an evolutionary advantage that is based on co-evolution within all kinds of intelligence.*

Living beings and human systems concentrate on the main function. Think, for instance, of the albatross, which can fly thousands of kilometres for several days without interruption or respite. Evolution of the fittest led to such abilities. This development is some kind of meta-evolution rule that uses principles and paradigms that we do not know for technical systems.

Thesis 2c: *Nature and evolution are oriented to the selection of the fittest and best accommodating system. AI system evolution is, at its best, based on meta-models of evolving systems that have to be developed in advance.*

2.3. Opportunities Offered by Artificial Systems

Technical systems are often far better than any human. Compare, for instance, a highly skilled chess player with a computation system that considers all potential moves and that adapts the position evaluation function after assessing the position quality and comparing it to different positions.

Thesis 3a: *AI and HI are different methods of solving certain difficult tasks. AI here does not include the four other kinds of human intelligence. In terms of intelligence in the sense of problem-solving quality in non-trivial subject areas, technology is already further ahead than humans in some areas today.*

AI systems can easily learn regular and smooth functions with a limited number of discontinuity points. Such functions can be considered as routine and approximative solutions. Beside this ‘normality’ we approach the unlearnable. Consider, for instance, a Cauchy function that is neuronally unlearnable. In general, non-learnability applies to all functions that require quite different results or actions for the smallest variations of parameters. So, a neural network of any depth cannot learn ‘ugly’ functions, but it can learn smooth ones [30].

Thesis 3b: *AI systems are better in the support of regular and well-understood application tasks but fail in complex domains which are badly understood, irregular, or insufficiently modelled.*

IT systems as well as AI are instruments that might ease human life whenever this is beneficial. As instruments, they can be used with proper intent, care and a belief in reliability and robustness. Modern life is impossible without all these instruments. However, we do not need instruments in every place, at every time, for every task, for everybody, and in every environment. On the other hand, often we cannot survive without our instruments.

Thesis 3c: *AI systems are instruments that are great in hostile-to-life environments, for the support of activities far beyond human skills, are more efficient and effective than humans and provide services that ease life.*

2.4. Risks and Threats of Artificial Intelligence

IT and AI have a high innovation rate that is far from being easy to understand. Society is unable to control the impact with rules, regulations or laws. Think, for instance, of the threats imposed by micro-trading or AI weapons. The insight into such systems is disappearing. Who can repair a car? Garages are also relying more and more on support instruments. Systems are built with ‘AI inside’. IT systems operate in such a form that Customer Service people have no chance to fully understand their operation. The ‘race between education and technology’ seems to have the artificial system as a winner. Look at the smartphone behaviour that substantially decreases human reasoning, understanding and reasoning, blindness to environment observation, and human social warmth. With the advent of internet and modern TV technology, we identified media competence as an essential skill for everybody (e.g. see [23]). Nowadays, we have to develop literacy and competence in AI systems.

Thesis 4a: *AI education is still in a fledgling stage. Similar to the impact of techniques and technologies on education and various disciplines, we have to develop a proper technology and disciplinary education for AI.*

Modern applications and environments are becoming more complex, advanced and sophisticated. This complexity is often far beyond the comprehension skills of humans. The reaction speed of technical systems beats human abilities. Programmers program whatever is requested without being limited by ethical restrictions. Technology

regulates human behaviour. [8] observes how AI is currently being intentionally misused to destroy democratic society and to nudge any human behaviour up to the current demands of daily life. In the near future, systems combining a group of humans and thousands of 'intelligent' computers are going to transform humanity into hybrid human-technical super-systems that will not be properly controllable or limitable.

***Thesis 4b:** Artificial systems pose a threat to human existence and must be changeable whenever systems start to 'command' humanity.*

Modern systems are making humans lazy. Nowadays, who does not rely on a navigation system rather than preparing in advance for a road trip with maps? Human intelligence regresses without being continuously challenged, since biological systems optimise themselves and thus reduce reasoning if it is not requested. However, evaluation algorithms and computational systems do not follow ethical principles. Computers and programmers have become ethics-free. Political systems are far too slow and too sloppy and cannot handle such challenges. Information overflow and pollution by senseless services make the human being a plaything for the big players (see [27]). The human 'laziness', loss of tacit background knowledge and resulting lack of education are resulting in AI dependence and debility similar to 'illiteracy'. The software crisis is a crisis of proper program and software development culture. Nowadays, we have a data crisis, a (large and embedded) system crisis, an infrastructure crisis and an energy crisis. The next crisis we can expect is an AI crisis. Sophisticated systems such as AI systems operate without feelings, without heart, without compassion, without conscience and without ethics.

***Thesis 4c:** The forthcoming AI crisis can only be tackled if we consider the end from the very beginning and if we develop a proper culture of co-evolving and collaborating symbiotic intelligent systems.*

2.5. The Qualia Question

HI and AI are two kinds of 'intelligence' that have to coexist. The former is not really well-understood, while the latter is human-made for the improvement of life. AI cannot mirror HI. AI is currently mainly based on programs and meta-programs made by humans within the human understanding of that moment when the programs were developed. Programs can be based on meta-programs that change the code according to change scenarios foreseen in advance. Therefore, HI and AI capabilities are different and will continue to be different in the future.

***Thesis 5a:** HI and AI coevolution and symbiosis are encouraging and are a resource for prosperity that should be used wisely. They will give us wings for a better life if properly designed, managed and handled with care and proper wisdom.*

We have determined that HI and AI are two very different kinds of 'intelligence'. Artificial 'knowledge' systems such as Google, Twitter and Wikipedia are also intentionally used for misguidance. They form their own ecosystem that goes beyond human understanding. On the other hand, HI is also based on model-based reasoning. Mental models are something like the 'third eye' in our human, emotional, experience-backed, intuition-guided and hormone-driven digestion of the observed environment. Models are also used for context-backed and culture-based human interaction. Of course, they are also a means for the language-based development of artificial or IT

solutions. HI is properly supported by models at any abstraction level. The mechanisms of model-based reasoning are not understood but we can consider models to be the fourth sphere of our understanding, alongside our understanding and handling of our natural environment, science and technology. The mismatch between model-based reasoning and AI model handling reminds us of the ‘lost in translation’ problem.

Thesis 5b: The mismatch between HI and AI is also caused by human model-based reasoning abilities that go far beyond what can be formally handled and managed.

Many researchers claim that machine intelligence and neural networks are going to cover human capabilities and might replace human reasoning. There are limits and boundaries of current computational approaches that are essentially state transformations and based on Turing-style computing. Human reasoning is far more advanced. Behaviouristic detection of brain activity uses rather naive models and assumptions of how the brain works. Furthermore, AI reasoning systems are bound by our current logic approaches. Logical deduction calculi already cover revisable and non-monotonic derivation. Human inductive, abductive, approximative, plausible and model-based reasoning is far more advanced. They do not have to be language-based. Humans use instruments beyond languages.

Thesis 5c: Neural networks used in AI are based on the neuron models developed in the 1950s. The next generation of neuro machines needs decades of advanced research in order to reach the maturity of a brain-based central nervous system with reasoning. The other human reasoning systems might be understood in the next century.

HI is one kind of natural intelligence. There is no reason why HI should be the only form of intelligence of living beings. Furthermore, the human body consists of many synergy-stimulated systems where human cells are the most essential part of the system. The human cell system cannot survive without the other systems. The other systems are currently very poorly understood. The interaction of such systems is a ‘black hole’ in medicine.

3. Artificial Intelligence and Intelligence

3.1. Evolution of AI and the Role of Enabling Technologies

AI is one of the technologies that has a *recurring appearance* in the role of emerging technologies. Emerging technologies are “*technologies that are perceived as capable of changing the status quo*” [10]. Emerging technologies have a radical novelty and potential for fast growth and impact, but under uncertainty, their progress may sometimes be different than expected (the hype phenomenon).

The evolution of AI has been highly dependent on the progress of *enabling technologies* (Jaakkola et al. 2017):

- *VLSI Technology – Processing Capacity* doubles every 18 months and the *Memory capacity* of computers every 15 months.
- *Mass memory capacity* (magnetic devices) increases by a factor of ten every decade, i.e. it doubles every 18 months.

- *Data transmission* capacity speed doubles every 20 months. This dimension is slightly complicated because of the heterogeneity of transmission channels and their role within the whole. Data transmission capacity is the key issue in the adoption of distributed solutions in information processing.

We agree that the forecasts above are not scientifically exact but they provide a rough trend about the progress in the key technologies related to AI. We have extrapolated the progress backwards from the late 1960s (invention of microprocessors and LSI) to the era of early computers (1950s). The progress is summarised in Table 1.

Table 1: Progress of AI-related enabling technologies.

Double capacity in months (m)	1955	1975	1990	2020	2030
Computing 18m	1	2^{13} 1 → $(2^{10}) = 1$	2^{23} → $(2^{19}) = 1$	2^{42} → $(2^{19}) = 1$	2^{50} → (157)
Memory 15m	1	2^{15} 1 → $(2^{12}) = 1$	2^{28} → $(2^{23}) = 1$	2^{51} → $(2^{23}) = 1$	2^{60} → (445)
Mass memory 18m	1	2^{13} 1 → $(2^{10}) = 1$	2^{23} → $(2^{10}) = 1$	2^{42} → $(2^{19}) = 1$	2^{50} → (157)
Transmission 20m	1	2^{12} 1 → $(2^9) = 1$	2^{21} → $(2^9) = 1$	2^{38} → $(2^{17}) = 1$	2^{49} → (97)

The years selected in the table represent the different eras of AI. In the next decade progress will continue and provide new means for the future of AI: 157-fold computing power, 445-fold memory capacity, 157-fold mass memory capacity and 97-fold higher data transmission capacity compared to the situation today.

People have been fascinated by AI throughout its existence. That is why new approaches are born cyclically in a kind of “*reincarnation cycle*”. These cycles can be explained (at least) by the following three factors:

- *Demand pull*: there is a continuous (hidden) demand for new (more) intelligent applications. People expect more and more intelligent applications to help in their daily life and to improve the productivity of their work.
- *Technology gap*: the performance of the existing technology limits the opportunities to implement applications that the users would like to have.
- *Technology push*: when technology allows, the demand pull “activates” the new type of applications and a new cycle starts.

Additional aspects of importance in the progress of AI applications come from general trends observable in ICT infrastructure: the transfer to mobile and wireless, distributed processing and data management, the transfer towards more complex user interface technologies, the growing interoperability of applications, the embedding of (AI) solutions, the growing role of “robotics” (IoT), etc.

3.2. Reincarnation Cycles of Artificial Intelligence

Figure 2 provides a general overview of the four reincarnation cycles of AI. To be exact, there should also be an additional one – the era of *Ancient AI*. AI has roots in antiquity in the form of myths, stories and rumours of artificial beings endowed with intelligence or consciousness by master craftsmen. However, the *Ancient AI* (Cycle 0) left the ideas at a theoretical and story level. As discussed at the beginning of this paper, real AI is based on the ability to cultivate (currently masses of) data for the users’ wisdom and help them to fulfil the goal set for (intelligent) data processing. This has been enabled by computers. However, a lot of ancient philosophical foundations (theories about the human mind and human way of thinking) are useful as a theoretical foundation in current AI research.

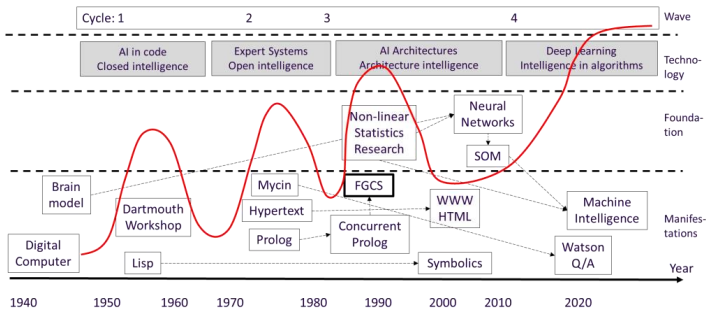


Figure 2. The four cycles of AI.

The four waves (cycles, eras) provide a view of the spread of AI in a new kind of application logic and ability to apply a new kind of intelligence in the systems developed. Typical for the wave-based approach is that every wave, as an emergent technology, has slowly grown from an *embryonic phase* at the beginning, followed by a phase of fast growth until reaching *the peak* (highest importance) and turning slowly to the *decline phase*, which makes it “part of the normal” without meaningful innovative power. Typically, the beginning of a new wave is based on new technology replacing an old one and taking its role as an emergent technology; this leads to the sequence of waves as in Figure 2. This aspect of technology analysis is handled in several papers by the authors, see e.g. [11; 12].

The First Wave – AI in program code, from the 1950s to 1970s

The term “Artificial Intelligence” was introduced by Professor John McCarthy in 1955. He was a key person in organizing the Dartmouth workshop. This was a summer school, which provided a brainstorming forum for a dozen of scientists about the novel technology, the research topic of “thinking machines”. The workshop proposal [16] was to make a study “of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be

made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves'.

Later in the 1950s, McCarthy introduced *Lisp language* (acronym of LISt Processor), which became the first tool to develop “real” AI applications. Lisp is based on the Lambda calculus and allows a code to modify itself in runtime. This creates a simple learning capability for the applications. Lisp has, since its birth, many dialect implementations, which have followed the general trends of improvements in programming languages.

Another remarkable finding in AI programming was the *logic programming language* Prolog, developed by Alain Colmerauer and Philippe Roussel. Prolog is based on first-order logic, a formal logic, in which the program is declared in the form of relations “if-this-then-this” (declarative programming). The execution of relations can create new relations, which for its part creates facilities for *learning* in the applications. The program is executed by applying relations (*reasoning*) in a parallel, instead of (typical for that era) a sequential, manner. Similarly to Lisp, Prolog has many different manifestations: one worthy of mention here is *Concurrent Prolog* used in the Japanese Fifth Generation Computer System Project (third wave in this paper) as a basis for the computer architecture.

The intelligence related to the first wave is that knowledge to solve the problem is “hard-coded” in the program and known by the programmer only (closed intelligence). For the user of the application, it is fairly difficult to see or understand the logics of the solution.

The Second Wave – Expert Systems, from the 1970s to 1980s

An expert system (ES) is a computer application that has “built-in” intelligence – knowledge in the form of a *rule base*. By definition, the expert system is a computer system emulating the *decision-making* ability of a human expert. Instead of programming language, the end-user defines his problems for the system by using the structures of the *problem-specific* user interface. Problem-solving is built-in to the implementation of the ES, which may be partially documented and understood by the end-user. This makes the built-in intelligence at least partially visible and understandable to the user (i.e. open intelligence).

The rise of ESs started in the 1970s. Two mainstream implementation technologies were rule-based and frame-based systems. In the former, the knowledge was presented to the system in the form of rules; the latter was based on the structured approach, in which the solution was found by matching the problem to the frames in the system’s frame base.

The first expert system was introduced by *Edward Feigenbaum* - the “father of expert systems” - from Stanford University. The *Mycin* system was developed for the diagnostics of infectious diseases to recommend medical treatment. The system was written in Lisp and had a knowledge base of 600 rules. Another early-stage ES developed at Stanford was *Dendral*, developed for hypothesis formation and discovery in science; it was first used to help organic chemists in identifying unknown organic molecules.

The currently best-known expert system is *IBM Watson*. It is a question-answering system capable of answering questions posed in *natural language* and used in a variety of application areas. Its knowledge resources are available via APIs to third parties to develop their own applications. Watson is an example of the rebirth of an idea connected to the ES outside their main era. The development effort can be timed as starting in the 2000s and is continuing, naturally applying a variety of technologies available today (compared to the situation in the 1970s).

In Figure 2, we have included Hypertext (Hypermedia) and WWW as technologies closely related to AI. These are of crucial importance in the foundations of current computing and information/knowledge management in the form of linked content structures; in a way this represents built-in structural intelligence in documents and document structures.

The Third Wave – AI in architectures, from the 1980s to 1990s

The traditional computers were designed to execute algorithmic programs in a sequential manner. Some trials about implementing parallel processing and parallelization of software were made as early as the 1970s; the supercomputers of the time were based on multi-processor architecture, in which mainly arithmetic operations could be executed in parallel. This allowed complex scientific calculations but was not useful for tasks executed by AI systems.

Knowledge engineering and AI systems are based on *reasoning and inference processing*, rather than algorithmic data processing. Rule-based knowledge management – like in Prolog – typically has no execution order for the rules and because of that is possible to parallelize. From the late 1960s, (V)LSI technology enjoyed rapid progress. Back in the 1970s, technology to develop *Application Specific Integrated Circuits* (ASIC; even processors) had provided the means for developing *specific computer architecture* to allow effective application-specific computing. This opened up the space for a new era in AI – implementing inference and reasoning support *directly in the computer architecture* to make processing in such tasks more effective.

The most famous activity in this area was the Japanese nationwide project called “*New (Fifth) Generation Computer System*” (FGCS) coordinated by the *Institute for New Generation Computer Technology (ICOT)*. This ten-year project was open for international collaboration and was focused on computer architectures, software, and (intelligent) applications like speech processing, natural language processing and language translation. The computer architecture was based on the *Concurrent Prolog* developed by Ehud Shapiro. On the architectural side, both computers for personal use (PSI – Personal Sequential Inference Machine) and massive processing (PIM – Parallel Inference Machine) were developed. The latter implemented parallel processing on a massive scale, having thousands of processors.

The direct commercial success of the project ultimately remained insignificant – inference machines did not remain a part of mainstream computing. Regardless of this, Japan made a giant step in two areas: *software engineering* and *intelligent applications* (image processing, speech recognition, natural language processing and online language translation) were developed in the project. The deep *knowledge of computer architectures* that the Japanese already had before the project and its consolidation

were obvious, too. However, this was evidence of the opportunity to build *intelligence in architecture (architecture intelligence)* as a new era in the history of AI.

At the same time as the Japanese effort, MIT was working towards the development of Lisp-based computer architecture. The commercial work was transferred to the MIT spin-off company, Symbolics Inc., which produced Symbolics computers for a while in the 1980s. These did not become a commercial success either but was evidence of the opportunity to implement intelligent architectures that provided effective processing capacity for knowledge engineering tasks.

Why did this kind of specialized architecture ultimately not stay on the market? We refer the reader to Table 1 which provides a view of the progress of key factors. In the 1980s, the slow processing of data in AI systems was a bottleneck. Fast progress in the enabling technologies changed the situation: instead of maintaining and further developing the specialized “niche architectures,” growth of computing power finally made the use of software-based solutions as effective as specialized implementations. In addition, AI systems are quite often not independent but a part of complex interacting systems of systems, implemented by mainstream tools.

The Fourth Wave – Learning-based AI, from the 2000s and continuing

Intelligent systems are based on a system’s ability to adapt (change behaviour, react to feedback) and to learn about the situation in which it is used. Learning might be first taught and then self-learning takes place during the use of the system. The traditional approach (in the three waves discussed above) is to build the learning capability (intelligence) into the code, rule base or architecture.

The current wave of AI is based on the effective use of learning algorithms. We have listed some concepts related to this in Figure 2: neural networks, self-organizing maps and deep learning. The *neural network* builds a model that resembles the structure processing of a human brain. It uses “what-if” based rules and it is taught (*supervised learning*) by means of examples. The network learns the non-linear dependencies between variables. An improved version of a neural network is the self-organizing map (SOM) that is based on *unsupervised learning*. A multidimensional input (learning) data set is organised into layered relationships, which are represented as a low-dimensional map. This can be used as an abstraction of the real data space. *Deep learning* theory is based on the *independent* learning of masses of data. The learning algorithms are based on the use of nonlinear statistics.

In this case, *intelligence is built in algorithms*, which themselves are application-independent and implement the learning capability of the system. Powerful learning algorithms and masses of data replace complex application-specific intelligent algorithms. For example, concerning its Translate application, Google reports that the earlier translation algorithm of 500,000 LOC was replaced by a learning algorithm of 500 LOC (and data). An additional benefit is the learning algorithms’ flexibility to learn new facts during use.

3.3. The intelligence of Artificial Intelligence – Analysis of the cycles

We have introduced four different approaches to Artificial Intelligence in the context of their birth:

- Intelligence in the software code. Closed intelligence, in which the details were known only by the programmer.
- Intelligence in the rules and the “knowledge engine” logics. Operational logics of the system are open to the user.
- Intelligence in the architecture. Intelligence is transferred to the computer architecture. Direct support for the efficiency of the applications.
- Intelligence in the (learning) algorithms (and data). Systems based on human learning. Algorithms are not known by the end-users. Key aspect is the *quality of the data*.

The cycles as a sequence are described in Figure 2. The idea is to view the importance of each AI cycle in the time it was born and its role as an emergent technology – i.e. its innovation power. In reality, all of these technologies are still valid and in active use in a wide variety of applications (Figure 3).

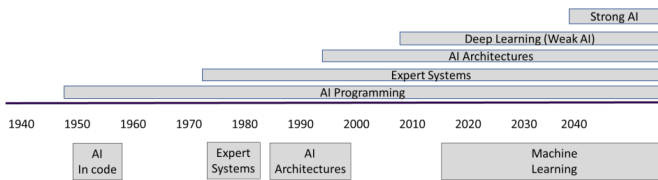


Figure 3. AI application space – the growth of AI applications.

The current AI wave is known as *weak (or narrow) AI*. The applications are task-oriented, in which the knowledge is not transferrable to a new application context. Learning algorithms themselves are general. Weak AI does not have its “own sense” related to the data it handles, nor its own will about how it should be handled. We are moving towards *strong AI*. It can handle facts and their relationships and has the characteristics of human beings, like *common sense*, but it does not yet have its own will, rather a kind of understanding of its surroundings.

4. Learning and Intelligence – Are computers like humans?

4.1. Creating new Human-like Texts

Currently we see a growing number of new AI applications based on the use of natural, human language in various industries, including banking, recruitment, healthcare, agriculture, transport, etc. Advances in AI in creating human-like communication and replicating natural language patterns used by humans are based on large language corpora – a collection of human-produced texts in various encodings, first of all written text, but also spoken, signed, etc.

Large corpora with billions of words are used to create text models, i.e. algorithms, which can parse input text and ‘understand’ it, in other words, answer some simple questions concerning the input. The abilities of corpora are often demonstrated with

text generation – the text model continues a given seed (start of a story) and produces believable output, i.e. new text which looks like it has been created by a human.

The main problem of language creation is prediction of the next word (or character). A text/corpus model is a collection of conditional probabilities of the next word in the text.

Suppose we already have a sequence of words:

$$w_1, w_2, \dots, w_{i-2}, w_{i-1}$$

The next word w_i could not be arbitrary as it depends on the preceding words. The next word could be guessed by maximizing the relative probability (the Bayesian inference [13]):

$$\max(w_i \in V) P(w_i | w_1, \dots, w_{i-1})$$

Here $P(w_i | w_1, \dots, w_{i-1})$ is the conditional probability that after words w_1, \dots, w_{i-1} follows the word w_i . In practice, probabilities are estimated from real word frequencies, i.e. the relative probability of the word 'students' after the previous words 'all our' could be calculated from the frequencies of use of these words in a (large) corpus of text where these words were already used:

$$P(w_i | w_1, \dots, w_{i-1}) \approx \frac{Fr(w_1, \dots, w_{i-1}, w_i)}{Fr(w_1, \dots, w_i)}$$

The probability of the whole phrase is the product of probabilities (the naive Bayes rule), i.e. the probability of the beginning phrase $P(w_1, \dots, w_{i-1})$ and the conditional probability that it is followed by word w_i :

$$P(w_1, \dots, w_i) = P(w_1, \dots, w_{i-1}) P(w_i | w_1, \dots, w_{i-1}) \quad (*)$$

Human language is often considered as a process with limited memory (the Markov process) – assuming that the meaning of the next word depends only on a limited number of preceding words. This is generally not true; we often expect that the reader/listener already knows the meaning of many words which have been used. However, applying the Markov process assumption that the '*probability of the word depends only on a few numbers of previous words*' simplifies programs and this is used in natural language processing (NLP) everywhere. Thus, for prediction of the next word only a sequence of fixed length k is used (the naive Bayes assumption, i.e. k is the length of the sliding cutout of the last k words) and the search goal is

$$\arg \max(w_i \in V) P(w_i | w_{i-k}, \dots, w_{i-1})$$

To simplify the notations, shift $i - k \rightarrow 1$, thus we are looking for

$$P(w_1, \dots, w_i) \approx P(w_1, \dots, w_{i-1}) P(w_i | w_1, \dots, w_{i-1})$$

The probability of the first phrase could be expressed the same way or those words are given as a seed.

In practice (to speed up calculations), the last formula is simplified even more. Using the naive Bayes conditional independence assumption that the probabilities $P(w_j | w_{j-1})$ are independent, in language models often only binary probabilities are used (a very rough assumption), thus:

$$P(w_1, \dots, w_n) = P(w_1) \prod_{i=2}^n P(w_{i-1} | w_i)$$

The argmax formula (*) presented above can be used to create new texts based on probabilities occurring in the text corpus – give some words w_1, \dots, w_{i-1} as a seed and find a word w_i which maximizes probability $P(w_i | w_1, \dots, w_{i-1})$, then shift the 'action window' one step to the right and repeat the process starting with the sequence w_2, \dots, w_{i-1}, w_i .

To obtain the influence from words used farther back in the text, in NLP words are used in contexts of lengths > 50 (words or characters – text processing has often also done this on the level of characters), but this 'far influence' does not come from n-grams.

Longer contexts allow prediction of the next words, e.g. in the above text, if we already see the words '*are using several programming*' then the next word could/should be '*languages*'; if we see '*are popular social*', then the next word should be '*networks*'. The longer the context, the better the prediction, but the prediction is always that the next word would be some word which has already followed it in the corpus – there would never be pairs of words which had not already occurred in this order in the corpus. A perfect parrot.

Modern digital methods can substantially increase the influence of previous context on the next word, but for this, everything is converted to digital and instead of n-grams (exact fragments of input text) functions are used, which calculate inferences from 'farther' contexts.

For calculations, words are first replaced by their numeric code in the vocabulary list and then are used in two functions (inverse to each other):

- for a list of words (the bag-of-words) find a word which can probably occur together with these words,
- for a word find its contexts, i.e. words which with high probability occur near this word (skip-gram).

In both cases for every occurrence of a word, a vector of probabilities of nearby words is calculated (the `word2vect`) – words are represented by their contexts.

Thus, words are considered as elements of vectors of probabilities of their context words. Mapping from words to vectors is called 'word embedding'. The dimensions of these vectors can be rather large, e.g. the dimensions of the Stanford collection of pre-trained word embeddings [22] vary from 50 to 600 (may be more, but always fixed).

The word vectors are not unique – they depend on the text corpus and even with the same text corpus different NLP packages (different methods for creating a text model) produce (somewhat) different results.

Representing words as vectors with real-valued coordinates allows us to calculate from the cosine's product of their vectors' distance (i.e. similarity) between words. The vocabulary of the whole corpus becomes a 'cloud' of dots in a multidimensional space. Methods to create word vectors and to use them for new text creation belong to the research area of machine learning (ML).

4.2. What is ML?

Learning is a process to improve, change learners' behavior so that learners can better respond to their environment, and achieve their tasks better. Computers are deterministic devices whose behavior never changes – if it does, then the computer is severely damaged. When the same text corpus is re-used (with the same model structure), a computer creates the same model and if it is used for text creation (with the same seed) the same text will appear. From here it follows that the term and acronym 'machine learning, ML' is a misuse of the word 'learning'.

In order to understand each other, we should have some common understanding of the terms that we use, but there is lot of dissension over the use of the terms 'information', 'knowledge' and 'learning'. Would you say that Newton learned the Law of Gravity or that Einstein learned the Theory of Relativity? They did not 'learn' those laws, they discovered them by setting up totally new frames of thought, performing experiments that nobody had thought of before. They first created a new mental approach, a new framework, then observed, collected data in this framework and then generalized their observations data as a new Law of Nature – something, what is impossible in ML.

When composing a new text, only a sequence of fixed length k is used for prediction of the next word (the naive Bayes assumption, i.e. k is the length of the sliding cutout of the last k words) of already created words and the search goal is:

$$\arg \max(w_i \in V) P(w_i | w_{i-k}, \dots, w_{i-1})$$

When humans speak/write, the next word also depends on all the already produced words, i.e. they use a procedure similar to rule (*) that computers use, but the process begins in their consciousness (denoted with " "):

$$\arg \max(w_i \in V) P(w_i | " , w_1, \dots, w_{i-1}) \quad (**)$$

The rule that computers use is only an approximation of the tail of the human procedure. The premise w_{i-k}, \dots, w_{i-1} used in rule conditional probability is only a small tail of the premise " , w_1, \dots, w_{i-1} used by humans, thus the consequence w_i is less exact (i.e. its probability is smaller) and thus also the entropy (information content) of the whole produced phrase is smaller.

Word vectors (however long) cannot express the meanings of words in the way we know them – we change them constantly. Depending on our mood, previous events, time of year/day etc. we can use the same words with quite opposite meanings: "John,

"you did well!" may mean ("Good, we expected you to fail") or ("You failed, we expected you to win!"). Current NLP research is trying to analyze sentiments (positive or negative) and some researchers even try to analyze more feelings [15; 2; 33]. However, this (and many problems connected with memory) are difficult forms of verbal expression and difficult to reproduce – computers do not have feelings (as yet) and do not know what to remember - is the word *Hamburg* the name of a student, bird, virus or programming language and should it be stored in the memory?

Thus, here lies the main, most important difference between machine learning (ML) and human learning (HL). Machine learning in NLP approximates the tail (visible) part of human communication. Both are based on memory search, but in ML the depth of the search is fixed in the ML program, but with HL we do not even know how deep can the search go – we can remember some episodes even from our pre-school life.

4.3. Disentangling the hype from reality

When speaking about neural algorithms, 'deep' learning, data science etc., it is often mentioned that none of the methods used here are mathematically proven. For many practical problems – how many 'hidden' layers, how many nodes in each layer, what kind of activation function to use, etc., there exist only some suggestions [7]. The design decisions are stated, not explained [9]:

- Input layer will have 784 nodes.
- Hidden layer 1: we have decided to reduce the number of nodes from 784 in the input layer to 128 nodes.
- Hidden layer 2: we have decided to go with 64 nodes.
- Output layer: we are reducing the 64 nodes to a total of 10 nodes.

Many approaches which have become nearly standard do not have any reasonable explanation. For instance, use of the sigmoid function as an activation function:

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

This function is computationally expensive – it uses power and division and can produce values close to zero, but its use is explained with "*The main reason why we use the sigmoid function is because it exists between (0 to 1)*"[29] – any function can be normalized to have values between any two constants. ML has an overabundance of 'ad hoc' methods and quasi-mysterious ways of producing 'deep' inference models - you start with a TensorFlow model and then a follow-on screen, how the main parameter – loss – first decreases, but then increases, i.e. the model is overfitting and should be re-organized:

```
58/987 [>.....] - ETA: 2:01 - loss: 3.9472
59/987 [>.....] - ETA: 2:00 - loss: 3.9366
60/987 [>.....] - ETA: 2:00 - loss: 3.9261
...
808/987 [=====>.....] - ETA: 23s - loss: 0.9714
809/987 [=====>.....] - ETA: 23s - loss: 0.9715
...
984/987 [=====>.] - ETA: 0s - loss: 0.973
```

985/987 [======>.] - ETA: 0s - loss: 0.9738
986/987 [======>.] - ETA: 0s - loss: 0.9738

Humans do not like such unexplained 'black magic' and thus establishing trust in NLP, ML and AI technologies may be one of the most important skills data scientists have. This has created a new research direction: explainable AI (XAI) – i.e. developing tools and frameworks to help you understand and interpret predictions made by your machine learning models [36; 32]. Yet XAI is trying to explain, not to prove anything.

To 'prove' ML or NLP is in principle not possible. Proving something (in mathematics) is possible only if we have a formal system in which all our statements can be formalized. Machine learning extracts the information that an input random variable $X \in X$ contains about an output random variable $Y \in Y$, thus if we have their joint distribution $P(X, Y)$ and a precise (i.e. mathematical) definition of input-output structures X, Y - we know what the properties are and all possible values of probabilistic variables X, Y .

Of the many (mathematical) results about neural nets, the key ones are the universal approximation theorems [36], which state that a neural net can approximate (i.e. calculate with whatever precision) any continuous (the graph is smooth continuous line) function $X \rightarrow Y$ (for a simple explanation see e.g. [4; 19]).

However, these theorems rely on the precise mathematical properties of the inputs-outputs. For NLP this means that we should have a formal description of human language, but a formal description for any human language is impossible in principle.

It is impossible to check whether there is an understanding common to all speakers of even our own mother language, in that we all always understand all our utterances in the same way – if there were, most of our social systems, courts, laws, advocates etc. could be cancelled and replaced with computers (and humans would be obsolete and superfluous). In fact, the whole of progress would vanish – progress happens if somebody interprets established facts and common beliefs in a different way.

Human languages change constantly in just the same ways as the whole of humankind – the next generations will constantly renew our language. For instance, more than 1400 new words were added to the Oxford English Dictionary in March 2021 [20]; another source reveals, that a new word is added to the English language every 98 minutes [5] and every living, natural language behaves in the same way.

All neural nets are inference algorithms, which can find consequences from given facts, but cannot create new facts which do not follow from given data. Mathematicians have long ago devised a precise definition for 'provable'. All ML algorithms are inferences on a given set of facts (called a database or text corpora). An inference algorithm $|=$ (e.g. TensorFlow) on a database or text corpora KB is provable if, for every sentence α inferred from KB, i.e.

$$KB \models \alpha$$

all interpretations in which sentences in the KB are true, α is also true (see any textbook on formal logic, e.g. [6]).

The interpreters of text are humans. We all know from our everyday experience that the database KB need not be very big in order for different meanings to appear and that the bigger the database KB, the more different interpretations there will be, i.e. for many people KB does not contain only true statements and the truth of inferred sentences α is (generally) a very rare event – nearly equally as rare as a production of Shakespeare's opuses using the knowledge engine.

Thus 'proving' NLP text models or ML inference algorithms is impossible in principle. NLP text models can make everything looking like the truth. One of the (currently) biggest models, the GPT-2, accepted the following fable [21]:

"In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley in the Andes. Even more surprising to the researchers was the fact that the unicorns spoke perfect English."

The GPT-2 system continued the fable to look like a true story from some news agency:

"The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science..."

If a program can fluently explain *four-horned silver-white English-speaking unicorns* then it can certainly also prove that the *Earth is flat, vaccines and 5G are evil* etc. - a perfect creator of 'fake news', but is this news 'fake' (or in modern terms 'alternative truth') just because it is not provable? The system can also explain contradicting statements, e.g. that those unicorns had five horns, could fly and were speaking Putonghua, the Mandarin dialect of Beijing, but some were also fluent in Russian.

The situation with 'proving' ML is rather similar to the claim that "Computers today are Turing-complete, i.e. they can represent any computable algorithm" [17, p.4]. This seemingly very forceful statement is non-provable – the concept 'computable algorithm' (or in more mathematical terms 'effectively calculable function') cannot be formalized, and thus cannot be used in mathematical formalized proof – any formalization will be subjective, will cover only the forms of algorithm that the author considered meaningful but nobody can know whether some other forms of algorithms exist.

Natural language may just be an example of an algorithm which cannot be formalized. NLP assumes that in a (long enough) sequence of words the next word is predictable.

Thus if the sequence w_1, \dots, w_{i-1}, w_i can occur in a natural language (accepted by native language users), then a function/algorithm exists for producing the next word:

$$f(w_1, \dots, w_{i-1}) = w_i$$

However, numerous efforts and continuing research have not yet managed to produce such a function/algorithm – they may produce an explanation for '*silver-white four-horned English-speaking unicorns*' or Chomsky's famous phrase "*Colorless green ideas sleep furiously*" but cannot prove the truth of these statements.

Human messages are created by our consciousness and feelings. Trying to make the source of meaningful messages the messages themselves (text corpora, however big) is exactly what Baron Munchausen did - pulling himself out of a swamp by his own hair.

Shouting: "NLP Cracked Transfer Learning" is adding the horse to the load (the baron did this – he was riding).

A neural net is a multi-variable function from input space to output space. All proven statements about neural nets assume a precise formal description of input-output spaces, otherwise we could not prove anything. The universal approximation theorems (for a popular introduction see e.g. [19]) have established that neural nets can approximate any continuous functions between Euclidean spaces; there are also variations for non-Euclidean spaces, algorithmically-generated function spaces, etc. In practice, some aspects of these theorems are often overlooked.

Firstly, the theorems do not say how to organize an approximating neural net – how many layers, how many units in every layer, what kind of activation function to use – the best values have to be found with practical experiments.

Secondly, many problems are not continuous functions, for example all classification problems, image recognition problems, etc. For non-continuous problems, a neural net may not converge at all and the researcher has to start experimenting.

There are no mathematical results of the type:

$$ML: Input_text \rightarrow Output_text$$

The *Input_text*, *Output_text* are not mathematical structures, every natural language model (e.g. neural net) creates (i.e. makes a mathematical approximation) in its own way. Ambiguity and misunderstanding have created a lot of frustration among data scientists [1; 34; 18]. This has both deeper causes and also deeper consequences.

According to research [31]: "Sixty-six percent of data scientists describe themselves as self-taught", thus most probably they have not learned the (elementary) facts about proofs discussed in the previous paragraph. As a consequence, they are uncertain about the meaning and value of their activities (see e.g. [26]) and every week are "spending 1-2 hours a week looking for a new job" [31].

5. Conclusion

AI, as well as any kind of computation, has its own merits. HI is oriented to the needs and challenges with a human face. It also supports human societies. However, there are many tasks that cannot be handled by humans and living beings. For example, [24] compares abilities to fly. As already mentioned, consider the albatross that can stay in the air for days and cover thousands of kilometres. If we need to carry hundreds of tons from Europe to Japan, then artificial devices such as planes can manage this task much better. Routine, heavy or complex tasks can be handled better by artificial devices.

Our artificial systems do not really produce anything new, in reality. They bring, however, a great, purely practical improvements in everyday life. They increase speed, effectivity and performance for everybody who has access to them. They enable a comfortable life for many people. Whether we call them 'intelligent' is a matter of the definition of intelligence.

Looking to the future, we need an approach to limit the abuse and maldevelopment of technical systems. Already the scare with the atomic bomb has brought us the insight that any kind of weapon – whether it be chemical or biological - must be wisely limited with a worldwide moratorium, lest humanity be brought to the brink of destruction by power-obsessed elites. This is just as true for AI systems today. We also need containment against such abuses of AI.

The history of AI proves that humans are attracted by human-like computer applications. The superiority of AI is based on its ability to handle *big amounts of data* mined from a *variety of distributed sources* at an extremely *high speed*. Ultimately, this superiority is based on “brute force” controlled by *algorithms*. In a way, even HI is based on algorithms. Researchers have tried to adopt these algorithms and apply them in AI. In the current wave of AI (data-driven, context-dependent), *learning* is the key element, based on training with masses of data. This is a good start towards HI, but still a lot is missing: human sense, human criticism, emotions, sensitivity, psychological and physiological reaction, and human ethics are examples of the missing elements. In connection with *semantic computing* even these elements are present⁴; machines can calculate and *simulate* them.

Current AI (weak, narrow) is still context-dependent and not transferrable to new application areas. The next step leads towards human-like strong (general) AI. An interesting topic to think about is computer-brain integration. Gartner⁵ defines it as “a type of *user interface*, whereby the user voluntarily generates distinct brain patterns that are interpreted by the computer as commands to control an application or device”. In the Gartner’s hype slope of Emerging Technologies 2020⁶ it is located in the first segment of five (Innovation trigger) but its appearance is expected during the next decade. The same segment in the curve includes a lot of promises in the AI area: self-supervised learning, adaptive machine learning, composite AI (variety of AI techniques combined) and generative AI (ability to create new content).

Can AI beat HI some day? In AI research there are still a lot of new developments to look forward to, but human intelligence is unattainable. AI is based on algorithms, and algorithms are developed by humans, not by other machines. Algorithms mirror the values and selections of the developers. The largest current neural networks include some 20 million neurons. To put this in perspective, in spite of this capacity, the networks’ intelligence is still lower than the intelligence of insects that have learned to live successfully in their environment with only hundreds of thousands of neurons.

References

⁴ The Mathematical Model of Meaning (MMM) models these elements, which are covered by the Japanese word “*kansai*”. See T. Kitagawa and Y. Kiyoki, “A mathematical model of meaning and its application to multidatabase systems,” Proceedings of RIDE-IMS ‘93: Third International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems, 1993, pp. 130-135, doi: 10.1109/RIDE.1993.281933.

⁵ <https://www.gartner.com/en/information-technology/glossary/computer-brain-interface>.

⁶ <https://www.gartner.com.au/en/articles/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020>.

- [1] Ackoff, Russell L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis* 16, pp. 3–9. Retrieved April 24th, 2021 from <http://www-public.imtbs-tsp.eu/~gibson/Teaching/Teaching-ReadingMaterial/Ackoff89.pdf>.
- [2] Brooks-Bartlett J. (2021). Here's why so many data scientists are leaving their jobs. Retrieved May 3rd 2021 from <https://towardsdatascience.com/why-so-many-data-scientists-are-leaving-their-jobs-a1f0329d7ea4>.
- [3] Explained AI (2021). Understand AI output and build trust. Retrieved May 3rd 2021 from <https://cloud.google.com/explainable-ai>.
- [4] Fortuner B. (2021). Can neural networks solve any problem? Retrieved May 3rd 2021 from <https://towardsdatascience.com/can-neural-networks-really-learn-any-function-65e106617fc6>.
- [5] Global Language Monitor (2021). Number of Words in the English Language. Retrieved May 3rd 2021 from <https://languagemonitor.com/number-of-words-in-english/no-of-words/>.
- [6] Hauskrecht M. (2021). Propositional logic. Inferences. Retrieved May 3rd 2021 from <https://people.cs.pitt.edu/~milos/courses/cs1571-Fall2013/Lectures/Class12.pdf>.
- [7] Heaton Research (2017). The Number of Hidden Layers. Retrieved May 3rd 2021 from <https://www.heatonresearch.com/2017/06/01/hidden-layers.html>.
- [8] Hesse, Wolfgang (2020). Das Zerstörungspotential von Big Data und Künstlicher Intelligenz für die Demokratie (The destructive potential of Big Data and Artificial Intelligence for democracy). *Informatik Spektrum* 43(3), 339-346.
- [9] IBM Developer (2021). Neural Networks from Scratch. Retrieved May 3rd 2021 from <https://developer.ibm.com/technologies/artificial-intelligence/articles/neural-networks-from-scratch/>.
- [10] Jaakkola, H., Henno, J., Thalheim, B. and Mäkelä, J. (2017). The educators' telescope to the future of technology. In P. Biljanović (Ed.), *MIPRO 2017 - Proceedings of the 40th Jubilee International Convention*. May 22-26 2017, Opatija, Croatia. (pp. 766-771). Opatija, Croatia: Mipro and IEEE.
- [11] Jaakkola, H., Brumen, B., Henno, J., and Mäkelä, J. (2013). Are We Trendy? In P. Biljanović (Ed.), *36th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2013 - Proceedings (Vol. ISBN 978-953-233-074-8; ISBN 978-953-233-075-5, pp. 649-657)*. Opatija: Mipro and IEEE.
- [12] Jaakkola, H., Henno, J., & Mäkelä, J. (2017). Technology and the Reincarnation Cycles of Software. In Z. Budimac (Ed.), *SQAMIA 2017 - Proceedings of the Sixth Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications*. Belgrade, Serbia, September 11-13, 2017. (Vol. Vol-1938, pp. 5:1-10). Belgrade, Serbia: CEUR Workshop Proceedings.
- [13] Jurafsky D., Martin J.H. (2020). *Speech and Language Processing*. Retrieved May 3rd 2021 from <https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf>.
- [14] Kandel, Eric R. (2007). In search of memory: The emergence of a new science of mind. WW Norton & Company.
- [15] Kerkeni L., Serrestou Y., Mbarki M., Raoof K., Mahjoub M.A. and Cleder C. (2019). Automatic Speech Emotion Recognition Using Machine Learning. DOI: 10.5772/intechopen.84856
- [16] McCarthy J., Minsky M.L., Rochester N., Shannon C.E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Retrieved April 25th 2021 from <http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf>.
- [17] Mikolov T. (2021). Statistical Language Model Based on Neural Networks. Retrieved May 3rd 2021 from <https://www.fit.vut.cz/study/phd-thesis/283/en>
- [18] Moutard C. (2021). Data Science and AI: From Promises to Frustrations. Retrieved May 3rd 2021 from <https://medium.com/data-by-saegus/data-science-and-ai-from-promises-to-frustration-cd0effc60933>
- [19] Nielsen M. (2021). Neural Networks and Deep Learning, A visual proof that neural nets can compute any function. Retrieved May 3rd 2021 from <http://neuralnetworksanddeeplearning.com/chap4.html>
- [20] OED (2021). Updates to OED. Retrieved May 3rd 2021 from <https://public.oed.com/updates/>.
- [21] OpenAI (2021). Better Language Models and Their Implications. Retrieved May 3rd 2021 from <https://openai.com/blog/better-language-models/>.
- [22] Pennington J., Socher R., D. Manning C.D., GloVe (2021). Global Vectors for Word Representation. Retrieved May 3rd 2021 from <https://nlp.stanford.edu/projects/glove/>
- [23] Postman, N. (1985). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. Penguin.
- [24] Radermacher, F. J. (2017). Die Zukunft der digitalen Maschine. Was kommt auf uns zu? (Die future of the digital machine. What lies ahead?). *Denkströme. Journal der Sächsischen Akademie der*

- Wissenschaften, 2017, No. 17. Retrieved March 20th 2021 from http://www.denkstroeme.de/heft-17/s_134-159_radermacher
- [25] Rogers, H. Jr. (1987). *Theory of Recursive Functions and Effective Computability*. Cambridge, MA: MIT Press.
- [26] Saroufim M. (2021). *Machine Learning: The Great Stagnation*. Retrieved May 3rd 2021 from <https://towardsdatascience.com/machine-learning-the-great-stagnation-3a0f044e17e0>
- [27] Schirmacher, F. (2015). *EGO – The Game of Life*. *Polity*, Munich.
- [28] Schäfer A.M., Zimmermann H.G. (2006). *Recurrent Neural Networks Are Universal Approximators. Artificial Neural Networks – ICANN 2006, 2006, Volume 4131, ISBN : 978-3-540-38625-4*
- [29] Sharma S. (2021). *Activation Functions in Neural Networks*. Retrieved May 3rd 2021 from <https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6>.
- [30] Siegelmann, H. T. and Sontag, E. D. (1995). *On the Computational Power of Neural Nets*. *Journal of Computer and System Sciences* 50 (1995), 132–150.
- [31] Smith Hanley Associates (2020). *Why Are Data Scientists Frustrated?* Retrieved May 3rd 2021 from <https://www.smithhanley.com/2020/03/31/data-scientists-frustrated/>.
- [32] Turek M. (2021). *Explainable Artificial Intelligence (XAI)*. DARPA. Retrieved May 3rd 2021 from <https://www.darpa.mil/program/explainable-artificial-intelligence>.
- [33] Wadhwa M., Gupta A. and Pandey P.K. (2020). *Speech Emotion Recognition (SER) through Machine Learning*. Retrieved May 3rd 2021 from <https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine-learning/>.
- [34] Waters R. (2021). *How machine learning creates new professions — and problems*. *Financial Times*. Retrieved May 3rd 2021 from <https://www.ft.com/content/49e81ebe-cbc3-11e7-8536-d321d0d897a3>.
- [35] Wiedermann, Jan (2012). *A Computability Argument Against Superintelligence*. *Cognitive Computation* 4(3), 236-245.
- [36] Zhang D., Wu L., Sun C., Li S., Zhu Q. and Zhou G (2019). *Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations*. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 5415-5422.

On Semantic Spatiotemporal Space and Knowledge with the Concept of “Dark-matter”

Xing Chen¹ and Kiyoki Yasushi²

¹*Department of Information & Computer Sciences
Kanagawa Institute of Technology, Japan*

²*Graduate School of Media and Governance, Keio University, Japan*

Abstract. It is highlighted for machine learning models implementing functions based on data training without program coding. Artificial neural network is one of the efficient machine learning models. Different from the other machine learning models like artificial neural network, we have presented semantic computing models which represent “meaning” of machine learning results. In our model, semantic spaces are created based on training data sets. Data calculations are performed on the space. Data are mapped to semantic spaces and presented as points in semantic spaces. The mapped positions of data represent the “meaning” of data. In this paper, we first present our new discovery in the formation of semantic spaces. We use the word “matter” to represent features of semantic spaces which are related to the non-temporal data. As the same time, we use the word “dark-matter” to represent the features of semantic spaces which are temporally changed. We use the word “energy” to represent matrixes which are used in the semantic computations to generate output data. We reveal that the “dark-matter” is the spatiotemporal matrix and present a mechanism of “memory” for implementing the semantic computation. The most important contribution of this paper is that we developed a new mechanism for implementing machine learning with “knowledge” in the “memory.” In the paper, we use case studies to illustrate the concepts and the mechanism. At the beginning, we present an example on creating a semantic space from a “chaotic state” to an “ordered state.” After that, we use examples to illustrate the mechanism of the “memory” and the semantic computation. The space expansion and the space division are also illustrated by examples.

Keywords. Artificial intelligent, semantic space, spatiotemporal space, multiple semantic space transmitting, machine learning

1. Introduction

Program coding is generally required for implementing functions on computers. Although many program languages and developing tools are proposed, it is still a hard work for program coding. Moreover, it is not always possible to create models to implement required functions. Machine learning is one of the methods to implement functions and models without the requirement of program coding. The functions or models can be implemented or created through training process. During the training

¹ Xing Chen, 1030 Simo-Ogino, Atsugi-shi, Kanagawa 243-0292, Japan; chen@ic.kanagawa-it.ac.jp

process, training data are used, which are constructed by input data and expected output data. Based on the training data, today's machine learning techniques are applied to wide application areas, such as image recognition, object classification, data retrieval, etc.

Different from the other machine learning models like artificial neural network, we have presented semantic computing models which can present "meaning" of learning results [1, 2, 3, 4]. In our model, input data are mapped through mapping matrices into semantic spaces and presented as points in semantic spaces. Distances of the points present the "meaning". In our models, semantic calculation is transmitted to calculate Euclidean distances of those points. For example, in the case for implementing semantic query, a query data set is mapped into a semantic space and summarized as a point in the space. Retrieval candidate data are also mapped into the semantic space and summarized as other points. Euclidean distance is calculated between the query point and each retrieval candidate point. When the distance of a retrieval candidate is shorter than a given threshold, its relative retrieval candidate is extracted as the query output.

In our method, creating a semantic space is a basic operation. This space can be created by applying one of the following two methods, Mathematical Model of Meaning (MMM) [3, 4] and Semantic Feature Extracting Model (SFEM) [1, 2]. The difference of the two methods is that, in MMM, a common data set, for example an English-English dictionary is used to create the space. But in SFEM, a different data sets are used to create different spaces according to the requirement of applications.

After the semantic space is created, input data will be mapped to the space and the Euclidean distance calculation between the mapped data points in the space will be performed. Mapping matrices are required to map input data into the semantic space. In SFEM, different mapping matrices are required when the semantic space model is applied in different application areas [5-13]. Therefore, we developed many methods to create mapping matrices and apply the model in the areas of semantic information retrieving [8, 9, 13], semantic information classifying [10], semantic information extracting [11], and semantic information analyzing on reason and results [12], etc. We further developed a method to create the mapping matrices through deep-learning [14]. Same as the semantic computation model, the multiple matrix calculation is also the basic computation for implementing artificial neural networks and deep-learning computation [15-18].

As true and false judgement are the basic computation required by machine learning, we present a mechanism to implement the basic logic computation based on the semantic space model [19]. Furthermore, we designed and conducted experiments for simulating unmanned ground vehicle control based on the mechanism [20]. Based on our studies presented in [19] and [20], we noticed that time factors are important in the semantic computation model.

In this paper, we first present our new discovery in the formation of semantic spaces. We use the word "matter" to represent features of semantic spaces which are related to the non-temporal data. At the same time, we use the word "dark-matter" to represent the features of semantic spaces which are temporally changed. We use the word "energy" to represent matrices which are used in the semantic computations to generate output data. We reveal that the "dark-matter" is the spatiotemporal matrix and present a mechanism of "memory" for implementing the semantic computation. The most important contribution of this paper is that we developed a new mechanism for implementing machine learning with "knowledge" in the "memory." In the paper, we use case studies to illustrate the concepts and the mechanism. At the beginning, we

present an example on creating a semantic space from a “chaotic state” to an “ordered state.” After that, we use examples to illustrate the mechanism of the “memory” and the semantic computation. The space expansion and the space division are also illustrated by examples.

In the following, we first briefly review the semantic computation model and the mechanism to implement logic computations based on the semantic computation in section 2. After that, we illustrate the concept of the “matter,” “dark-matter,” “energy” and “knowledge” with examples in section 3. In this section, we also illustrate how to create the semantic space from the “chaotic state” to the “ordered state” with an example. In this example, the concepts of the “chaotic state” and the “ordered state” are illustrated. In the same time, the mechanism of “knowledge” for data retrieval on the semantic space and the relation between “knowledge” and “dark-matter” are illustrated. In section 4, we illustrate the concept of the space expansion from subspaces and dividing a space into subspaces. Finally, we will present our conclusions in section 5.

2. The semantic computation models and their applying to logic calculations

In this section, we first briefly review two semantic computation models, the Semantic Feature Extracting Model (SFEM) and the Mathematical Model of Meaning (MMM). After that, we review the mechanism to implement logic computations based on the semantic computation.

2.1. The semantic computation models

In the semantic space model, we create a semantic space by a pre-selected training data set into several clusters. In SFEM model [1, 2], a data set is used where each of the clusters has a feature that some common features of data frequently appear among the data sets in the same cluster but rarely appear in the data sets of the other clusters. The common features which frequently appear in a cluster C_i are referred to as C_i 's key features. By using the training data clusters, we construct a matrix, which is referred to as K - C matrix. In the K - C matrix, each of the rows corresponds to a key feature set K_i , and each of the columns corresponds to a cluster. The ij^{th} entry of the matrix is the number of the key features in the set K_i appearing in the cluster C_j . Because key features in the set K_i only appear in the cluster C_i , therefore, if i is not equal to j , $i \neq j$, the value of the ij^{th} entry is 0. Therefore, the K - C matrix is a diagonal matrix. That is to say that an orthogonal space is created. The value of the ii^{th} entry of the matrix is the number of the elements of the set K_i , $|K_i|$.

Next step in the semantic space model is to map data into the semantic space. By using the K - C matrix, each cluster is represented as a q dimensional vector. We use C_i to represent the vector of the cluster C_i . We use a unit vector c_i , where its norm is 1 ($|c_i|=1$), to represent the cluster vector C_i as $C_i=|K_i|c_i$. When the data are classified into q clusters, we obtain q cluster vectors. Therefore, we define q unit vectors c_1, c_2, \dots, c_q , to represent the q cluster vectors. We refer the vector space constructed by the q unit vectors to as the “space”. Because the inner product of two different unit vector is 0, $(c_i \cdot c_j)=0, i \neq j$, and there are q unit vectors, the space is a q dimensional orthogonal space.

When data d_j is vectorized according to the key features, the count of the occurrences of the C_i 's key features in the data d_j is defined to be $e_{i,j}$. We set the value of $e_{i,j}$ based on the following rule: *when a key feature in the key feature set K_i appearing in the data d_j , it is counted only once*. If the number of the key feature sets is q , the data d_j is vectorized to a q dimensional vector. We represent the vector of the data d_j as \mathbf{d}_j :

$$\mathbf{d}_j = \begin{bmatrix} e_{1,j} \\ e_{2,j} \\ \vdots \\ e_{q,j} \end{bmatrix}$$

We use v_t to represent the counted value of a key feature t . In the following, we use the expression $\sum_{t \in K_i} \{v_t\}$ to represent the sum of $v_{t_1} + v_{t_2} + \dots + v_{t_a}$, where, t_1, t_2, \dots, t_a are the elements of the key feature set K_i :

$$\sum_{t \in K_i} \{v_t\} = \{v_{t_1} + v_{t_2} + \dots + v_{t_a} \mid t_1 \in K_i, t_2 \in K_i, \dots, t_a \in K_i\}.$$

In this way, the calculation for $e_{i,j}$ is represented by the following formula:

$$e_{i,j} = \sum_{t \in K_i} \{v_t \mid \text{if } t \in d_j \text{ } v_t = 1 \text{ else } v_t = 0\},$$

where, K_i is the set of the C_i 's key features and v_t is the counted value of one of the C_i 's key feature t . If the data d_j contains the key feature t , v_t is set to 1. If the data d_j does not contain the key feature t , v_t is set to 0.

With the definition of the retrieval space, we express the data vector \mathbf{d}_j as

$$\mathbf{d}_j = \sum_{i=1}^q e_{i,j} \mathbf{c}_i.$$

In this way, data are mapped onto the q dimensional space.

The third step is to calculate Euclidean distances for data retrieval, classification or recognition. Take the data query as an example. In the processing of data query, the retrieval candidates are mapped in the semantic space and summarized as retrieval candidate points. A query is also mapped in the semantic space and summarized as a query point. We use two methods to implement the Euclidean calculation. The first method is to calculate the distances between the retrieval candidate points and query point. The second method is to select a subspace by a given query and mapped the query into the original point of the subspace and calculate the length of each retrieval candidate points from the original point. That is to calculate the norms of the retrieval candidate points. By ranking the retrieval candidates based on the norms of their relative points, the query result is obtained.

When a subspace is selected from the semantic space based on queries, the subspace is a v -dimensional space which is a part of the q -dimensional space, where v is smaller than q . The v -dimensional subspace correlates to v clusters. The subspace is selected by the following steps.

- (1) When a query Q is given, the data which contain the same features in the query is searched. Data that have the same feature as those in the query are

extracted.

- (2) From all the component items of the selected data vector, the cluster vector \mathbf{c}_i is extracted where the related component item $|e_{i,j}\mathbf{c}_i|$ has the maximum value.

$$|e_{i,j}\mathbf{c}_i| = \text{MAX}(e_{1,j}, e_{2,j}, \dots, e_{q,j})$$

We use a vector \mathbf{q} referred to as the query vector to represent the extracted clusters. We add the extracted cluster vector \mathbf{c}_i to the vector \mathbf{q} ,

$$\mathbf{q} = \mathbf{q} + \mathbf{c}_i; \text{ where the initial value of } \mathbf{q} \text{ is } \mathbf{0}.$$

- (3) A retrieval subspace S corresponding to the query is selected from the entire retrieval space by calculating the inner product of \mathbf{c}_i and \mathbf{q} . If the value of the inner product $\mathbf{c}_i \cdot \mathbf{q}$ is greater than or equal to a threshold ε , which is referred to as the subspace selection threshold, \mathbf{c}_i is added to S .

These steps (2) and (3) are repeated for each extracted data vector.

When the subspace S is obtained, the data vector on the subspace is projected and represented as

$$\mathbf{d}_j = \sum_{i=1}^q \{e_{i,j}\mathbf{c}_i | \mathbf{c}_i \in S\}.$$

In the second step, the data are ranked by calculating the norms of the data vectors on the selected subspace:

$$|\mathbf{d}_j| = \left| \sum_{i=1}^q \{e_{i,j}\mathbf{c}_i | \mathbf{c}_i \in S\} \right|.$$

In MMM, the semantic interpretation is performed as projections of the semantic space dynamically, according to contexts, as shown in Figure 1.

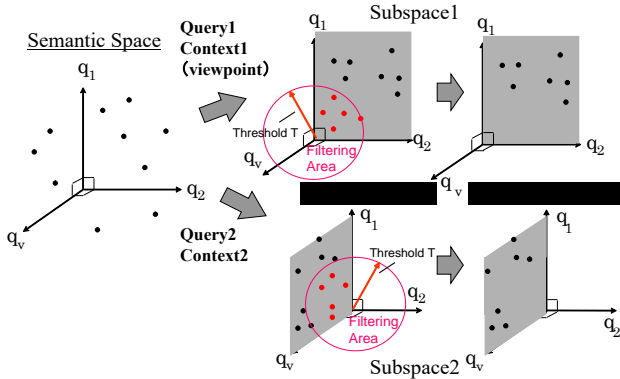


Figure 1. Semantic interpretation according to contexts in MMM

In the Mathematical Model of Meaning (MMM) [4, 7], an orthogonal semantic space is created for semantic associative search. Retrieval candidates and queries are

mapped onto the semantic space. The semantic associative search is performed by calculating the correlation of the retrieval semantic space.

In MMM, the acquisition of information or knowledge is performed by semantic computations. Context-dependent interpretation means that information is dynamically extracted by a semantic computation with context-recognition. The method realizes the computational machinery for recognizing the meaning of contexts and obtaining the semantically related information to the given context. MMM is essentially different from those methods. The essential difference is that this method provides dynamic recognition of the context. That is, the “context-dependent interpretation” is realized by dynamically selecting a certain subspace from the entire semantic space. The other methods do not provide the context dependent interpretation, that is, their space is fixed and static. The outline of MMM [4, 7] is summarized as the following:

The semantic associative computing algorithm is extended to include a deep-learning process in the MMM semantic space in the following steps:

- (1) A set of m words is given, and each word is characterized by n features. That is, an m by n matrix M is given as the data matrix.
- (2) “**Context words**” and “**image**” are characterized as “**context**” by using the n features and representing them as n -dimensional vectors.
- (3) The context words and “**image**” are mapped into the orthogonal semantic space by computing the Fourier expansion for the n -dimensional vectors.
- (4) A set of all the projections from the orthogonal semantic space to the invariant subspaces (eigen spaces) is defined. Each subspace represents a phase of meaning, and it corresponds to “**context**.”
- (5) A subspace of the orthogonal semantic space is selected according to the given “**context**” expressed in n -dimensional vectors, which are given as “**context**” represented by “**a sequence of words**” and “**image**.”
- (6) The most correlated information resources to the given “**context**” are extracted as the selected subspace by applying the metric defined in the semantic space.

2.2. Implement logic computations based on the semantic space model

Logical design is performed based on truth tables. In the truth table, all output values are given to all possible input values. The input data values and output data values are Boolean values. That is, the value can only be ‘1’ or ‘0’. For example, if there are two logical input data, x_1 and x_2 , all the possible input data values of the input data x_1 and x_2 are (0, 0), (0, 1), (1, 0) and (1, 1), in which (a, b) means the value of x_1 is ‘a’ and the value of x_2 is ‘b’. As shown in Table 1, each input data pair is given a corresponding output value in the truth table. The output values ‘0’, ‘0’, ‘0’ and ‘1’ are given to the input values (0, 0), (0, 1), (1, 0) and (1, 1) in the “and” logic truth table. Based on the truth table, logical formulas are derived. For example, for the “and” logic, an equation

$$y = x_1 * x_2,$$

is derived, where ‘*’ presents logic “and”. In the same way, an equation

$$y = \sim x_1 * x_2 + x_1 * \sim x_2$$

is also derived for the “xor” logic, where ‘*’, ‘+’ and ‘~’ present logic “and”, “or” and “not”, respectively.

Table 1. Truth Table of “and,” “or” and “xor” logic

"and" logic			"or" logic			"xor" logic		
x ₁	x ₂	y	x ₁	x ₂	y	x ₁	x ₂	y
0	0	0	0	0	0	0	0	0
0	1	0	0	1	1	0	1	1
1	0	0	1	0	1	1	0	1
1	1	1	1	1	1	1	1	0

A logical system is constructed by the derived formulas from the truth table. Boolean algebra is applied to simplify the derived formulas in order to reduce the complexity of the designed system. For example, the formula

$$a * b + a * c + \sim a * \sim c + b * c$$

can be simplified as

$$b + a * c + \sim a * \sim c.$$

In the following, we use an example to illustrate how to implement “and,” “or” and “xor” combination logical calculation based on the semantic calculation model. First, we present a data set with two inputs x₁ and x₂ and three outputs corresponding to “and,” “or” and “xor,” as shown in Table 2.

Table 2. State transition table

x ₁	x ₂	Y _{and}	Y _{or}	Y _{xor}
0.2	0.1	0.1	0.2	0.1
0.1	0.9	0.2	0.8	0.9
0.9	0.1	0.1	0.9	0.8
0.9	0.8	0.9	0.8	0.2

The values in data set are set as: when a value is close to 0, it might be logic ‘0;’ when a value is close to 1, it might be logic ‘1.’ If a data value is 0.5, it might be logic ‘0’ or logic ‘1.’ For the given data set, x₁ = 0.9 and x₂ = 0.1, it means that the input might be x₁ = 1 and x₂ = 0. The “and” output of it might be ‘0.’ The “and” output of it might be ‘0.’ The “or” output of it might be ‘1,’ and the “xor” output might be ‘1.’

Set a data set as a matrix M. A well-known method of the principal component analysis is the Singular Value Decomposition (SVD), which is a matrix computation widely used in spectral analysis, eigenvector decomposition and factor analysis. The computation is performed on a matrix with different entities on the rows and the columns. When SVD is performed on the matrix M, this matrix is decomposed into three other matrixes that contain “singular vectors” and “singular values”. We call these three matrixes as U, S and V:

$$M = U * S * V'$$

where, S is a diagonal matrix that contains singular values, matrixes U and V are left and right matrix of S, respectively. V' is the transposed matrix of V. The matrix V has orthonormal columns, that is

$$V^T * V = I$$

where I is the identity matrix.

We call the space $U * S$ is the semantic space created by the matrix M. As

$$\begin{aligned} M &= U * S * V^T \\ M * V &= U * S * V^T * V \\ M * V &= U * S * I \\ M * V &= U * S, \end{aligned}$$

we call the matrix V as the *space mapping matrix*. That is, any matrixes of data sets which have the same number of columns of the matrix M can be mapped to the semantic space through the mapping matrix V.

When SVD is performed to the matrix of the given data set,

$$M = \begin{bmatrix} 0.2 & 0.1 & 0.1 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0.2 & 0.8 & 0.9 \\ 0.9 & 0.1 & 0.1 & 0.9 & 0.8 \\ 0.9 & 0.8 & 0.9 & 0.8 & 0.2 \end{bmatrix},$$

we get three matrixes:

$$\begin{aligned} U &= \begin{bmatrix} -0.1 & -0.1 & 0.1 & -1.0 \\ -0.5 & 0.6 & -0.6 & 0.0 \\ -0.5 & 0.3 & 0.8 & 0.1 \\ -0.6 & -0.7 & -0.2 & 0.1 \end{bmatrix} & S &= \begin{bmatrix} 2.5 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.9 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} & V &= \begin{bmatrix} -0.5 & -0.4 & 0.6 & 0.1 & -0.5 \\ -0.4 & 0.0 & -0.7 & -0.2 & -0.5 \\ -0.3 & -0.6 & -0.3 & 0.5 & 0.5 \\ -0.6 & 0.1 & 0.1 & -0.6 & 0.5 \\ -0.4 & 0.7 & 0.1 & 0.6 & 0.0 \end{bmatrix} \end{aligned}$$

The semantic space $U * S$ is

$$\begin{bmatrix} -0.3 & 0.0 & 0.1 & 0.0 & 0.0 \\ -1.3 & 0.6 & -0.5 & 0.0 & 0.0 \\ -1.3 & 0.2 & 0.6 & 0.0 & 0.0 \\ -1.6 & -0.7 & -0.2 & 0.0 & 0.0 \end{bmatrix}$$

This is a five-dimensional space. Each row of the matrix represents a mapping point of the data set. As it is a four rows matrix, that means four points are mapped to the semantic space. It is worth to notice that the values of the last two columns of the matrix is zero, therefore, we remove the last two rows and get a new matrix P,

$$\begin{bmatrix} -0.3 & 0.0 & 0.1 \\ -1.3 & 0.6 & -0.5 \\ -1.3 & 0.2 & 0.6 \\ -1.6 & -0.7 & -0.2 \end{bmatrix}$$

The meaning of removing the last two rows of the matrix is that the semantic space is compressed from a five-dimensional space to a three-dimensional space.

Strictly speaking, that is each of the value of the last two rows is smaller than a threshold. In this example, the absolute value of the threshold is set to 0.003.

As $U^T * U = I$, that is

$$U^*U = \begin{bmatrix} 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix},$$

the data set is mapped to an orthogonal space. This orthogonal characteristic is not changed to the compressed space:

$$P^*P = \begin{bmatrix} -0.3 & -1.3 & -1.3 & -1.6 \\ 0.0 & 0.6 & 0.2 & -0.7 \\ 0.1 & -0.5 & 0.6 & -0.2 \end{bmatrix} * \begin{bmatrix} -0.3 & 0.0 & 0.1 \\ -1.3 & 0.6 & -0.5 \\ -1.3 & 0.2 & 0.6 \\ -1.6 & -0.7 & -0.2 \end{bmatrix}$$

$$= \begin{bmatrix} 6.2 & 0.0 & 0.0 \\ 0.0 & 0.8 & 0.0 \\ 0.0 & 0.0 & 0.6 \end{bmatrix}.$$

As the data set is mapped to an orthogonal space, Euclidean distance calculation can be applied to calculate the distance of a new mapped point to the points already in the space.

For example, give a new input set $x_1 = 0.7, x_2 = 0.3$. As is different from the given data set, we should calculate four outputs Y_{and}, Y_{or} and Y_{xor} for the given input data set. As the output value is not known, we set the values of the three outputs as 0.5, that means it might be logic '0' or logic '1.' Thus we get a vector $[0.7 \ 0.3 \ 0.5 \ 0.5 \ 0.5]$.

Mapping the vector to the semantic space,
 $[0.7 \ 0.3 \ 0.5 \ 0.5 \ 0.5] * V,$

we get a mapping point p_x represented as a three-dimensional vector,
 $p_x = [-1.1 \quad -0.2 \quad 0.2].$

Dividing each rows of matrix P, we get four mapping points of the given data set in the semantic space, p_1, p_2, p_3 and p_4 , represented as four vectors:

$$p_1 = [-0.3 \quad 0.0 \quad 0.1],$$

$$p_2 = [-1.3 \quad 0.6 \quad -0.5],$$

$$p_3 = [-1.3 \quad 0.2 \quad 0.6],$$

$$p_4 = [-1.6 \quad -0.7 \quad -0.2].$$

Calculating the Euclidean distances of p_x to p_1, p_2, p_3 and p_4 , we get four values: 0.64, 1.01, 0.44 and 0.58. Among them, the smallest value is the third one, 0.44. That is, point p_x is most close to the point p_3 . The value of p_3 in the data set is

$$[0.9 \ 0.1 \ 0.1 \ 0.9 \ 0.8],$$

thus, we set the output value as $Y_{and} = 0.1, Y_{or} = 0.9$ and $Y_{xor} = 0.8$ for the input value $x_1 = 0.7$ and $x_2 = 0.3$.

3. The concept of the “matter,” “dark-matter,” “antimatter,” “energy” and “knowledge”

In this section, we use a case study to illustrate the concept of the “matter,” “dark-matter,” “antimatter” and “energy.” For the logical computation as reviewed in Section 2, we use two matrixes M and E to implement it,

$$Y = M * E,$$

where M represents an input space, E is a mapping matrix to map input data to the output space, “*” represents matrix multiply and Y is the matrix represents output data.

As shown in Figure 2, if input data is a two-dimensional logical value matrix, all the possible input data are (0, 0), (0, 1), (1, 0) and (1, 1), which is a four-rows and two-columns matrix. Take it as an example, we use four data 0.0, 0.1, 0.2 and 0.3 as an index of the input matrix. As the input data are that we use to obtained output results, that is, the input data are “visible” to us, we refer to the data in the index as “matter.” For example, the data “0.2” in first column of the space matrix X is a kind of “matter” which indicates the input data “10.” In order to calculate the invers matrix of X, we use random data to fill the other three columns of X. We refer to the data in these columns as “dark-matter.” As these columns are filled with random data, we refer to the space X as “chaotic space.”

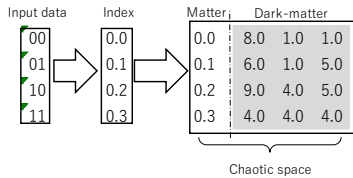


Figure 2. Creating a “chaotic space”

Before change the space from “chaotic space” to ordered state, let’s introduce a concept of “time” used in this paper. We refer to the “time” as state transition. Considering our system as a state machine, the state of the system will be changed as the time lasted. If the states of a system are not changed, the time of the system is stopped. That is, as the time lasted, states transition is happened continuously. Figure 3 is an example of state transition diagram. In the figure, a circle and the number in the circle represents a state, an arrow represents a state transiting from one to another one. The numbers by the arrows represent the condition for the state transition. For example, if the number is “0.0,” it means if the index number if “0.0,” the state will be transited from state “0.0” to the state “0.1.” In the example, we use the numbers of the index to indicate states.

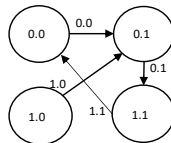


Figure 3. An example of state transition

Next, we introduce how to create the “ordered space” from the “chaotic space” as the time lasted. The example of the state transition diagram shown in Figure 4 is used to illustrate it. In the example, state “0.0” transits to the state “0.1” at the first step. Figure 4 (b) shows this state transition. In the next step, the state “0.1” transits to the state “0.3.” Figure 4 (c) shows this state transition. At last, when the state transition diagram shown is represented as Figure 4 (d), we say that we created an “ordered state” from the “chaotic state” shown in Figure 4 (a). In the figure, the “matter” is represented in the first row of the matrix, and the “dark-matter” is represented in the second, third and fourth row of the matrix, painted gray.

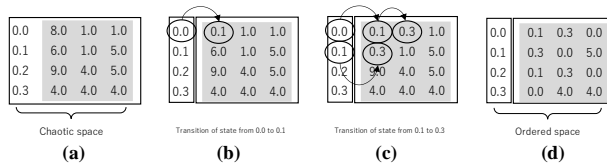


Figure 4. Changing the “chaotic space” to “ordered space”

From Figure 4, we can find that the “dark-matter” is a space represents state transition. It determines the state transition rules of the “matter” in space which we created. Based on the “ordered space,” we refer to the inverse matrix of the “ordered space” as the “anti-matter space.” If X^{-1} is an inverse matrix of an “ordered space” X , data in the matrix X^{-1} are referred to as “anti-matter.” An example of the “anti-matter space” is shown in Figure 5 (b), which is the inverse matrix of the space represented as a matrix X shown in Figure 5 (a).

By applying antimatter to the desired result, we create a matrix. We refer to this matrix as “energy.” Taking the “and,” “or” and “xor” logical calculation results as shown in Table 1 as an example, for all possible input data values, the related output result values are shown in Figure 5 (c), represented as Y_{and} , Y_{or} and Y_{xor} . Three different kinds of “energy”, E_{and} , E_{or} and E_{xor} are represented as three vectors in Figure 5 (d), where “energy” is created by the matrix multiplying result of “antimatter” and calculating results:

$$E = X^{-1} * Y .$$

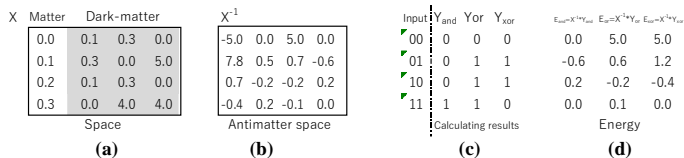


Figure 5. The concept of the “matter,” “dark-matter,” “anti-matter” and “energy.”

For a given input, we use a function *mem* which we refer to as “memory” to extract its relative row vector from the space,

$$M = mem(\text{Input}).$$

The row vector M is composed by “matter” and “dark-matter,” we simply call it “matter.” For example, if the input value is a vector [0, 1], its relative row vector [0.2, 0.1, 0.3, 0.0] is extracted from the space matrix X,

$$M = mem([0,1]),$$

$$M = [0.2 \ 0.1 \ 0.3 \ 0.0] \begin{bmatrix} 0.0 \\ -0.6 \\ 0.2 \\ 0.0 \end{bmatrix}$$

$$M = 0.$$

We implement the semantic computation by apply energy to the matter,
 $Y = M * E.$

As an example, when energy E_{and} is applied to the matter [0.2, 0.1, 0.3, 0.0], we get a value “0” which is the logic “and” calculation, “0 and 1.” In Figure 6, we summarize the two bits logical “and,” “or” and “xor” calculations. When the matter shown in Figure 6 (a) is multiplied by the energy shown in Figure 6 (b), the logical calculations “and,” “or” and “xor” are performed. The calculation results are shown in Figure 6 (c).

M				E_{and}	E_{or}	E_{xor}	$M * E_{and}$	$M * E_{or}$	$M * E_{xor}$
0.0	0.1	0.3	0.0	0.0	5.0	5.0	0	0	0
0.1	0.3	0.0	5.0	-0.6	0.6	1.2	0	1	1
0.2	0.1	0.3	0.0	0.2	-0.2	-0.4	0	1	1
0.3	0.0	4.0	4.0	0.0	0.1	0.0	1	1	0

(a)
(b)
(c)

Figure 6. Implementing calculation by applying “energy” to “matter”

As “energy” E is calculated with invers of matrix X,
 $E = X^{-1} * Y,$

the calculation of $M * E$ can be written as

$$M * E = M * (X^{-1} * Y).$$

In the case that the matrix M calculated through the memory function mem is the same as the matrix X, we obtain the equation,

$$\begin{aligned} M * E &= M * (X^{-1} * Y) \\ &= X * (X^{-1} * Y) \\ &= X * X^{-1} * Y \\ &= I * Y \\ &= Y, \end{aligned}$$

where, I is a unit matrix. It is required that the memory function should recall the original matrix X based on the input.

A simple way to implement the memory function is to use a list with the columns of input data and the matrix M as shown in Figure 7 (a). In the list, each input data

corresponds a row of matrix M. As shown in the figure, if the input data is [1, 0], the output of the memory function $mem([1, 0])$ is a row vector [0.2, 0.1, 0.3, 0.0].

However, in some cases, one input value may correspond two or more rows of matrix M. It will happen if the input data is obtained from a sensor, but the resolution of the sensor is not high enough to provide the required input data value. As shown in Figure 7 (b), although two-bit input data values are required, but the efficient input values are only one-bit values. As we can not tell the difference for the first, second and the third, fourth rows, we use the medium value between 0.0 to 0.1, and that between 0.2 to 0.3, that is, "0.05" and "0.25," respectively as the index of "00" and "10," as shown in Figure 7 (b).

Input	M			
00	0.0	0.1	0.3	0.0
01	0.1	0.3	0.0	5.0
10	0.2	0.1	0.3	0.0
11	0.3	0.0	4.0	4.0

(a)

Input	M			
00	0.05	0.1	0.3	0.0
00	0.05	0.3	0.0	5.0
10	0.25	0.1	0.3	0.0
10	0.25	0.0	4.0	4.0

(b)

Figure 7. Implementing calculation by applying "energy" to "matter"

We add the "dark-matter" in the memory function as "*knowledge*" to implement the recall function. As shown in Figure 8 (a), at the beginning of the system, we know that the next state is "0.1." If the input is "00," the "dark matter" vector [0.1, 0.3, 0.0] is added to the "matter" [0.05]. Thus, the output the memory function is a row vector, [0.05, 0.1, 0.3, 0.0]. If the input is "10," the "dark matter" vector [0.1, 0.3, 0.0] is added to the "matter" [0.25]. Thus, the output the memory function is a row vector, [0.25, 0.1, 0.3, 0.0]. Applying "energy" to these output vector, as shown in Figure 8 (b), the logical calculations are correctly implemented.

Input	M			
00	0.05	0.1	0.3	0.0
00	0.05	0.3	0.0	5.0
10	0.25	0.1	0.3	0.0
10	0.25	0.0	4.0	4.0

(a)

M^*E_{and}	M^*E_{or}	M^*E_{xor}
0	0	0
0	1	1
0	1	1
1	1	0

(b)

Figure 8. Implementing calculation by applying "energy" to "matter"

As described above, "dark-matter" represents the rule of the space that we created. Based on the rule, we know that when the state transition happens, which row of the matrix M should be extracted. As the example of Figure 8 (a), at the beginning, the state of the system is "0.0." If the input data is "10", the third and fourth rows relate to the input data. As the next state of the state "0.0" is "0.1," the third row of the matrix M is extracted. That is, "knowledge" is the mechanism of the memory function which extract rows as the output from the space M based on the rule of the space. We refer to the space M as the "semantic space." We refer to the process of extracting rows from the matrix M related to the input data and current state as the "semantic retrieval." We refer to the processing of applying the matrix E to the matrix M as the "semantic computation."

4. The concept of “space expansion” and “space division”

In this section, we use a case study to illustrate the concept of “space expansion” and “space division.” We use a virtual agent in this case study. The agent moves on a plane surface with obstacles. As shown in Figure 9 (a), the start point of the agent is marked as “S” and the goal is “G.” The row number is indicated by two-bit numbers same as the column. The agent has five actions, “Non,” “Left,” “Right,” “Up” and “Down,” which are assigned with the number “0.0,” “0.1,” “0.2,” “0.3” and “0.4,” as shown in Figure 9 (b). The agent moves from the position “01” row and the “00” column, as shown in Figure 9 (a). We use four bits number to indicated the position of the agent. Thus, the start position is “0100” and the goal position is “1111.” When the agent starts to move, its next position is “0101.” We use arrows to show the movement and the actions of the agent. Taking the positions as the states, we get a state transition diagram through the movement of the agent. That is, the “rule” of the semantic space is defined. Based on the “rule,” we create the semantic space.

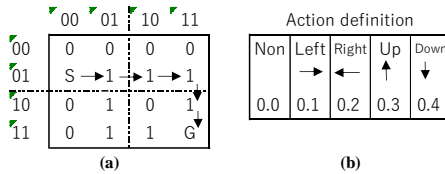


Figure 9. The plane surface for an agent moving and the definition of the agent action

In this example, there are 16 positions. Although we can create a 16 states space, but we prefer to use the 4 states space in order to introduce the concept “sub-space.” As shown in Figure 9 (a), we divide the plain surface into four blocks, each part has 4 positions. As mentioned above, we use row-column number (abbr. RC) to indicate the positions. In the first block, the four positions are “0000,” “0001,” “0100” and “0101.” As shown in Figure 10 (a), we set an index to the four positions as “0.0,” “0.1,” “0.2” and “0.3,” respectively. The agent can move from the position “0100” to the position “0101.” Taking the index number to represent the four possible states, we can draw a state transmission diagram, which has four states, as shown in Figure 10 (b). As the start position of the agent is at position “0100,” the initial state is “0.2.” As the agent is required to move left, the calculation result is set to “0.1” as defined in the Figure 9 (b). When the agent is moved left, the state transmission is happened from the state “0.2” to the state “0.3,” as shown in Figure 10 (b). Based on the state transmission diagram, we create a space for the first block as shown in Figure 11 (a).

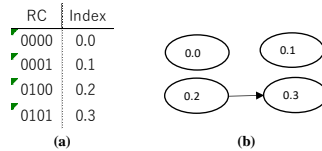


Figure 10. Implementing calculation by applying “energy” to “matter”

As shown in Figure 11, four spaces X_1 , X_2 , X_3 and X_4 are created for the plain surface. Each space corresponds to a block of the surface. We refer to these spaces as sub-spaces. A sub-space X_i is created based on the states and following the “rule” defined on the space. “Energy” E_i of the sub-space is created by applying the action matrix defined based on the “rule” of the space to the inverse matrix of the sub-space, X_i^{-1} . In Figure 11, four sub-spaces are created. As the shown in Figure 11 (d), no “rule” is defined on the fourth block, the created sub-space X_4 is a chaotic space. As no action is required on this block, “energy” is a zero vector.

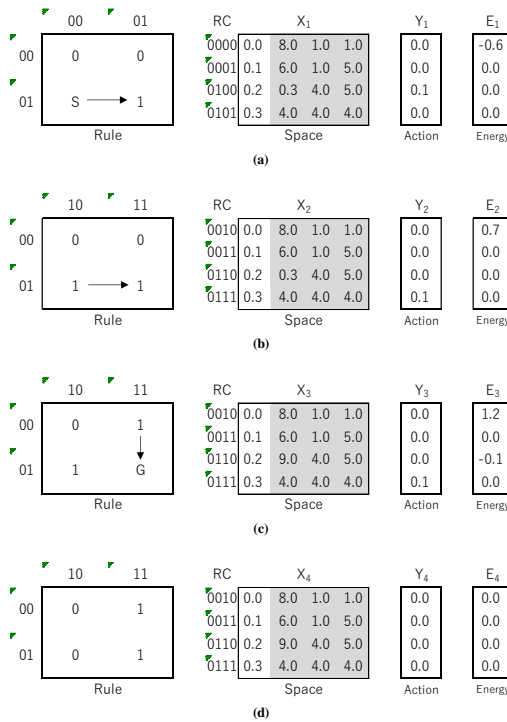


Figure 11. Creating four sub-spaces based on the movement of the agent

The semantic space of the plain surface is created through the creation of the sub-spaces. As shown in Figure 11, the semantic space is the same as the sub-space X_1 . When the sub-space X_2 is created, the semantic space is created with X_1 and X_2 , and the number of the state of the semantic space are increased from 4 to 8. After the sub-space X_3 is created, the number of the state of the semantic space are increased to 16. That is, a sub-space must be created when a new state is added to the semantic space

and the number of the state of the semantic space are increased exponentially by 2. We refer to this process as the “*space expansion*.”

Based on the sub-spaces, we can calculate actions of the agent on each block. However, we can not calculate the actions of the agent from one block to the others. For the four divided blocks, we define four positions, “0i0j,” “0i1j,” “1i0j” and “1i1j,” where i and j are one-bit values “0” or “1.” Using the four positions, we define relative index and states as shown in Figure 12. The agent starts at the position “0i0j,” and moves to the position “0i1j.” The goal of the agent is at the position “1i1j.” Based on the “rule” of the movement of the agent, a space X with four states are created and “energy” E is also created by applying the agent action definition matrix Y to the inverse matrix of X, as shown in Figure 12.

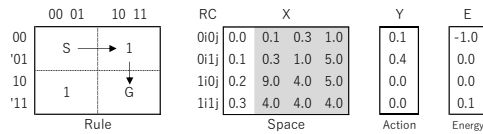


Figure 12. Creating the semantic space with four states

For the action of the agent in each block, we can create relative sub-space as shown in Figure 11. In summary, if the number of states for creating the semantic space are more than four, we divide them into two parts. In each part, if the number of states is still more than four, we divide them again until the number of the states in each part is smaller or equal to four. In the example shown in Figure 9 (a), the surface is divided twice into four blocks as shown in Figure 12. We refer to this process as “*space division*.”

“Space division” is a process from global to local as shown in Figure 13. The results of the semantic computation on the global space shows the movement of the agent between blocks. Detail movements of the agent in each block are calculated on the local space.

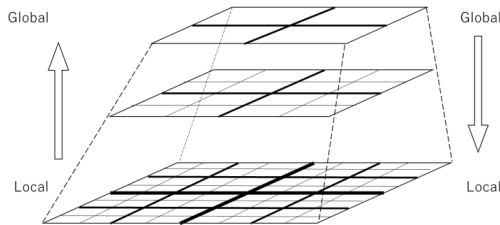


Figure 13. “Space expansion” and “Space division” among global and local

We use “space division” in the case that we have “knowledge”, or in other words, we have “rules” from global to local. In the example shown in Figure 9 (a), the actions of the agent are shown from the start point to the goal point. That is, we have all

required “rules” for the agent moving on the surface. Therefore, we use “space division” to create sub-spaces for the agent movement between blocks and those in each block.

If we do not have all the required “rules,” the semantic space creation is performed from local to global as shown in Figure 13. For example, if reinforcement learning is performed to the agent, “rule” and states are generated during the learning process. During the “rules” and states are generated, sub-spaces are created. Using reinforcement learning in a period of time, we can only get part of “rules” and states. All the other “rules” are remained to be found. Therefore, for the semantic space with all the possible states, only small parts of the space are in ordered state, the remained parts are in chaotic state as show in Figure 14.

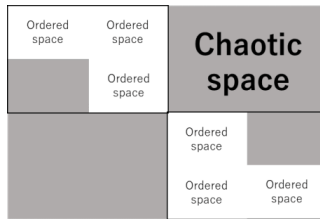


Figure 14. A semantic space creating through reinforcement learning

5. Conclusion and future work

In this paper, we presented that the semantic space is a spatiotemporal space. We introduced the concept “dark-matter” to reveal the temporal characteristic of the space. We defined state transmission as the “time” in a system performing semantic computation. In the paper, we illustrated following concepts and computations: A matrix X created with the states and state transmission diagram is referred to as the semantic space. The elements of the matrix representing states are referred to as “matter” and the elements representing the state transmission diagram are referred to as the “dark-matter.” Its inverse matrix X^{-1} is referred to as the “anti-matter.” The result of multiplying the “anti-matter” by the defined output is referred to as “energy.” Memory function are defined to exact vectors representing “matter” and “dark-matter.” “Knowledge” is required to be added to the memory function in the case that the input data cannot definitely indicate relative states. By applying “energy” to the output of the memory function, we get the semantic computation result related to the giving input. The “space expansion” is happened as the increasing of the “matters.” When a state transmission diagram is given, we divide the diagram into sub-diagram and create relative sub-spaces. We refer to the process as the “space division.” The divided semantic space is a pyramid shaped structure. Each layer of the pyramid is composed by sub-spaces. The top layer represents global state transmission diagram. The bottom layer represents details of the divided sub-diagram. Based on these concepts and mechanism, we developed a new mechanism for implementing machine learning. In this mechanism, machine learning is to calculate “energy” matrixes and “knowledge” is used to extract vectors from the semantic space based on the input data during the

semantic computation. Furthermore, we revealed that the semantic computation must be performed dynamically with the state transmission. We also revealed that the semantic computation must be performed from global to local to give rough and detailed calculation results. As our future work, we will use this new machine learning mechanism to implement artificial intelligence systems and furtherly confirm the effectiveness of the mechanism in practice.

References

- [1] Chen, X. and Kiyoki, Y., "A query-meaning recognition method with a learning mechanism for document information retrieval," Information Modelling and Knowledge Bases XV, IOS Press, Vol. 105, pp.37-54, 2004.
- [2] Chen, X. and Kiyoki, Y., "A dynamic retrieval space creation method for semantic information retrieval," Information Modelling and Knowledge Bases XVI, IOS Press, Vol. 121, pp.46-63, 2005.
- [3] Kiyoki, Y. and Kitagawa, T., "A semantic associative search method for knowledge acquisition," Information Modelling and Knowledge Bases, IOS Press, Vol. VI, pp.121-130, 1995.
- [4] Kitagawa, T. and Kiyoki, Y., "A mathematical model of meaning and its application to multidatabase systems," Proc. 3rd IEEE International Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems, pp.130-135, April 1993.
- [5] Chen, X., Kiyoki, Y. and Kitagawa, T., "A multi-language oriented intelligent information retrieval system utilizing a semantic associative search method," Proceedings of the 17th IASTED International Conference on Applied Informatics, pp.135-140, 1999.
- [6] Chen, X., Kiyoki, Y. and Kitagawa, T., "A semantic metadata-translation method for multilingual cross-language information retrieval," Information Modelling and Knowledge Bases XII, IOS Press, Vol. 67, pp.299-315, 2001.
- [7] Kiyoki, Y., Kitagawa, T. and Hitomi, Y., "A fundamental framework for realizing semantic interoperability in a multidatabase environment," International Journal of Integrated Computer-Aided Engineering, Vol.2, No.1(Special Issue on Multidatabase and Interoperable Systems), pp.3-20, John Wiley & Sons, Jan. 1995.
- [8] Kiyoki, Y., Kitagawa, T. and Hayama, T., "A metadatabase system for semantic image search by a mathematical model of meaning," ACM SIGMOD Record, Vol.23, No. 4, pp.34-41, Dec. 1994.
- [9] Kiyoki, Y., Chen, X. and Kitagawa, T., "A WWW Intelligent Information Retrieval System Utilizing a Semantic Associative Search Method," APWeb'98, 1st Asia Pacific Web Conference on Web Technologies and Applications, pp. 93-102, 1998.
- [10] Ijichi, A. and Kiyoki, Y.: "A Kansei metadata generation method for music data dealing with dramatic interpretation," Information Modelling and Knowledge Bases, Vol.XVI, IOS Press, pp. 170-182, May, 2005.
- [11] Kiyoki, Y., Chen, X. and Ohashi, H.: "A semantic spectrum analyzer for realizing semantic learning in a semantic associative search space," Information Modelling and Knowledge Bases, Vol.XVII, IOS Press, pp.50-67, May 2006.
- [12] Takano, K. and Kiyoki, Y.: "A causality computation retrieval method with context dependent dynamics and causal-route search functions," Information Modelling and Knowledge Bases, ISO Press, Vol.XVIII, pp.186-205, May 2007.
- [13] Chen, X. and Kiyoki, Y.: "A visual and semantic image retrieval method based on similarity computing with query-context recognition," Information Modelling and Knowledge Bases, IOS Press, Vol.XVIII, pp.245-252, May 2007.
- [14] Nitta T., "Resolution of singularities introduced by hierarchical structure in deep neural networks," IEEE Trans Neural Netw Learn Syst., Vol.28, No.10, pp.2282-2293 Oct. 2017.
- [15] Wiatowski, T. and Bölskei, H., "A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction," IEEE Transactions on Information Theory, PP(99) · Dec. 2015.
- [16] Hochreiter, S., Bengio, Y., Frasconi, P. and Schmidhuber, J. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In Kremer, S. C. and Kolen, J. F. (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*," IEEE Press, 2001.
- [17] Hochreiter, S. and Schmidhuber, J., "Long short-term memory," Neural computation, Vol.9, No.8, pp.1735-1780, 1997.
- [18] Kalchbrenner, N., Danihelka, I. and Graves, "A. Grid long short-term memory," CoRR, abs/1507.01526, 2015.

- [19] Chen, X. and Kiyoki, Y., "*On Logic Calculation with Semantic Space and Machine Learning,*" Information Modelling and Knowledge Bases XXXI, IOS Press, Vol. 321, pp.324-343, 2019.
- [20] Chen, X. Prayongrat, M. and Kiyoki, Y., "*A Concept for Control and Program Based on the Semantic Space Model,*" Information Modelling and Knowledge Bases XXXII, IOS Press, Vol. 333, pp. 26-44, 2020.

System Formed by Combining Means of Mobility and Facility That Support Sensitivity to Context in Mobility Routes — Emotional MaaS —

Koichiro Kawashima* Yasuhiro Hayashi** Yasushi Kiyoki*** Tetsuya Mita*

*JR East R&D Center Frontier Service Laboratory 2-479 Nissincho, Kita-Ku, Saitama City, Saitama 331-8513

**Keio University Graduate School of Media and Governance 5322 Endo, Fujisawa City, Kanagawa 252-0882

***Keio University, Faculty of Environment and Information Studies 5322 Endo, Fujisawa City, Kanagawa 252-0882

E-mail: *{kouichirou-kawashima,t-mita}@jreast.co.jp, **yasuhiro.hayashi@keio.jp, ***kiyoki@sfc.keio.ac.jp

Prologue This paper introduces a method of realizing a system of forming combinations of means of mobility and facility that supports sensitivity to context in mobile routes. MaaS refers to a service for integrating mobility in actual space and information spaces and linking various types of mobility. The essence of this is that leisure can be designed whilst moving, based on reducing the effort required by the user themselves for mobility. This makes it possible for humans to think about something other than the mobility itself while moving.

The system presented in this paper returns the optimal means of moving to a destination and a means of facility that exists on the route based on a “sensitivity to context” that supports the intentions of the moving user, and the situation. This system applies a “mathematical model of meaning”, as a semantic associative search model that dynamically calculates semantic associations. The calculation method proposes a “sensitivity to context file” generated by a query generation operator group for synthesizing and expressing “everyday intention” and “mobility situations”, the distance calculation of “feature value vectors” for means of mobility within the DB and facility spots, and the output values are the optimal means of moving towards the destination and the forms of facility existing on the route.

Keywords MaaS, sensitivity, context, query generation operator

1. Introduction

1.1 Background

One of the social issues identified in recent years is the concentration of the population in urban areas, and it is predicted that approximately 70% of the world’s population will be concentrated in cities by 2050, with a steady increase in large cities containing 10 million people or more [1]. A wide variety of issues arise from this over-concentration including environmental pollution, electricity/energy shortages, traffic jams, the spread of viral infections, and deterioration of regional areas. Smart cities are receiving

resolving these issues in one go using technologies such as IoT and sensing etc.[2]. The concept of linking these issues was taken up by the UN summit of September 2015, and was included as an international goal in the SDGs from 2016 to 2030, which has been a common language for linking disparate ideas up to this point [3]. According to a survey by the Statistics Bureau of the MIAC (Table 1), there is a diverse range of forms of mobility by region in Japan, and many different issues are faced [4]. Therefore, mobility services that suit regional characteristics and transportation systems are required.

Table 1 Regional mobility issues and mobility service business characteristics

City type	Summary	Japan ratio (population) (municipalities)	Senior traffic sharing rate	Examples of mobile tasks
Large city (Over 500,000 people)	Ordinance-designated cities, etc.	32.4% (41.22 million)	1.7% (29)	<ul style="list-style-type: none"> ● Daily road congestion and public transport congestion ● Cumbersome connectivity between travel modes ● Limited means of transportation in the last mile
Medium-sized city (50,000 to 500,000 people)	<ul style="list-style-type: none"> ● private car sharing rate Less than 50% ● Near a large city Bedroom town, etc. 	19.3% (24.52 million)	8.9% (153)	<ul style="list-style-type: none"> ● Limited means of transportation to the city center. ● Congestion of public transportation during commuting, etc. ● Mobility for the elderly in Old Town and other areas is an issue.
Suburbs / depopulated areas (50,000 people or less)	<ul style="list-style-type: none"> ● private car sharing rate Over 50% ● The location of the local prefectural office Company town, etc. 	32.5% (41.28 million)	19.8% (340)	<ul style="list-style-type: none"> ● Decline in convenience and business potential of public transportation ● Elderly people face difficulties in securing transportation ● Transportation is mostly by private car, unable to maintain public transportation. ● Elderly people face difficulties in securing transportation ● Expansion of transportation gap areas
	Local suburbs, etc.	15.8% (20.07 million)	69.6% (1,197)	69.6%

Attempts at creating new mobile business formats that can resolve these issues include multimodal services for integrating, linking and optimizing the various means of mobility, sharing services for using them in a mutual way, demand-based transportation to match the behavior of mobile users, and a hybrid cargo/passenger service for transporting and carrying both

64 attention as a mechanism of connecting and

cargo and passengers in an integrated way. Businesses and verification testing are being rolled out in the various regions of Japan. Additionally, in order to cope with diversified mobility needs, micro mobility, green slow mobility, and mobility services with new characteristics, such as automated driving etc., are being promoted and spread [5]. The extension of these mobility worlds will lead to the general construction of smart cities.

1.2 Positioning of this study

Mobility as a Service (hereinafter, referred to as MaaS) is a concept in which mobility based on various mobility services is integrated into one service, and the mobile users can perform one-stop route searches, transportation facility reservation/arrangement, and settlement via payment. In other words, it means the spread of a world in which the “actual space” and “information space” in mobility is integrated. Therefore, we considered that “the essence of MaaS is the ‘design of mobile users’ leisure’ created by lightening the burden of effort in mobility”. Previously, mobile users investigated the means of mobility themselves, and followed the operating convenience of the transportation facilities, sometimes driving themselves, and with these various restrictions etc., this consumed much of their attention and time in their daily lives. However, with the spread of the previously-described new mobility services, such as automated driving and micro-mobility, multi-modality and sharing, and promotion of new mobility business formats, such as seamless transfer etc., humans are being released from such limitations. Therefore, these services enable humans to turn their thoughts to matters over than mobility when moving. They can concentrate on satisfying themselves in their everyday lives. That is to say, even while they are mobile, they can contribute to achieving their original goals as human beings.

In this study, with the aim of creating a lifestyle in which mobile users are able to satisfy the desires of ordinary life even in a situation in which they are moving, we have proposed a format and attempted to create a system

(“Emotional MaaS” (Figure 1) for such a format that presents means of mobility that support a sensitivity to context based on “ordinary intentions” and “mobility situation” of the searching party, from point of departure to destination, and provide guidance that combines everyday means (facilities, stores, events etc.) on this mobile route.

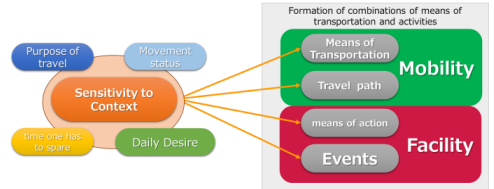


Figure 1 Image of Emotional MaaS

1.3 Related Studies

1.3.1 Mathematical model of meaning (MMM)

Meaning, sensitivity to identity, similarity, and association between data are not determined based on static relationships, but are considered to change dynamically according to the context and situation, and there is the “Mathematical model of meaning”[6], which is a computational model that dynamically calculates the semantic and sensitive equivalence, similarity, and association between data, based on “situation and context”. By applying this model in this paper, an optimized proposal based on the sensitivity to context of mobile users has been proposed.

1.3.2 Query generation system using multiple images

There have been studies focusing on systems of generation by efficiently combining multiple images to generate image queries that express user intentions [7]. This system is calculated dynamically using query generation operators corresponding to the features (color/form) of the respective images. Using this method, it is possible to express the creativity and intention of the users using multiple images. In this paper, we generate a sensitivity to context vector based on a query generation operator group to merge and express the “mobility situations” and “everyday intentions” of the mobile users,

1.3.3 Creation of diverse driving/travel plans

Table 2 Feature space structure for means of mobility

dimension	-1	0	+1	
Price	low price	No relation	expensive	
Time required	Fast	No relation	Slow	...(a)
Amount of exercise consumed	Small	No relation	Large	
Crowding level	Low	No relation	High	...(b)
Regularity	High	No relation	Low	
Motor/operation control	Unnecessary	No relation	Necessary	
Desk environment	Not available		Available	
Communication environment	Not available		Available	
Charging facilities	Not available		Available	
Exercise facilities	Not available		Available	...(c)
View facilities	Not available		Available	
Conversation space	Not available		Available	
Eating and drinking space	Not available		Available	
Luggage space	Not available		Available	

Table 3 Qualitative information feature correspondence

	Crowding level			punctuality	Operation and driving maneuvers 2
	public transportation	walking	uncongested traffic 1		
+1					● Awareness required
+0.75	● 7:00 < Moving time < 9:00			● Use public roads	
+0.25	● 18:00 < movement time < 21:00			● Use of independent track and public road	
0	● Travel time excluding rush hours			● Self-powered, such as on foot or by bicycle	● No awareness required
-0.5		○			
-1			○	● Use independent orbit	

table

*1 Mobility services that ensure private space and social distancing
 *2 If the required time (total mobility time) divided by the ratio of this mobility time is t_i , this can be calculated as $v_i(t_i)$ in previous ①.

that use the results of sensitivity assessments of people’s motives when traveling

With the aim of encouraging people to travel, a method that uses the results of sensitivity assessments regarding user trips to create multiple and diverse driving travel plans that reflect sensitivity in user tourism has been proposed [8]. Here, the motives for people traveling have been classified into five elements, and the expected elements of each tourist site introduced. In this paper, we construct a sensitivity space using these five motive elements as a characterization of means of facility.

2. Proposal format

2.1 Feature Space Generation

A feature space (Figure 2) that supports the feature vector for means of mobility and facility is generated based on the procedures shown in the following steps 1-3.

Step 1 : n features in horizontal axis (f_1, f_2, \dots, f_n)

Table 4 Feature space structure for means of facility

dimension	0	+1	
desire to enrich knowledge	No relation	satisfaction	
desire for general developmental growth	No relation	satisfaction	...(d)
desire to alleviate tension	No relation	satisfaction	
desire to do something fun	No relation	satisfaction	
desire to shape human relationships	No relation	satisfaction	
Desk Environment	Not available	Available	
Communication environment	Not available	Available	...(c)
Charging facilities	Not available	Available	
Exercise facilities	Not available	Available	
View facilities	Not available	Available	
Conversation space	Not available	Available	
Eating and drinking space	Not available	Available	
Luggage space	Not available	Available	

Table 5 “Everyday intentions” and “mobility situation” items

Intention	Situation
To focus and concentrate	Commuting to work or school
To relax	Returning home
To exercise	Outing (day trip)
Sightseeing	Travel (more than 2 days)
I want companionship	Other travel
No particular intention	

Step 2 : m means of mobility or facility in the vertical axis (O_1, O_2, \dots, O_m)

Examples of means of mobility: walking, regular route buses, railways, taxis

Examples of facility: favorite cafes, parks where you can exercise

Step 3 : Apply metadata based on groups corresponding to the various means of mobility or facility.

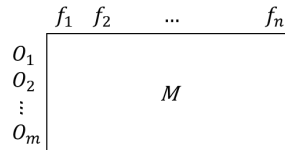


Figure 2 Expression of metadata based on matrix data M

2.1.1 Means of mobility feature value vector expression

The structure of the feature value space for means of mobility is shown in Table 2.

(a) **Definition of “price, required time, exercise consumption” from which quantitative information can be obtained**

If there are m target means of mobility for arbitrary interval

Actual measured value vector A_{column}
 $= (a_1, a_2, \dots, a_m)$

Feature value vector $V_{column} = (v_1, v_2, \dots, v_m)$

At this time, feature vector elements v_i for each means of mobility are defined as Eq. (1). Each of these are normalized in a non-linear way. This can be expressed as the feature value based on the market of an arbitrary interval. Further, by setting a threshold value, the impact of the outliers in the relative calculation to be described later can be minimized.

$$v_i(a_i) := \frac{a_i - \mu_a}{\sigma_a} \quad (-1 \leq v_i \leq 1) \quad (1)$$

$$\text{Mean value } \mu_a = \frac{1}{m} \sum_{i=1}^m a_i$$

$$S.D \sigma_a = \sqrt{\frac{1}{m} \sum_{i=1}^m (a_i - \mu_m)^2}$$

(b) Definition of “human congestion, punctuality, action/operation manipulation” vector data, based on qualitative information

This is defined in the feature correspondence table shown in Table 3.

(c) Definition of other {desk environment, communication environment, charging facilities, exercise facilities, observation facilities, conversation space, eating/drinking area, luggage space} vector data

Vector elements are indicated in Table 2 as 1.for “Yes” and 0 for “No”.

2.1.2 Means of facility feature value vector data expression

From the related studies described in 1.3.3, we can see that according to a survey by the Public Relations Office of the Cabinet, there are the five facility goal factors set as objectives to be achieved through mobility, including trips taken by people: (1) desire to enrich knowledge, (2) desire for personal (mental/physical) growth, (3) desire to alleviate tension, (4) desire to do something fun, and

(5) desire to deepen human relationships [9]. The structure of the feature value space for means of activities including these are shown in Table 4.

(d) Definition of facility goal factor vector data

The vector elements for each of the respective means of facility were set to 1 for “Satisfied”, 0.5 for “Quite satisfied”, or 0 for “Unrelated”, using information from Google Maps[10], Jalan[11], and Trip Advisor [12]for reference.

(c) is the same as in 2.1.1.

2.2 Sensitivity to context and query generation operators

Here, we shall demonstrate the query generation method in this system. When humans pass their days accompanied by mobility, if we categorize these into intrinsic motivation and extrinsic motives, the former can be replaced by “everyday intentions” of “wanting to achieve an arbitrary goal”, and the latter as “mobility situations” in which you are already placed. Therefore, for the queries we used a “sensitivity to context vector2” that merges these two.

“Everyday intentions” can be separated by desires in daily activities, whereas “mobility situations” can be separated by the goal of the mobility, and these are shown in Table 5.

To give some examples, queries can be expressed as “commuting to work or school for purpose of exercise”, “coming home to relax”, “going out to interact with friends” or “traveling for sightseeing”.

At this time, the size of the elements in each vector are set to their related size corresponding to (a),(b),(c),(d) in Tables 2 and 4.

Here, if we set the elements of each vector to *Daily intentions(Intention)* $I = (i_1, i_2, \dots, i_n)$,

Movement situations(Situation)

$$S = (s_1, s_2, \dots, s_n)$$

the query generation operators can be defined as in Eq. (2).

$$\text{query} = F(S, I) \text{ OR } F(I, S) = (q_1, q_2, \dots, q_n) \quad (2)$$

$$F(A, B), A = (a_1, a_2, \dots, a_n), B = (b_1, b_2, \dots, b_n)$$

At this time,

$$IF a_k \cdot b_k \geq 0 THEN q_k = \max_{0 \leq k \leq 1} \{|a_k|, |b_k|\}$$

$$ELSE a_k \cdot b_k < 0 THEN q_k = a_k$$

The merged sensitivity to context vector has the maximum feature values for everyday intentions and mobility situations. However, for (a),(b) in Tables 2,4, in case the symbols are reversed, a_k , depending on the situation in which the user is placed, may be selected dynamically as prioritized i_k or s_k .

2.3 Distance calculation

We shall perform a distance calculation for the “feature value vector” for means of mobility and facility stored in the database in 2.1 and “sensitivity to context file” generated in 2.2. At this time, the relative quantities for the means of mobility and facility are calculated respectively based on the subspace selection (Figure 3) from one query [13].

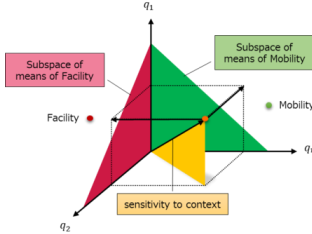


Figure 3 Means of mobility and facility subspace selection

The relative calculation quantity corresponds to feature value weight as a vector size, and as this is calculated with consideration for the orientation of the vector at this time, the inner product is used, and this is defined as in Eq. (3). In an arbitrary service ($service_i$) with means of mobility and facility, if

$$Sensitivity\ context\ V_{query} = (v_{q1}, v_{q2}, \dots, v_{qn})$$

$$Service\ feature\ value\ V_{service_i} = (v_{s1}, v_{s2}, \dots, v_{sn}),$$

$$Relative\ quantity\ C_{service_i}(V_{query}, V_{service_i}) := \frac{1}{68} \sum_{j=1}^n v_{qj} v_{sj} \quad (3)$$

3. Implementation Method

Here, we shall demonstrate the implementation of this system. With the point of departure fixed at the Tohoku Shinkansen South transfer ticket gate of Tokyo station, examples of 4 destinations in the last mile from Tokyo station (Suitengu, TokyoTower, Tsukiji Honganji, Nihon Budokan) were used. The means of mobility were set as the mobility service and mobility route by each destination. Figure-4 shows the example of means of mobility from Tokyo station→Suitengu.

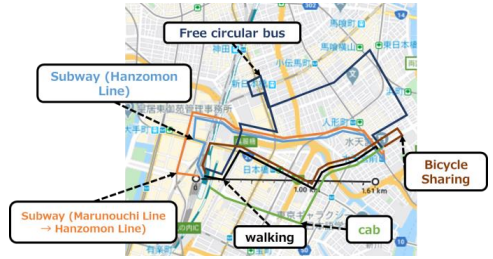


Figure 4 Means of mobility from Tokyo station→Suitengu

3.1 System structure

The structure of this system is as shown in Figure-5, and the flow of this process is indicated by steps 1-7.

- Step 1 :** Acquire input information from UI
- Step 2 :** Generate query based on input information
- Step 3 :** Filter means of mobility candidates based on input information (excess time), business hours, and weather information
- Step 4 :** Perform distance calculation and select optimal solution for means of mobility
- Step 5 :** Filter candidates for means of facility based on the means of mobility and business hours information from the optimal solution in Step 3
- Step 6 :** Perform distance calculation and rank “facility spots” in mean of facility from 1st to 5th place
- Step 7 :** Output combination of means of mobility and facility as information

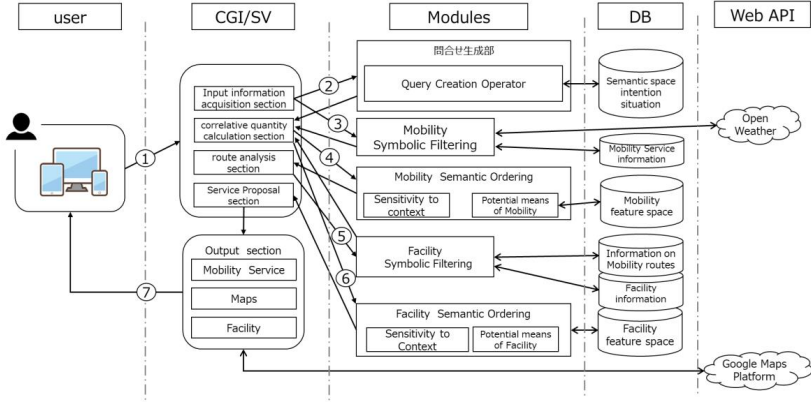


Figure 5 System configuration diagram

Additionally, the flow for the method from query generation shown in the previous chapter to distance calculation, corresponding to this system, is as shown in Figure 6.

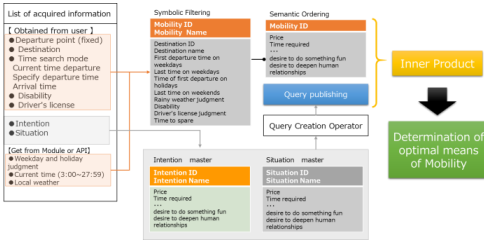


Figure 6 Flow of query generation and distance calculation (means of mobility)

This shows the objective of each filtering processing section. Filtering of the means of mobility in Step 3 prevents the return of unrealistic results to users (example: exclude transportation facilities outside of business hours and bicycle sharing not feasible on rainy days etc.), and lessens burden on the later processing section. In the same way, filtering of the facility spots in Step 5 sets points that are easy to stop by, and so are limited on the mobility routes in Step 4. Commonly, the mobility routes are connected in a straight line. By applying a fixed width to this straight line, this connects in a rectangle shape. Here, by collating with a facility spot

information DB that includes longitude and latitude information, we can exclude facility spots outside of the rectangular area. This method is the Rectangle Mobility Scope Model (Figure 7). In the case of “boarding-type mobility services (trains and buses)”, as it is not possible to stop by on the mobile route, this method is only applied to the transport hub.



Figure 7 Image of Rectangle Mobility Scope Model

The method of calculation in this method is defined as follows. All of it is treated as a plane orthogonal coordinate system. Let us set

$$\text{Start coordinate within route } S = (ng_1, lat_1)$$

$$\text{End coordinate within route } G = (ng_2, lat_2)$$

At this time, the Range is set to variable based on the selection of “everyday intention (Intention)”.

$R[m] = \{\text{"want to dedicate/concentrate"}: 320, \text{"want to relax"}: 240, \text{"want to exercise"}: 400, \text{want to sightsee}: 320, \text{want to interact with friends}: 320, \text{"no particular intention"}: 320\}$

The distance at a longitude/latitude of 1° around Tokyo station is $90.4219[\text{km}]$

That is to say, $r = \frac{R}{90421.9}$. At this time, the coordinate points in Figure-8 are

$$\begin{aligned} S_1 &= (lng_1 - r(\cos \theta + \sin \theta), lat_1 - r(\sin \theta - \cos \theta)) \\ S_2 &= (lng_1 - r(\cos \theta - \sin \theta), lat_1 - r(\cos \theta + \sin \theta)) \\ G_1 &= (lng_2 + r(\cos \theta + \sin \theta), lat_2 + r(\cos \theta - \sin \theta)) \\ G_2 &= (lng_2 + r(\cos \theta - \sin \theta), lat_2 + r(\cos \theta + \sin \theta)) \end{aligned}$$

At this time, $\theta = \tan^{-1} \left(\frac{lat_2 - lat_1}{lng_2 - lng_1} \right)$

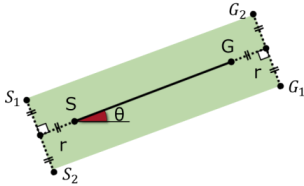


Figure 8 Acquisition of Rectangle Coordinates

Next, we judge the inside and outside of the facility spot points. With the rectangle at the peak $[S_1, S_2, G_1, G_2]$, when we can set an arbitrary facility point to $F = (lng_f, lat_f)$, the side made with F is set to $(l_1 \dots l_4)$. At this time, with the declination in Figure-9 as θ_i , number of rotations wn is defined as in Eq.(4).

$$wn := \frac{1}{2\pi} \sum_{i=0}^3 \theta_i \quad (4)$$

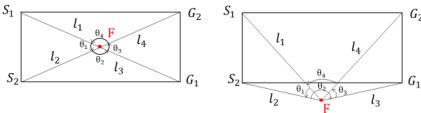


Figure 9 Internal/external judgement

However, with regard to the orientation of declination θ_i , if the respective cross products

$$C_1 = \overrightarrow{FS_1} \times \overrightarrow{FG_1}, C_2 = \overrightarrow{FS_2} \times \overrightarrow{FG_1}, C_3 = \overrightarrow{FG_1} \times$$

$\overrightarrow{FG_2}, C_4 = \overrightarrow{FG_2} \times \overrightarrow{FS_1}$ are negative, the signals are treated as reverse rotations. Therefore,

$IF |wn| > 0$, judged to be inner side ... *true*
 $IF |wn| = 0$, judged to be outer side ... *false*

3.2 Languages/libraries used

The languages used are JavaScript, jQuery, Python3.8.2, and PostgreSQL12.2; for the display section, HTML, CSS, Python libraries used are flask, psycopg2, numpy, pandas, matplotlib, scipy, googletrans, json. For the Web API, we used OpenWeatherMap(weather information at a particular point)[14], JapanHoliday(public holidays etc. announced by the Cabinet)[15], and for depicting map information, Google Maps Platform[16] was used.

3.3 Input/output data

The input values are “everyday intentions”, “mobility situations”, “time restrictions”, “destination”, and commands issued with these. “Everyday intentions” and “mobility situations” are selected from a list. Additionally, for “time restrictions”, when the arrival time is specified, the list is filtered down to means of mobility that allow you to arrive on time. “Destination” is selected from the map.

The optimal mobility service and mobility route to the destination is returned as map and text information. Additionally, the recommended facility spots that exist on the mobility route are returned in ranking format corresponding to the values in Eq.(3), as map and text information.

The input screen and output screen are as shown in Figure-10.

4. Evaluation

In this section, we shall show the results and observations from three evaluation tests conducted with the objective of verifying the effectiveness of this method and system.

4.1 Coverage rate of search targets

Here, we measure the coverage rate of the search targets in relation to expressed sensitivity to context (generated queries and selected subspace). The aim of this verification is to confirm, in relation to the queries that can be expressed in this system, whether th



Figure-10 Input screen (left) and output screen (right)

ere is any polarization in the search targets in the DB, and its appropriateness as a test collection, making it easy to perform the evaluations in 4.2 and 4.3 below. The coverage rate is as defined in Eq.(5).

$$Coverage\ rate\ (Coverage) := \frac{Conformity}{q} \quad (5)$$

q : Number of queries that can be expressed (29 in this system)

The coverage rate and data space distribution for each destination is shown in Table 6. This indicator, when performing the correlation amount calculation, is used to confirm that data with a high degree of similarity is output.

Table 6 Coverage rate in the data space

Destination	Mobility	Facility1st	Facility2st	Facility3st	Facility4st	Facility5st	Facility Total
Sutenju Shrine	1.000	1.000	1.000	0.931	0.931	0.897	0.952
Tokyo Tower	0.966	1.000	1.000	1.000	0.966	0.759	0.945
Tsukiji Honganji Temple	0.966	1.000	1.000	0.966	0.931	1.000	0.979
Budokan	1.000	1.000	0.966	1.000	0.966	0.828	0.952
Average	0.983	1.000	0.991	0.974	0.948	0.871	0.957

4.2 Fill-rate for queries

This system generates queries corresponding to the sensitivity to context of mobile users. We surveyed, using a questionnaire survey, to what extent the generated queries represented user sensitivity. The questionnaire included respective items (Table 7) for “everyday intentions” and “mobility situations”, and for the vector data, the ordinary data of (a),(b),(c),(d) in Tables 2, 4 were measured. The average value for 34 samples is shown in Figures 11, 12. The question items for intentions were answered with a 4-stage rating level {think so, somewhat think so, don’t really think so, don’t think so}, and the situation question items were answered with the 4 levels of {applies, somewhat applies, does not really apply, does not apply}.

Table 7 “Everyday intention items (left) and “mobility scenario” question items (right)

Questionnaire	Questionnaire
desire to reduce travel costs.	Low travel costs
desire to shorten my travel time.	Short travel time
desire to exercise	High exercise consumption
desire to arrive on time	High punctuality
desire to drive and control	Driving control
desire a desk	Has a desk
desire Wi-Fi	Wi-Fi is available
desire charging facilities.	Charging facilities
desire exercise equipment.	Exercise facilities
desire an environment with a view.	Environment with a view
desire an environment where I can eat and drink.	Environment for eating and drinking
desire Luggage space	Luggage space
desire to enrich knowledge	
desire to alleviate tension	
desire to do something fun	
desire to deepen human relationships	

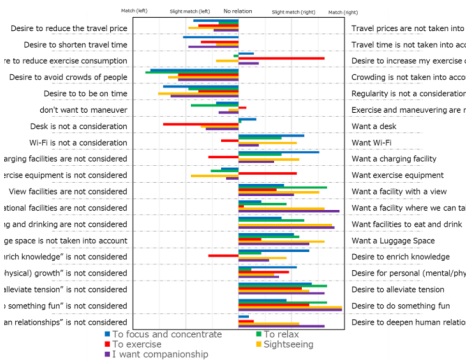


Figure 11 “Everyday intention” common data

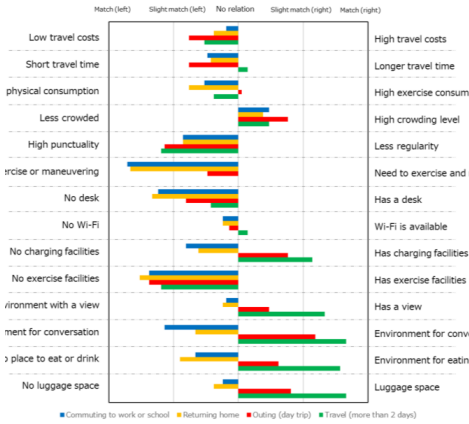


Figure-12 General data of “mobility situations”

4.3 Fill rate in regard to the output value

We surveyed the appropriateness of the calculation method in Eq. (2).(3) with regard to the search results in this system. This was used 3 times per person by 4 user test subjects, and a comparison of the query and output values each time were rated, as shown in Table 8. Each query, name of means of mobility, and facility spot genres 1-5 are displayed in Table 9. A comparative evaluation in relation to these search results is shown in Figures 13, 14, and 15.

Table 8 Fill-rate survey items in search results

Grading target	Rating Item			
The results of the transportation system	satisfied	match	slight match	unsatisfactory
Recommended spots (1st to 5th place respectively)	Query and genre comparison (reference and location comparison)	easy to approach	slight discrepancy	discrepancy
Operability	easy to understand	easy to approach	slight discrepancy	discrepancy
Diagram	easy	Somewhat easy	Somewhat difficult	difficult
Characters	easy	Somewhat easy	Somewhat difficult	difficult

Table 9 Queries and search results during experiment

Number of experiments	Subject A Male in his 30s			Subject B Female in her 30s			Subject C Female in her 60s			Subject D Male in his 60s		
	1	2	3	1	2	3	1	2	3	1	2	3
Intention	Sightseeing	To relax	companionhip	To exercise	Sightseeing	To exercise	To relax	companionhip	Sightseeing	To exercise	To relax	Sightseeing
Situation	Outing	Returning home	Returning home	Outing	Travel	Outing	Returning home	Returning home	Outing	Returning home	Returning home	Outing
Destination from Tokyo Station	Tokyo Tower	Need urgent train	Suitengu	Need urgent train	Suitengu	Suitengu	Budokan	Tokyo Tower	Suitengu	Tokyo Tower	Suitengu	Tokyo Tower
Means of transportation Result	Share Cycle	cab	cab	scheduled bus	walking	subway	Share Cycle	Share Cycle	Free circular bus	cab	Share Cycle	Share Cycle
Facility Spot Genre Result No.1	park	cafe	cafe	cafe	park	cafe	park	message	park	message	park	park
Facility Spot Genre Result No.2	tourist attraction	cafe	message	cafe	cafe	cafe	park	park	tourist attraction	cafe	cafe	cafe
Facility Spot Genre Result No.3	park	library	cafe	cafe	cafe	cafe	park	sports gym	tourist attraction	cafe	cafe	tourist attraction
Facility Spot Genre Result No.4	tourist attraction	Shared Office	cafe	cafe	tourist attraction	library	cafe	sports gym	tourist attraction	tourist attraction	cafe	park
Facility Spot Genre Ranking: 5th	tourist attraction	sports gym	cafe	cafe	sports gym	cafe	tourist attraction	sports gym	message	Shared Office	cafe	park

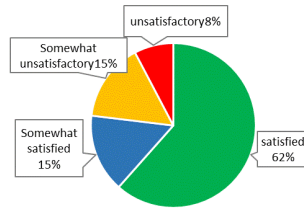


Figure-13 Means of mobility evaluation results

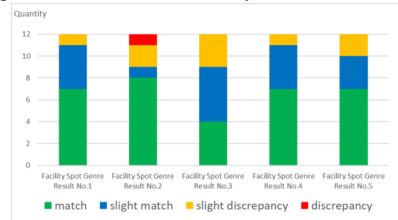


Figure-14 Facility spot evaluation results (Facility genres related to queries)

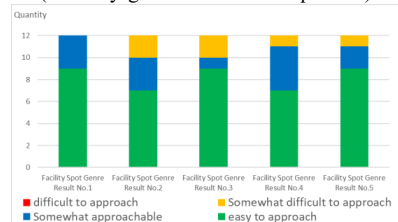


Figure-15 Facility spot evaluation results (Locations related to mobility routes)

4.4 Observations

Based on the results of the coverage rate related to the search results, it is suggested that there is little polarization in the distribution of the data space. In particular, with regard to the means of mobility, the intention and situation context is reflected, and if the types of mobility services increase moving forward, it is considered that the coverage rate will further increase. For the facility spots, genres with low association were seen in some of the lower ranking spots. When filtering

g this in the Rectangle Mobility Scope Mode l, it is thought that data with a high similarity to the sensitivity to context is discarded or depleted due to mobility routes with few facility spots. Measures that promise to improve this are increasing the facility spot data quantity and setting a threshold for correlation amount.

Next, in terms of the results of the fill-rate for queries, we show the extent to which sensitivity of mobile users can be expressed, as queries, from ordinary data. For items where a particular difference can be seen, it is considered that with a distance calculation this is an effective dimension/axis. Additionally, by applying ordinary data to “everyday intention” and “mobility situation” vector data, this contributes to the generation of highly objective queries.

Finally, with regard to the result of the fill-rate for the output value, we were able to achieve a constant rating for both the means of mobility and the facility spot. Based on results, the effectiveness of the series of calculation methods including the query generation operator group was demonstrated. Additionally, some opinions were received from the test subjects. Favorable opinions included that “it was interesting how means of mobility and tourist sites I was not aware of were proposed”, “going out has become more enjoyable”, and “I want to use it when traveling to regions I don’t know.” On the other hand, points that could be improved on were that there was a desire for personal and hobby-related factors matters to be reflected in the means of mobility, such as “being difficult to ride a bicycle due to being in one’s 60s”, and “when dedicating oneself to, or concentrating on, something, it is acceptable to take a longer way around so a more coherent description of movement time would be desirable.” For matters related to means of facility, there was also a desire for psychological reassurance, stating that “I want to be conducting the activity around the destination area before meeting somebody in that location” and “the facility method changes according to how much spare time you have”. It was considered that more information was desirable, and it was stated that “they would like to know more about the feel of

it and what kind of thing it is through photos and review information.”

5. Summary

In this study, we have been able to propose multi-modal means of mobility and means of facility on mobility routes that go beyond the framework of public transportation facilities, and that could be known through transfer navigation apps, corresponding to the sensitivity to context of mobile users. By forming this combination, it is possible to merge “mobility” and “everyday” in a temporal, spatial, and semantic way, and through the seamlessness of “mobility \leftrightarrow everyday”, it is possible to devote oneself to intellectual activities aimed at accomplishing what we want to achieve as human beings.”

In terms of the future outlook, we aim to make it possible to combine route searches with mobility, not only in the form of “one last mile”, but also compatible with free movement over a wider range, and allow mobile users to select from a variety of mobile services and mobility routes. Additionally, proposals for the means of facility should not be limited to facility spots, but by combining them with event information (e.g.: blood donation, picking up garbage, roadside live music performances etc.), it will be possible to contribute to achieving SDGs and construct this as a platform linking various issues.

We are promoting this study, with the aim of achieving a world in which there is sustainable mobility, people capable of mastering this, and where mobility situations required in everyday life can be fulfilled in a natural way.

Reference Literature

- [1] Nissay Asset Management Co., “70% of the world’s population in 2050 shall be city dwellers”, Market Report, May 24, 2018
- [2] Ministry of Land, Infrastructure, Transport and Tourism, “Government/Private Enterprise Collaboration Platform for Smart Cities” <https://www.mlit.go.jp/scpf/>.
- [3] Katsuji Imada, Forum for a collaboration platform with pollution materials utilized in SDGS through the power of citizens in Tokyo, December 15, 2018.
- [4] METI : Study Group on New Mobility Se

services that can make IoT and AI possible - "Activation of new mobility services", https://www.meti.go.jp/shingikai/mono_info_service/smart_mobility_challenge/20190408_report.html.

- [5] Ministry of Land, Infrastructure, Transport and Tourism, Round-Table Discussion on New Mobility Services for Urban and Regional Areas, "Overview of Interim Report", https://www.mlit.go.jp/sogoseisaku/transport/sosei_transport_fk_000089.html.
- [6] Yasushi Kiyoki, Takashi Kitagawa, Takamasa Hayama: "A meta-database system for semantic image search by a mathematical model of meaning," ACM SIGMOD Record, Vol.23 Issue4, December 1994.
- [7] Hayashi, Y., Y. Kiyoki, and X. Chen. "A Combined Image-Query Creation Method for Expressing User's Intentions with Shape and Color Features in Multiple Digital Images." *Frontiers in Artificial Intelligence and Applications*. Vol. 225. IOS Press, 2011. 258-277. Web.
- [8] Atsushi Tsuya, Creating diverse driving plans using sensitivity rating results of people's motives when traveling, 2011, Vol. 10, No.3 p. 433-443.
- [9] Cabinet Public Relations Office : Free time and tourism, Monthly Public Opinion Survey, No.11, 2004, pp.3-84.
- [10] Google Maps, <https://www.google.co.jp/maps/>.
- [11] Jalan Tourism Guide, <https://www.jalan.net/travel/>.
- [12] Trip Advisor, <https://www.tripadvisor.jp/>.
- [13] S. Kurabayashi, N. Ishibashi and Y. Kiyoki, "A Multidatabase System Architecture for Integrating Heterogeneous Databases with Meta-Level Active Rule Primitives", *Proceedings of the 20th IASTED International Conference on Applied Informatics*, pp478-387, 2002.
- [14] OpenWeatherMap, <https://openweathermap.org/>.
- [15] JapanHoliday, <https://github.com/suzuki-shunsuke/japanese-holiday-api>.
- [16] Google Maps Platform, <https://cloud.google.com/maps-platform/>.

Modeling the Variety of Trip Opportunities in Recreational Route Networks

András J. MOLNÁR

*SZTAKI, Institute for Computer Science and Control,
Kende u. 13-17. H-1111 Budapest, Hungary;*

*Viator Association for Hiking and Culture,
Horánszky utca 20. H-1085 Budapest, Hungary;*

E-mail: modras@sztaki.hu, molnar.andras.jozsef@gmail.com

Abstract.

In this paper, we deal with the question of how the variety of trip opportunities can be modeled in - possibly complex - recreational trail networks (such as hiking paths or cycling ways). In order to quantify the variety of possible loop trips starting from specific trailheads (starting nodes accessible from outside the network) and the variety of connecting trips between specific origin-destination pairs, two novel measures of Loop Trip Variety Index (LTVI) and Connecting Trip Variety Index (CTVI) are proposed preliminarily and informally in [12], respectively, in the frame of assessing the impacts of some recent trail network developments. This paper establishes the formal definitions of improved variants of these measures, shows their well-definedness, presents the algorithms of their computation, investigates on their properties and benefits, and gives reasons of how and to what extent they can be treated as models of trip variety. Possible uses, application areas and future improvements are sketched especially for visitor management planning and profile-based trip recommendation systems.

Keywords. tourism and recreation, route network graphs, knowledge and information modeling, route planning, facility management, trip variety, network analysis

1. Introduction and Related work

Trails, such as hiking paths or cycling ways are used for outdoor activities and are available both for local recreation as well as reaching or exploring tourist destinations or being destinations themselves. They usually do not exist in isolation, as they form networks. Trip opportunities are determined by the network structure of specific modes of activity (hiking, biking, etc.) and the possible starting points (trailheads, such as parking places, bus stops, railway or ferry stations), trip destinations, and points of interest (POIs).

Current apps or systems provide either selected edited, prepared trips without any or minor variations, such as [14,2,1,10], or a free navigation over the network, as in most mapping systems. If a user does not know the area or is not professional on maps, it might be difficult to find or plan a suitable trip route if the prepared, recommended trips (usually

chosen with some subjectivity and with random overlaps) do not fit completely to one's profile. Based on these information, especially, planning a location for a longer holiday with multiple overnight stays can be difficult. Furthermore, if something changes in the network, its impact on trip opportunities of different target groups have to be assessed.

The field of recreational ecology has been emerged in the recent decades for studying the opportunities and impacts of field activities. The Recreational Opportunity Spectrum is one of the most utilized concept, which is based on local features and distances of specific types of features [6]. These works usually focus on local features and layout, environmental impacts at a global scale or along a longer trail, but not on assessment of structural properties of networks [16].

Works such as [15,4,11] focus on conceptual modeling of mobility data and infrastructure. [9] analyses patterns in a network based on user trajectories. Prototypes of recommendation or assessment systems has been developed mainly for urban context but also for outdoor activities, some of them involving the social dimension [13,17,19].

Graph theory has been extensively applied for transportation networks. Centrality measures such as node degree, closeness, betweenness centrality and others are effectively applied for them [8,5]. For outdoor trails, the situation is different. Trail usage differs from transport networks, and needs specific metrics to assess network plans and changes, and assist users to find attractive opportunities. It is a somewhat 'reversed situation' compared to transport networks. The purpose here is not to bring people from their particular original locations to (regular) desired destinations in the most effective way possible, and - in most cases - not to help users find the shortest way between two arbitrary locations, but to provide a pleasurable activity for a specified time frame, while considering the impact of these activities on the environment as well.

It has been revealed [3] that 60-64% of hikers prefer returning trips over linear hikes, and if we add the higher attractiveness of variation over repetition (at least for most users), minimizing necessary back-and-forth (dangling) sections and detours is reasonable by scoring circle trips and parts higher than those which need to be walked along the same way back. As different users have different aims and characteristics in their activities, developing a profile-based model seems to be beneficial.

Quantifying trip variations around a location, set of locations or between (sets of) locations can be trivially done by counting the number of possible trips in a length range. However, this number will show a combinatorial explosion, and because two different trips may have common sections, it will not be a useful metric to reflect on the actual variety of trip opportunities. If the network layout offers only trips without shared sections (such as a flower graph), counting these independent trips can be ideal. When, however, sections are shared, they must be somehow downgraded, and ideally, the measure of independent trips should be generalized, with looking at the maximal number of 'covering' trips of the reachable subgraph. Actual trip length should not make a difference if it falls into the range of the user's preference.

Modeling the variety of trip opportunities at nodes or sets of nodes of a trail network will provide a better understanding by an explicit expression of the knowledge related to possible trips contained by the network structure. It gives an instrument to help answering the following questions in particular:

1. For trail users with specific profiles: Where to go outdoors? Which is the variety of opportunities given by (a) specific trailhead(s) or origin-destination relations? Where can I find more variety if I consider staying somewhere for a couple of

days? How many days shall I there to get a more or less full 'picture' of the area by a series of trips and how many repetitions of path sections I am likely to face?

2. For trail network managers: What is the impact of the existence or non-existence of a particular trail section (link) or set of sections in a network? Where and how shall we invest (to make new or upgraded trail sections) if we want to give more opportunities to trail users with specific profile(s)? In which locations shall we invest (trailheads, lodges) to improve facilities, transportation and their capacity if we want to focus on hubs with the best connectivity and variety (potential attractiveness for returning visitors)? How will the trip opportunities change or degrade across a larger area if we close or reroute a particular trail section?

Similar initiatives have recently begun in an urban context, such as [7], in which the effect of possible short additions to the bicycle network is investigated in terms of global city reachability. However, variety is not yet considered there.

The article [12] is proposing two novel metrics: the *Loop Trip Variety Index (LTVI)* and the *Connecting Trip Variety Index (CTVI)* along with two respective auxiliary measures of the *Maximal Covering Loop Trips (MCLT)* and the *Maximal Covering Connecting Trips (MCCT)*, in order to quantify the variety of possible loop trips starting from specific trailheads (starting nodes accessible from outside the network) and the variety of connecting trips between specific origin-destination pairs, respectively. These are proposed preliminarily and informally, in the frame of assessing the impacts of some recent trail developments in a real-life network scenario with some promising results.

This paper aims to contribute to the thorough discussion and formalization of these measures as models of trip variety, as proposed in [12]. We put them into a user-profile-based setting, and extend their applicability to sets of network nodes instead of single nodes. We add an index to reflect on back-and-forth sections (STVI), which were not considered at all, and include profile-based preferences such as the necessity of preferred POI visits in trips, which have substantial impact on the way of computation. A simple informal method of LTVI calculation is replaced by a full-fledged index computation algorithm, giving results of all 3 index components (CTVI,STVI,LTVI).

2. Modeling User Profiles, Trail Networks and Trips

Example 1 See Figure 1 taken from [11] with some modifications. Assume these are all walking paths. Line styles indicate difficulties, and the dotted lines are proposed extensions and improvements. Numeric labels denote section length in kms. *TNG* denotes the graph without these additions, while *ExtTNG* includes the planned sections as well.¹ Connected points of interests and trailheads are shown with pictograms, while capital letters denote nodes. We define two user profiles by personas, where both have no restrictions on the access mode or the format of their trips:

Adam or simply, *Pa* or *a*, who walks 5-20 km trips along both easy and difficult trails, has a preference for natural and landscape-related POIs (e.g. scenic spots).

Betty or simply *Pb* who walks shorter and easier trips, 3-10 km along easy sections only, but has no POI preference, likes walking for its own sake.

¹The dotted line in parallel to section *QK* means the difficult section is to be upgraded as an easy section, so *QK* counts as an easy section in *ExtTNG*.

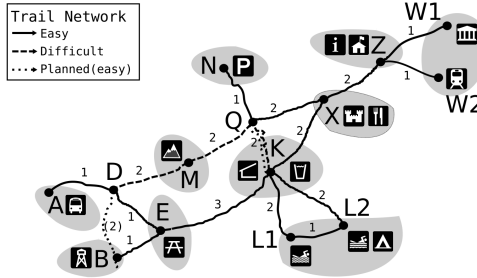


Figure 1. Network example with distances

Questions like the following may be asked for either profile, with or without the planned sections (how beneficial those will be for different profiles and cases):

- The variety of trips starting from a given node and returning there, such as N (parking lot), K (lodge) or A (bus stop).
- The variety of trips connecting two specific nodes or sets of nodes, such as $A \rightarrow W2$ (public transport), or $W = W1, W2 \rightarrow L = L1, L2$ (city to lake & camp).

We characterize the needs and preferences of trail users by the following fields:

Activity mode Hiking or biking, or any other modality of activity exercised along the trails. A finite set $ActModes$ is assumed to be given with the possible values.

Trail and trip type/difficulty What type of trail(s) is the user ready to go along a trip, in accordance to one's equipment and preparedness (technical difficulty level). We assume the finite set of possible values (which may be a direct product of multiple relevant trail section parameters) are given by $TrailTypes_{act}$ for each $act \in ActModes$. Furthermore, the set $TripTypes_{act}$ for each $act \in ActModes$ proposes a set of typical trip types defined by a function $allowedTrailTypes_{act}(tripType) \in 2^{TrailTypes_{act}}$. The universal trip type set is $TripTypes = \bigcup_{act \in ActModes} TripTypes_{act}$.

Trip length range the user is ready to take (min, max), in km. We assume a finite set of typical ranges $TripRanges_{act}$ is given for each $act \in ActModes$, and $TripRanges = \bigcup_{act \in ActModes} TripRanges_{act}$. The actual min and max distances of range are available using functions $min, max : TripRanges_{act} \rightarrow \mathbb{R}^+$.²

Returning trips only A boolean value determining whether the user wants to take trips only which return to their starting point, or ending possibly somewhere else.

Access mode How the user gets to the trailhead (the starting point of the trip). Mainly car or public transport. Possible values are given as the finite set $AccModes$, including

²In practice, the distance is usually enhanced by the ascent and descent in meters by section and forms a basis for calculating the ideal time for completing the trip. We push the elevation and time issues under the trail/trip type/difficulty part and keep the distance range in order to keep simplicity here; which is, on the other hand, independent of the direction of moving along a trail.

a special value \top meaning no access mode is given and the trip can start and end in any node of the network.

POI preference A finite set of types of points of interest relevant for the user and wanted to be included in one's trip. Types are defined by $PoiTypes$, and a subset is given for each profile, meaning the user wants to undertake such trips only which include visiting of at least one point of interest whose type is in the set (it can be the starting or ending point of the trip as well). Detours are taken only to such types of POIs. The empty set means no preference, i.e. no restriction on the user trips.

Definition 1 *The characteristics of trips a user is willing to take is defined by the **Trail User Profile**: a tuple $p = (actMod, tripType, tripRange, retOnly, accMod, poiTypes)$ over $ActModes \times TripTypes \times TripRanges \times \{\perp, \top\} \times AccModes \times 2^{PoiTypes}$ where $tripRange \in TripRanges_{actMod}$ and $tripType \in TripTypes_{actMod}$.*

Example 2 *The user profiles of Example 1, given the single activity walk and trip types by allowedTrailTypes = {easyTrip \mapsto {Easy}, anyTrip \mapsto {Easy, Difficult}} are:*

- $Pa = (walk, anyTrip, (5, 20), \perp, \top, \{mountain, spring, lake, lookout, in\}ocentre)$
- $Pb = (walk, easyTrip, (3, 10), \perp, \top, \emptyset)$

A single user may have multiple profiles being active depending on the actual situation (e.g. time availability).

We assume the trail network is given in form of an undirected graph (assuming most trails are bidirectional, adaptation to directed sections is a future issue), a classical geospatial routable graph structure enhanced by specific labels corresponding the user profiles in the following way.

Definition 2 *A **Trail Network Graph** is a labeled graph $TNG = (TN, TS^+, link^+, pn, ps)$ where $TN = tnodes(TNG)$ is the finite set of trail nodes, $TS^+ = tsections^+(TNG)$ is the finite set of (directed) trail sections, $link^+ : TS^+ \rightarrow TN \times TN$ is a function defining the linkage of nodes by the sections. Furthermore, $pn = (geoCoords, accModes, poiTypes)$ and $ps = (geoPath, length, actModes, trailType)$ are tuples of functions defining properties of nodes and sections, respectively.*

The sections are considered to be bidirectional (undirected graph) with each section represented in a specified direction in TS^+ . Therefore, we assume a natural extension of the above notations to a symmetric setting where each section and its reversal is present with a reversed linkage extension. We omit the $^+$ upper indices for this extension.

The function rev over TS defines the reversal of each section. The link function can be specified as a tuple of functions (start, end). With these notations, we can denote a TNG with a tuple $(TN, TS, (start, end), rev, pn, ps)$ as well.

Properties ps of each trail node n: $geoCoords(n)$ is a pair of geospatial coordinates (latitude-longitude), $accModes(n) \subseteq AccModes \setminus \{\top\}$ defines the access modes for trail-head nodes (possible starting point of trips), $poiTypes(n) \subseteq PoiTypes$ defines the types of points of interests located at n.

Properties ps of each trail section s: $geoPath(s)$ is the linestring geometry, $length(s) \in \mathbb{R}^+$ is the length in km, $actModes(s) \subseteq ActModes$ defines the possible activity types, $trailType(s, a) \in TrailTypes_a$ defining the trail type/difficulty properties for

each $a \in \text{actModes}(s)$.³ For topological consistency, we require the coordinates of start and end points of sections match the nodes they connect.

Furthermore, we assume the nodes and sections are uniquely identified by their geographical coordinates, so one can 'point out' a node or segment by its coordinate position(s). Besides that, a name and/or a technical identifier can be given but it is not relevant for our discussion.

For simplicity, we use the nodes in our examples to identify the trail sections, as we do not have parallel edges in our network graph.

Given a set of activity and/or trail types, a trail network graph can be restricted by filtering its sections valid for the given activity and trail types, respectively. Given a set of access modes and/or poi types, the respective nodes of a trail network graph can be selected. The **Mode-specific Trailheads** for an access mode $accm$ are denoted by $\text{Trailheads}(TNG, acc)$. The **Locations of POIs (or simply, POIs)** are denoted by $\text{Poi}(TNG, pts) = \{n \in \text{tnodes}(TNG) \mid \text{poiTypes}(n) \cap pts \neq \emptyset\}$ and, in general, $\text{Poi}(TNG) = \{n \in \text{tnodes}(TNG) \mid \text{poiTypes}(n) \neq \emptyset\}$.

Based on the trail network graph-related definitions, we give definitions of *trips*, which users can undertake in a trail network graph:

Definition 3 A *trip* t in a trail network graph TNG is a sequence (s_1, s_2, \dots, s_n) of connected (directed) trail sections of TNG where $\text{start}(s_{i+1}) = \text{end}(s_i)$ for each $i \in \{1, \dots, n-1\}$.

- The **length** of the trip is the sum of lengths of its sections along the sequence.
- The starting and ending nodes of the trip are defined as $\text{start}(t) = \text{start}(s_1)$, $\text{end}(t) = \text{end}(s_n)$, respectively.
- The set of all possible trips of TNG is denoted by $\text{Trips}(TNG)$. Since repetitions in trips are allowed, this is not a finite set if the graph contains any sections.
- The subgraph of TNG defined by the nodes and sections of a set of trips TripSet is denoted by $\text{TripG}(TNG, \text{TripSet})$.
- Consistently with referring trail sections by their start and end nodes (see above) we use a sequence of trail nodes to describe trips if no parallel edges are present.
- We define specific types of trips or parts of trips in the following.

Definition 4

- Two (or more) trips are called **independent trips** if they have no sections (pair-wise) in common, and no reversal of a section occurs in (any of) the other trip(s).
- A section or sequence of adjacent sections in a trip is called a **repeated part** if it occurs at least twice in the same trip (in the same direction).
- A **simple trip** is a trip without any repeated sections. Where not stated explicitly otherwise, we are considering simple trips only from now on.
- A **back-and-forth part** of a trip is a section (or sequence of adjacent sections) whose reversal occurs along the same trip (not necessarily directly following it).
- A **dangling part** in a trip, a.k.a. **spike**, is a sequence of sections in a trip where the hiker walks back the same way in reversed direction immediately. The turning point of a dangling part (spike) is its middle node.

³Properties ps of a reversed trail section are equal to the properties of the original section except $\text{geoPath}(\text{rev}(s))$ which is the reversed sequence of $\text{geoPath}(s)$.

- A **circle** is a single section or a connected series of sections along a trip (not necessarily following directly one another) if it returns to its starting point at its end and no node or section (or its reversal) is repeated along it.
- A **loop** is a single section or a series of subsequent sections in a trip, or a full trip, if it returns to its starting point at its end and contains at least one circle.⁴ A loop may contain circles and back-and-forth sections (even including spikes as well).
- Two specific types of loops are identified: an **8-shaped-loop** has a circle in it, which being removed from the loop results solely another circle. A **P-shaped-loop** is a loop having a circle in it, which being removed from the loop results a spike leading originally to the circle.
- A trip is a **returning trip** if $start(t) = end(t) = n$. A returning trip is either a loop trip or a **non-loop returning trip**, containing only back-and-forth sections.
- A non-returning trip is called a **connecting trip**.
- A **direct trip** is a connecting trip not containing any loops or spikes in it.
- A **visiting trip** in reference of a specific set of preferred nodes $VN \subseteq TN$ denoted by $VTrip(t, VN)$ is a trip having at least one node being a member of VN . It can be the starting or ending point of the trip as well.
- Similarly, a **strict visiting trip**, denoted by $StVTrip(t, VN)$, is a visiting trip in which a dangling section is allowed only if its turning point is a member of VN .
- Given $StartTN, DestTN, VisTN \subseteq TN$, the **set of possible trips** between $StartTN$ and $DestTN$ is $TripsFromTo(TNG, StartTN, DestTN)$, is the set of trips t in TNG for which $start(t) \in StartTN, end(t) \in EndTN$ (not necessarily simple). $STripsFromTo(TNG, StartTN, DestTN)$ is defined in a similar way, denoting simple trips only. In order to shorten the notations, if any of the sets $StartTN, DestTN$ is a singleton, it can directly be replaced by its member; and if it equals TN (thus meaning no restriction) it can be replaced with the marker \top .
- Given $minL, maxL \in \mathbb{R}^+ \cup \{\infty\}$, the **set of possible trips of length range** between $minL$ and $maxL$ is denoted by $LRangeTrips(TNG, (minL, maxL)) = \{t \in Trips(TNG) \mid minL \leq length(t) \leq maxL\}$. $SLRangeTrips(TNG, minL, maxL)$ is defined in a similar way, denoting simple trips only.
- Given a network (sub)graph TNG . A set of trips $TSet$ is a **cover** (or **covering trip set**) of TNG if each section s of TNG appears in at least one of the trips in $TSet$, and each trip in $TSet$ has at least one section not occurring in all other trips of the set. A **maximal cover** is a cover with the maximal number of trips. This concept can be refined by a predicate P as a trip property where only trips of that property are allowed and sought for the (maximal) cover.

Example 3 With the above concepts:

- Trailheads of our network in Figure 1 are $\{A, N, W2\}$ (access points to transport), however, trips are allowed to start at other nodes as well, for instance, at nodes $K, L2$ (lodge or campsite) or $W1$ (city residential area).
- The trip $t1 = (N, Q, X, K)$ is a simple direct connecting trip (although not the shortest one), while $t2 = (N, Q, X, Z, X, K)$ is a simple connecting trip which is not direct. $t2$ has a back-and-forth part (a spike) (X, Z, X) and has no loops, is a strict visiting trip w.r.t. POI type in focentre, and a visiting trip (but not strictly) w.r.t.

⁴Note that loops usually defined as single reflexive links in graphs. We use the term *loop* in a wider meaning.

parkinglot, castle, restaurant, lodge, spring. t_1 is a strict visiting trip for these POI types and is not a visiting trip for type infocentre.

- The trip $t_3 = (K, Q, X, Z, X, K, L1, L2, K)$ is a simple returning trip with circle parts (K, Q, X, K) and $(K, L1, L2)$. The latter is a spike-free loop part of it, another loop part is (K, Q, X, Z, X, K) , they form a 8-shaped loop together, and the whole trip can be called as a loop trip. It is independent of the trip $t_4 = (K, E, B, E, K)$, which is a non-loop returning trip (it is a spike). If we add the planned section B, D to the network, the trip $t_5 = (K, E, B, D, E, K)$ becomes available as a P-shaped loop trip. All loops are simple loops.

Whether a trip corresponds to a (/set of) user profile, is defined in the natural way:

Definition 5 Trip-profile matching: Given a trail user profile $p = (\text{actMod}_p, \text{tripType}_p, \text{tripRange}_p, \text{retOnly}_p, \text{accMod}_p, \text{poiTypes}_p)$ and a trail network graph $TNG = (TN, TS, (\text{start}, \text{end}), \text{rev}, (\text{geoCoords}, \text{accModes}, \text{poiTypes}), (\text{geoPath}, \text{length}, \text{actModes}, \text{trailType}))$ and a trip t in TNG , we say t **matches** (is acceptable for) p , denoted by $\text{MatchTP}(t, p)$, iff

- its sections correspond to the activity mode(s), trail and trip type(s)/difficulty(/ies) of p : $\text{sections}(t) \subseteq \text{TrailTypeG}(\text{ActModeG}(TNG, \text{actMod}_p), \text{trailType}_p)$;
- its length falls between the trip length range of p : $t \in \text{LRangeTrips}(TNG, \min(\text{tripRange}_p), \max(\text{tripRange}_p))$;
- it is a returning trip if p requires that: $\text{retOnly}_p \rightarrow \text{start}(t) = \text{end}(t)$;
- its starting and ending point have the access mode of p : $\text{accMod}_p \neq \top \rightarrow t \in \text{TripsFromTo}(TNG, \text{Trailheads}(TNG, \text{accMod}_p), \text{Trailheads}(TNG, \text{accMod}_p))$
- it is a strict visiting trip for preferred POI types: $\text{poiTypes}_p \neq \emptyset \rightarrow \text{StVTrip}(\text{Poi}(TNG, \text{poiTypes}_p))$.

For a set of trail user profiles P , t corresponds to P , denoted by $\text{MatchTP}(t, P)$, if for any $p \in P$ t corresponds to p .

A relaxation (fuzzification) of the definition is given in order to allow shorter simple trips than requested - these offer something but not fully match -, with a match degree:

Definition 6 Given a trail user profile p and a trip t as above, we define the **match degree** of t for p , denoted by $\text{MatchTPD}(t, p)$ as being 1 if $\text{MatchTP}(t, p)$; otherwise, the value $\frac{\text{length}(t)}{\min(\text{tripRange}_p)}$ if $t \in \text{LRangeTrips}(TNG, 0, \max(\text{tripRange}_p))$ and obeys the other, non-length-range-relevant conditions of matching as defined above; and 0 for all others.

We can define the 'personalized' (profile-relevant) subgraph of a trail network graph for each (set of) trail user profile(s), which contains nodes and sections only reachable by a user with the given profile(s), and optionally a set of acceptable trip starting and destination nodes.

Definition 7 Assume a set of trail user profiles P , a trail network graph TNG with nodes TN and sections TS are given with a set of acceptable starting nodes $\text{StartTN} \in TN$ (trailhead locations), the set of acceptable destination nodes $\text{DestTN} \subseteq TN$.

The **P-matching (simple) trips** of TNG w.r.t. $(\text{StartTN}, \text{DestTN})$, denoted by $\text{ProfileMatchST}(TNG, P, \text{StartTN}, \text{DestTN})$, consists of all possible simple trips t in $TNP \subseteq TN$ for which $\text{MatchTP}(t, p) \wedge \text{start}(t) \in \text{StartTN} \wedge \text{end}(t) \in \text{DestTN}$ is true.

The *P-relevant (simple) trips* of TNG w.r.t. $(StartTN, DestTN)$, denoted by $ProfileRelevST(TNG, P, StartTN, DestTN)$, consists of all possible simple trips t in $TNP \subseteq TN$ for which $MatchTPD(t, p) > 0 \wedge start(t) \in StartTN \wedge end(t) \in DestTN$.

The *P-relevant trail network subgraph* of TNG w.r.t. $(StartTN, DestTN)$, denoted by $ProfileRelevG(TNG, P, StartTN, DestTN)$, consists of nodes $TNP \subseteq TN$ and sections $TSP \subseteq TS$ being part of any $t \in ProfileRelevST(TNG, P, StartTN, DestTN)$.

Example 4 Recall the user profiles and network from Examples 1, 2.

- Examples of *Pb-relevant trips* w.r.t. $(L2, L2)$ (returning) are: $(L2, K, X, K, L1, L2), (L2, L1, L2)$. The latter is not a match, as it is shorter than specified (3 km), but will count with relevance score 2/3.
- The *Pb-relevant subgraph* between nodes $N, W2$ is covered by the two matching trips $(N, Q, X, Z, W1, Z, W2), (N, Q, X, K, X, Z, W2)$.

3. Trip Variety Indices

The following subsections define the core indices of our contribution. The main motivation is to quantify the length proportion of not yet visited sections of different types (connecting, loop/circle, dangling/spike) in a sequence of trips and look at/between certain locations how high this number can be if a user visits all reachable sections combined into trips matching one's profile.

Definition 8 Assume a (finite) series of arbitrary, subsequent trips $SeqT$ a user is taking one after another.

- The **trip novelty ratio** of the i^{th} trip t_i in $SeqT$ is the ratio of summed section lengths of t_i not yet visited before (not being part of any $t_j, j < i$), divided by the total length of t_i .
- The summed novelty ratio of $SeqT$ is the sum of novelty ratios for all t_i in $SeqT$.

When a location or multiple locations (nodes) are given with a user profile, we consider all possible trips from that location and try to maximize these values. That means, in a maximized case, we assume a user takes trips in a sequence with the minimum possible change than the previous one, and accumulate the length proportions of the newly visited sections(s).

We, however, give priority to (direct) connecting trips first, then remaining loops and then, the rest is are those dangling sections, which can only be visited in a back-and-forth manner (spike in a trip); this way, we can differentiate the trip variety of such trip formats and index them separately. Such ordering is called a **C-L prioritized ordering** of possible (P-relevant) trips.⁵

⁵Different approaches could be resulted by evaluating the variety of connecting trips with/without loops and/or spikes, returning loop trips with/without spikes, and returning trips with spikes independently of each other. However, we wanted to construct a composite index whose components are additive and give a single value of (prioritized) variety under flexible circumstances, as we consider the start and destination nodes as sets of nodes and whether connecting or returning trips are allowed, is determined by their intersection and (symmetric) set difference.

It is possible that a user can explore the reachable part of the network in less number of trips fitting to one's profile. If the variety index equals the number of trips maximized case above, it means all possible trips (or remaining parts of the respective format - such as loop or spike) are independent. A reversal of a trip is considered as the same, adding no more variety. If we considered non-simple trips as well, it is obvious that their novelty ratios will not be higher than simple trips.

Examples and more discussion will be given after the definitions.

Definition 9 Let $TSet$ be a set of simple trips in a trail network graph TNG .

For each section s in $TripG(TSet)$, we specify a set of trips in order to determine the characteristic role of s in the given set of trips, and then assign a weight to s based on that.

- Let $DirectTrips(s, TSet)$ (**Direct trips of s**) be the set of trips in $TSet$ in which s is not a member of a loop (circle or back-and-forth section).
- Let $CLoopTrips(s, TSet)$ (**Circle-Loop Trips of s**) be the set of trips in $TSet$ in which s is a member of a circle.
- Let $BFTrips(s, TSet)$ (**Back-and-Forth Trips of s**) be the set of trips in $TSet$ in which s is a member of a back-and-forth section.

When defining the variety indices, one may give a priori weights to the trips being evaluated. A function $tw : TSet \rightarrow \mathbb{R}_0^+$ is used for this. By default, it is constant 1 and in this case, it can be omitted from the notations.

The weights and indices:

- The **General Trip Variety Index** of $TSet$ given their a priori weights tw , denoted by $GTVI(TSet, tw)$ is a weighted sum of all sections in $TripG(TSet)$, where the index weight of each network section s , denoted by $w_{GTVI}(s, TSet, tw)$ is determined by the ratio of its length and the weighted length of the shortest trip with the highest a priori tw value of $TSet$ containing s : $w_{GTVI}(s, TSet, tw) = \frac{length(s)}{tw(t) * length(\bar{t})}$ where $s \in ts(t)$, $t \in argmax_{tw}(T)$ with $T = argmin_{length}(TSet)$.⁶ The notation is extended to sections not in $TSet$ by assigning a weight of 0. We omit tw if it is constant 1: $GTVI(TSet, 1) = GTVI(TSet)$.
- The **(General) Connecting Trip Variety Index**, denoted by $GCTVI(TSet, tw)$ is a similar weighted sum over each section, where the weight $w_{GCTVI}(s, TSet, tw) = w_{GTWI}(s, DirectTrips(s, TSet), tw)$. It is also called simply $CTVI$, and w_{CTVI} .
- The **(General) Loop Trip Variety Index**, denoted by $GLTVI(TSet, tw)$ is a similar weighted sum over each section, where the weight $w_{GLTVI}(s, TSet, tw) = w_{GTWI}(s, CLoopTrips(s, TSet), tw)$.⁷ A strict variant of $GLTVI$ is the simply called $LTVI$, which acts as a supplement to $CTVI$, where only the non-connecting sections are counted here: $w_{LTVI}(s, TSet, tw) = 0$ if $w_{GCTWI}(s, TSet, tw) \neq 0$, otherwise it equals w_{GLTVI} .
- The **(General) Spike Trip Variety Index**, denoted by $GSTVI(TSet, tw)$ is a similar weighted sum over each section, where the weight $w_{GSTVI}(s, TSet, tw) =$

⁶The functions $argmin$ and $argmax$ are assumed to return subsets because multiple items may have the same max/min value.

⁷The term loop refers to the phenomenon that sections of circles are parts of loop trips or loop parts of connecting trips. We quantify the variety of circles of possible loops here.

$w_{GTWI}(s, BFTrips(s, TSet), tw)$.⁸ A strict variant of GSTVI is the simply called STVI, which acts as a supplement to CTVI and LTVI, where only the non-connecting and non-circle sections are counted here: $w_{STVI}(s, TSet, tw) = 0$ if $w_{GTWI}(s, TSet, tw) + w_{GLTWI}(s, TSet, tw) \neq 0$, otherwise it equals w_{GSTVI} .

- The **Composite Trip Variety Index** is denoted and defined by the triplet $CompTVI(TSet, tw) = (CTVI(TSet, tw), LTVI(TSet, tw), STVI(TSet, tw))$. It reveals more details of the structure of trips in TSet than the GTVI.
- If all trips start and end at the same node, $CTVI(TSet, tw) = 0$ and the Composite Returning Trip Variety Index can also be used instead: $CompRTVI(TSet, tw) = (LTVI(TSet, tw), STVI(TSet, tw))$.
- The **Summed C-L prioritized Trip Variety Index** is defined by $SumTVI(TSet, tw) = CTVI(TSet, tw) + LTVI(TSet, tw) + STVI(TSet, tw)$.

Note that each section is calculated only once in each index, and in only one of the components of $CompTVI$, regardless of how many trips the section is contained by.

The main step in the definition follows. We intend to give variants of the above variety indices based on visitor profiles and locations in the network where the index represents all possible trips given that context.

We do not allow longer trips than requested by the profile, but shorter ones are allowed partially, as their length is divided by the minimum allowed length. This approach not only gives an indication of partial relevance to a trail user (and reflects the option for the user to repeat a shorter trip as a non-simple trip to reach the min range length value), but will also help in the effective computation.

Definition 10 Let TNG be a trail network graph with nodes TN , P a trail user profile, $StartTN \in TN$ (possible starting locations) and the set of acceptable destination nodes $DestTN \subseteq TN$.

The **Profile Based (Composite) Trip Variety Index** is denoted and defined by the triplet $CompTVI(TNG, P, StartTN, DestTN) = (CTVI(TNG, P, StartTN, DestTN), LTVI(TNG, P, StartTN, DestTN), STVI(TNG, P, StartTN, DestTN))$ where

- $CTVI(TNG, P, StartTN, DestTN) = CTVI(ProfileRelevST(TNG, P, StartTN, DestTN), length/\min(tripRange_P))$ is the connecting trip variety index,
- $LTVI(TNG, P, StartTN, DestTN) = LTVI(ProfileRelevST(TNG, P, StartTN, DestTN), length/\min(tripRange_P))$, is the loop trip variety index,
- $STVI(TNG, P, StartTN, DestTN) = STVI(ProfileRelevST(TNG, P, StartTN, DestTN), length/\min(tripRange_P))$, is the spike trip variety index.

Usually, the index is applied to singleton sets where specific nodes (mostly trail-heads) are evaluated by the variety of possible trips the network offers. In this case, $StartTN = \{start\}$, $DestTN = \{end\}$, and the members can be written directly instead of the sets: $CompTVI(TNG, P, start, end)$, etc.

⁸The term spike refers to the phenomenon that if circles are omitted from loop trips or sections, the remaining sequence becomes a composition of spikes (dangling sections). This index includes values for all sections in back-and-forth parts of loop trips even if they do not form a spike in strict manner in their trip originals.

If $StartTN = DestTN = \{tn\}$ then we may omit the first item and call it the **Composite Returning Trip Variety Index** instead:
 $CompRTVI(TNG, P, tn) = (LTVI(TNG, P, tn), STVI(TNG, P, tn))$.

Proposition 1 *The Profile Based (Composite) Trip Variety Index (and all the above defined indices) are well defined.*

Well-definedness is ensured by, first of all, the well-definedness of GTVI, since if a section s is in multiple trips, their minimum length is unique (lengths are positive and we consider only simple trips), and even if there are multiple min-length trips of a specific format (connecting, returning with/without loops) containing s , the weight depends only on their, unique, length value, and the *a priori* trip weight which is maximized among them. An actual choice of the next shortest trip (if there are multiple available) will have no influence on the later assigned weight values of sections of the other shortest-trip-options, since trips with a given (min) length are exhausted before any of the longer trips are considered for section weighting. Furthermore, $CompTVI$ for a set of trips explicitly assigns a nonzero weight to s only at most one of its component indices. The allowed ordering of specific formats (direct connecting, remaining loop, remaining spike) ensures each section will be assigned to a specific component index with a nonzero value, regardless of the actual ordering of the trips in the sequence, as each section being part of any allowed direct connecting trip will be exhausted before any other loops are considered, and all possible circle sections are exhausted before the remaining possible spike sections are scored with weights. For given parameters, $ProfileMatchST$ and $ProfileMatchG$ are finite, and each section $s \in ProfileMatchG$ has at most one component index in P-based $CompTVI$ with a nonzero value, and it is not dependent on the order of trips which component index is nonzero for a specific section.

In order to provide an even better insight on the variety, namely, to see the scale of the above indices, the number of trips covering the opportunity space is useful to add.

Definition 11 *With the above notations, a set of trips $TCov$, w.r.t. $(TNG, P, StartTN, DestTN)$ is a*

- **Covering Connecting Trip set (CCT)** if it contains connecting trips of $ProfileRelevST(TNG, P, StartTN, DestTN)$, covering the direct connecting trips between $(StartTN, DestTN)$ in $TripG(ProfileRelevST(TNG, P, StartTN, DestTN))$, where each trip has at least one section not occurring in all others, and each trip has a minimum possible length with these conditions.⁹ A Maximal CCT (MCCT) is a CCT with the highest number of trips. The number $MCCT(TNG, P, StartTN, DestTN)$ is the weighted count of trips in an MCCT, each multiplied by their $MatchTPD$ value.
- **Covering Loop-containing Trip set (CLT)**, if it contains (non-direct) connecting trips or or loop trips of $ProfileRelevST(TNG, P, StartTN, DestTN)$, covering the circles of $TripG(ProfileRelevST(TNG, P, StartTN, DestTN))$ not being part of any direct connecting trips between $(StartTN, DestTN)$, where each trip has at least one section in those circles not occurring as a circle section in

⁹Note that a direct connecting trip of $TripG$ between $(StartTN, DestTN)$ is not always a P-matching trip (if a preferred POI must be visited outside of it) but it is covered by a P-matching trip of the CCT with the minimum length possible.

all others, and each trip has a minimum possible length with these conditions.¹⁰ A Maximal CLT (MCLT) is a CLT with the highest number of trips. The number $MCLT(TNG, P, StartTN, DestTN)$ is the weighted count of trips in an MCLT, each multiplied by their MatchTPD value.

- **Covering Spike-containing Trip set (CST)**, if it contains (non-direct) connecting or returning trips of $ProfileRelevST(TNG, P, StartTN, DestTN)$, covering the spikes (dangling sections) of $TripG(ProfileRelevST(TNG, P, StartTN, DestTN))$ not being part of any direct connecting trips between $(StartTN, DestTN)$ or circles, where each trip has at least one section in those spikes not occurring as a section in spikes in all others, and each trip has a minimum possible length with these conditions. A Maximal CST (MCST) is a CST with the highest number of trips. The number $MCST(TNG, P, StartTN, DestTN)$, the number of is the weighted count of trips in an MCST, each multiplied by their MatchTPD value.

A union of a CCT and a CLT is called a **Combined Covering C-L Trip set (Comb-CLT)**, with its maximal size of $MCombCLT(TNG, P, StartTN, DestTN) = MCCT(TNG, P, StartTN, DestTN) + MCLT(TNG, P, StartTN, DestTN)$. A union of a CombCLT and a CST is called a **Combined Covering Trip set (CombCT)**, with the size of a maximal one being $MCombCT(TNG, P, StartTN, DestTN) = MCombCLT(TNG, P, StartTN, DestTN) + MCST(TNG, P, StartTN, DestTN) - DupST$ where $DupST$ is the number of trips occurring in both $MCombCLT$ and $MCST$ (having a back-and-forth section not covered by other trips of the covering set).

The summed novelty ratio of a trip sequence of an $MCombCT$ in the order of increasing trip length is called the **Maximal C-L-prioritized summed novelty ratio** w.r.t. $(TNG, P, StartTN, DestTN)$.

Example 5 The set of direct connecting trips between D, X (in the graph without planned sections) are $\{(D, E, K, X), (D, M, Q, X), (D, M, Q, K, X), (D, E, K, Q, X)\}$. Either 3 of them forms a MCCT. Additionally, the singleton trip set $\{(D, E, K, L1, L2, K, X)\}$ is a MCLT, and an MCST is the trip set $\{(D, A, D, E, K, X), (D, E, B, E, K, X), (D, M, Q, N, Q, X), (D, E, K, X, Z, X), (D, M, Q, X, Z, W1, Z, X), (D, E, K, X, Z, W2, Z, X)\}$.

The following important statements follow from the respective definitions, due to the fact that if two different (shortest) trips with the same length share a section, the weight of that section will be the same if either one of the trips is selected for forming the index value; and the fact that if a section appears in multiple trips then its weight will be counted only with the(/a) shortest one for the index, and the first occurrence in a sequence for the summed novelty ratio.

Proposition 2 Trip variety (in terms of either index) of all possible trips for a profile (weighted by MatchTPD) equals to the trip variety (of the same index) of a maximal combined cover (for each of the variety indices).

Proposition 3 Trips being generated in the following (C-L-prioritized) order result in a combined covering trip set whose values of each trip variety indices are equal to that of

¹⁰Although the definition contains circles instead of loops, there is a matching with the loops (as each loop must contain a circle in our terms) and this way our current definition remains consistent with [12], and puts more emphasis on the loops as circle-containing returning trips or parts than the circles themselves.

a maximal combined cover. First, connecting trips without loops and with the necessary spikes only (for preferred POI visit where needed), in the increasing order of their length, starting with the shortest one(s) possible. Second, trips with loops: returning trips or connecting trips containing at least one circle and with the necessary spikes only, in the increasing order of their length. Third, trips with spikes not having been covered by the previous trips, in the increasing order of their length.¹¹

Proposition 4 *The maximal C-L-prioritized summed novelty ratio is an upper bound for a possible summed trip novelty ratio for any sequence of ProfileRelevST($TNG, P, StartTN, DestTN$) trips where each connecting trip precedes each returning trip, the first appearance of each non-returning-part-section of any possible connecting trip is not in a returning part of any trip, and the first appearance of each section of any circle in the network does not appear in a returning (back-and-forth) part of its first trip in the sequence. It is a strict bound if there is no partial P-matching trip (shorter than the range minimum) and no POI-preference given.*

If a respective index equals to its covering number counterpart, it means all the respective trips / loops / spikes are independent (pairwise disjoint), so the network provides the maximal variety in that respect with the given number of (independent) opportunities. If a variety index is relatively small for a large covering number, it means there are quite a few trip / loop / spike part variations but with many overlaps.

Definition 12 *Including the above defined counts, the Extended Composite Trip Variety Index is a 6-tuple (parameters omitted for the sake of simplicity): ExtCompTVI = (CTVI/MCCT, LTVI/MCLT, STVI/MCST), where the slash is used as a notational separator marker and does not denote division (however, treating it as a divider and counting the rations is also meaningful for measuring relative variety of the amount of trip opportunities provided, see later in Section 4).*

Further properties of the indices are explored in the next section.

Next we consider the computation of the indices. Based on the last three propositions in the previous subsections, an algorithm is sketched for computing the combined trip variety index, by simulating generation of a proper (C-L prioritized) combined covering trip set, which might not be maximal, but assigns the same weights to sections as a maximal cover. The actual implementation may vary.

The following concepts are needed for the algorithm:

Definition 13 *The POI-closeness of a trip (or a node) x in a network TNG regarding a set of preferred POI types $poiTypes_p$ specified by a profile p , denoted by $PoiDist(TNG, x, poiTypes_p)$, or if the context is obvious, in short $PoiDist(x)$, is defined as:*

- 0 if $poiTypes_p = \emptyset$ or it contains a node (or is by itself, in case of being a node) of a preferred POI type,
- the length of the shortest spike trip leading to a preferred POI if reachable in the allowed max range - minus the trip length of x - given by p .

¹¹In some cases, covering trips generated in the increasing order of their length does always not form a maximal cover, see, for instance, part a2 on Figure 2. But the variety index values will be the same as for a maximal cover.

- ∞ if such POIs are not reachable using a spike trip having its length in the allowed range max given by p (minus the trip length of x).

Note that the POI-closeness is twice the distance of the closest preferred POI of the respective trip or node.

Definition 14 The **(POI-neutral) trip-distance** of a node n in a network TNG , denoted by $TripDist(TNG, n, P, StartTN, DestTN)$, regarding a profile P (with the usual notations) and node sets $(StartTN, DestTN)$ is the length of a shortest trip with the following conditions: it is either a connecting trip between $(StartTN, DestTN)$ or a returning trip from $StartTN \cap DestTN$, visiting the node n (no preferred POI visit is necessary). If this length is larger than $\max(tripRange_p)$ it will be ∞ . A similar definition can be given for sections s , for the shortest trip with the given condition including s .

The **POI-visiting trip-distance** of a node or section x , denoted by $TripPoiDist(TNG, x, P, StartTN, DestTN)$, is similar to the above with the extra condition for the trips visiting a preferred POI if $poiTypes_p \neq \emptyset$.

The **(POI-visiting) Loop-length-distance** of a loop l , denoted by $LoopLengthPoiDist(TNG, l, P, StartTN, DestTN)$ is the length of the shortest trip with the following conditions: it is either a connecting trip between $(StartTN, DestTN)$ or a returning trip from $StartTN \cap DestTN$, containing the loop l and visiting a preferred POI if $poiTypes_p \neq \emptyset$. If this length is larger than $\max(tripRange_p)$ it will be ∞ .

Proposition 5 Let TNG be a trail network graph with nodes $TN, P = (actMod_p, tripType_p, tripRange_p, retOnly_p, accMod_p, poiTypes_p)$ a trail user profile, $StartTN \subseteq TN$ (possible starting locations) and the set of acceptable destination nodes $DestTN \subseteq TN$. Assume all elements of $StartTN$ and $DestTN$ are compatible with $accMod_p$.

The value of $CompTVI = (CTVI, LTVI, STVI)$ for the above parameters can be computed with the following **algorithm**:

1. Assign the value of $PoiDist$ as $\min p_n$ to each node n in the network (at least to the reachable nodes for P - this can also be done dynamically during further steps).
2. Generate a cover of P -matching connecting trips $CTSet$ in their increasing order of length, by BFS (breadth-first search), utilizing the $\min p_n$ values so that each of these trips will be a direct connecting trip with an optional spike to the closest preferred POI. If $StartTN \cap DestTN \neq \emptyset$ than it must be separately done for each $a \in StartTN \cap DestTN$ towards $DestTN \setminus \{a\}$, so that no returning trips are included.
3. Assign to each non-spike section contained by trips of $CTSet$ the length of the shortest trip of such (denoted by $\min D_s$). The value $length(s) / \min D_s$ will be the candidate for the eventual w_{CTVI} of that section.
4. Assign the value of $TripDist$ as $\min t_n$ and $TripPoiDist$ as $\min t_{p_n}$ w.r.t. P to each node n in the network (its reachable part for P) and similarly, $\min t_s, \min t_p$ for sections s respectively.¹²

¹²This step can be done by generating an optimized cover of direct connecting trips $DCTSet$, regardless of POI visits, and (optional) spikes connected to these trips from each node. Note that in some cases, the two values $\min t_n$ and $\min t_{p_n}$ will be based on different trips, and neither of them must be based on a direct connecting trip in which node n appears. Furthermore, if $retOnly_p$ is true then this step must be done separately for each $a \in StartTN \cap DestTN$, storing possibly multiple values for each (a, n) .

5. Find and generate covering loops $LTSet$ each having at least one section not yet scored by w_{CTVI} , in their increasing order of $LoopLengthPoiDist$, which is the value $mintL_l = \min(\{mint p_k + length(l) | k \in nodes(l)\} \cup \{mint_m + length(l) + min p_n | m, n \in nodes(l)\})$. This is going to be the length of the trip giving the weight value candidates for $LTVI$ of the circle sections in the loop. Assign the value $mintL_l$ to each loop l .¹³
6. Assign to each circle section s contained by loops of $LTSet$ the value of $mintL_l$ by a containing loop for which it is minimal. The value $length(s)/mintL_l$ will be the candidate for the eventual w_{LTVI} of that section.
7. Remaining sections not covered yet by Steps 3 and 6 can only be parts of back-and-forth sections (as parts of spikes or loops) of profile-relevant trips. Prune them if $poiTypes_p$ is non-empty: remove sections (set their $mint_s, mint p_s$ values as ∞) not being part of any loops of Step 6 or spikes leading to a preferred POI.
8. Assign to each remaining section having finite $mint p_s$ the value $minS_s = \min(\{mint p_s\} \cup \{mintL_l | s \in sections(l)\})$. The value $length(s)/mintS_s$ will be the candidate for the eventual w_{STVI} of that section.
9. Upscale the weights for short trips: if any value assigned to sections in Steps 3, 6, 8 is lower than $\min(tripRange_p)$, then set it explicitly to $\min(tripRange_p)$ so that the weight will be the maximal possible value (c.f. *MatchTPD*).
10. Aggregate (sum by the respective trip variety index type $CTVI, LTVI, STVI$) and output the computed section weight values $w_{CTVI}, w_{LTVI}, w_{STVI}$ for each relevant section as values of the combined trip variety index.

What the algorithm does is exactly how the indices are defined. There are three types of sections in the (reachable part) of the graph, getting their weights differently: A *section of a direct connecting trip* gets a $CTVI$ weight, as the proportion to the length of the shortest profile-relevant connecting trip (including a preferred POI visit if necessary, by a spike). A *Circle section of a trip containing a loop not being part of any direct connecting trip* gets an $LTVI$ weight, as the proportion to the length of the shortest profile-relevant trip with a loop containing it in a circle. A *Back-and-forth section not being part of any of the above* gets an $STVI$ weight, as the proportion to the length of the shortest non-direct/returning trip (may be a non-circle part of a loop).

The algorithm is deterministic by counting the same weight for each section, and giving the same results, regardless of the order of the particular nodes and trips taken. It is guided by the length and trip format, and any variation beyond that yields the same weighting for each graph section (c.f. the reasoning for well-definedness of the indices).

4. Examples and Discussion

Example 6 *Different simple trail network topologies (with no trip length restrictions or POI preferences) are shown on Figure 2 with their (approximate) values of $ExtCompRTVI = (LTVI/MCLT, STVI/MCST)$ (for returning cases) and*

¹³Circles can be detected, for instance, by BFS, starting from each node being an endpoint of a section not covered yet by Step 3, or directly during the operation for Step 4, in parallel to assigning the values to nodes. Furthermore, if $retOnly_p$ is true then this step must be done separately for each $a \in StartTN \cap DestTN$.

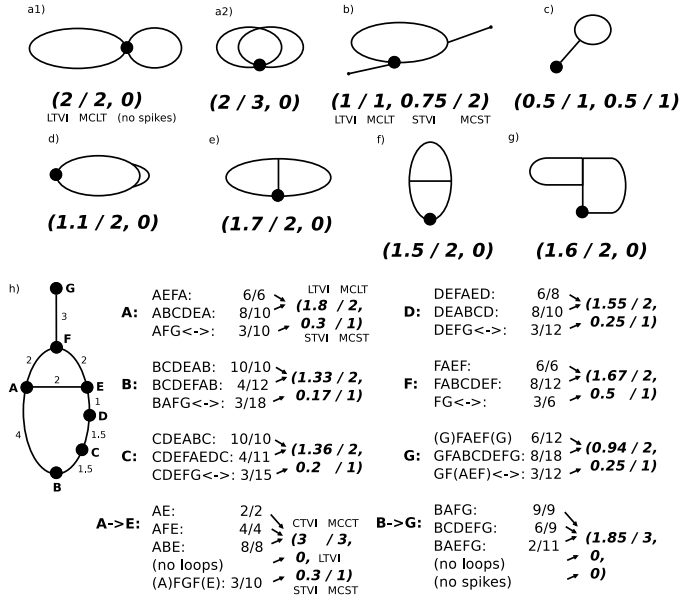


Figure 2. Sample cases with the Composite (Returning) Trip Variety Index, with two cases of connecting trips (extended variant of the figure in [12])

$ExtCompTVI = (CTVI/MCCT, LTVI/MCLT, STVI/MCST)$ (for the last two, connecting cases).¹⁴ We are referring these cases as illustrations for the properties listed below.

Proposition 6 The following general properties hold (and follow directly from their definitions), illustrated by the examples of Figure 2. They reveal more details of the meaning, the characteristics and benefits of these measures:

1. For returning trips, if the first (shortest, matching) loop trip is a circle, it gets a $LTVI=1$. Independent circle trips are counted as 1 each in the LTVI, therefore in this case, the LTVI is simply the number of them, and in general, it is a generalization of this number. Flower graphs show the maximal variety in loop trips (see case a1). If there are independent circle trips, LTVI is least the number of them.
2. If all (direct) connecting trips are independent, CTVI equals the number of them, so it is a generalization of the number of independent direct connecting trips. CTVI is not lower than the number of possible independent connecting trips.

¹⁴If any of them is 0 or not applicable for a case, we simply put a 0 instead of the formal 0/0, which may have caused confusion. Recall that the / marker in the index tuples is a separator and not a division mark, although an alternative interpretation of it as a division gives meaningful, relative values.

3. *For returning trips, if the second shortest (matching) loop trip share 0.5 of its length with the first one, it gets the value of 0.5, and results a total of 1.5 in the LTVI (case f). If there is only a minor variant for a short section of a loop, its addition to the LTVI will be proportionally smaller (case d). If the two loops share only a shorter section, their LTVI will be near to 2 (cases e and g).*
4. *The indices are invariant of the length of the actual trips (if they fall into the profile range), they only depend on the ratio of shared sections among them.*
5. *Each trail section is counted only once in the weighted lengths, in the shortest possible trip of the respective format (direct connecting, circle part, spike part) it is contained by. Namely, if a section is part of a direct connecting trip, it is added to the CTVI. Otherwise, if it is part of a circle, added to the LTVI. Sections used only in back-and-forth parts of trips are counted into the STVI (cases b and c).*
6. *If a circle (not part of a direct connecting trip) is cut (separated) at a point, its remaining sections will be counted into other circles they are part of, and the rest of the sections are transferred to STVI. After two independent spikes are joined to form a circle, their weights are transformed from STVI to LTVI.*
7. *Circles of P-shaped loops are downgraded proportionally to the distance of the circle from the starting/ending point (or shortest connecting trip). Their circle parts are valued by LTVI, and their back-and-forth parts by STVI. The more distant a circle is, the less relevant it is for LTVI (overdominated by closer sections).*
8. *Weights of sections constituting the index values may differ for different starting (or destination) nodes for the same profile, reflecting their local relevance.*
9. *Index values are continuous w.r.t. modifications in the network graph without topological changes (c.f. cases d and f). A topological change may cause value transfers between components of composite indices.*
10. *If a section is removed by gradually decreasing its length and merging the two of its ending nodes, continuity remains if it does not eliminate or transfer a trip of a certain format relevant to the profile (direct connecting, trip with loop, trip without loop but spike). The same holds when spitting a node into two and gradually enlengthening it (c.f. cases a1, e and g, where two loop trips remain in each).*
11. *It follows from the above that if a node with a degree of 2 (a.k.a. pseudo-node) is eliminated by merging the two sections it connects (having the same properties), and the node is not a preferred POI, it will not influence any of the indices (except for the case of STVI for spike parts when no POI preference is given).*
12. *The STVI is however, sensitive to the number of (pseudo)nodes placed along a trip part (if no POI preference is given) as each possible turning point generates a different trip if it does not exceed the max length of the profile range. If this is not the intention, an adaptation of the STVI definition is necessary to treat only the longest possible spikes as separate trips.*
13. *If the length of a circle (not part of direct connecting trips) is gradually decreased to become 0, the LTVI gradually disappears and the spikes connected to it (reachable through it) will have a continuous change in the STVI.*
14. *If a spike section (a section not being part of any circle or direct connecting trip) is removed, it has no effect on the CTVI or LTVI. If a dangling loop (a loop part not being part of any direct connecting trip) is removed from the graph, it has no effect on the CTVI (if no preferred POI types are given).*
15. *The indices are not sensitive to hubs with multiple nodes close to each other, when the minimum trip length of the profile is significantly bigger than the distances inside the hub (the local loops and variations will cause an increase, although not significant, in the index).¹⁵*

¹⁵An exception of this rule is when some circles starting and ending in such a hub actually do not return to

16. *the LTVI cannot distinguish between the cases of two independent loops of the same length, whether they are joined at their middle point or not (cases a1-a2),¹⁶ However, the MCLT will be different in the two cases (3 vs. 2). A similar effect can be observed when many but distant circles add up to 1 in the LTVI - their MCLT will be higher than 1, showing they do not form one single circular trip.*
17. *The 3 indices (CTVI,LTVI,STVI) have different meaning but they are additive and a summed index is also a characteristic measure (see the SumTVI definition).*
18. *After a user has taken at least MCLT number of different returning loop trips from the same, given node(set), where each has at least one section not visited before, one must have been visited all sections of the network reachable from that node(s) by loop trips available for her/his profile The same is true for direct connected trips between (sets of) specific origin-destination nodes with MCCT.*
19. *For any (set of) node(s), there exists at least $LTVI + 2 * STVI$ number of returning trips, each of which, taken by a user in a sequence, having some section(s) not visited by the user before in that sequence.*
20. *The value of GTVI correspond to the maximal summed novelty ratio of possible series of subsequent matching trips taken by a user from a specific (set of) node(s) of the network (towards specific destination(s) and/or taking returning trips as specified by the profile and the start-destination node sets).*
21. *The actual values of the variety indices (in general, value $SumTVI = CTVI + LTVI + STVI$) correspond the maximal summed novelty ratio of possible series of subsequent matching trips taken by a user from a specific (set of) node(s) of the network (towards specific destination(s)), with the C-L preference rule.¹⁷*
22. *A similar property as above holds for returning trips, with LTVI and STVI together, with the preference condition of each section being potentially part of a circle appears in a circle when it is first visited.*
23. *The ratio of $CTVI/MCCT$ (divided) for direct connecting trips can be interpreted as a relative variety index, giving an average trip novelty ratio value for any sequence of a maximal CCT. In relaxed terms, if a user takes any covering series of (connecting) trips which have only the necessary (POI-visiting) spikes (a CCT, with each trip having at least one section not yet visited before), their average trip novelty ratio will not be lower than this value.*
24. *A similar property as above is true for loop trips with $LTVI/MCLT$ for CLT, connecting trips with loop parts (with only the necessary, POI-visiting spikes) with $(CTVI + LTVI)/MCombCLT$ for CombCLT, and any covering set of trips with the conditions of Proposition 4 (CombCT) with $(CTVI + LTVI + STVI)/MCombCT$ for CombCT.*

the exact same node, but another one closer to it. In such a case, the (pseudo-)circle may be counted as part of a direct connecting trip (to CTVI), instead of a loop (to LTVI). To overcome this problem, the index must be improved by a minimum distance threshold for direct connecting trips (see future issues) or the network should be generalized before the indices are computed.

¹⁶Computation of the LTVI in such cases (as a2) can be done in two different ways (either from the two independent loops as 1+1, or starting with one of them and changing one of its parts with the other two trip part options in return as 1+1/2+1/2), resulting the same value.

¹⁷More precisely, assume a user takes any series of profile-matching connecting trips between two (sets of) nodes in the network with the following restrictions: each section potentially being part of a direct connecting trip appears in a non-back-and-forth part of a connecting trip when it is first visited, and each section not directly part of any direct connecting trip but is potentially a section in a circle appears actually in a circle when it is first visited. Then, the following will be true for any section of these trips: the length of the section divided by the length of the first trip it appears in is not higher than the weight of the section as counted into either of the CTVI, LTVI, STVI indices. Summing these up for each trip, we can state simply that the novelty ratio of a trip in such a(n almost arbitrary) series of subsequent trips can not be higher than the actual sum of weights of its novel sections contributing to the $CTVI + LTVI + STVI$ indices.

Table 1. Trip variety computation: a detailed example of composite returning trip variety index calculation

Variety of returning trips at node N			
Profile Pa 5-20 km, easy+difficult nature+landscape preference		Profile Pb 3-10 km, easy only no POI preference	
LTVI	STVI	LTVI	STVI
Without planned sections			
$NQKXQN : \frac{5}{8}$	$(NQ)XZ \rightleftharpoons: \frac{2}{10}$		$[NQ \rightleftharpoons] : \frac{1}{2} * \frac{2}{3}$
$NQKL_1L_2KQN : \frac{5}{11}$	$(NQK)EB \rightleftharpoons: \frac{1}{14}$		$(N)QX \rightleftharpoons: \frac{2}{6}$
$NQMDEKQN : \frac{10}{12}$	$NQ(K) \rightleftharpoons: \frac{2}{6}$		$(NQ)XK \rightleftharpoons: \frac{2}{10}$
			$(NQ)XZ \rightleftharpoons: \frac{2}{10}$
Σ TOTAL LTVI: 2.04	Σ TOTAL STVI: 0.60	Σ TOTAL LTVI: 0	Σ TOTAL STVI: 1.07
MCLT: 3	MCST: 3	MCLT: 0	MCST: 3.67 ¹⁸
With planned sections (change effect)			
$...DBE... : +\frac{3}{14}$	$del[EB \rightleftharpoons] : -\frac{1}{14}$	$NQKXQN : +\frac{6}{8}$	$(NQ)KL_1 \rightleftharpoons: +\frac{2}{10}$
			$(NQ)KL_2 \rightleftharpoons: +\frac{2}{10}$
			$del[QX \rightleftharpoons] : -\frac{2}{6}$
			$del[XK \rightleftharpoons] : -\frac{2}{10}$
Σ TOTAL LTVI: 2.25	Σ TOTAL STVI: 0.53	Σ TOTAL LTVI: 0.75	Σ TOTAL STVI: 0.93
MCLT: 4	MCST: 2	MCLT: 1	MCST: 3

Intuitively, the value of the various trip variety indices is an idealized number of corresponding ‘full’ trips of a respective format (direct connecting, with loops, and the rest) a hiker could enjoy without repetitions for given start (and destination) nodes. The minimum number of actual exhaustive trips may be lower, because some sections belonging to multiple trips in a maximal cover can be combined into longer, more complex-shaped trips. The index is therefore not a minimum value, nor a maximum value of possible covering trips matching the given profile, but a value stating approximately how many independent trips the total ‘novelty’ will be equivalent to, if a user fully explores the profile-matching possibilities with any number of trips.

Referring back to our running example with the sample network graph, computations of the indices in different settings have been carried over, with the following results:

Example 7 Based on the trail network graph on Figure 1 and the profiles in Example 2, trip variety indices for the following nodes/sets are computed and analyzed, before and after the planned extensions (for each $G \in \{TNG, ExtTNG\}$), for each profile $P \in \{Pa, Pb\}$. Results are shown on Figure 3. Sample computations are shown in Table 1.

1. Returning trips starting at node A (bus stop): $CompTVI(G, P, A, A)$;
2. Returning trips starting at node N (parking lot): $CompTVI(G, P, N, N)$;
3. Returning trips starting at node K (lodge): $CompTVI(G, P, K, K)$;
4. Connecting trips from A to W2 (public transport): $CompTVI(G, P, A, W2)$;
5. Connecting trips from nodes $W = \{W1, W2\}$ to nodes $L = \{L1, L2\}$ (from the city to the lake & camp): $CompTVI(G, P, W, L)$.

Remarks below contribute to a better understanding of the meaning and behaviour of the trip variety indices, from an empirical point of view:

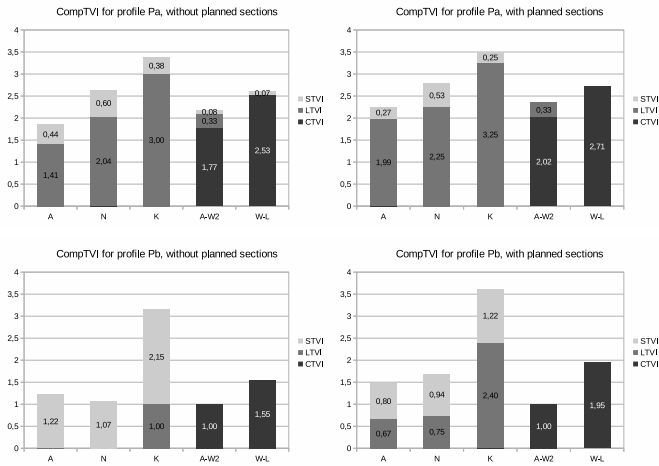


Figure 3. Trip variety index values for the network in Figure 1 with different profiles P_a and P_b of Example 2, for different nodes and trip formats (returning for A, N, K , connecting for $A \rightarrow W2$ and $W = \{W1, W2\} \rightarrow L = \{L1, L2\}$), showing the impact of planned network improvements for each case

- As expected, implementing the planned sections in the network has a bigger impact on profile P_b . While a P_b -hiker was only able to take a loop trip from K in the original network, the indices show a significant increase in the $LTVI$ for the nodes A, N, K . The whole trail network became much more attractive for easy loop-trail-fans. Moreover, the lake with nodes $L1, L2$ became directly reachable from N for P_b -hikers, without the need of an overnight stay at K . The $STVI$ value remains relatively high for node N even after loop trip opportunities came into the picture.
- At the same time, the planned sections have no impact on the variety of connecting trips for profile P_b at the observed node relations. This is because the direct connecting trips were already long for P_b -hikers and no extra loops or longer variants became acceptable. An increase is, however, caused by the new improvements for profile P_a , as all sections for the connecting trips can be incorporated into the direct parts of the trips or into a loop (to the lake, but that has not been changed).
- The difficulty upgrade of section QK has no effect on profile P_a since it has an effect on difficulty only, with the same length, so any change for profile P_a in the trip variety indices is caused by establishing the new section DB . The highest impact of it - as expected - can be observed at the returning trips from A , and a significant change is also caused at the connecting trips between A and $W2$. Interesting to see at the changes in varieties of other node relations for profile P_a that there are minor, but far-reaching impacts of this addition. The reason for this

is the capacity of *Pa*-hikers to walk long hikes, so that this far addition slightly increases the variety for them even between *W* and *L*, for example.¹⁹

These trip variety indices are intended to *model* the trip variety and opportunities for users in recreational trail networks. So far, we have analyzed and observed some of their quantitative properties. Their initial variants - without their formal establishment yet - have been used in a real-life, pilot study case in [12] with promising results. For a thorough assessment and evaluation of their usefulness more experimental computations in real-life scenarios are needed. However, if we want to make sure we are on the right path, both with the definitions as they are introduced in this paper, and further, what kind of further evaluations and how they should be conducted, to see a broad view on the preferred ways of applicability of these, it is useful at this point to look at the *model-being* of these indices, in particular the composite *CompTVI* and its extended form. So we take a step back now, and look at it from a broader perspective and make a reflection about how and at what extent this can be taken as a model.

According to [18] a *model* is an *instrument* with the characteristics listed below. Although a detailed analysis is out of the scope of this paper, these aspects are briefly addressed in the following for the concept of the trip variety index:

Instrument for a purpose: recommendation and network management (change impact / design assessment). By having a sound definition and computation method, the indices can contribute to answering the question sets 1. & 2., proposed in Section 1. A more specific utilisation portfolio shall be elaborated by looking at the properties and behaviour of the indices (see also sufficiency below).

Well-formedness: definitions and conventional notations introduced above ensure well-formedness of *ExtCompTVI*.

Adequacy: Compared to the simple combinatorial counting of the possible trips, or other conventional measures, the above presented properties support the adequacy of the index as being focused and purposeful. In terms of *analogy*, the measure itself is analogous to the centrality measures used in traditional graph theory and to the number of independent trips of respective formats. Furthermore, the index values (extended with the covering trip counts) show a correspondence between some intuitive min/max values for novelty ratios of series of trips the user may take, without representing the complexity of the trips themselves. Typical topology patterns and their variety indices have also been shown.

Justified: Index values are, by their definition, corroborated, coherent and conform, and falsifiable. Stability is achieved by several properties, including being continuous (small changes in the network, the from-to node sets or the profiles cause small changes in the index values), not sensitive to local hubs (with some exceptions), being independent of particular trip lengths, etc. Plasticity is achieved by the generality of the definition as being adaptable to any network topology, and the flexibility of the profile definition. Coherency and conformity can be improved for specific cases by proper adaptation of the definition(s), such as lifting the C-L preference condition for the price of the 3 indices not being additive any more.

¹⁹The paper [12] gives a real-life example when a 'winner' trailhead was identified with the most benefit (loop trip variety gain) of the changes in the network, while the institution initiated the changes was actually located at a different node of the network. For connecting trips, it has shown some far-reaching variety effects of local changes when relatively longer trips were allowed. Such an effect can be seen in our example as well.

Sufficiency: To assess whether our indices are of firm quality and evaluated, such as of correctness, generality, usefulness, comprehensibility, parsimony, robustness, novelty, tolerance, modality, confidence and restrictions, specific experimental computations are needed for real-life networks, and verification with trail professionals. It is also required for assessing the possible acceptance by the community of practice. Other measures may supplement these indices to get a more profound insight, as our indices are based solely on the network structure and user profiles.

Grounding: There are surely non-explicit elements of grounding, but the network being modeled as a graph, or the trips as paths along the graph, the users as profiles with some preferences, etc. and the initial questions related to the purpose of the index (whether variety is a relevant measure and is related to the novelty ratio) can be identified as parts of the (indisputable) grounding.

Basis: The (disputable, adaptable) basis is formed by the particular way of our definitions. How a user profile looks like in particular, or whether the network is uni- or bidirectional, do direct connecting trips have priority over loops, do circles have priority over spikes etc. these are the adjustable parts of the background, but each adjustment needs a re-definition of the respective indices.

Context: It is a really important question in what network scenarios these indices are useful. Many areas of the world do not have such a complex trail network which would require these computations. However, there are countries and regions with complex networks having been emerged during the last century which provide many variations with a sometimes unclear or not easily comprehensible structure, partly managed by different stakeholders. In such contexts, we do believe these indices are useful for the given purposes. Local properties of paths, visitor counts, the level of service at facilities, personal relations to particular places are not considered, which may substantially modify the user preferences.

Utilization of the index needs more consideration. It can be viewed as a comparative value when different nodes or relations, or profiles are in question and the relative variety of the trip opportunities of the impact of a network change is to be assessed. It can also be viewed as a section-level measure, and a way of evaluating a section is to calculate some variety indices of the same network with the section included and without it.

Some current limitations and promising future directions are considered next.

Adapting the indices for directed graphs - as a possible future option - is closely related to the issue of giving, for instance, walking times based on elevation profiles instead of single distances as labels to the trail sections (c.f. [20,4]). This will lead to a more precise quantification, however, it is not sure such modification will worth the effort of redesigning the indices. The trip variety indices serve mainly as an approximate guide which can be refined according to other measures not represented here. The effects of these factors is likely to override the precision gained by a refined trip variety index.

Although the user profiles include preferences for POI types, the indices do not tell any information about the actual number of reachable POIs. The RPOI used in [12] seems to be a proper complementing measure.

The most important future work is to have more evaluations in real-life networks and experiment on how specific types of network change are reflected in the index values, and how these changes correlate to the changes in the index values to users' perceptions. As [7] presented a method for improving connectivity in urban cycle networks, these indices can be utilized in a similar manner with trail networks for outdoor recreation activities.

Furthermore, a more profound analysis on user preferences regarding the types and properties of trip routes may reveal that different user profiles require different variety indices. Our composite trip variety index has a meta-directive as it prioritizes direct connecting trips over trips with loops or spikes, and loops over spikes.

5. Conclusion

In this paper, graph measures have been defined and formally established for the use in geographical information systems for modeling and quantifying the variety of user trip opportunities in - possibly complex - recreational trail networks (such as hiking paths or cycling ways), based on visitor profiles. Each index gives a score to either a single node, sets of nodes, a pair of nodes (origin-destination places), or a pair of sets of nodes reflecting on the variety of possible trips of specific formats around or between them. The indices can be used for ranking of these as potential locations of their activities - the larger the value of an index for a specific trip format is, the more variety is offered for possible trips by the network for specific user needs and preferences.

A formal establishment is given based on undirected graphs and simple trips (without repeating sections in the same direction), and a formal user profile model is proposed. Different forms of trip variety index values are can reflect on how different are the possible connecting trips between two nodes or node sets in the network, the possible loop trips or loop parts of connecting trips, or detour sections which can only be taken in a back-and-forth manner(spikes). Some of these indices were partially and informally introduced in [12], with a proposed agenda of formal definitions and assessment. This paper intends to serve that purpose, including the implementation-ready definitions of network, user profile and trip models as well as the algorithms of computing the indices.

The result is a solid, well-defined and theoretically verified index construction, with some initial computations on simple networks with promising results. A brief insight is also given on the model-being of these indices, revealing further properties and directives for future use and improvement. The main purpose or application area of the indices is the personal trip recommendation in complex systems (where to go for a greater variety of specific types of possible trips), and network change impact assessment (how does addition, deletion or upgrading a section effects on the various trip opportunities). This is a novel approach as being solely dependent on the network structure, similarly to graph centrality measures, enhanced with preferred nodes for visit (POI nodes).

Further experiments and computations in real-life trail networks must be taken in order to assess the comprehensibility and usefulness in the community of practice, as well as to develop effective visualisations and other methods of utilization. According to our knowledge, these indices reveal a new type of knowledge not having been commonly extracted and utilized from geospatial route network graphs in such a manner and for such purposes.

Acknowledgement

The research was supported by the Hungarian Ministry of Innovation and Technology NRD Office within the framework of the Artificial Intelligence National Laboratory Program.

References

- [1] Wikiloc - trails of the world, <https://www.wikiloc.com/>, [Online; accessed 26-Jan-2021]
- [2] OuterSpatial - a stewardship-first approach to data, maps and apps for recreation (2019), <http://www.alaska-trails.org/2019-trails-conference-presentations.html>, [Acc. 25-Jan-2021]
- [3] Brämer, R.: Profilstudie Wandern (2008), Deutsches Wanderinstitut
- [4] Calbimonte, J.P., Martin, S., Calvaresi, D., Zappelaz, N., Cotting, A.: Semantic Data Models for Hiking Trail Difficulty Assessment, pp. 295–306 (01 2020)
- [5] Cheng, Y.Y., Lee, R.K.W., Lim, E.P., Zhu, F.: Measuring Centralities for Transportation Networks Beyond Structures, pp. 23–39 (05 2015). https://doi.org/10.1007/978-3-319-19003-7_2
- [6] Clark, R.N., Stankey, G.H.: The recreation opportunity spectrum: A framework for planning, management, and research. Tech. rep., Department of Agriculture, Forest Service, Pacific Northwest Forest and Range Experiment Station, (1979)
- [7] Guillermo, L., Orozco, N., Battiston, F., Iñiguez, G., Szell, M.: Data-driven strategies for optimal bicycle network growth. *Royal Society Open Science* 7(12) (2020)
- [8] Hong, J., Tamakloe, R., Lee, S., Park, D.: Exploring the topological characteristics of complex public transportation networks: Focus on variations in both single and integrated systems in the seoul metropolitan area. *Sustainability* 11(19) (2019), <https://www.mdpi.com/2071-1050/11/19/5404>
- [9] Lera, I., Pérez, T., Guerrero, C., Eguiluz, V.M., Juiz, C.: Analysing human mobility patterns of hiking activities through complex network theory. In: *PloS one* (2017)
- [10] Márkus, Z., Wagner, B.: GUIDE@HAND: digital GPS based audio guide that brings the past to life. In: Pavlov, R., Stanchev, P. (eds.) *Digital Preservation and Presentation of Cultural and Scientific Heritage*, pp. 15–25. Bulgarian Academy of Sciences, Sofia (2011). <http://eprints.sztaki.hu/6712/>
- [11] Molnár, A.J.: Trailsigner: A conceptual model of hiking trail networks with consistent signage planning and management. *Information Modelling and Knowledge Bases XXXII* (2021)
- [12] Molnár, A.J.: Synergistic planning of long-distance and local trails: A twin case study of trail network development in Northern Transdanubia. *Tourism Planning & Development* (2021). <https://doi.org/10.1080/21568316.2021.1936148>
- [13] Natera Orozco, L.G., Deritei, D., Vancso, A., Vasarhelyi, O.: Quantifying life quality as walkability on urban networks: The case of budapest. In: Cherifi, H., Gaito, S., Mendes, J.F., Moro, E., Rocha, L.M. (eds.) *Complex Networks and Their Applications VIII*, pp. 905–918. Springer I.P., Cham (2020)
- [14] Outdooractive: Outdooractive - Data Model, [Online; accessed 20-Jan-2020]
- [15] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., Theodoridis, Y., Yan, Z.: Semantic trajectories modeling and analysis. *ACM Computing Surveys* 45(4) (2013)
- [16] Peterson, B., Brownlee, M., Marion, J.: Mapping the relationships between trail conditions and experiential elements of long-distance hiking. *Landscape and Urban Planning* 180 (2018)
- [17] Semenov, A., Zelensov, V., Pimanov, I.: Application suggesting attractive walking routes for pedestrians using an example of saint-petersburg city. *Procedia Computer Science* 156, 319–326 (2019), 8th International Young Scientists Conference on Comp. Sci., YSC2019, 24–28 June 2019, Heraklion, Greece
- [18] Thalheim, B., Nissen, I., Allert, H., Berghammer, R., Blättler, C., Börm, S., Brückner, J.P., Bruss, G., Burkard, T., Feja, S., Hinz, M., Höher, P., Illenseer, T., Kopp, A., Kretschmer, J., Latif, M., Lattmann, C., Leibrich, J., Mayerle, R., Wolkenhauer, O.: *Wissenschaft und Kunst der Modellierung (Science and Art of Modelling)*, Philosophical Analysis, vol. 64. De Gruyter (05 2015)
- [19] Vias, J., Rolland, J., Gómez, M., Ocaña, C., Luque, A.: Recommendation system to determine suitable and viable hiking routes: a prototype application in sierra de las nieves nature reserve (southern spain). *Journal of Geographical Systems* 20 (2018)
- [20] Witt, P.J.: The development of a predictive hiking travel time model accounting for terrain variations. In: T. Jekel, A. Car, J. Strobl, & G. Griesebner (Eds.), *GIForum 2012: Geovisualization, Society and Learning*, pp. 102–112. Salzburg (2012)

Deep Learning for Knowledge Extraction from UAV images¹

S. Brezani ^a, R. Hrasko ^a, D. Vanco ^a, J. Vojtas ^a, P. Vojtas ^{a,b,2}

^a*Globesky ltd. Zilina, Slovakia*

^b*Dpt. Software Engineering, Charles University, Prague, Czechia*

Abstract. We study possibilities and ways to increase automation, efficiency, and digitization of industrial processes by integrating knowledge gained from UAV (unmanned aerial vehicle) images with systems to support managerial decision-making. Here we present our results in the secondary wood processing industry. First, we present a deployed solution for repeated area and volume estimated calculations of wood stock areas from our UAV images in the customer's warehouse. Processing with the commercial software we use is time-consuming and requires annotation by humans (each time aerial images are processed). Second, we present a partial solution where for computing areas of woodpiles, the only human activity is annotating training images for deep neural networks' supervised learning (only once in a while). Third, we discuss a multicriterial evaluation of possible improvements concerning the precision, frequency, and processing time. The method uses UAVs to take images of woodpiles, deep neural networks for semantic segmentation, and an algorithm to improve results. (semantic segmentation as image classification at a pixel level). Our experiments compare several architectures, backbones, and hyperparameters on real-world data. To calculate also volumes, the feasibility of our approach and to verify it will function as envisioned is verified by a proof of concept. The exchange of knowledge with industrial processes is mediated by ontological comparison and translation of OWL into UML. Furthermore, it shows the possibility of establishing communication between knowledge extractors from images taken by UAVs and managerial decision systems.

Keywords. automation of industrial processes; decision support; knowledge and information modeling and discovery; deep neural learning; modeling multimedia information and knowledge; content-based multimedia data management; UAV; photogrammetry; semantic segmentation

1. Introduction

Our long term research project is focused on studying possibilities and ways to increase automation, efficiency, and digitization of production, technological and logistic processes in the automotive industry using autonomously controlled UAV (unmanned aerial vehicles) means combined with ICT equipment for real-time processing and evaluation of acquired data according to Industry 4.0.

¹ This publication was realized with support of the Slovak Operational Programme Integrated Infrastructure in frame of the project: Intelligent systems for UAV real-time operation and data processing, code ITMS2014+: 313011V422 and co-financed by the European Regional Development Fund

² Corresponding Author, KSIMFF UK, Malostranske nam. 25, 118 00 Prague1, Czech Republic; E-mail: vojtas@ksi.mff.cuni.cz

There are numerous UAVs applications in managing civil infrastructure assets, such as routine bridge inspections, disaster management, power line surveillance, and traffic monitoring. This article describes our experience with an internally developed and deployed solution that uses a commercial photogrammetric product in the wood processing industry. Furthermore, we design new methods and prototypes in the mentioned above Industry 4.0 direction. That is, we increase automation of all processes, decrease the need for human expert intervention and interconnect our application with a decision support system via an ontology.

Industry 4.0 is the digital transformation of manufacturing and related industries and value creation processes in organizations, including logistics, supply chain, finance, accounting, and human resources. It helps manufacturers with current challenges by becoming more flexible and reacts easier to changes in the market. It can increase the speed of innovation and is very consumer-centered, leading to faster design processes. Implementation of this trend in an organization focuses on creating detailed digital models of reality, optimally real-time. This digital model (digital twin, see [22] for a framework reducing reality to a model) makes it much easier to oversee, control, and actively manage all production and manufacturing processes. A critical prerequisite is the acquisition of detailed data that can be processed and transformed into the knowledge needed for qualified management decisions by enriching the classic high-level data of the ERP system (e.g., orders and deliveries, accounting, plant management) with little-detailed operation data. It is commonly achieved using barcodes, QR codes, and scanners or using different IoT sensors.

Nevertheless, some data cannot be obtained, collected, or measured automatically. Appropriately equipped workers are necessary for manual collection, processing, and visual or sound data transformation. Humans' processing of visual or sound data means high costs and very long data update intervals. For example, inventory of externally stored material, such as containers, coal, wood stockpiles, or freshly made cars, can take several hours and days and often requires more personnel with adequate equipment — measuring equipment, dedicated software, or a protective kit. After an inventory check, the data is entered manually into the basic ERP systems, far from real-time processing. Based on this data, no real-time correction is possible. Only subsequent actions can be performed.

The research project aims to automatically collect outdoor visual data using pre-programmed UAVs and automatically process and transform them into knowledge using advanced computational tools such as machine learning based on deep neural networks. Deploying this solution to a real production facility can bring the capability of automatic data collection and processing of visual data regularly, with direct integration to core ERP systems in the form of alerts or data transfer. This way, the outdoor reality could be manageable almost in real-time. Our ambition is to deploy such a solution in the automotive production plant or its suppliers for logistics, warehousing, security, or maintenance. We start our research with automatic measurements of wood stockpiles in the wood storage facility. Slovakia is the fourth largest forest-covered country in the EU (with about 41% of the area, after Sweden, Finland, and Austria). The wood processing industry characterizes lower profit margins than in other sectors. Therefore, it is necessary to create value-added products in Slovakia and not just export raw wood abroad. For this reason, we consider it essential to bring new procedures and solutions using UAVs and intelligent image processing.

Our company has a research and development department, where we prepare prototypes in knowledge modeling and processing high-quality images from different

sources, such as UAV aerial images. Our starting point for this paper is a deployed solution to calculate wood stockpiles volume in the customer's warehouse using UAV images. The photogrammetry software we use to process UAV images requires annotating the area of interest from an expert. It can be challenging when multiple customers need to be served. Developing a generic solution that makes this annotation automated can be exciting in many practical applications.

2. Use case description

The customer is active in secondary wood processing, also known as value-added wood products manufacturing, generally defined as continuous manufacturing beyond lumber production. This customer needs information about the temporal development in wood stockpiles.

The on-site process begins by setting calibration points, placing them manually, and marking a known length. It is necessary for precise photogrammetry processing. The next step is to manually set the flight route data so the drone (UAV) is ready to fly and take appropriate area pictures. Afterward, pictures are taken and collected to fit automatic processing by a commercial tool, Pix4D, which creates an orthophoto map [17]. The created orthophoto map trained user manually annotates the areas of interest, which took the trained user about 20 minutes (depending on the area shape). Finally, Pix4D can calculate the woodpiles' area and volume.

Our goal is to eliminate manual processing steps to have a generic solution (with an API) that automated this process. Thus, our solution could be deployed for more customers without the need for a trained human expert annotator. We hope this can be very interesting in many practical applications also beyond the wood industry.

The solution we present here works using neural network-trained solutions for semantic segmentation and domain ontology in multimedia/spatial environments. Our main contributions are:

- Experiences and data from a deployed system using UAVs and the professional/commercial photogrammetry software.
- A new system integration solution interconnecting UAV aerial images from a wood log warehouse with the decision support system mediated by an ontology and customers' requirements. Depending on the application need, we can tune our solution up along several axes (e.g., precision, execution time, amount of human expert activity, frequency).
- An experimental prototype of the UAV aerial image processing system, based on several alternative deep neural network architectures and several pre-trained backbones. We improve semantic segmentation with a new algorithm.
- Experiments with calculations of the area of the wood logs pile base with real-world data from several UAV flights during 9 months and their comparison with the results from Pix4D.
- The extracted knowledge can be sent to the decision support system using a general external ontology equipped with the respective domain extension.

Other extensions of this use case can classify the type or quality of wood or work in places where manual calibration point settings are impossible. For example, the idea is to use a car catalog with known car dimensions. An alternative task may be to decide

only whether the warehouse wood reserve has increased or decreased. Another option would be to estimate the amount of wood delivered for a given time compared to the declared invoice for delivery (combined with vehicle weighting and license plate reader).

3. First deployed application and experiences

The UAV systems can quickly and efficiently collect data and capture vast areas of the globe from the surface in various spectra. The most commonly used spectrum for imaging is the orthophoto layer (geometrically corrected ("orthorectified") such that the scale is uniform). The data captured from the UAV also contains field data, which in conjunction with the orthophoto layer, we can process photogrammetrically post-process and thus provide data with higher added value. In addition, photogrammetry provides optical and mathematical methods and tools for calculating spatial/dimensional coordinates based on digital photography from the scanned area.

The main issue is how to use these data to get optimal information for specific tasks. First, let us focus on calculating estimates of wood stockpiles volumes.

- Several parameters affect the measurement results:
- Ability to obtain the most accurate information for 3D processing
- Processing time (higher accuracy takes longer)
- Degree of automation (how much manual work of an expert is required)

At first sight, it is clear that these parameters conflict with current technology standards in the field. However, we already successfully deployed our first solution in the customer warehouse, where a large amount of wood was processed. The primary task is to estimate the volume of regular wood stockpiles and their changes in time. As we mentioned above, the necessary process of collecting aerial pictures using drones is in place, so we already have calibrated system using drones, which collects the aerial area images.

In further processing, Pix4Dmapper[17] transforms the geodetic coordinates of the images' common points into a single 3D model of the scanned area into a point cloud. Such an approach can achieve high accuracy, but it is time-consuming for processing. The whole process displays a red arrow procedure in Figure 1. The rising demands on precision require more processing time in the range of hours to days.

To address processing time the automation is necessary. The automated data processing process follows the green arrow procedure in Figure 1. That shows how our new solution works. However, the only difference is which steps are manual and which are automated.

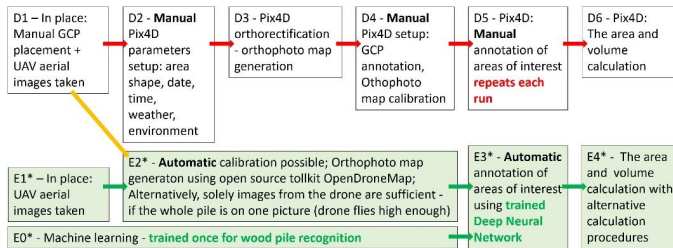


Figure 1. Procedure for processing the already deployed solution (red arrows) using GCP (ground control points) and our new approach (green arrows) with an alternative without GCP and functional area calculation and proof of concept of volume calculation

The already mentioned product process (red arrows procedure in Figure 1) takes place almost automatically, except for the three manual steps:

1. The first manual step (D2 in Figure1) is manual program (Pix4D) processing settings. The program has several attributes that are necessary to select before the process according to the parameters of the monitored environment and the subject of measurements, such as environment type, measured object shape, surrounding environment, the quality of the collected data (e.g., the influence of current weather on local brightness), required quality and precision of the results and other specifications.
2. The second manual step (D4 in Figure 1) is calibration using ground control points (GCPs) marked and measured at the beginning of the scan. The calibration itself consists of entering these parameters into the program based on their identification from the evaluated area's images. Again, we use state-of-the-art satellite technology with a deviation at the centimeter level to target GCP reference points. This step is performed on-site just before the drone takes off and is a manual annotation in the program itself (Figure 2).



Figure 2. Ground Control Points (GCPs) annotation in Pix4D

3. The third manual step (D5 in Figure 1) is to annotate the edges precisely to demarcate and define the wood base's shape (in all directions). This step is critical for accurate area and volume calculation. It consists of manually marking edges of a convex area directly in the orthophoto map, between which we want to measure the exact value of dimensions in space (see Figure 3). By the way, it is the most time-consuming step, and the prerequisite is trained professionals.



Figure 3. Manual annotation of the area of interest

After these three manual steps, the program (Pix4D) can calculate the area of annotated surface and volume of the corresponding wood stockpiles.

4. Model training and data annotation

The Pix4D photogrammetric method has proved to be valid and gives satisfactory results for our customers. However, in this method's background, we still find many activities requiring the manual work of highly trained users. Moreover, it would be necessary for ontology engineering and connection to a decision support system to have complete control over the application (e.g., using API).

Suppose we created a system that can proceed automatically without intervention. In that case, we could streamline the entire process of regular daily inventory measurements and at the same time effectively evaluate and monitor the movements (increase and decrease in the area) of material in the warehouse. We would potentially gain an overview of materials' movement over large areas of one or more warehouses of different customers. It would find justification in several industries by extending today's limited capabilities to almost unlimited use with automated UAVs for regular inventories. The idea to create an automated system for identifying the content (storage space occupancy) of stored wood in the warehouses of a wood processing company has been our quest.

For localizing the woodpiles in the images, we used deep neural networks (DNN). DNNs are used to solve several types of tasks such as image classification, object detection, etc. In our case, we use DNNs to solve an image segmentation type task with two classes (woodpile, background on pixel level). Since training models from scratch is very computationally and data-intensive, we used a transfer learning method to train our models. In this method, an existing model - which has been trained on a different dataset (e.g., an image segmentation model with a backbone trained to discriminate 80 object types) is used, and this model is then trained on a new task and data sample - identifying woodpiles. The advantage of this approach lies in the fact that the original model with a backbone was already able to identify basic shapes and their combinations. Thus, subsequent training just adapted this model to the new task and data.

The image segmentation task is of supervised machine learning type and hence needs annotated data (ground truth) for training. The annotated image represents the image itself and its metadata, defined as the relevant objects' location in the image. We

used the *Label Studio tool*³ to annotate the images. We used it to mark the wood stockpiles on a sample of pictures. For annotation, we used original images from 3 flights in 4K resolution. These images have been scaled down and split into smaller 480x480px images. We manually annotated 1000 images. Subsequently, we split the data images into a training dataset (800 images) and a validation dataset (200 images).

We used augmentation within the training cycle to prevent overfitting in neural networks to increase the size and diversity of annotated data. The *imgaug library*⁴ provides a wide range of transformations in order to transform both image and segmentation data. In our case, we used affine transformations (rotation, shift, zoom), contrast adjustment, noise generation, and the like.

As a baseline implementation of image segmentation models, we used the *Segmentation Models library*⁵. This library implements 4 model architectures for binary and multi-class classification (U-Net, PSPNet, Linknet, and FPN). In addition, the library uses the transfer learning method and allows using one of the 25 pre-trained networks (trained to classify the *2012 ILSVRC ImageNet dataset*⁶) as a backbone for the semantic segmentation architecture. This method makes it possible to use a trained neural network, or part of it, for another (related) category of tasks. In our experiment, we used 3 architectures (U-Net, PSPNet, and FPN) and 2 backbones (ResNet-18 and VGG-16). These models were trained using our annotated training and validation dataset.

Several factors affect the performance and accuracy of the trained model and the speed of training and inference during deployment in the production environment. These include:

- The chosen architecture - determines the model's performance, stability, the time required for its training and inference, and other aspects of neural network models. The development of neural network architectures for semantic segmentation is an area of intensive scientific research and development. An overview of current architectures can be found on the link⁷.
- The selected type of pre-trained neural network backbone - is essential when the transfer learning method is applied. As with architecture selection, the backbone neural network selection affects model performance, stability, training time, and inference.

³ <https://github.com/heartexlabs/label-studio>

⁴ <https://github.com/aleju/imgaug>

⁵ https://github.com/qubvel/segmentation_models

⁶ <http://image-net.org/challenges/LSVRC/2012/>

⁷ <https://paperswithcode.com/methods/category/segmentation-models>

5. Methods, tools, and experiments

Figure 4 shows the progression of our experiment, which consists of multiple phases: First phase <1. ODM> shows the generation of an orthophoto map from the input images using the OpenDroneMap⁸ tool.

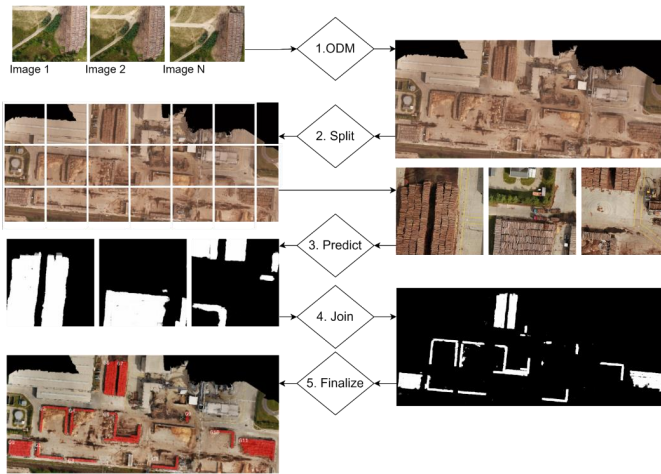


Figure 4. The progression of our experiment. See details below in the following text.

These images are in 4K quality, and their number is in the order of hundreds. This step is necessary because it is not always possible to have the whole woodpile on a single image (Figure 5).

The generated orthophoto map shows the entire monitored area in a single image. However, the orthophoto map is too large for further direct processing (~ 18000x12000 px) and therefore needs to be processed. Thus, the next phase <2. Split> splits the input orthophoto map into N images with a resolution of 480x480 px.



Figure 5. The input image shows three different parts of different woodpiles, none of which is visible as a whole.

⁸ <https://www.opendronemap.org/>

The next phase <3. Predict> is the actual prediction/localization of the woodpiles. We used trained deep neural network models for the image segmentation task (described in section <4. Model> training and data annotation). The input for such a model is a 3-channel (RGB) image with a resolution of 480x480px. The output is a 1-channel (gray-scale) image with the same resolution, where each pixel determines the probability with which a woodpile is or is not present at a given location. Thus, the output of the whole phase is N gray-scale images with a resolution of 480x480px. Since we chose several image segmentation models within the experiment, section 5.2 *Experiments* presents the results of each model.

For further processing, these N predictions need to be combined. These images are merged in the next phase <4. Join>. For the join, an analogous procedure as in the phase <2. Split> is used.

After the predictions join, an image analogous to the orthophoto map containing the predicted positions of the woodpiles is produced. Since the result from the prediction is only a probabilistic map, these results still need to be processed. The final processing takes place in phase <5. Finalize>. For a more detailed description of this processing, see the next section, 5.1 *Finalize*.

5.1. Finalize

The output from the model is a segmentation mask. The segmentation mask represents the probability that each pixel of the input image is part of the wood stockpile. The post-processing task is to convert such a segmentation mask of probabilities into a set of contours. Subsequently, it is possible to transform these regions into a set of polygons. This procedure is shown in Figure 6.

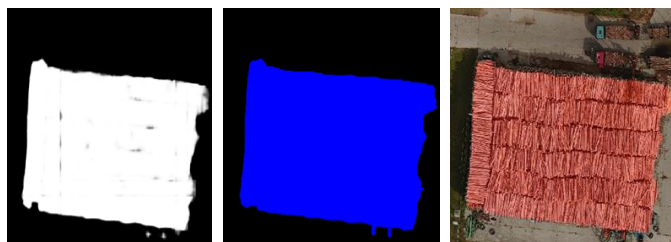


Figure 6. Segmentation mask with probabilities (black and white image with shades of gray), detected area by our algorithm (blue-black image) and projected detected area on orthophoto map (in red)

The left image shows the probability segmentation mask contains empty spaces (places without woodpiles), protrusions, and areas with different probabilities (places where the algorithm located woodpiles). The middle image represents the mask processed by our algorithm. It clearly defines where the stored timber is and where it is not.

The conversion of a segmentation mask with probabilities to a set of polygons directly impacts the prediction quality. The conversion algorithm can be influenced by several hyperparameters, which can significantly impact the final accuracy of the model prediction. We searched for the values of these parameters manually when solving

subtasks. The algorithm uses the OpenCV⁹ library. Initially, the segmentation mask is converted to binary using the threshold parameter. Subsequently, the algorithm performs the close morphological operation on the binary mask using a kernel with the shape *kernel_shape* and size (*kernel_size*, *kernel_size*). The morphological operation serves to close the "holes" which are visible in Figure 6, left, for example. First, contours (contiguous areas of similar color and intensity) are searched for on such a modified binary mask. Next, the algorithm iterates the contours. If the contour area is smaller than the *min_area*, the algorithm excludes such a contour from further processing. Otherwise, the algorithm tries to approximate the contour using a polygon (*epsilon* parameter). Code converting a segmentation mask to a set of polygons:

```

INPUT: mask, threshold, epsilon, kernel_size
SET result TO []
SET mask_bin TO mask > threshold
CALL cv2.getStructuringElement WITH kernel_size, cv2.MORPH_ELLIPSE RETURNING
kernel
CALL cv2.morphologyEx WITH mask_bin, kernel, cv2.MORPH_CLOSE RETURNING mask_mod
CALL cv2.findContours WITH mask_mod RETURNING contours
FOR EACH contour IN contours
  CALL cv2.contourArea WITH contour RETURNING area
  IF area < min_area THEN
    CONTINUE
  END IF
  CALL cv2.arcLength WITH contour RETURNING arcl
  SET econtour TO arcl * epsilon
  CALL cv2.approxPolyDP WITH contour, econtour RETURNING cpoly
  APPEND cpoly TO result
END FOR
OUTPUT: result

```

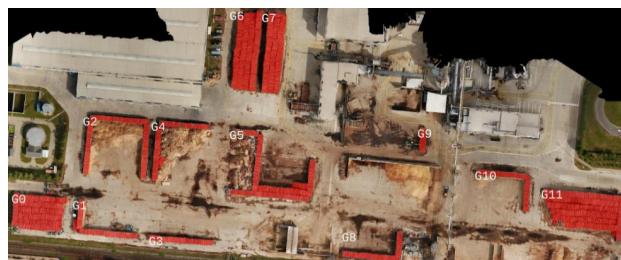


Figure 7. Areas with stored wood that were the survey subject, marked in red, with the group identifier as a recognized area

5.2. Experiments

Figure 7 shows the area where we performed our experiments. The woodpiles are annotated (marked in red) by our algorithm. There are 11 woodpiles labeled G0 to G11 in the figure. To evaluate the performance of the models, we also need to know the actual area of the woodpiles - the ground truth. We obtained these using the Pix4D tools. The ground truth of the stack area ranges from 98 m² (G9) to 1794 m² (G11).

⁹ <https://opencv.org/>

The following table (Table 1) shows the results of the model predictions and GT areas of G0 - G11.

Table 1. Comparison of the results of our predictions of models for surfaces G0 to G11 with the ground truth (GT).

	Model prediction m ²						GT m ²
	FPN ResNet18	FPN VGG16	PSPNet ResNet18	PSPNet VGG16	UNet ResNet18	UNet VGG16	Pix4D
G0	1084	1080	1096	1078	1087	1083	1104
G1	334	314	344	317	344	311	325
G2	536	534	540	528	518	523	597
G3	218	214	207	214	214	211	234
G4	431	350	440	512	467	383	368
G5	915	867	898	890	929	909	948
G6	1147	1149	1156	1177	1150	1155	1138
G7	886	890	903	894	886	891	887
G8	293	281	283	292	290	282	298
G9	87	86	97	83	88	86	98
G10	281	278	272	263	281	276	305
G11	1685	1765	1708	1720	1755	1752	1794

Let M be the set of models, G be the set of areas $\{G0, \dots, G11\}$. A_{mg} is the model prediction for $m \in M$ and area $g \in G$, and GT_g is the ground truth area for $g \in G$. We evaluated the performance of the models using the MAE, CAE, and MAPE.

We can define the MAE (*mean absolute error*) metric as:

$$MAE(m) = \frac{\sum_{g \in G} |A_{mg} - GT_g|}{|G|} \quad (1)$$

The CAE (*cumulative absolute error*) metric is defined as:

$$CAE(m) = \sum_{g \in G} |A_{mg} - GT_g| \quad (2)$$

The MAPE (*mean absolute percentage error*) metric is defined as

$$MAPE(m) = \frac{100}{|G|} \sum_{g \in G} \left| \frac{A_{mg} - GT_g}{GT_g} \right| \quad (3)$$

Table 2. Comparison of individual models predictions using MAP (*mean absolute error*), CAE (*cumulative absolute error*), and MAPE (*mean absolute percentage error*) metrics

	MAE / m ²	CAE / m ²	MAPE / %
UNet VGG16	25.39	304.72	5.63
FPN VGG16	26.41	316.93	5.64
UNet ResNet18	28.98	347.74	6.88
FPN ResNet18	30.24	362.87	5.87
PSPNet ResNet18	33.55	402.63	6.48
PSPNet VGG16	42.48	509.71	9.19

The table shows that the model with U-Net architecture and VGG16 backbone is the best in all metrics. The MAE metric expresses the average absolute error of the area prediction. In the case of the best model, this value is $\pm 25.39 \text{ m}^2$. Because the areas of the measured stacks are diametrically different in size 98 m^2 vs. 1794 m^2 (G9 / G11), the MAPE metric, which abstracts the area size, is also of interest. This metric expresses the average absolute percentage deviation from GT. Again, the best model has this value $\pm 5.63\%$.

The last CAE metric expresses the total absolute deviation of all G areas from GT. The CAE of the best model is 304.72 m^2 . If we compare this value with the total area of woodpiles in the area $GT_{all} = \sum_{g \in G} GT_g = 8096.141 \text{ m}^2$, we get a deviation of 3.76% ($304.72 \div 8096.141$). The error rate from the point of view of individual areas $g, g \in G$ shows the graph in Figure 8.

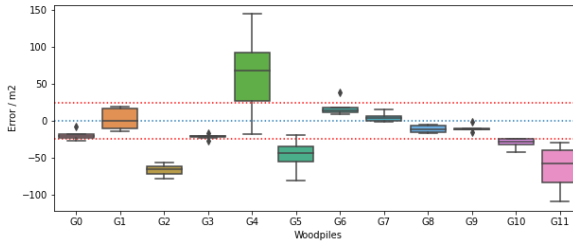


Figure 8. The graph shows the error rate of model predictions for each woodpile.

The graph shows G2, G5, G10, and G11 areas, for which all models show a high error rate. Therefore, these areas can be described as "difficult". Similarly, the G4 area in which most models showed a high error rate. We assume that these areas contain patterns/structures of wood that the trained models cannot recognize. Figure 9 shows

images of the areas G4 and G11 with the marked problem parts. An ongoing analysis of these areas can therefore help increase the accuracy of the solution.



Figure 9. The red squares are the problematic parts in areas G4 and G11, where the models systematically fail.

6. Possible future development – proof of concept

Next, we describe our experience with further developments. These were neither fully implemented in our experimental tool nor fully evaluated so that we can describe their status only as proof of concept.

6.1. The backbone retrain

Our following motivation is to increase the accuracy of a neural network by the backbone improvement and its convergence to error loss elimination and a progression towards a network state where the network has learned to appropriately respond to a set of training patterns within some margin of error.

The backbone in the neural network serves as a features extractor. It is often trained on different tasks such as image classification and different data than we use. An interesting approach may be to retrain the backbone on a task similar to the required objective. For example, we need more annotated training data to retrain the backbone to image classification. However, data annotation is time-consuming and economically demanding; therefore, it is not worth using this exercise.

A self-supervised learning task, known as Contrastive Learning, does not require annotated data, maybe a suitable approach. An example is SimCLR method[2]. Using this method, we can train the backbone on a large set of real data (in the order of 10,000 images) and thus adapt it to our needs and then use it in our semantic segmentation models.

6.2. Increasing the diversity of the training dataset

We aim to improve the following research training dataset incrementally by manually annotating the images where the model showed the highest error rate (see Figure 10).



Figure 10. Example of images with incorrect prediction - candidates for training set extension

6.3. Data input enrichment

In our research, we mainly focused on identifying wood stockpiles from UAV aerial images. Implemented models currently use these images as a prediction input, with each image being 3-channel (RGB).

Since the wood stockpiles are spatial objects, extending the models' inputs by the 4th channel is an elevation map (see Figure 11).

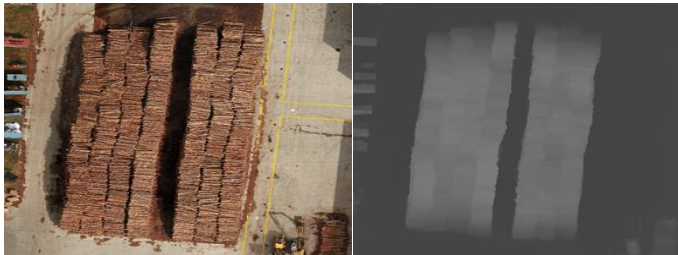


Figure 11. Example of wood stockpiles with the corresponding elevation map

An elevation map can be obtained directly from a LIDAR device as an elevation map or as a side output of the orthophoto map generation. Furthermore, the elevation map (see Figure 12) could extend the input of the prediction model by another channel - the elevation map. Thus, the input can be a 4-channel image (RGB + elevation map) for a composite model that improves semantic segmentation prediction. In the future, using an elevation map could be an important step in volume estimation. Our proof of concept shows it is feasible as envisioned.

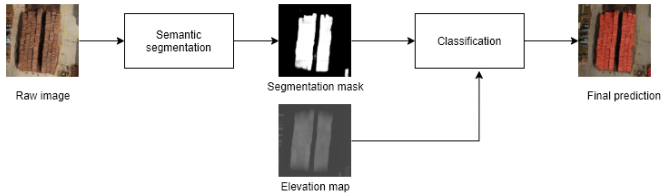


Figure 12. Schematic use of elevation map for improving semantic segmentation

Even under the assumption that the use of elevation maps can be a significant contribution, there are situations where its contribution can be debatable or misleading, for example, when the timbers merge with the surroundings, as in Figure 13 below.

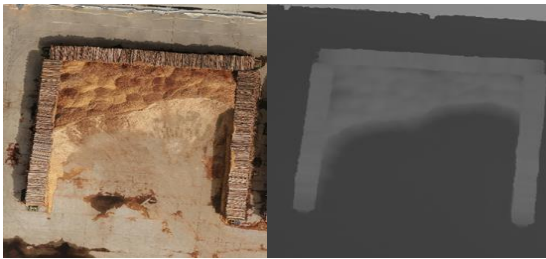


Figure 13. Example of an elevation map where the wood mass overlaps with the environment (beams vs. sawdust or wood chips)

6.4. Other ideas

In paper [14], the authors describe the usage of car catalogs for object recognition. UAV aerial image is compared to one in the catalog with a similarity measure. In this way, it would be possible to estimate real dimensions in the future without the need for ground calibration (known vehicles have known dimensions). Some other objects can have known dimensions, e.g., track gauges. However, similar considerations are, so far, only a future work.

7. Decision support system enriched by knowledge extracted from UAV aerial images

We briefly describe how knowledge extracted from UAV aerial images can be sent to a decision support system. Start with a flat general ontology (e.g., DBPedia¹⁰, Schema¹¹),

¹⁰ <http://dbpedia.org/ontology/>

¹¹ <https://schema.org/>

then we extend it with a Spatio-temporal model [16], and finally with a domain ontology (e.g., here [15]). The connection between OWL and UML can be made by [1] (or by owl2uml, which is a Protege plugin¹²)

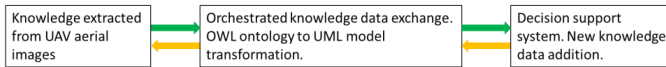


Figure 14. Knowledge exchange

Communication between our knowledge extractor and managerial decision system (inner company processes) goes both ways, as indicated in Figure 14. The right-to-left direction is first dedicated to automated communication of user requirements. There are several conflicting possible user preferences.

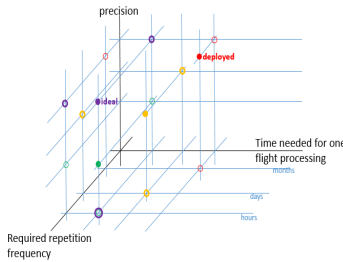


Figure 15. Different optimization strategies

We mention three axes here: the time required for one flight processing, required accuracy and required repetition frequency.

Figure 15 illustrates interrelations between several alternatives: in **red**, our deployed solution with high precision, the considerable time needed for processing, and required repetition in months (solid bullet is the 3D position and bullets with no shape fill are respective projections to 2D planes).

Orange and **green** are alternatives we

consider in our experiments. **Violet** is a position of an ideal point for a user that wants all criteria of maximal benefit. This is a clear multicriterial situation, and we use our learning of aggregation function (to have an FLN-class preference model), see [11]. Many architectures, backbones, and other hyperparameters allow us to move almost continuously along with coordinates within a reasonable range. Trained preference models can then find an optimum in the area.

Our next plans are devoted to more general specifications of customer user requirements in natural language. Using our NLP techniques (see [4]), we can parse sentences into dependency trees, and after automated annotation, we can learn, e.g., a new domain of interest, which can be an entry into web search. In our example, these Datalog rules are pretty simple. In more complicated domains, this learning can be more involved. Communication is based on semantic models on both sides. Figure 16 shows an application diagram for data acquisition and subsequent knowledge extraction.

¹² <https://protegewiki.stanford.edu/wiki/OWL2UML>

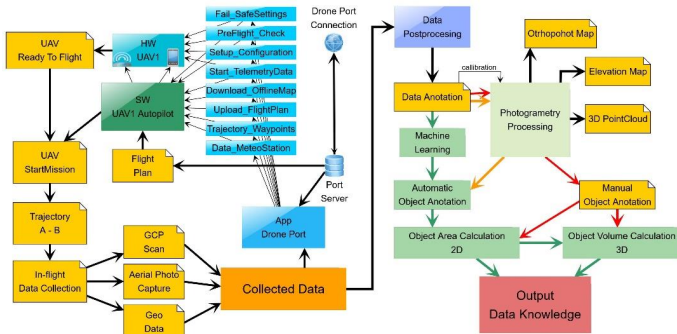


Figure 16. Data collection on the left; Knowledge extraction on the right

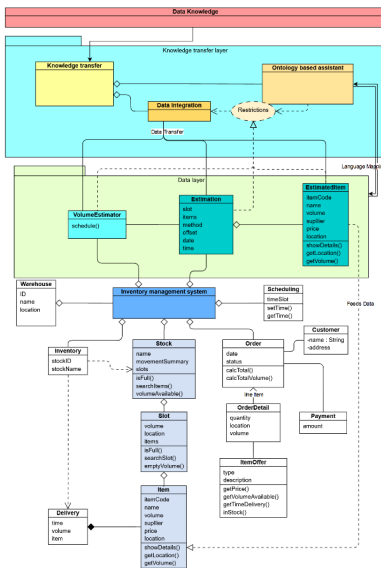


Figure 17. Example of integration with any general inventory management modul (illustrated snapshot)

The extracted data knowledge is possible to transfer into any inventory management system negotiated by ontologies and contribute to the decision process. Inventory management modules worldwide provide processes of maintaining the appropriate level of stock in a warehouse. Inventory management activities involve identifying inventory requirements, setting targets, providing replenishment techniques and options, monitoring storage item usages, reconciling the inventory balances, and reporting inventory status. Integrating inventory management modules with other ERP modules (sales, purchase, and finance modules) allows ERP systems to generate vigilant executive-level reports.

One of the most crucial parts of collecting inventory data information nowadays is obtaining and processing vivid stock data. Real-time data provides unique opportunities to increase the capability of production efficiency in manufacturing environments. Real-time data collection is on the rise in every industry using very advanced approaches to data

collection - machine learning (ML), artificial intelligence (AI), deep learning (DL), unmanned aerial vehicle (UAV), and many more.

Using these advanced technologies, we can develop faster and reasonably precise solutions to obtain warehouse wood mass in real-time. Every warehouse dealing with wood mass has procedures (advanced procedures) on measuring volume wood mass in stock. Most of the time, it is based on volume estimation by experienced workers checking the stock regularly.

Our approach based on using drones to monitor and execute volume estimation can provide data knowledge to transfer to any inventory management system with an ontology assistant's help. Of course, several restrictions and rules must be applied to map the local domain model and data. However, such a knowledge management ontology tool can smartly integrate the resources into a coherent corpus of interrelated information as an inventory management data addition. Figure 17 shows a brief example of such a model.

Ontologies offer an alternative way to cope with heterogeneous representations of customer internal inventory management models. The domain model implicit in an ontology can be considered a unifying structure for giving information a common representation and semantics. The idea is that collected and calculated data can transparently transform from the UAV data processing model by introducing the knowledge transfer ontology-based assistant.

8. Related work

There are many commercial products and many more UAV aerial image processing applications: [17], [5], and, e.g., use of GNSS, ArcGIS reported by [13], to mention a few. [13] noticed that ArcGIS was 12-20 times faster than the use of GNSS, with comparable precision. On the other side, we can see that their camera and imaging gave a smaller density of points than ours. The main difference between our method and that of [13] is the increased precision of measurements achieved by a camera with higher resolution and a new UAV hardware generation. Secondary differences are twofold: A) we used the precise targeting of space using GCP, the reality against the virtual model, which created an even more accurate digital projection of the terrain. B) We adapted the data processing process in the Pix4D process settings, intending to achieve maximum accuracy.

Wood or forest-related studies, where objects may be under treetops, often use geographic or geological tools (see, e.g. [3]) to calculate accurate volumes using Lidar data.

Great motivation for us was papers [21] and [7]. Further ideas from [14] and [18] were also very inspiring. A big help was a lot of public domain software – detailed references are in footnotes on the appropriate place in this paper.

There is yet a broader context of our work, namely integrating neural (subsymbolic) and symbolic AI, which can be considered as a fifth dimension (added to 3D + time) as used in [10] and [19]. While machine learning has advanced thanks to deep neural networks rapidly, the trial and error approach it uses is similar to the way humans learn. It is sometimes failing due to the lack of data or context. However, humans developed language and other systems that make it possible to pass on knowledge directly to others who integrate that into their knowledge. [8] looks at the rules which enable knowledge transfer to work best in AI and integrate them with existing machine learning approaches.

[9] aims to bridge between the two paradigms. Authors discuss neural-symbolic integration in relation to the Semantic Web field, focusing on promises and possible benefits for both, and report on some current research on the topic. The particular added value of [10] and [19] is the highly elaborated user interface, which increases intuitiveness.

In our approach on the symbolic AI side, we were motivated by [16], [15], [1], [6], and our former work, e.g. [4] and [20]. We used our [11] approach to fuzzy multicriterial systems based on the Fagin-Lotem-Naor FLN class of models. In one direction, we learn the preference model; in the opposite direction, we use the trained model to send users useful information.

Connections to the automotive industry are our long-term interest. In the paper [12], the authors describe a questionnaire for Spain's industry and its conclusions. Maybe we could go that way as well.

9. Conclusions and future work

Our long-term interest in studying possibilities and ways to increase automation, efficiency, and digitization of industrial processes using autonomously controlled UAV means interconnected with managerial decision support systems (especially in the automotive industry). In this paper, we took the first step to master the necessary methods in a much simpler domain. We consider two-way communication of knowledge and requirements between our system and an industry managerial system.

Here we presented our results in the secondary wood processing industry. First, we presented a deployed solution for calculating woodpiles volume from our UAV flight images during a time frame of 9 months. Processing is based on commercial photogrammetry software with some manual human intervention necessary. Second, we developed alternative automated solutions based on deep neural network learning and experimented with several deep neural network architectures, several backbone variants, and hyperparameters on real-world data.

Future work considers the use case extensions, e.g., concerning wood quality, monitoring the whole process from the forest via sawmills, transportation, various warehouses, and a more significant number of users.

In future work (hopefully in the final version after the conference), we would like to extend our experiments by studying the influence of loss function, training set, learning rate, learning scheduler, and regularization on final results.

10. References

- [1] Brockmans S. et al. Visual Modeling of OWL DL Ontologies Using UML. In – ISWC 2004, McIlraith S.A. et al. eds. LNCS 3298, Springer 2004, 198-213
- [2] T. Chen et al. A Simple Framework for Contrastive Learning of Visual Representations. In Proc. 37th International Conference on Machine Learning, PMLR 119:1597-1607, 2020 also <https://arxiv.org/abs/2002.05709v3>
- [3] F. Chudy et al. Identification of Micro-Scale Landforms of Landslides Using Precise Digital Elevation Models, *Geosciences* 2019, 9, 117; doi:10.3390/geosciences9030117
- [4] J. Dedek. Semantic annotations. 2012 Ph.D. thesis, Charles University, Faculty of Mathematics and Physics, Dpt. Software Engineering, advisor P. Vojtas <https://dspace.cuni.cz/handle/20.500.11956/41689>
- [5] Volume Measurement with Drones. <https://support.dronedeploy.com/docs/volume-measurement>,
- [6] J. Euzenat, P. Shvaiko. *Ontology Matching*, Springer 2013
- [7] G. Ghiasi et al. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation, arXiv:2012.07177v1
- [8] G. Gottlob - Knowledge Processing, Logic, and the Future of AI - World Logic Day 2021, Vienna Center for Logic and Algorithms, January 14th, 2021 <https://www.youtube.com/watch?v=i-HeIYBJEw>
- [9] P. Hitzler et al. Neural-symbolic integration and the Semantic Web. *Semantic Web* 11(2019)1-9
- [10] Y. Kiyoki et al. System with SPA-Based Semantic Computing for Integrating and Visualizing Ocean-Phenomena with "5-Dimensional World-Map", *Information Modelling and Knowledge Bases XXXII*, IOS Press, pp. 76-91, January 2021
- [11] Kopecky, M., Vojtas, P.: Visual E-Commerce Values Filtering Framework with Spatial Database metric. *Computer Science and Information Systems*, 17,3 (2020) 983–1006
- [12] A. Martínez Sánchez, M. Pérez Pérez. Supply chain flexibility and firm performance: A conceptual model and empirical study in the automotive industry *INTERNATIONAL JOURNAL OF OPERATIONS AND PRODUCTION MANAGEMENT* 25,7 (2005) 681-700
- [13] M Mokroš et al. Unmanned aerial vehicle use for wood chips pile volume estimation. *ISPRS - The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41 (2016) 953-956
- [14] T. Moranduzzo and F. Melgani, "Detecting Cars in UAV Images With a Catalog-Based Approach," in *IEEE Transactions on Geoscience and Remote Sensing*, 52,10 (2014) 6356-6367
- [15] A. Öhgren. Developing an Ontology for Wood-related Industry: An Experience Report. Research Report 04:6, School of Engineering, Jönköping University access December 10th, 2020
- [16] Ch. Parent et al. Spatio-temporal conceptual models: Data structure + space + time. in *ACM-GIS'99*
- [17] Pix4D: Professional photogrammetry&drone mapping <https://www.pix4d.com/>
- [18] Radovic, M.; Adarkwa, O.; Wang, Q. Object Recognition in Aerial Images Using Convolutional Neural Networks. *J. Imaging* 2017, 3, 21
- [19] S. Sasaki et al. Global & Geographical Mapping and Visualization Method for Personal/Collective Health Data with 5D World Map System, *Information Modelling and Knowledge Bases XXXII*, IOS Press, pp. 134-149, January 2021
- [20] V. Vanekova, P. Vojtas. Comparison of Scoring and Order Approach in Description Logic (EL(D)). In *SOFSEM 2010*, J. van Leeuwen et al. eds. LNCS 5901, pp. 709-720, Springer 2010
- [21] Hengshuang Zhao et al. Pyramid Scene Parsing Network, arXiv:1612.01105v2
- [22] {BV} Brezani, S., Vojtas, P. (2021). Aggregation for flexible Challenge-Response. to appear in *FQAS'21*

A Knowledge-Model for AI-Driven Tutoring Systems

Andreas Baumgart^{a,1} and Amir Madany Mamlouk^b

^a*Engineering and Computer Science, HAW Hamburg, Germany*

^b*Neuro- and Bioinformatics, University of Lübeck, Germany*

Abstract. A powerful new complement to traditional synchronous teaching is emerging: intelligent tutoring systems. The narrative: A learner interacts with a digital agent. The agent reviews, selects and proposes individually tailored educational resources and processes – i.e. a meaningful succession of instructions, tests or groupwork. The aim is to make personal tutored learning the new norm in higher education – especially in groups with heterogeneous educational backgrounds. The challenge: Today, there are no suitable data that allow computer-agents to learn how to take reasonable decisions. Available educational resources cannot be addressed by a computer logic because up to now they have not been tagged with machine-readable information at all or these have not been provided uniformly. And what's worse: there are no agreed conceptual and structured models of what we understand by „learning“, how this model-to-be could be implemented in a computer algorithm and what those explicit decisions are that a tutoring system could take. So, a prerequisite for any future digital agent is to have a structured, computer-accessible model of “knowledge”. This model is required to qualify and quantify individual learning, to allow the association of resources as learning objects and to provide a base to operationalize learning for AI-based agents. We will suggest a conceptual model of “knowledge” based on a variant of Bloom’s taxonomy, transfer this concept of cognitive learning objectives into an ontology and describe an implementation into a web-based database application. The approach has been employed to model the basics of abstract knowledge in engineering mechanics at university-level. This paper addresses interdisciplinary aspects ranging from a teaching methodology, the taxonomy of knowledge in cognitive science, over a database-application for ontologies to an implementation of this model in a Grails service. We aim to deliver this web-based ontology, its user-interfaces and APIs into a research network that qualifies AI-based agents for competence-based tutoring.

Keywords. Competence-Based Learning, Knowledge Concept, Ontology, Web-Service, Taxonomy of Knowledge, Web-Application

1. Introduction

Digital learning is seen as a new and promising approach in creating a more effective and efficient teaching environment, in delivering higher returns on education. One line of development has been to provide structural elements like educational resources and courses online – many of them being freely accessible or openly licensed as in HOOU [17] – or to provide environments for testing domain-specific knowledge – as in MINTfit [22].

¹ Corresponding Author, Corresponding author, Book Department, IOS Press, Nieuwe Hemweg 6B, 1013 BG Amsterdam, The Netherlands; E-mail: bookproduction@iospress.nl.

The other line of development addresses processes required to implement educational resources into active learning, including pedagogies as e.g. blended/hybrid learning or gamification. A distinguishing feature between these processes is the level of individuality that a teaching environment can offer – from addressing the full group of students with the same interventions to addressing one individual student with specific, tailor-made interventions. It is assumed that an individual offer to a student creates improved learning opportunities in a significant number of settings (Wood [30]). Such an individual approach could be incorporated by a tutoring agent, an intelligent tutoring system (Graesser [12]).

This agent must – in order to add value to any teaching environment – have a conceptual understanding of the teaching-subjects and their dependencies. And such a conceptual model in computer accessible form does not exist yet.

A prototype for such a model is being presented in this manuscript.

2. Structures and Taxonomies for Knowledge

With knowledge being a substantial competitive factor for individuals and national economies alike, scientists have long developed models to make learning-outcomes measurable and comparable. This section describes, how existing concepts have been integrated into our approach.

2.1. Learning Analytics

Learning Analytics and its scientific research field Educational Data Mining describe ways to employ data - e.g. test-results - generated by students when using e-learning platforms. These data can be used – subject to data protection regulations - to monitor learning behavior and progress of an individual or a group of students, the data can be analyzed with statistical methods, visualized and can be employed to compare between groups, between teaching methodologies employed for learning and between individual teacher personalities or could drive computer adaptive testing. However, these data are usually not appropriate for an AI approach to learning:

Data acquired by Learning Management Systems (LMS) are mostly points that students achieve in tests. These tests usually have no metrology - no measurement system - for knowledge in place. A common example is a succession of tests (quizzes) in a course implemented in Moodle. Each test delivers an aggregated point-value which stands implicitly for a certain achieved learning outcome. The point-values have no meaning beyond the test.

As a remedy, the LMS “Moodle” offers Competency Frameworks. These are hierarchical tree structures with a parent / child -structure. So Mechanical Engineering could have “Solid-State Mechanics” and “Fluid-Mechanics” as children, “Solid-State Mechanics” could be the parent of “Homogeneous Materials” and “Heterogeneous Materials”. The test-performance in child-concepts can then be employed to indicate performance in parent-concepts. This approach creates structure – a semantic tree - but not a meaningful context between knowledge items. The tree might assign “Vector-Algebra” as a child-node of “Geometry” in the Competency Framework “Mathematics” – but it is also the pre-requisite for understanding equilibrium conditions in statics – which is engineering or physics.

As a remedy, Halloun [13] designed tests some 40 years ago to measure students' understanding in the application of physical concepts, e.g. "force". Questions in these tests are not confined to a singular knowledge-item like "force" but aim to find indicators for the degree of conceptual-understanding that students have.

All questions in such a Concept Assessment Test (CAT) are subject to a rigorous scrutiny, they are therefore called evidence- and research-based questions. To give reproducible and meaningful results, CATs must follow a strict standardized approach: no question must be altered or replaced, no questions must be added and the time available to solve all questions is fixed. In engineering, the CAT for statics, the "CATS", consists of 27 questions for students in the first semester of engineering mechanics. It tests for 8 concept-categories with "static equilibrium conditions" being one of them. So, CATS can be used to measure mastery of a set of learning concepts from engineering e.g. after a course when comparing different teaching approaches as has been described successfully by Direnga [10] or with a similar intent by Adams [1].

Another approach is the item response theory (IRT) which comes with a dedicated statistical analysis approach to the problem (Hambleton [14]). In contrast to CATs, questions can be designed as needed and can have different weighing factors when contributing to the overall result.

Common for both approaches though is that the concepts or items tested for have no explicit meaningful connection to each other – their interrelation can only be established by a discussion between the designers of the tests. And both approaches have no objectively defined level of reference-knowledge defined – the skill-level achieved when obtaining all points in a test is not

Example: As the author of a test, I will have a subjective understanding of the conceptual knowledge that students have attained, and the points not awarded indicate the distance to the intended learning objective. I could pass on this test to a colleague and explain my understanding of the test and the targeted learning objectives.

For an AI, this approach is "incomprehensible", it would need more structure, a schema of knowledge to adhere to. Knowledge must therefore be decomposed into distinct and explicit subjects of suitable complexity – as in the concept-categories of CAT. But these concepts should not be confined to be stand-alone items in one individual course at one university. Example: "static equilibrium conditions" is not a sufficiently distinct concept as it may refer to an approach with force equilibriums or with analytical methods as in the Principle of Virtual Work.

And the structure should incorporate an objective skill-level that indicates the learner's performance in employing the subject to solve a problem. Example: a student may compute the forces required to achieve "static equilibrium conditions" in a practice-and-drill approach or analyze geometric conditions under which no solution to the problem exists, with the latter being significantly more demanding.

2.2. A Schema for Learning Objects

Similar to the approach of structuring knowledge is the idea to structure learning objects to an agreed schema. In IEEE [19] a schema is proposed for the representation of interoperability conditions of learning objects using learning objects metadata (LOM). This standard uses XML Schema definition to describe and define an aggregation relationship between data items to considerable detail: it specifies e.g. roles, resource types and interactivity levels of LOM. Like Moodle Competency Frameworks, it creates a tree-structure of parent/child elements. But in this approach, a child-element can have

a specific composition attribute, e.g. can be a “Requirement” or an “OrComposite”. And it adds difficulty levels – from “very easy” to “very difficult”.

2.3. Consolidated Approach

Both Moodle and IEEE approaches add a text-book-like structure to knowledge-elements and LOM. Additionally, IEEE offers a rough taxonomy (from “easy” to “difficult”) e.g. for an exercise. They both provide a qualitative structure of a body of knowledge.

Krathwohl [4] creates a detailed taxonomy of knowledge for one individual knowledge-base aimed at quantifying skills. But its concept includes no relations between the knowledge-bases and was not designed to be implemented in the context of information technology.

Instruments like Concept Inventories and IRT describe in considerable technical detail a confined element of knowledge to be tested for and focus on the analytics to assess a learners’ skills. The target of the analysis is a statement as “concept understood / not understood” and does not include a rationale for the quantitative manifestation of knowledge so it could be “read” by a computer.

And none of the above account for the semantic relationship between learning objects.

In Chi [8] and Paquet [23], the focus is tuned to the modelling of semantic networks for knowledge, the latter targeting computer applications. Paquet’s work provides an excellent overview and describes the generic approach to engineering ontologies and the abstract objects to be employed. However, this approach does not explicitly incorporate Krathwohl’s taxonomy (though it refers to Bloom’s taxonomy) and the models are not built as an interactive Web-application to be used by tutoring agents.

What we aim for is a composition of key features from these references. Our representation of knowledge must be tailored to provide structured access to human knowledge for a computer application so it can mimic decisions in tutoring scenarios. Our approach is to

- breakdown a knowledge domain into discrete items,
- represent inferences between items that allows for an abstraction of dependencies between items,
- include an explicit continuous taxonomy that quantifies for each knowledge item a student’s performance or skill to solve specific problems,
- describe this taxonomy for discrete, uniform, hierarchical taxonomy level prototypes – synonym to Krathwohl’s categories – represented by one or several statements describing a learner’s capability to perform a task and
- incorporate these items in a collaborative, web-accessible platform.

This approach to knowledge representation will not – as it may appear – degrade an agent to adhere to formal rules (Alven [2]) or simple “if-then-else” approaches. But in a situation where no structured data from students are available that represent a learning process, a tutoring algorithm can start teaching. It can derive learning tracks from the links in the knowledge representation, propose to either deepen the understanding in one knowledge-base or broaden the knowledge by adding new neighboring bases instead. This, however, is pure guesswork or a random approach. We do not think that this is a major drawback for a tutoring agent, since my strategy as teacher is similar – even though I may not be aware of the concrete decisions that I take in a specific scenario.

Without data being available on “success-tracks” in our learning landscape, an algorithm cannot decide, which decision is best – neither statistically nor individually. In other words: there is no Dijkstra- or shortest-path-algorithm for this problem. But any student consenting to having his or her data of the learning process to be recorded adds a meaningful set of data to the collection. Heuristics as “first deepen a subject, then broaden” may emerge and develop to providing meaningful, individually relevant teaching interventions at specific branching points. With each individual learning experience recorded and made available to other agents, an agent may see a “beaten” track of successful learning emerging. This will be the kick-off for a novel recommendation system. And this is where AI comes into play.

3. Concept of Knowledge

Our model of knowledge shall be an agreed, comprehensive and complete representation of abstract subjects to know. We use “mechanical engineering” as an example for a knowledge domain because it is well suited for modelling conceptual knowledge and has a consolidated, agreed body of knowledge. If the complete landscape of structured knowledge exists today in any form, it is locked in the minds of university professors and teachers. What we intend is to project this mind-model into something that finds its symbolic representation in a computer system and is sufficiently standardized so that different people would come up with similar knowledge-models.

So far, we have been using the catchy expression “knowledge” to name the intended outcome of learning. Knowledge, however, is not a standardized and not an unambiguous expression. To make things more complicated, we address different science domains – educational psychology, the modelling of mental structures, web-application frameworks, instructional methodology and engineering mechanics – which introduce their own subject-specific nomenclature. We will therefore “integrate” expressions from these domains and attempt to create a consolidated system of terms.

The connecting element in this approach shall be our vision of how AIs would employ the knowledge-model-to-be. As a teacher, I like the idea that our graduates learn to address professional problems in engineering by sketching and discussing a subject on a paper napkin over lunch. My picture implies that these engineers have a good command of a common technical language, need not to retreat for days of consultation to come up with a first justifiable opinion and understand the basics of their business, e.g. coordinates, principles, axioms (like the equilibrium of force) intuitively. A teaching approach which explicitly aims at these practical skills is “competency-based learning”. Therefore “competence” shall be the lead-term in the development of this model.

Competency-Based Learning is characterized by properties, that we are addressing with a tutoring system in a similar way: its concept focuses on learning outcomes, the learner itself and a differentiation of appropriate learning approaches.

Foremost, learning outcomes are described as clearly observable performances of a student. This does not imply a particularly (high) level of mastery nor a point rating in terms of “8 out of 10”, but rather an approach that describes concrete levels of achievements.

The student in this context is seen as an individual – having a unique personal, social and methodological background. Teaching should account for this background in the sense that any knowledge and any qualified perception of this background should be accounted for in teaching.

Finally, differentiation implies to account for this individual background in a most appropriate way, e.g. in the proposal of group- or single-work, the choice of learning objects or interventions as status-tests.

In this sense, competency-based learning phrases a concrete and realistic strategy for tutoring by agents using our knowledge-model. And it created an agenda of how to translate the above picture of engineers debating over lunch into a concrete definition of learning interventions.

As in the design of a curriculum, the learning objectives have two dimensions: the subjects like “static equilibrium of forces” and a depth of their understanding.

In our model for cognitive knowledge, we call a learning-subject

- a “competence-base”,

their inferences

- a “competence-relation” and

the depth of understanding of this subject

- a competence-level.

Our competence-level-prototypes will describe distinct skill-levels and are identical to Krathwohl’s categories. The new expression used intends to strengthen the consistency of terminology around the “competence”-concept.

3.1. The Knowledge Map

We approach this model with a hiking-analogy to learning: knowledge is seen as a landscape, our model is a map that represents the topology of the landscape and our student hikes in this landscape.

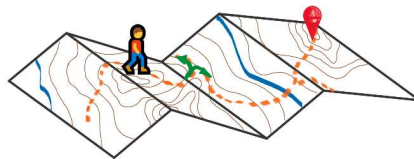


Figure 1. The Imaginative Learning Landscape.

An AI-agent would know from online-tests, “where” the student is, would know the topology of the map and the next target from the curriculum specification. Using information from the map, an AI would guide the student in this landscape employing waypoints and routes. So we first need to provide the map of a knowledge-landscape, including topology and infrastructure information.

Initial step to our model is a discretization and mapping of the learning-landscape to a network of competence-bases:

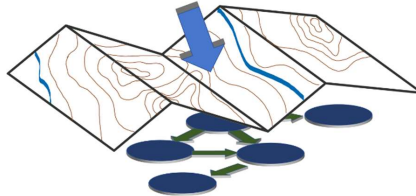


Figure 2. From the Map to a Discrete Network of Competence-Items and Relations.

Each competence-base (node) is connected to neighbors by relations (edges), these relations express directed conceptual inferences. The learning-items in a curriculum-design using our knowledge-model would thus be defined by selecting competence-bases. Then, the level of practical application skills shall be quantified as competence-levels.

3.2. A Semantic Network for Knowledge

This definition of learning-targets as competence-levels is distinct and unambiguous – if the knowledge-model is consistent. An agent can now follow all relations in the knowledge-model and identify all pre-requisites for achieving a learning target within a course and – by employing the conceptual inferences between the competence-bases - break it down to learning sessions, educational resources, and tests.

In this way, an agent will be able to assess the competence-level associated with a competence-base of an individual learner by appropriate tests. In our hiking-analogy, the agent queries all possible learning paths to the “GPS” coordinates of the learning objective. By weighing the current level of knowledge of a learner and the available learning material, the agent with AI can then optimize a “path” towards the learning goal.

3.3. Capabilities Associated with a Competence-Base

For each competence-base, we need to describe depths of understanding. In his revised version of Blooms taxonomy, Krathwohl [4] works with six discrete categories as attributes in a learning-taxonomy: remember, understand, analyze, apply, evaluate, and create. We adopt this approach and depict learning as adding value to a competence base.



Figure 3. Learning Seen as Adding Value to a Competence-Base.

Please note also the implicit statement here: learning in category four (apply) is only possible when building upon category three competence (analyze)!

Krathwohl made some effort to create a definite and clear description of these categories. They are key to a concise application of the model so that different parties share the same understanding. For each category, he listed verbs that should be employed when describing one category – here named competence-level. In the category “apply” for competence-base “static equilibrium of forces”, the statement describing the competence of a learner could be:

“The student can build the system of ordinary equations for the bearing forces that describe the equilibrium condition of the body.”

A selection of other verbs suggested by Krathwohl describing this category are

- apply
- choose
- construct
- develop
-

Along this line of thought learning and learning strategies can be tutored by an AI driven tutoring system: the knowledge-model that we create provides distinct decision opportunities, e.g. in terms of adding value to a competence-level or addressing a neighboring competence-base.

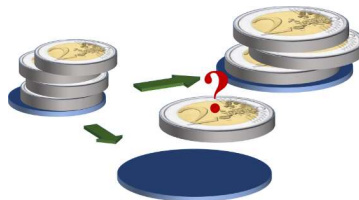


Figure 4. A Decision Opportunity: Should Learning Add Value to a New Base or Increase the Value of a Previously Built Competence-Level?

We now have the three ingredients in place that are required to represent our knowledge-concept in a computer algorithm:

1. we have proposed “competence-based learning” as a concept for the teaching methodology that our knowledge-model intends to support by conveying structured information of knowledge,
2. we have the competence-bases and their inferences defined that represent the qualitative items in the learning landscape, and
3. we have the competence-level adopted from Krathwohl [4] so that we can consistently describe a depth of understanding for each competence-base.

4. Representation of a Concept of Knowledge

Our concept of knowledge employs associated quantifiable objects. Learning and knowledge are linked by interpreting learning as an increase in competence-level. Our concept should be able to represent subjects of very different complexity – not only in engineering mechanics – to be useful. But at the same time, it should build upon a very limited number of classes of objects in order to be manageable and user-friendly.

As for all models, it serves a specific use in its abstraction of the world. It simplifies real physical, social or economic systems and there can be different appropriate models to serve one purpose. The challenge is usually to make the model as simple as possible and as complex as necessary.

4.1. Usability Criteria

Criteria for the usability of our knowledge-model have been clearly phrased in Paquet [23] as

- simplicity: it should be intuitive and easy to operate by untrained users,
- generality: it should not be tailored for one individual thematic area only (as mechanical engineering),
- completeness: it should not prohibit the representation of essential elements of knowledge due to missing types of competence-bases or relations,
- ease of interpretation: it should have a concise and plain structure so that the meaningful context of the model is self-explanatory without ambiguity,
- standardization and communicability: it should support the intuitive interpretation and the exchange of contents among users, and
- computability: it should allow an AI or servers for educational resources to be able to readily access the information contained.

4.2. Nodes

In adopting elements of a particular approach, we aim for simplicity rather than philosophical rigor. First, we follow Romiszowski [26] in his approach from instructional design to assign knowledge-items to be either factual or conceptual (abstract).

E.g. in mechanical engineering, an essential knowledge-item is the “freebody-diagram” – a pictorial representation of the forces and moments acting on a body. It is usually attributed to Joseph-Louis Lagrange who addressed it in part I (Statics), Section IV, §1 of his *Mécanique Analytique*. In engineering, the “freebody-diagram” would be considered conceptual knowledge, the information about author, publication etc. would be considered factual. For our knowledge-model, we will not account for factual knowledge. It only accounts for conceptual knowledge. This implies that it holds e.g. instances of principles, axioms, and rules. Thus, it is a container for the understanding of a knowledge-domain but cannot be employed to solve actual engineering problems that require factual knowledge of the systems as mass, dimensions, gravitational acceleration etc. Following Romiszowski [26] and Paquet [23], we further subdivide conceptual knowledge into

- concepts: they represent key objects in the knowledge domain, here coordinates, equilibrium conditions, forces, strains, stresses, eigenfrequencies etc.
- procedures: they represent operations on objects, here the solution process for problems in statics, the computation of principal stresses in a continuum etc. and
- principles: they represent undisputable statements, here axioms, principles, nomenclature, naming or color-conventions including ISO-norms etc.

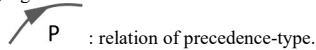
These three are accounted for and will represent elementary types of nodes in our model. Also note that sometimes (e.g. in engineering literature, the “freebody-diagram”,

in German: “Lagrangesches Befreiungsprinzip”) is either attributed as an axiom or a principle. We will not differentiate between these two in our model and use “principle” as a representation for both. Later we will graphically represent the three competence-bases as a circle with the circle-color indicating its type, i.e.

- : concept,
- : procedure and
- : principle.

4.3. Edges

Relations between our competence-bases – the edges – will be represented by a directed arc, e.g.



Following Paquet [23], the relational objects between nodes have each one out of six possible types, namely

- I-Link / Instantiation: the Principle of Virtual Displacement is an instance of the Principle of Virtual Work,
- C-Link / Composition: the concept of “force-equilibrium” has the “force”-concept as a component,
- S-Link / Specialization: a 2D-problem is a specialization of a generic static equilibrium problem,
- P-Link / Precedence: a “freebody-diagram” precedes the “static force equilibrium” in a solution process,
- I/P-Link / Input/Product: a quantity and vector are input to the procedure of “scalar multiplication”, the product is a vector, and
- R-Link / Regulation: the axiom on “equilibrium conditions for a rigid body subject to two forces” regulates the concept “static force equilibrium”.

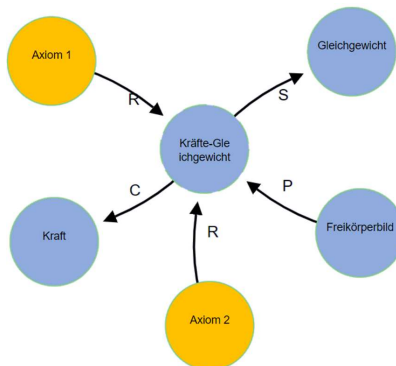


Figure 5. A Small Section of the Model.

In AI-literature, this approach to represent knowledge as linked graphic elements has been introduced under the name "semantic networks" (Hulpus [18]). Built upon the nodes of this network – or the competence-bases – is a taxonomy of detailed descriptions of skills in the form of prototypes for discrete competence-levels – from “remember” to “create”. In the same way, the competence of a student can be diagnosed and quantified from appropriate tests providing a measure for the mastery of a certain subject.

5. The Web-Interface for Knowledge-Domain Experts and Tutoring Agents

This section describes the requirements and consequently implemented features of an application that we propose as an exchange-platform between teachers and AI. Due to the interactive nature of its key-function – to provide an exchange of structured information between multiple actors – this application is web-based.

It serves two purposes:

- provide an interface for knowledge-domain experts to facilitate the collaborative creation of a consistent knowledge-model,
- provide an interface for the interoperability with other web-based services on the Internet.

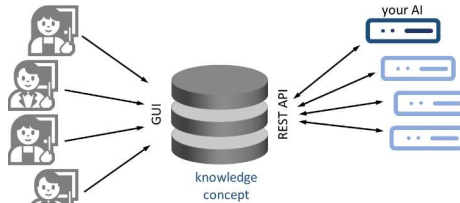


Figure 6. The database as exchange-medium between teachers and Web-based clients.

5.1. User-Interface

This section will describe the features of a user-interface to our proposed knowledge model for teachers. Our main success-scenario is that a teacher

- identifies, generates and describes a set of clearly demarcated competence-bases including their relations,
- describes each competence level in detail – including references to e.g. text-“books” (wikis) or literature and
- links the additions made to already established and agreed competence-bases.

To facilitate this scenario, the software development has been based on several use-cases. We started from the point of view of engineering mechanics – the knowledge domain that serves as the crystallization point for this application – which employs graphic content and mathematical formulas in abundance. It also turned out that users spend most of their effort on declaring and editing prototypes of competence-level-definitions. A key requisite of the software is therefore to support the unambiguous declaration of these levels, i.e. to create a distinct manifestation of a competence on explicit prototype levels. The goals that were identified are

- facilitate a simple and transparent collaboration of individuals,
- make for an intuitive, easy editing experience that allows for text-formatting, images, formulas (LaTeX), and links,
- provide featured support for teachers in describing competence-levels - e.g. by proposing appropriate verbs or examples per level-prototype,
- implement a roles & rights concept to protect and allow ownership of the created contents
- allow individuals and organizations to control the extent of model-transparency and collaboration,
- have a staged approval process installed that allows content classed from “sandbox”- to “master”-level,
- aim for a generic knowledge-model that can be applied to many domains – not only engineering mechanics,
- provide a visual graph of the conceptual model that can be controlled w.r.t depth of representation (number of relations from reference-base) / contributors and allow for a standardized exchange of the knowledge-model (upload / download of XML-files)

```

1 package de.knowledge.concept.domain.competence
2
3 import de.knowledge.concept.domain.Application
4 import de.knowledge.concept.domain.Level
5 import de.knowledge.concept.domain.reference.Information
6 import de.knowledge.concept.domain.reference.Section
7 import de.knowledge.concept.domain.security.User
8 import de.knowledge.concept.enums.CompetenceTag
9
10 abstract class Competence {
11
12     String name
13     String description
14     String shortDescription
15
16     User user
17     Section section
18     CompetenceTag competenceTag
19
20     static hasMany = [
21         applications: Application,
22         levels: Level,
23         informations: Information,
24
25         sources: CompetenceRelation,
26         targets: CompetenceRelation
27     ]
28
29     static mappedBy = [
30         sources: 'target',
31         targets: 'source'
32     ]
33
34     :
35     :
36     :
37     :
38     :
39     :
40     :
41     :
42     :
43     :
44     :
45     :
46     :
47     :
48     :
49     :
50     :
51     :
52     :
53     :
54     :
55     :
56     :
57     :
58     :
59     :
60     :
61     :
62     :
63     :
64     :
65 }

```

Figure 7. Abstract Grails Domain for Competence-Bases.

A prototype of the application was first done in Protégé [24] to define the basic classes, their dependencies and understand procedural interactions. The Web-application was then implemented in Grails.

Core of the application is the definition of the abstract competence-base domain (see Fig. 7). The individual types of competence-bases extend this abstract domain, e.g. “Principle” extends “Competence”. The competence-relations are declared by a domain employing different types, e.g. type “specializes”.

5.2. Features

Creating a knowledge model is a very time and resource consuming process. Accordingly, it is of crucial importance that models, once they have been created, can be saved, exchanged and further developed. Therefore, as a basic feature, our editor allows to import (upload) a competence model in XML-form or export (download) the competence-model presently stored in the database.

Furthermore, as formulas are essential ingredients to describe some knowledge-domains, we have also created the option of embedding formulas directly in the textual descriptions using LaTeX with a `%{\latex(.)}`-tag, e.g. `“%{\latex(“\vec{F}”)}”` as depicted in the edit-page of a Principle:

Name *

Description

Thus, in a list of concepts, the “force-equilibrium” would appear like this:

<u>Name</u>	<u>Description</u>	<u>Short Description</u>
Kräfte- Gleichgewicht	Aufstellen der Gleichgewichtsbedingungen für alle Kräfte und Momente: $\sum \vec{F}_i = \vec{0}$ $\sum \vec{M}_i = \vec{0}$	Kräfte- und Momentengleichgewicht

6. An Excerpt from the Ontology

Complex conceptual models could explain the whole world – but would make them very difficult to implement and populate. Therefore, we have explicitly targeted an as basic as practical model which provides only few node- and edge-types. As a result, it is rather simple to elaborate the knowledge-model for a single knowledge-domain. Though basic, the software GUI allows for an intuitive approach. And thus, models usually grow rapidly.

6.1. Building your Knowledge-Domain

Assume you are an editor of the ontology and start from scratch with an empty database. You would choose a central competence-base as e.g. “force-equilibrium” and add it as competence-base #1. What other competences depend on “force-equilibrium”; what competences are needed to understand it?

Obviously, “force-equilibrium” requires “force” and we add “force” as a concept to the database. Since “force” is part of “force-equilibrium” we employ the “composition” or C-Link between the two. In our user-interface, you would see two blue circles connected with an arrow, all three elements being linked to objects in the database. Next comes the concept of “freebody-diagram” – which visualizes the impact of bearings or gravitational forces on a body by drawing respective forces and moments. Both – forces and moments – are represented as red arrows in the diagram in accordance with specific rules. Without the freebody-diagram, forces would not be “visible” and can not be employed in a force-equilibrium – so the freebody-diagram is a pre-requisite to the force-equilibrium. We connect the two concepts with a precedence or P-Link from freebody-diagram to force-equilibrium. The red arrows in the (factual) freebody-diagram are a special representation of a concept that we have defined already: force. We define “graphical representation” (of force) as a component of the freebody-diagram – again using the C-Link. And the graphical representation of force is a specialization of the concept of force.

So far we are only using concepts as competence-bases. A principle becomes a necessary ingredient when we deploy knowledge-items that regulate the force-equilibrium. That principle shall be “axiom 1” in statics and we link it to the force-equilibrium using a R-Link.

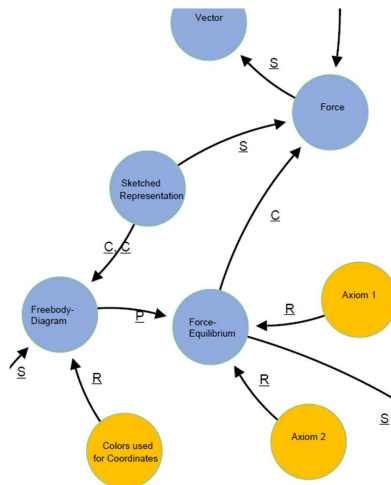


Figure 8. A Section from Our Knowledge-Model.

And another aspect enters our considerations as we construct “force”. For engineers, a force has a magnitude, direction and orientation (in space). In mathematics, it would have its equivalent in a “vector”, so force specializes vector. Let us assume that vector is a concept already defined in our database by a colleague from mathematics. We link it by applying an S-Link from “force” to “vector”. We would learn that “vector” is already defined as the scalar multiplication of a “quantity” and a “unit vector”. The new green circle “scalar multiplication with a unit-vector” in our model is thus a process taking “quantity” and “unit vector” as input and which delivers a generic “vector” as its product. In turn, a quantity would be composed of concepts “number” and “unit of measurement”; and so forth.

What appears to be a straightforward approach to create a model of knowledge in mechanical engineering has its hidden difficulties. In the example from Fig. 8 we have not included a principle that explains what “force” is. You could turn to the concept of stresses to explain force – but then it is advisable not to explain stress using force to avoid circular references.

And the precedence-links also creates an order, a process to be followed as described using “freebody-diagram” and “force-equilibrium” above. Is that a process-competence-base (a solution schema for mechanical engineering) or does it come “automatically” from an agent following the precedence-links in the ontology?

Note that guidance for the creation of a knowledge-model could also come from known student-misconceptions (Streveler [29]) which should be reflected in the structure of the model.

Finally, the main task following the creation of our semantic network, is to describe prototype-levels (Krathwohl’s categories) of competence per base. Following the concepts of Competence-Based Learning, the levels must be described so they can be identified unambiguously. The software supports this process e.g. by providing a choice of appropriate verbs from Krathwohl per competence-level and the possibility to add pictures, formulas, links etc. The description must identify learning objectives as clear and concise as possible and allow for the observability of the respective actions of a student on the respective level.

6.2. Consolidation of Models

Along this line, the ontologies from different contributors grow into a structure of significant unconsolidated complexity. This raises the new challenge of elaborating a consensus-model: Just as two lectures on the same topic are never identical, the semantic models of two knowledge-domain experts will never be exactly identical either. Nevertheless, a common structure for the knowledge should be present in both approaches that can be formalized in the model. A process – and a supporting technology – is needed to facilitate this consolidation of the model. In our approach, each competence-base is therefore created with a “sandbox”-status. The database may hold incongruent nodes and these can be merged and edited by selectively displaying the ontologies of different users. Then, the role of a “competence-master” is to moderate a consolidation of these sub-models.

The representation of a consolidated ontology composed from different contributors could look like this:

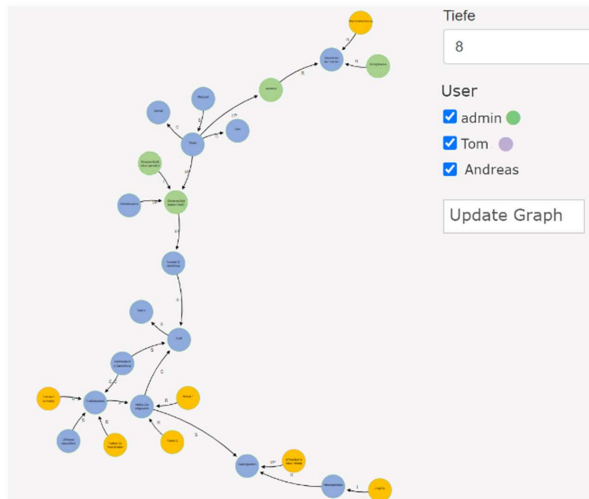


Figure 9. More Bases Shown from the Model.

7. Next Steps

Our knowledge model provides a consistent and structured model to collect and represent competence across knowledge domains. It provides a web-based user-interface suitable for phrasing competencies which allows for collaboration between individuals and organizations. Our implementation of the ontology as a semantic network in a web-application can provide a generic interface for AIs aiming to operationalize tutoring as proposed by learning experience platforms.

These models are uniform and interchangeable, i.e. as part of a community project. Basic ontologies could be created that future users can continue to develop. Catalogues are now able to add generally agreed annotations to their resources so that AI applications can parse effectively through their archives, like advanced search engines using the semantic web. Knowledge-servers as ours could also be key to future learning approaches that help tutoring agents in filtering out those bits of information that really help students to solve a difficult task (Herrman-Werner [16]), competently feed-back the findings of tests, create an individual task list to master a given competency, or even to check a curriculum for consistency.

Thus, the next step will be to employ this ontology in tutored learning environments and collect data – based on the objective progress in learning or on the subjective satisfaction of a learner (Edström [11]). Obviously, Learning Analytics using tests that address a specific competence level or simply ask for the current state of mind of a student will have an integral part in this.

This “generation zero”-tutor has to be based on explicit rules because initially, no data are available to train an AI-based agent. But as data is being collected from real learning processes, an AI could enter the stage – with a basic recommendation algorithm to begin with.

But how would they address learning? Will they aim to facilitate effective learning? Will they aim to motivate students to learn? What opportunities emerge to provide qualified feedback to students (Hartung [12])? How can the now transparent structure of the network be used to create incentives for computerized learning environments (Bartholomé [5])? Questions that can be addressed differently by different tutoring systems but based on an identical knowledge-model.

With our ontology in place, tutoring-AIs would also decide upon the fitting pedagogical approach (e.g. Schunk [27]) to teach a specific competence depending on the competence-type: a concept requires a different approach than a principle or a process! And a concept that specializes another concept may be “taught” first before offering the more generic approach – or the other way round. Also, the type of competence regulates the type of test strategy to follow, and the inference-type between bases regulates, how to address the subject. Finally, established learning materials as “Tutorials” (Riegler [25], Steinberg [28]) are an excellent basis to build a tutoring approach upon.

Our ontology provides all the pre-requisites to facilitate these developments.

References

- [1] Adams WK, Wieman CE. Development and Validation of Instruments to Measure Learning of Expert-Like Thinking. *International Journal of Science Education*, (2011) 33(9):1289–1312.
- [2] Alevin V, McLaren B, Roll I, Koedinger K. Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence in Education*, 16 (2), 101-128 (2006).
- [3] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington D.C.
- [4] Anderson LW, Krathwohl DR. A taxonomy for learning, teaching, and assessing. A Revision of Bloom's Taxonomy of Educational Objectives, Pearson new international ed. Harlow: Pearson Education (2014).
- [5] Bartholomé T, Stahl E, Pieschl S, Bromme R. What matters in help-seeking? A study of help effectiveness and learner-related factors. *Computers in Human Behavior*, (2006) 22(1), 113-129.
- [6] Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives: The classification of educational goals. *Handbook I: Cognitive domain*. New York: David McKay Company, (1956).
- [7] Brose A, Kautz C. Identifying and addressing student difficulties in engineering statics. (2011) In *Proceedings of the 2011 ASEE Annual Conference and Exposition*, Vancouver.
- [8] Chi MTH, Slotta JD. The Ontological Coherence of Intuitive Physics. *Cognition and Instruction*, 10(2/3):249–260. (1993) Taylor & Francis.
- [9] Ding L. Theoretical perspectives of quantitative physics education research. *Physical Review Physics Education Research*, (2019) 15(2):020101.
- [10] Direnga J. Assessing the Effectiveness of Research-Based Active Learning Materials for Introductory Engineering Mechanics. Dissertation; Technische Universität Hamburg-Harburg, Institut für Abteilung für Fachdidaktik der Ingenieurwissenschaften Z-2, 2021.
- [11] Edström K. Student feedback in engineering: a disciplinespecific overview and background. In *Enhancing Learning and Teaching Through Student Feedback in Engineering*, p 1–23 Elsevier (2012).
- [12] Graesser AC, Conley MW, Olney AM. Intelligent tutoring systems. In S. Graham, & K. Harris (Eds.), *APA Educational Psychology Handbook: Vol. 3. Applications to Learning and Teaching* (pp. 451-473). Washington, DC: American Psychological Association (2012).
- [13] Halloun IA, Hestenes D. Common sense concepts about motion. *American journal of physics*, 53(11):1056–1065 (1985).

- [14] Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications (1991).
- [15] Hartung AB. Studie zum Einsatz von Mentoring-Programmen als Instrument struktureller Förderung für Studierende an deutschen Universitäten. Arbeitspapier, 243. Düsseldorf: Hans-Böckler-Stiftung (2012).
- [16] Herrmann-Werner A, Loda T, Junne F, Zipfel S, Madany Mamlouk A. "Hello, my name is Melinda" – students' views on a digital assistant for navigation in digital learning environments; a qualitative interview study. *Frontiers in Education*, (2021) Vol. 5, pp. 291-297)
- [17] [HOOU] Hamburg Open Online University. <https://www.hoou.de/>.
- [18] Hulpus I, Prangnawarat N. "Path-Based Semantic Relatedness on Linked Data and Its Use to Word and Entity Disambiguation". *The Semantic Web – ISWC 2015*, Springer International Publishing. p. 444 (2015)
- [19] IEEE Approved Draft Standard for Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata," in *IEEE P1484.12.3/D2*, November 2019 , vol., no., pp.1-0, 5 March 2020.
- [20] IEEE: Learning Technology Standards Committee (LTSC). *Standard for Learning Technology - Data Model for Reusable Competency Definitions*. IEEE, 2008.
- [21] Madany Mamlouk A, Geick C, Lämmerrmann K. From Zero to Hero – New Methods for Motivating Students. In Jansen-Schulz B, Tantau T (Edt.). *Excellent Teaching - Principles, Structures and Requirements, Blickpunkt Hochschuldidaktik* (2018) 134, 99-111.
- [22] MINTfit. Online tests for math, physics, chemistry and computer science, <https://www.mintfit.hamburg/en/>
- [23] Paquette, Gilbert: *Visual Knowledge Modeling for Semantic Web Technologies: Models and Ontologies*, Information science reference, Hershey, 2010
- [24] PROTÉGÉ: *Ontologie-Modellierung*, <https://protege.stanford.edu/>
- [25] Riegler P, Simon A, Prochaska M., Kautz C, Bierwirth R, Hagendorf S, Kortemeyer G. Using Tutorials in Introductory Physics on circuits in a German university course: observations and experiences. *Physics Education* (2016) 51(6):065014.
- [26] Romiszowski AJ. *Designing Instructional Systems*. New York: Kogan Page and Nichols Publishing (1981).
- [27] Schunk DH. *Learning theories: An educational perspective*. Macmillan Publishing Co, Inc. (1991).
- [28] Steinberg RN, Wittmann MC, Redish EF. Mathematical tutorials in introductory physics. *AIP Conference Proceedings*, (1997) volume 399, pages 1075–1092, College Park, Maryland (USA).
- [29] Streveler RA, Brown S, Herman GL, Montfort D. Conceptual Change and Misconceptions in Engineering Education. In Johri A and Olds BM, eds, *Cambridge Handbook of Engineering Education Research*. Cambridge University Press, New York (2014).
- [30] Wood, D. (2001). Scaffolding, contingent tutoring, and computer-supported learning. *International Journal of Artificial Intelligence in Education*, 12.

Visual Programming for Artificial Intelligent and Robotic Application (VPAR) Framework

Goragod PONGTHANISORN ^a, Waranrach VIRIYAVIT ^{b,c}, Thatsanee CHAROENPORN ^d, and Virach SORNLERLAMVANICH ^{d,e,1}

^a*Department of Computer Engineering, Faculty of Engineering, Thai-Nichi Institute of Technology, Thailand.*

^b*School of ICT, Sirindhorn International Institute of Technology, Thammasat University, Thailand.*

^c*Graduate School of Science and Engineering, Chiba University, Japan.*

^d*Asia AI Institute (AAII), Faculty of Data Science, Musashino University, Japan.*

^e*Faculty of Engineering, Thammasat University, Thailand.*

Abstract. Computer programming is popularized in 21st century education in terms of allowing intensive logical thinking for students. Artificial Intelligent and robotic field is considered to be the most attractive for programming today. However, for the first-time learners and novice programmers, they may encounter a difficulty in understanding the text-based style programming language with its special syntax, semantic, libraries, and the structure of the program itself. In this work, we proposed a visual programming environment for artificial intelligent and robotic application using Google Blockly. The development framework is a web application which is capable of using Google Blockly to create a program and translate the result of visual programming style to conventional text-based programming. This allows almost instant programming capability for learners of programming in such a complex system.

Keywords. Artificial Intelligent, Robotic, Web Application, Visual Programming Framework, Block-based Programming, Blockly

1. Introduction

Since the starting of human history, communication is an effective tool which leads to emerging of ancient civilization and a foundation of the modern society. The purpose of communication is a knowledge sharing in many disciplines by utilizing a verbal communication which is comprehensible by each other. However, a verbal communication has a flaw in which information may be altered before reaching a destination. To preserve the correctness, a writing system was invented where a verbal expression is described in a formal set of characters creating a pronunciation guide for a

¹ Virach Somlertlamvanich, Asia AI Institute (AAII), Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan; and Faculty of Engineering, Thammasat University, 99 Moo 18, Patholyothin Road, Klong Nueng, Klong Luang, Pathumthani 12120, Thailand; E-mail: virach@gmail.com

reader. These inventions have been used since then. The writing system has made a great contribution to many innovations while the knowledge transfer is continuously improved. Later, human has transcended in a way of communication, from just only to themselves, to a silicon-based device (i.e., computer). The language of computer, called machine code, is a series of binary numbers which represents the instructions that the computer have to execute accordingly. In early age of programming, the programmers are used to communicate in such an almost-incomprehensible language to create a computer program. Nonetheless, the primitive way of the communication with computer obstructs higher feature-rich applications. This leads to a birth of programming language.

Rather than writing a computer program directly in machine code, a programming language compromises human language and the machine language. Programming language is a combination of a language alphabets (mainly in English) and special characters to create instructions. These instructions, later, are going to be compiled by a compiler, a programming language to machine code translator, to create a computer program. The development of programming language allows more complex and newer features to be programmed by an aid of provided tools that can be used by programmers. Nowadays, there is almost 600 programming languages for each type of application ranging from a computer hardware interfacing to a cloud computing and a web-service.

Currently, Python is one of the most popular programming languages [1], which its application lies from an education to research area. Not only its simplicity, but also its community where various modules and libraries for applications are shared. As in the field of Artificial Intelligence, which an intensive computation is required, Python provides a framework for ease of programming in complex mathematic related field, for instance, Keras and Pandas. While in robotic applications, conventional programming languages (e.g., C/C++) are widely used because its simplicity and low-level feather for hardware related programming. However, the recent emergence of MicroPython has caused a trend of microcontroller application to be written in Python. This leads to rapid prototyping and being more user-friendly for a first-time learner as the Python syntax is simple. Including its Object-oriented paradigm, the new era of microcontroller is opening.

Although the design of Python is simple and introduced in object-oriented style, just like other programming languages, it is still problematic from being a text-based language. Firstly, a text-based programming contains a considerably amount of mathematic sign to express its instruction. This reflects the original purpose of programming, to perform a complex mathematic computation. While application of programming in modern day has covered wide range of field already. Secondly, as born from mathematic, programmer must pose an ability to transform real-world problem into a logical and instruction representation in programming. Not to mention the commonly found mistake for a novice, a syntax violation. Programming languages must strictly follow the syntax otherwise execution of a program is not initiated.

The complexity of text-based programming language may hinder learners' interest in programming [2]. When novice programmers try to write their first program, the first lesson is always a simple program, such as a print function (the most empirical output from their program). This is a process of making a learner get familiar with a style (syntax) and specific function provided by each programming language. However, the naiver, the more mistakes are likely to be found, e.g., missing terminating character (; semi-colon) in C language, style of writing a conditional statement, and scope of for-loop operation defined by curly brace. Not only syntax, but novice programmers may also encounter difficulty in trying to evaluate a logical error. A logical error is also a common mistake, no exception to most experience programmers, and is not easy to

evaluate since a logical error requires some moderate competence in a programming including a functionality of currently used the programming language. While the core function of programming language is to solve a problem by creating a set of instructions so called algorithm, a programmer tends to waste time for this process. These difficulties obstruct the main purpose of programming.

Recently, visual programming has become popular for the novice and first-time learner. This paradigm of programming, instead of text, uses a block or another notation to represent a logical flow. It is also called a block-based programming style. This enables more intuitive of a computer programming. Moreover, a graphical representation of visual programming evaluates information in the closest manner to human mental representation of real-world problems [3]. There are multiple well-known visual programming styles in broad range, for example a model-based design of MATLAB [4] which represents an equation in a block and flow of logic using a flow-based design. LabView is used for an embedded application that implements a graphic of an electronic device and sign for the representation the system [5], Scratch, MIT Block and Google Blockly [6, 7, 8] which employ a concept of representing computer instruction into a block called block-based programming. Visual programming seems promising for a new programming paradigm as multiple applications employed the idea and concept. For instance, the works of [9] and [10] implemented a visual programming for a machine learning application through a web application. Especially, in [10], the broad of application using MIT block is introduced. [11, 12, 13, 14, 15] have selected a Google Blockly, an open-source block-based programming which are developed on web application as a tool for a visual programming and apply to a variety of application ranging from robot to Augment Reality (AR) application. However, [11, 12, 13, 14, 15] have some limitations. These works require a user to install and setup a required tool before. Although [9] and [10] are accessible via online using a general web browser, the work of [10] which can use as robot programmer, requires an installation of external tool (e.g., mLink before connecting to a physical robot). The researchers of this present study found the flaw of the robotic application on such a feather-rich framework and saw its potential for AI application. Thus, we have designed and developed the visual programming framework which is able to develop AI and robotic application without any additional tool installation. For a visual programming and translation of text-based, we have selected Google Blockly and Python.

Table 1. Comparison of Robot for Programming [15]

Product	Processor	Processor Clock Speed (MHz)	Supported Programming Language
Mindstorms EV3	ARM 92EJ-S (32-bit)	300	<ul style="list-style-type: none"> • The EV3 on Brick Programming
Robot-PICA	PIC16F887 (8-bit)	8	<ul style="list-style-type: none"> • Assembly • BASIC • C Language
Scribbl3	Multi-core Propellor P8X32A (32-bit)	80	<ul style="list-style-type: none"> • C Language • Blockly • Scribbl3 S3 GUI
Robot-CreatorXT	ATmega644P (8-bit)	16	<ul style="list-style-type: none"> • C/C++ Language
IPST-MicroBOX	ATmega644P (8-bit)	16	<ul style="list-style-type: none"> • C/C++ Language
mBot1.1	ATmega328 (8-bit)	16	<ul style="list-style-type: none"> • Scratch 2.0 • C/C++ Language

As for a robotic application, the main application is to retrieve information from robot sensory system and evaluate a robot action. There are many robot programming frameworks currently placed in the education market. Table 1. provides information of well-known programming robot products from various manufacturers. Three key attributes are used to describe and distinguish the robot system.

Information in Table 1. implies that a robot programming mostly supports text-based programming language. Although some of them (Mindstorm3, Scribbler3 and mBot1.1) employ a visual programming concept to extend a range of suitable age for playing. Nonetheless, all mentioned robots are not able to perform a heavy computation application (e.g., image processing) since their performance is limited by equipped processor. These robots may categorize in application-specific robot which a range of applicable tasks are limited while the need of a general-purpose robot is rising. The term of general-purposed robot refers to a robot which is capable of execute general trivial task similar to human. For instance, a trivial task for human, object detection. Human brain and our visual sensory system are effective enough and allow us to distinguish an object shape and color then correctly describe them. Yet, for a robot, it is considered one of toughest tasks. To perform an object detection in logic computation, first, the robot needs to obtain an image (which is normally represented in multi-dimensional array). Once image is obtained, robot executes a heavy computation using high-level mathematic to extract a necessary piece of information from the image. Then, the robot executes another heavy computation to identify an object. With all said, it is impossible to execute that task in such a low performance processor on the mentioned robot.

In order to archive such a complex system, a multiple component of computation software and an effective sensory system equipped including high-performance processor to handle a heavy computation. Thus, this work selects a Temi to implement the proposed system. Temi is a personal assistant robot which embedded with high performance ARM HEXA core processor and multiple sensory system (i.e., LiDAR, microphone, 2 RGB depth camera, 5 proximity sensor, 6 time of flight sensors and IMU sensor) [16]. Equipped with such a high-performance processor, Temi robot is capable of executing a complex task in a real-world environment, for example, a navigation through a dynamical environment, an interaction via speech and conversation context awareness with the built-in NLP (Natural Language Processing) unit, a trajectory free path planning using an image and point cloud data from LiDAR.

Temi robot programming is different from typical other robot programming in term of a software structure. The robots, mentioned in table 1., is equipped with low-end processor that easily to program in a low-level language (C/C++, assembly). Not for Temi robot that equipped with high-performance processor ARM HEXA core. Since complexity of low-level language for non-embedded-field-programmer hinders a full potential, Temi robot's manufacturer decide to provide SDK (software development kit) for develop an application for Temi robot. The SDK is developed on Android framework using both Java and Kotlin language [17]. Temi SDK allows a programmer to create robot application called *skill*.

While the provided SDK soothes a difficulty of implementation of a software component directly into the robot's hardware by providing an inherited method as a callback for each event. For instance, when the robot starts speaking, the information regarding to a state of the speech can be obtained from derived method call *onTtsListener*. Although program implemented by the provided SDK is easy, the program structure is becoming more complicate. Since multiple components are required in single application which is commonly found. The more codes and processes for a program, the harder

program evaluation becomes. A visual programming paradigm should help mitigating the matter. Rather than describe the robot response of behavior in text-based programming, visual-based block seems far easier for novice.

In addition, the robot requires a multiple step of a configuration on Android Studio and ADB (Android Debug Bridge) in order to perform an installation of the developed program. This task's complexity is comparable to understanding of the robot's program flow.

In this work, not only AI application but also such the feature-rich robot, Temi is going to be programmed by visual programming Google Blockly on an online IDE which requires minimal configuration. This enables even a novice programmer to create a simple application for Temi. The detail of implementation of both AI and robotic application will be discussed later.

The paper is organized in the followings. Section 2 outlines the proposal of a framework for AI and robotic application development using visual programming style. In this section, we introduce a motivation and design concept of the system and describe an overview architecture including the flow of the system. Section 3 discusses the implementation of block-based concept into a simple Neural Network (NN) programming. The section presents a basic parameter of NN, mathematic and block representation. Section 4 discusses a concept designed for programming Temi, personal assistant robot via visual programming language. We implement an additional component to the proposed system including Temi application environment allowing such a new paradigm of program to the robot. After the key idea of design and implementation is introduced, we include a prototype of web-application of the proposed framework with functions like an IDE which resides remotely on server. In Section 5, we discuss the limitations of our proposed system. Lastly, a conclusion is drawn, and some on-going and planned work is shown.

2. Proposal of Visual Programming for AI and Robotic Application via Online IDE

In order to create a computer program, an integrated development environment (IDE) is required. There are two basic components for every IDE, 1) text editor and 2) compiler or interpreter. Text editor is used for a text input to create a programming instruction. After the program is written, a compiler or interpreter (depending on which type of programming language) translates the human comprehensible language into a hardware native machine code. One of the drawbacks is when starting learning a program, programmers do not only learn syntax and semantic of the language but also have to learn how to perform an installation and configurations of the additional tools. Luckily, many modern IDEs have included all the required components in a single installation package. Yet, another issue is raised, with a different environment of each computer (e.g., OS, version, and previously install program). These can affect the operation of an IDE. In order to avoid this matter but still provide programming environment, a web-application online IDE seems to be a promising solution. A general idea of online-IDE is to provide a programming service via a web-browser without any installation of specified tool set. Moreover, variant of environment, compared to locally installed IDE, is relatively small (concerning on web-browser and its version).

The proposed visual programming for AI and Robotic Application system is designed for beginners in the world of programming. The key concept of the proposed system is the visual programming for representing a complex computation and logic flow

while providing an insight of a program by translating the visual to text-based language. Users should be able to visualize a program flow via visual-based approach. Once they are competent in the flow of logic, a text-based approach is introduced to them. Thus, our system uses block-based for visualizing of program and providing a translation of that block-based to text-based code. In addition, the proposed system was designed to be web-application. The concern is on the complexity of tool setup and installation when writing a program. In that sense, the proposed system should be designed in ready-to-use manner without further installation.

As for block and text-based programming language, Python is selected because of its simplicity and suitability for novice programmers. For block-based programming, we use Google Blockly. Google Blockly is developed using JavaScript and executed on a client-side web-browser. Google Blockly has built-in translation unit from block-based to text-based conversion (JavaScript, Python, Lua, Dart and PHP) [8].

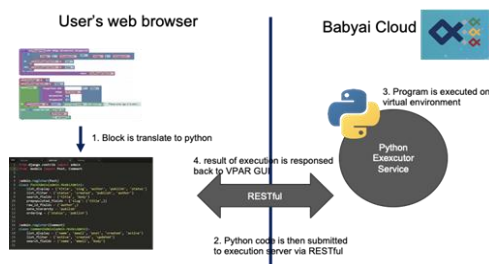


Figure 1. The Architecture of the proposed Visual Programming for AI and Robotic Application system.

The proposed system is comprised of web-application which is developed by React framework. Then the developed web-application is executed on the web browser of client side. The key role of the web-application is to provide user-friendly GUI to users, block code development using Blockly framework and a translation of block-base programming to Python programming language. Nonetheless, a typical web-browser poses strict security protocol which an arbitrary script execution is prohibited. The solution is that the translated program should be executed somewhere else. This leads to the second component of the system, Python Execution Server (PES). PES is a web-service hosted on cloud and exposed an available service through HTTP REST API. The primary features of the system are to 1) execute a program code, 2) export the current workspace, and 3) send a result of execution by HTTP response. PES is considered as a backbone of our proposed system. PES executes Python code in docker container and virtual environment so that any harmful operation will not directly affect the server itself. PES is used by both AI and robotic application, despite variant on module and library usage. PES is able to retrieve necessary package through Python package installer, *pip*, before execution. In addition, PES also supports multiple language input and display, as mandatory to higher programming language as well. As for web-application, this feature may be provided by web-browser, but there is not on the server script, such as PES. However, the translated program must be executed by PES. Thus, it is equipped with

Unicode character encode/decode service to avoid incorrect output when dealing with Unicode character.

The proposed system work process starts with a block-based program which is translated into a text-based programming language result a code in Python. The translated program is then transferred to Python Execution Server (PES) via REST API. Once the code is transferred, PES determines the code program and executes the following code. The result of execution is then retrieved and sent back via the response of HTTP request to client. The communication between web-application and PES is on REST API fashion which PES exposes only necessary commands for security purpose.

3. A Design of Block-based Programming for AI Application Case Study: Neural Network

One of the purposes of the proposed system is to create Artificial Intelligent (AI) application using Blockly. With a broad field of AI application, the proposed system aims for the fundamental of AI application, Neural Network (NN), which is a great starter for understanding. The principle of NN starts with a mimic of human perceptron which connects and creates a neural network by mathematic equation. The mathematic model derived from NN is illustrated by the simple model in Figure 2.

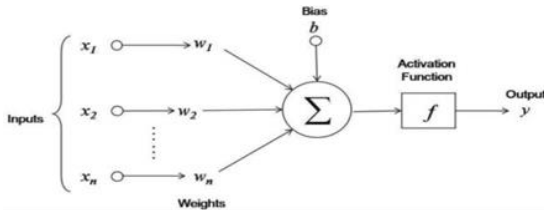


Figure 2. Single Perceptron Neural Model.

For a representation of neural to mathematics and computer programming language together with an activate function for an output mapping, the equation of the system is described in (1) as following.

$$y = f(\sum_1^n(x_n * w_n + b)) \quad (1)$$

For novices or first-time learners who are not familiar with programming, realization of a model and a program flow design is not easy tasks. Lack of understanding is potentially a great obstacle for further implementation or modification. An alternative way for providing better understanding of application is visual programming. The proposed system compromises an ease of understanding of model view and sematic of programming language by representation of neural unit into a block command.

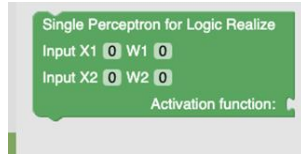


Figure 3. Block Representation of Single Perceptron.

For case study, the logic realization application of single perceptron (2 inputs and 1 output) is implemented. The block representation of a perceptron is illustrated in Figure 3.

As shown in equation (1), a neural network requires an activation function for an output mapping. Plug-and-play design principle is used to provide a wide range of experiment to examine the output of NN due to activation a certain function. Thus, the representation of activation function to block is designed and shown in Figure 4.



Figure 4. Block Representation of Activation Function.

The activation function block connector is different comparing with NN block since they are designed to connect to a NN block only. A parameters of activation block are projected by their equation of (2) and (3) as follows.

$$f(x) \begin{cases} 0 & \text{for } x < Zeta \\ 1 & \text{for } x \geq Zeta \end{cases} \quad (2)$$

$$f(x) = \frac{1}{1+e^{(zeta-\alpha)}} \quad (3)$$

Equation (2) is called binary step function which the output is determined by Zeta to be 0 or 1 (integer), while equation (3) describes a sigmoid function which the output is mapped between 0 and 1 (floating point).

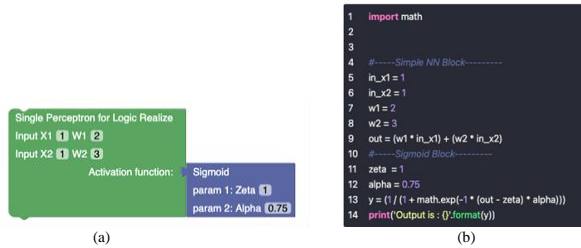


Figure 5. (a) Block Program (b) Translated Code

Instead of writing a NN program, the block code of NN provided by the proposed system is used. Figure 5 (a) and (b) demonstrates a NN application for logic realization using block program including translated code.

Not only a block-styled code, but an user can also start learning a concept of programming with Python by examining a translated code as well. As a plug-and-play design, an activation function can be replaced easily by drag-and-drop manner. The translated code and output are then changed due to the difference of activation function.

4. A Design of Blockly for Personal Assistant Robot (Temi)

Temi is a personal assistance robot, equipped with multiple useful functionalities, e.g., user-interaction via built-in natural language processing (speech to text and text to speech), a camera which enables facial recognition and person tracking, and internet connectivity for android applications such as YouTube and simple web-browser. The robot has a set of functions called skills which describe abilities of robot for certain tasks. Table 2 provides a detailed description of the robot abilities.

Table 2. Summary of the Robot Skills [16].

Skill	Description
Location and Map	Location saving, Map generation, and start-to-location navigation
Movement	Control direction of robot movement
Speech	Temi-user interaction via speech and simple conversation
User and Telepresence	Make a video calling using saved contact via application called Temi App
Face Recognition and Sequence	Face detection unit using built-in RGB camera
Detection and Interaction	Exploit a usefulness of Face Recognition ability to create a human detection
Follow	Exploit a usefulness of Face Recognition and Detection ability, enabling a follow-trackable-person navigation

Although a preliminary installed program can easily suffice a daily life usage, Temi developer provides an SDK for programmers to create their own application (called skill) as well. The provided SDK is based on Android and Kotlin which provide an API access to a variety of robot features [16].

The proposed framework aims to provide ease of programming framework of Temi robot for novice programmers by projection of Temi skill to a block-based programming,

Blockly. The block-based program is then translated into Python and transferred to execute server where the command is submitted to the robot through MQTT communication. The MQTT communication does not provide only the access to the robot for command issuing, but also robot status, e.g., battery level and current skill status in order to provide more insight information of robot on translated program. The overall operation of translated program and the robot communication is described in Figure 6.

The visual programming for Temi robot consists of multiple components which enable a feature and provide an alternative of the robot's program execution. The components are listed as follow

- VPAR, a web-application for a Blockly-to-Python translation
- PES, a code execution service
- MQTT broker, a mediator between the Python code and the robot
- The developed android application that bridges (using MQTT) between the Python code and the robot

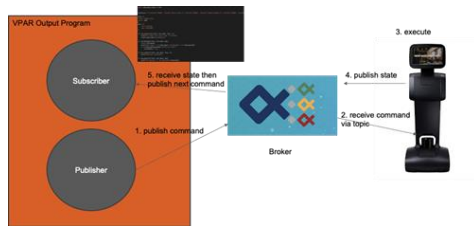


Figure 6. Overall Operation of Translated Robot Program Execution

In order to program the robot, the proposed system provides a block-based program which can be created using the web framework of proposed system. The robot skills are represented by a block type. These blocks can be connected to each other to create an application for the robot. The available block command is shown in Figure 7(a) and 7(b).

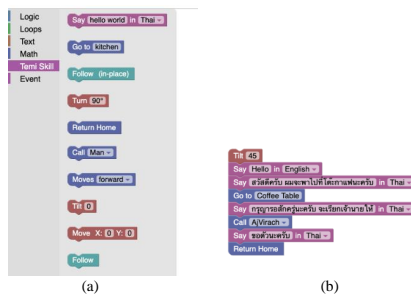


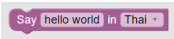



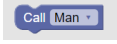

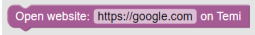
Figure 7. (a) Block Command for the Robot (b) Example Program

Figure 7 (a) shows a set of commands which can be used for creating a program for the robot while Figure 7 (b) demonstrates a program from blocks. Table 3 provides information of robot skill represented by blocks as follows.

The proposed system projects a skill set of the robot in a block representation which each block requires different parameter. For example, *Say* command requires two parameters, the text to be spoken by the robot and language option according to the previous option. Each block can be connected, and the program will execute a command accordingly. The translated program controls a timing of command transferred to the robot. If previous command has not yet finished, the next command must wait to create a sequence of block execution.

To achieve such a feature and communication of the proposed system, the environment for the robot programming comprises two essential units.

Table 3. Skill and Block Representation

Block	Skill	Description	Parameter
	Speak	Read out a text in the box, language must be agreed	Text, Language Option
	Go to	Tell the robot to go to saved location	Text (in all lower case)
	Track	Track a person but the robot will not move along with tracked target	-
	Turn	Robot turn itself specified by degree	-
	Call	Robot makes a video call to a person specified in dropdown menu	Selected person from dropdown
	Tilt head	Move head along its moving axis (between 37 – 53)	Degree 37 - 53
	Open website	Temi open the specific URL in built-in web browser	

4.1. Temi Command Actuator Android Application

The Android application for MQTT communication and the robot instruction decoder are installed into the robot, allowing a user to experience ready-to-use application development framework.

Table 4. API Overview for Location Related [17]

Return	Method	Description
Void	speak (TtsRequest ttsRequest)	Ask Temi to speak (play TTS)
Void	cancelAllTtsRequests	cancel TTS request
Void	wakeup()	Wake Temi up
String	getWakeupWord()	Get current wake-up word
Void	askQuestion(String question)	Temi speaks actively and waits for user reply
Void	finishConversation()	Finish a conversation (Stop recording for ASR)
Void	startDefaultNlu(String text)	Trigger default NLU service

The key functions of the application are 1) To bridge the Python program from VPAR and the robot 2) To translate instruction from JSON format into a function call accordingly to command 3) To issue a command received from the Python program to the robot's SDK. These works implement the application using provided SDK on Java programming language which contains an API for accessing Temi skill. Table 4 shows some API for Temi robot. The application is simple and provide only information for debugging purpose (at this state of development. Most of the application operation is to perform MQTT connection, to receive JSON package to an instruction conversion and to control the robot's action flow. Figure 10 shows some of an essential function inside the bridge application *i.e.*; MQTT communication and an instruction (in JSON format) decoder function.

```

/----- ACTION DECODER -----*/
public void actionDecoder(SOCKET actionId) {
    try {
        actionId.getAction().toString();
        case "speak"
            speak(actionId.get("content").toString());
            if(actionId.get("language").equals("TTS")) {
                ttsRequest = actionId.get("content").toString();
            }
            else if(actionId.get("language").equals("TTS")) {
                speak(actionId.get("content").toString());
            }
        case "wakeup"
            if(get(actionId.get("content").toString()) != null) {
                startTtsRequest("wakeup agent", actionId.get("content").toString());
            }
        case "ask"
            startTtsRequest("ask", actionId.get("content").toString());
        case "ask"
            startTtsRequest("ask", actionId.get("content").toString());
        case "speak"
            String url = actionId.get("content").toString();
            HttpServletRequest req = HttpServletRequestWrapper(new HttpServletRequest());
            HttpServletResponse res = HttpServletResponseWrapper(new HttpServletResponse());
            try {
                getHttpURLConnection().startConnect(url);
                catch (IOException e) {
                    e.printStackTrace();
                }
            }
        } catch (IOException e) {
            e.printStackTrace();
        }
    }
}

/----- MQTT -----*/
private void startMqttClientFace() {
    String urlTopic = "Temi" + "/" + "Temi" + "/" + "Temi-cmd";
    printLog(urlTopic);

    mqttInterface = new MqttClient(getApplicationContext(), urlTopic, "virachai@robotclinet");
    mqttInterface.setCallback(new MqttCallbackImpl());
    @Override
    public void connectToMqtt(MqttConnectOptions, String serverURI) {
        printLog("connected to " + serverURI.toString());
    }
    @Override
    public void connectToMqtt(Throwable cause) {
        mqttInterface.connect();
    }
    @Override
    public void messageArrived(String topic, MqttMessage message) throws Exception {
        printLog("arrived message: " + topic.toString());
        printLog("arrived message: " + message.toString());
        try {
            JSONObject jsonObj = new JSONObject(message.toString());
            actionDecoder(jsonObj);
        } catch (JSONException e) {
            log.e("Error", e.toString());
        }
    }
    @Override
    public void deliveryComplete(MqttDeliveryToken token) {
    }
}

```

Figure 10. Partial of the Bridge Application Java Code

When a command is issued by the program, it sends a requested command to robot internal system enabling non-blocking program paradigm. The internal system, then, raises a status associated with issued command to identify the current status via callback API. Table 5. shows some of callback API for built-in text-to-speech system, and Figure 11. shows pseudo code for overall application operation.

Table 5. Speech-related Callback [17]

Interface	Description
TtsListener	TTS status listener
WakeUpWordListener	Wake-up event listener
AsrListener	ASR result listener
ConversationViewAttachesListener	Conversation view attaches listener
OnTtsVisualizerWaveFormDataChangedListener	Listener for wave form data changes of TTS audio visualizer
OnTtsVisualizerFftDataChangedListener	Listener for FFT data changes of TTS audio visualizer

```

procedure: application(robot, mqtt, robot_status_callback,)
  robot ← RobotInstance()
  mqtt ← MQTTInstance()
  arrived_message ← Null
  loop forever :
    if arrive_message is not Null
      action ← decode(arrive_message)
      robot_execute(action):
        while status is not done:
          if status is not prev_status:
            mqtt_publish(status)
          end if
        end while
      end if
    end loop

```

Figure 11. Pseudo Code for Overall Application Operation.

The proposed system has exploited this programming paradigm, together with MQTT communication. Then, the robot skill and status can be realized through MQTT communication.

4.2. Robot module for Translated Program

Python execution server is a shared component between AI and robotic Python applications which are the result from block code translated into Python. Although an AI application may implement a built-in module or site package module (installed by pip), there is no module that provides an ability to control the robot. Thus, the proposed system for robotic application has integrated a special Python module for interacting with the robot via MQTT. As shown in Figure 12, the translated program for the robot controlling and block code are described.

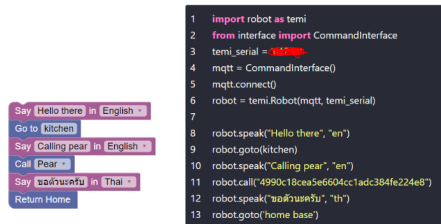


Figure 12. Block Program VS Translated Python Code from Block Program

The robot module is comprised of two units which are tightly coupling to each other. The first module, named *robot*, is written in Python. The main role of the module is to converse an issue command into predefined MQTT package for the robot in JSON format. As for delivery of the constructed message, including the robot status retrieval, the command is handled by a module named *interface*. This module is tightly coupling with robot module which creates a message and manages a program flow according to the robot status. Another role of this module is to program the robot directly with Python code.

5. Results and Limitations

The proposed system is deployed on a web-browser as web application using React framework. GUI of the web-application is shown in Figure 13. The web application has two variants. The first one is for a robotic application, and the second one is for AI application with the same GUI. The main difference is the provided blocks in the web application. The components of web application are described by numbering labels as follows:

- 1 - A block menu for desire block selection
- 2 - Canvas where the blocks program is created.
- 3 - Output result of block program
- 4 - Preview Python code for learning
- 5 - Control Menu

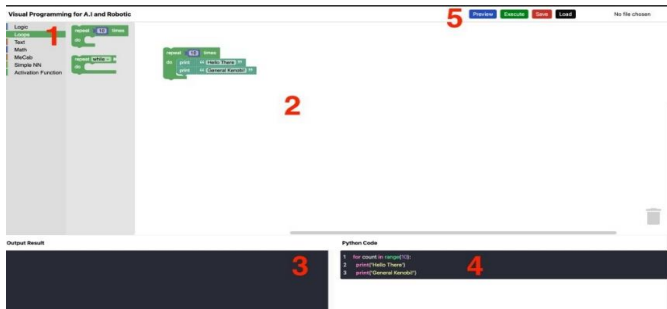


Figure 13. VPAR Web Application GUI.

The web application is considered as online-IDE for programming which provides editor, output result, and remote compiler. The block program is translated into Python code and then executed remotely. Once the execution is completed, the result is transferred back to be displayed on the web application. The web application also provides a save/load menu in which the users can import an existing program in Blockly XML file format to the web. The save button provides an export context to the web where the created block program can be saved and downloaded to a local computer for further usage.

As for the robot application, Temi, it's required to pre-install the developed bridge program to enable MQTT communication. The MQTT connection is used by the Python code (output from VPAR) which is executed by PES on cloud. Once the execution is finished, PES returns a result and displays on the web-application. The bridge application GUI, as shown in Figure 13, provides information of currently executing action of the Python code. There are two buttons for cancelling a current executing action and stopping MQTT communication. In other word, the robot doesn't receive any command at all

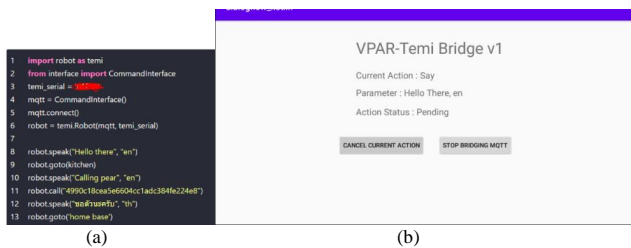


Figure 14. (a) The Python Code Executed on PES
(b) Bridge Application GUI

However, there are some drawbacks of the design of the proposed system. Firstly, as the translated code is compiled and executed remotely, the proposed system requires constant internet connection. There is no local application installed or the accessing of local Python interpreter for the security reason. In addition, the proposed system must finish the execution of translating program before the result can be retrieved to be displayed on the web application. It is because the proposed system relies on REST API which is basically the HTTP communication. As a result, the interpreter style of Python is not available in the proposed system as well as debugging. Moreover, the robotic application is also affected, so that the user cannot realize a communication in real-time since its execution of the program must be finished first.

6. Conclusion and Future Work

In this work, we proposed a visual programming for AI and robotic application (VPAR) framework. The system utilizes a Google Blockly for creating an AI and robotic application for novice programmers. The proposed system is fully executed remotely and provides an access to the system via the developed web application. The web application is based on React framework which serves as an online-IDE for programming. It enables save/load a workspace including previewing of translated-from-block-to-code program for better understanding. The proposed system allows a block-based usage to develop AI and robotic applications for Neural Network (NN) and Temi robot programming respectively. The result of program execution can be examined on the built-in web application as well.

For future work, we are going to provide more blocks for AI applications. We aim to provide the best experience for users by adding an interactive block-to-NN model which can dynamically generate a NN model when the block is connected. For robotic applications, we are going to introduce an event-specified action block that the robot will execute a block code according to self-defined event. This results in rapid prototype Internet of Robotic Things. As for web application, we plan to add programming mode for Python in case of more experienced users.

Acknowledgement

The authors gratefully acknowledge the financial support provided by Thammasat University Research Fund under the TSRI, Contract No. TUFF19/2564, for the project titled “AI Ready City Network in RUN”, based on the RUN Digital Cluster collaboration scheme.

References

- [1] Stephen C. Top Programming Languages. IEEE Spectrum 2020. Retrieve Jan 19, 2021. Available from <https://spectrum.ieee.org/at-work/tech-careers/top-programming-language-2020>.
- [2] Kris P, Stacey E, and Leanne MH. Through the looking glass: teaching CS0 with Alice. SIGCSE Bull. 39, 2007 Mar, 40(1). pp.213–217.
- [3] Brad AM. Taxonomies of visual programming and program visualization. Journal of Visual Languages & Computing. 1990 Mar, 1(1). pp.97 – 123.

- [4] MathWorks. Simulation and Model-Based Design. Retrieve Jan 19, 2021, Available from <https://www.mathworks.com/products/simulink.html>.
- [5] Bitter R, Mohiuddin T, Nawrocki M. LabVIEW. Advanced programming techniques. 2006. Crc Press, c2007. pp.1-65.
- [6] Mitchel R, John M, Andrés MH, Natalie R, Evelyn E, Karen B, Amon M, Eric R, J, Brian S, and Yasmin K. Scratch. Programming for All. Communication of ACM. 2009 Nov; 52(11), pp. 60–67.
- [7] S.C. Pokress, and J.J.D. Veiga. MIT App Inventor Enabling Personal Mobile Computing. 2013. Retrieve Jan 19, 2021, Available from <https://arxiv.org/abs/1310.2830>.
- [8] Blockly. A JavaScript Library for Building Visual Programming Editors, Retrieve Jul 10, 2020, Available from <https://developers.google.com/blockly>.
- [9] MLBlock, Retrieve Jan 10, 2021, Available from: <https://mlblock.org/>.
- [10] MBlock, Retrieve Jan 21, 2021, Available from: <https://mblock.makeblock.com/en-us/>.
- [11] Alexandru R and Ioana C. Open Cloud Platform for Programming Embedded Systems, 2013 RoEduNet International Conference 12th Edition, Networking in Education and Research. 2013, Iasi, pp.1-5.
- [12] Nguyen VT, Jung K, Dang T. BlocklyAR: A Visual Programming Interface for Creating Augmented Reality Experiences. Electronics. 2020; 9(8):1205
- [13] David W, David CS, Patrick F and Diana F, Blockly goes to work: Block-based programming for industrial robots, 2017 IEEE Blocks and Beyond Workshop (B&B), Raleigh, NC, 2017; pp. 29-36
- [14] Nguyen, V.T., Jung, K., Dang, T. BlocklyAR: A Visual Programming Interface for Creating Augmented Reality Experiences. Electronics 2020, International Conference on Information Technology and Electrical Engineering (ICITEE), Phuket, 2017; pp. 1-6
- [15] Matenat K, Natavut K, Kamol K and Kazuhiko F: MicroPython-based educational mobile robot for computer coding learning, 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES), Chonburi, 2017; pp. 1-6
- [16] Temi Personal Robot, Retrieve Nov 25, 2020, from <https://www.robotemi.com/>
- [17] Temi SDK Document Retrieve Nov 25, 2020, from <https://github.com/robotemi/sdk>

Towards Drug Repurposing for COVID-19 Treatment using Literature-based Discovery

Marina TROPMANN-FRICK ^{a,1} and Tobias SCHREIER ^a

^a*Hamburg University of Applied Sciences
Department of Computer Science
Hamburg, Germany*

Abstract. The ongoing COVID-19 pandemic brings new challenges and risks in various areas of our lives. The lack of viable treatment options is one of the major issues in coping with the pandemic. The development of a new drug usually takes approximately 10-15 years, time that we don't have during a progressing pandemic. As an alternative to the development of new drugs, the repurposing of existing drugs has been proposed. One of the scientific methods that can be used for drug repurposing is literature-based discovery (LBD). LBD uncovers hidden knowledge in the scientific literature and has already successfully been used for drug repurposing in the past. The aim of this work is to give an overview of existing LBD methods that can be used to search for new COVID-19 treatment options. We compare the three existing LBD systems Arrowsmith, BITOLA and SemBT concerning their suitability for this task. Our research shows that semantic models appear to be the most suitable for drug repurposing. However, Arrowsmith produces the best results, despite the fact that it uses a co-occurrence model instead of a semantic model. But it achieves the good results at the moment due to the fact, that both BITOLA and SemBT currently do not allow for COVID-19 related searches. Once this limitation is removed, we assume that SemBT, which uses a semantic model, will be the better choice for the task.

Keywords. Literature-based discovery, drug repurposing, COVID-19, Arrowsmith, BITOLA, SemBT

1. Introduction

Almost a year after the COVID-19 pandemic started, there is still a lack of viable treatment options. Although a few promising drug candidates have been proposed, including remdesivir, chloroquine and hydroxychloroquine [25], their efficacy and safety is still under investigation. What some of the proposed drugs have in common is that they have been repurposed for the treatment of COVID-19. Remdesivir, for example, was initially developed for treating Ebola virus disease and Marburg virus disease, but was found to be ineffective against these viral infections [25]. However, antiviral activity was demon-

¹Corresponding Author: Marina Tropmann-Frick, Hamburg University of Applied Sciences, Germany; E-mail: marina.tropmann-frick@haw-hamburg.de

strated against SARS and MERS, two diseases caused by coronaviruses closely related to SARS-CoV-2 [2]. Chloroquine and hydroxychloroquine were initially developed for treating malaria [25].

It comes at no surprise that promising candidates for the treatment of COVID-19 are existing drugs. The development of a new drug takes approximately 10-15 years and costs between \$500 million and \$2 billion [1,11,6]. While the cost of developing a new drug for COVID-19 should be of secondary concern, considering 191 countries/regions around the world are currently affected by the virus (Johns Hopkins COVID-19 Dashboard², 13 November 2020), the development time of 10-15 years is not. The rapidly evolving situation demands new treatment options as fast as possible. An alternative to the development of new drugs is the use of existing drugs for new indications, a process known as drug repurposing or drug repositioning.

Literature-based discovery (LBD) is a cost- and time-efficient method that can be used for drug repurposing. It automatically or semi-automatically generates hypotheses for scientific research by finding hidden links in existing scientific literature [14,13]. LBD operates on large literature datasets such as MEDLINE³. Discoveries have the form of relations between two previously unrelated concepts, for example a disease, and a drug that treats the disease. Such relations are discovered by uncovering a third concept, like a physiologic function or gene expression, that relates to both the drug and the disease [14]. The discovery of the linking concept leads to the assumption, that there may also be a link between the two primary concepts, which can then be investigated further.

In this work we give an overview of LBD methods and systems that could be used to search for possible COVID-19 treatments. Our objective is not to give a comprehensive overview of existing LBD methods. Such an overview can be found in [13]. Instead, we discuss the LBD methods that seem to be the most suitable regarding the search for COVID-19 treatments. Our method selection is based on a literature research about LBD with focus on drug repurposing. Section 2 presents related work on LBD and its application to drug repurposing. Section 3 gives an overview of LBD methods that may be used to search for drug candidates. Section 4 introduces three LBD systems and explains how they operate. Based on examples, we demonstrate how they could be used to search for COVID-19 treatment options. Section 5 concludes with a summary of our findings and discusses the future work.

2. Related work

The Literature-based discovery was first proposed by Swanson for uncovering hidden knowledge in scientific literature [28]. He read about Raynaud's disease increasing blood viscosity and platelet aggregation in one set of articles, and fish oil reducing blood viscosity and platelet aggregation in another set of articles. But he found no studies that reasoned that fish oil could treat Raynaud's disease. As a result, he proposed fish oil as a new treatment option for Raynaud's disease. Another previously unknown relationship he found, was that magnesium may help against migraine [29]. These relationships were discovered manually by researching the literature. To automate the LBD process, Swanson and Smalheiser initiated the Arrowsmith project, a co-occurrence based LBD system

²<https://coronavirus.jhu.edu/map.html>

³<https://pubmed.ncbi.nlm.nih.gov/>

for automatic knowledge generation [30]. To overcome some of the limitations of co-occurrence based models, the system was later improved with the integration of MeSH [31] and the UMLS [35]. Another co-occurrence based LBD system named BITOLA was developed by Hristovski et al. [15]. To search for drugs that may be repurposed for different indications, Hristovski et al. proposed the use of discovery patterns [14]. To apply the discovery patterns, semantic knowledge had to be derived from the literature. Hristovski et al. used the semantic parsers BioMedLEE and SemRep for this task. Later they integrated SemRep with BITOLA and named their system SemBT (Semantic BITOLA). Ahlers et al. adapted the discovery patterns proposed by Hristovski et al. and developed their own discovery pattern [3]. While the discovery patterns proposed by Hristovski et al. are aimed at drug repurposing, the discovery pattern derived by Ahlers et al. is focused on investigating hitherto unknown mechanisms of action involved in existing drug applications. Henry and McInnes provide an comprehensive overview of current and past LBD methods and systems [13]. Thilakarathne et al. [34] give a systemic review on existing LBD literature.

3. Methods

This section explains methods and steps involved in the LBD process. The structure and the discussed topics are derived in part from Henry and McInnes overview on LBD methods [13]. Section 3.1 explains semantic models and the semantic predications they operate on. Section 3.2 introduces SemRep and BioMedLEE, two semantic parsers for semantic predication extraction. Section 3.3 proposes different methods for removing uninformative terms. Section 3.4 describes relations from BioMedLEE and SemRep used to extract semantic predications from biomedical text. Section 3.5 introduces Swanson's ABC model and explains the difference between open and closed discovery. Section 3.6 presents discovery patterns for drug repurposing with LBD. Sections 3.7 and 3.8 briefly address term ranking and results display. Section 3.9 discusses different techniques for evaluating the performance of LBD systems.

3.1. Semantic models

The most promising LBD models for drug repurposing seem to be semantic models. Other LBD models include co-occurrence models and distributional models. They will not be discussed here, but an explanation of these models and literature for further reading is provided in [13]. Other than co-occurrence models, which directly use co-occurrences to determine what constitutes a relationship, semantic models use semantic predications that are extracted from biomedical literature using semantic parsers [13]. Semantic predications reflect relations assumed between two terms in text [3], for example, *chloroquine treats malaria*. The two terms in this case are *chloroquine* and *malaria*, the relation assumed between them is *treats*. Using semantic predications increases the quality of the extracted relations at the risk of missing relations. Thus, it increases the models precision at the cost of recall. Another benefit of using semantic predications is, that the extracted relations are labeled, which allows for the removal of uninteresting relations. This reduces the amount of reading by the user required. The precision of the model can be increased further by manually eliminating relations that have been wrongly

identified (false positives). Furthermore, using predications can possibly explain potential discoveries [14]. Other benefits of semantic predications include normalization, stop word removal, and identification of multi-word terms [13].

3.2. Semantic parsers

Semantic parsers extract semantic predications from text. The most popular semantic parser in the biomedical domain is SemRep⁴ [24]. Semantic predications in SemRep are three-part propositions, which consist of a subject argument, an object argument and a relation that binds them. For example, from the sentence in 1, SemRep extracts the predications in 2.

1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
2.
 - (a) Hemofiltration-TREATS-Patients
 - (b) Digoxin overdose-PROCESS_OF-Patients
 - (c) hyperkalemia-COMPLICATES-Digoxin overdose
 - (d) Hemofiltration-TREATS(INFER)-Digoxin overdose

SemRep is based on the UMLS⁵, a comprehensive collection of biomedical vocabularies and standards [5]. The UMLS consists of three knowledge resources, the Metathesaurus, the Semantic Network, and the SPECIALIST Lexicon and Lexical Tools. The Metathesaurus⁶, the biggest component of the UMLS, is a large biomedical vocabulary organized by concept or meaning. It links similar names for the same concepts from nearly 200 different vocabularies and identifies useful relationships between them. The Semantic Network⁷ consists of semantic types that provide a categorization of concepts from the Metathesaurus and useful semantic relations that exist between them. The SPECIALIST Lexicon and Lexical Tools⁸ include a large syntactic lexicon of biomedical and general English terms and tools for NLP tasks such as normalizing strings, generating lexical variants and creating indexes. SemRep has been used to extract about 94 million semantic predications from all 27.9 million PubMed articles (31 December 2017), which are stored in SemMedDB⁹ [17]. The subject and object arguments in SemRep semantic predications are concepts from the Metathesaurus, while the relationships between them are semantic relations from the Semantic Network. Hristovski et al. use SemRep together with another semantic parser, BioMedLEE, in order to gain knowledge from biomedical literature. BioMedLEE is a knowledge-based phenotype organizer system that extracts genotype-phenotype relations from biomedical text [19].

⁴<https://semrep.nlm.nih.gov/>

⁵<https://www.nlm.nih.gov/research/umls/index.html>

⁶https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html

⁷<https://semanticnetwork.nlm.nih.gov/>

⁸<https://lexsrv3.nlm.nih.gov/Specialist/Home/index.html>

⁹<https://skr3.nlm.nih.gov/SemMedDB/>

3.3. Uninformative term removal

Uninformative terms in biomedical text include English stop words such as *the* or *and* and general or broad biomedical terms like *drug* or *disease* [13]. They provide no new or interesting information and correlate with most other terms [21]. Removing uninformative terms is a vital step in the LBD process in order to decrease the number of extracted relations and to increase their quality. Different methods may be used to restrict the extracted terms, including stop word removal, hierarchical filtering, and semantic type and relation filtering.

Stop word removal SemRep automatically eliminates English stop words by mapping terms to UMLS semantic types using MetaMap [8]. Therefore, automatic stop word generation techniques in the biomedical domain only have to focus on general biomedical terms. Stop word lists may be generated automatically by searching for terms that produce many linking terms. Preiss identifies and removes highly connected terms by repeatedly and randomly finding hidden knowledge [23].

Hierarchical filtering Pratt et al. and Hu et al. eliminate overly broad biomedical terms by removing concepts on the first, second or third level of the UMLS hierarchy [21,16]. The UMLS hierarchy may also be used to remove terms that are too similar to either the start or linking terms. Although UMLS maps synonymous terms to the same concept, the concept distinctions are often too fine grained. For example, *migraine* and *common migraine* [37] are very similar but map to different UMLS concepts. Pratt et al. remove terms that are parents and children of the start term [21]. Yildiz expands this to include grandparents and siblings [37]. Preiss et al. use a list of synonymous concepts provided by the UMLS to identify similar terms [23,22].

Semantic type filtering Uninformative terms may also be eliminated by restricting linking and target terms to specific UMLS semantic types. Selecting appropriate semantic types is challenging and requires an understanding of both the UMLS and medical terminology. A too restrictive approach may lead to the loss of important linking and target terms, while an approach not restrictive enough will produce too many uninformative terms [13].

Relation type filtering Similar to semantic type filtering, uninteresting relationships between terms may be removed using relation type filtering. SemRep provides 58 predefined relation types, which are assigned to relationships extracted from biomedical text. Those can be used to remove irrelevant relations for the respective task such as *prevents* relations when the objective is to find *stimulates* relations, or to eliminate negative relations such as *neg.causes*. The discovery patterns introduced in section 3.6 use both semantic type and relation filtering to remove uninformative terms.

3.4. Semantic predication extraction

Following the approach of a semantic model, the first step in the LBD process is the extraction of semantic predications from the literature. Semantic predications reflect known facts, that are contained explicitly in the literature. Hristovski et al. use the *Associated_with_change* relation from BioMedLEE to extract predications where one concept (e.g. a disease) is associated with a change in another concept (e.g. a pathological func-

Num	System	Extracted Relations	Sentence (or fragment)
1.	BL	Associated_with_change(oxidative stress, iron, increase)	reducing the oxidative stress associated with increased iron levels
2.	SR	Treats(coenzyme Q10, Huntington Disease)	Oral administration of CoQ10 significantly decreased elevated lactate levels in patients with Huntington's disease.
3.	BL	Associated_with_change(Raynaud's, blood viscosity, increase)	Local increase of blood viscosity during cold-induced Raynaud's phenomenon.
4.	BL	Associated_with_change(Raynaud's, viscosity, increase)	Increased viscosity might be a causal factor in secondary forms of Raynaud's disease, ...
5.	BL	Associated_with_change(eicosapentaenoic acid, blood viscosity, decrease)	We recently reported that eicosapentaenoic acid (EPA) also reduces whole blood viscosity.
6.	BL	Associated_with_change(eicosapentaenoic acid, blood viscosity, decrease)	A statistically significant reduction in whole blood viscosity was observed at seven weeks in those patients receiving the eicosapentaenoic acid rich oil.
7.	BL	Associated_with_change(Huntington's disease, insulin, decrease)	Huntington's disease transgenic mice develop an agedependent reduction of insulin mRNA expression and diminished expression of key regulators of insulin gene transcription, ...

Table 1. Relations extracted from biomedical literature with the BioMedLEE (BL) relation *Associated_with_change* and SemRep (SR) relation *Treats* [14].

tion). In addition to the *Associated_with_change* relation from BioMedLEE, they use the *Treats* relation from SemRep to extract drugs that are known to treat certain diseases. Table 3.4 shows semantic predications extracted from biomedical literature using the *Associated_with_change* relation from BioMedLEE and the *Treats* relation from SemRep [14].

The use of BioMedLEE is not required to extract semantic predications that reflect some kind of change in a concept provoked by another concept. Ahlers et al. entirely rely on SemRep for the extraction of semantic predications from biomedical literature. To represent the inhibitory action of one bioactive substance on another, they use the INHIBITS relation. Etiological relations between a bioactive substance and a pathological process are represented with the relations CAUSES, PREDISPOSES, or ASSOCIATED-WITH. Known drug-disease relationships are extracted using the TREATS and PREVENTS relations [3].

3.5. Swanson's ABC model

After extracting explicit semantic predications that reflect known facts in the literature relevant for the respective discovery task, the next step in the LBD process is the retrieval of previously unknown implicit knowledge. Basically all LBD systems rely on Swanson's ABC model to discover new knowledge from research literature [28]. It builds on the assumption, that when a term A is connected with a term B, and a term B is connected with a term C, it can be assumed, that there also is a link between terms A and C. Terms

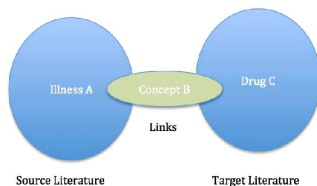


Figure 1. Swanson's ABC-model in the biomedical domain [32]. A terms from the source literature are linked to C terms in the target literature through B concepts shared between both distinct literature sets.

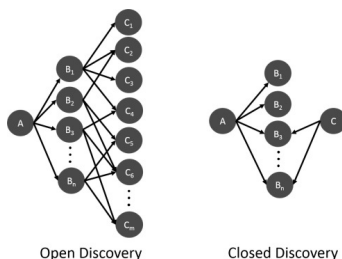


Figure 2. Two forms of Swanson's ABC model, open and closed discovery. Open discovery searches for linking terms (B) that are related to a start term (A) and based on the linking terms target terms (C) that are connected to the linking terms. Closed discovery searches for linking terms (B) that are connected to both a start (A) and target term (C) [13].

A and C in this model exist in distinct literature sets that are linked through B concepts. Figure 1 shows Swanson's ABC-model in the biomedical domain.

For example, Swanson proposed fish oil as a new treatment for Raynaud's disease. The A term in his discovery was Raynaud's disease. Two of the B terms that co-occurred with Raynaud's disease were blood viscosity and platelet aggregation, which are both increased in Raynaud's disease (AB literature). Blood viscosity and platelet aggregation were then in turn found to co-occur with a C term, fish oil (rich in eicosapentaenoic acid), which reduces blood viscosity and platelet aggregation (BC literature). Consequently, fish oil was proposed as a new treatment for Raynaud's disease (newly found AC relationship). Swanson's ABC model can be used for open and closed discovery, which are depicted in figure 2.

3.5.1. Open discovery

Open discovery is used to generate new hypotheses [13]. An initial start term A is given as input to the system by the user. The system then generates a list of linking terms B, that co-occur with the start term. Outgoing from the linking terms B, the system generates a list of target terms C, that co-occur with the linking terms B. The result is a list of previously unknown relations between the start term A and target terms C.

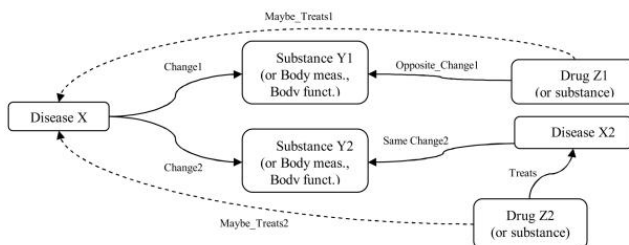


Figure 3. Two forms of the *Maybe_Treats* discovery pattern *Maybe_Treats1* and *Maybe_Treats2*. They can be used for both open and closed discovery. For open discovery, they take a disease as input. *Maybe_Treats1* searches for Y concepts changed by the disease. From that, drugs are identified, that elicit an opposite change of these Y concepts and thus may treat the disease. *Maybe_Treats2* searches for similar diseases X2 that cause the same changes of Y concepts as disease X. Drugs that treat disease X2 are considered to possibly also treat disease X. For closed discovery, both a drug and a disease are provided as input. The patterns then search for Y concepts that may explain the relationship between the drug and the disease [14].

3.5.2. Closed discovery

Closed discovery, on the contrary, is primarily used to explain correlations or observations [13]. For example, it may be used to investigate hypotheses previously generated with open discovery. To perform closed discovery, the user inputs a start term A and a target term C to the system. The system then generates a list of linking terms B, that are related to both the start term A and the target term C. The result is a list of linking terms B that could explain the relationship between the starting term A and the target term C.

3.6. Discovery patterns

Different discovery patterns have been proposed based on Swanson's ABC model to extract potentially new treatments from biomedical literature, or to explain existing drug-disease-relationships by identifying previously unknown pathways.

3.6.1. *Maybe_Treats*

Hristovski et al. [14] proposed the *Maybe_Treats* discovery pattern, which has the forms *Maybe_Treats1* and *Maybe_Treats2* depicted in figure 3. Hristovski et al. use X, Y and Z to denote the start, linking and target terms instead of A, B and C.

Following the *Maybe_Treats1* discovery pattern, a drug Z is considered to maybe treat a disease X, if there is a change of a Y concept, which might be a substance, or body measure or function, that is associated with the disease, and if the drug provokes an opposite change of this substance. To consider Z as a possible treatment for X as a new discovery, co-occurrences of X and Z in the literature have to be absent. In case of Swanson's discovery of fish oil as a new treatment for Raynaud's disease, drug Z refers to fish oil, while Raynaud's disease represents disease X. The Y concept that links Raynaud's disease to fish oil is blood viscosity, a body measure, which is increased in patients with Raynaud's disease and reduced by fish oil.

The *Maybe.Treats2* discovery pattern follows a different approach. Instead of searching for drugs that cause an opposite change of some kind of Y concept that is changed by the disease X, the pattern searches for different diseases X2, that evoke similar changes of a Y concept as disease X. Drugs that treat the disease X are assumed to possibly treat disease X as well. As with the *Maybe.Treats1* discovery pattern, articles that mention both disease X and drug Z cannot exist in the literature for the discovery to be new. Hristovski et al. observed, with the use of this approach, that insulin levels are decreased in patients with Huntington disease. Insulin levels are also decreased in patients with Diabetes Mellitus (type 2 diabetes). Therefore, drugs for the treatment of Diabetes Mellitus were proposed for treating Huntington disease.

3.6.2. *May_Disrupt*

Ahlers et al. use the discovery pattern *May_Disrupt* shown in 1 to investigate the mechanisms underlying drug therapies that are currently used but poorly understood [3]. Other than the *Maybe.Treats* discovery pattern, which searches for a wide range of linking concepts that may connect a drug to a disease, the *May_Disrupt* discovery pattern concentrates on pharmacogenomics, the relationship among drugs, genes, and diseases.

1. Substance X <inhibits> Substance Y
Substance Y <causes> Pathology Z
Substance X <may_disrupt> Pathology Z

To apply the pattern, first the relations depicted in 1 have to be extracted from the literature. Other than Hristovski et al., who use a combination of BioMedLEE and SemRep, Ahlers et al. solely rely on SemRep for that task. Relations where drug X inhibits substance Y are extracted using the INHIBITS relation. Relations where substance Y causes disease Z are extracted using the CAUSES, PREDISPOSES and ASSOCIATED.WITH relations. Relations where drug X may disrupt disease Z are extracted using the TREATS relation. When used for open discovery, the pattern states, that if drug X inhibits substance Y and if substance Y causes disease Z, drug X may disrupt and thus prevent or treat disease Z. When used for closed discovery, the pattern states, that for a drug X that treats or prevents a disease Z, if drug X inhibits substance Y and if substance Y causes disease Z, then substance Y is involved in the mechanisms of action of drug X treating or preventing disease Z.

3.7. *Term ranking*

Term ranking can be used for ordering and displaying linking and target terms. It can also be used for removing uninformative terms by applying a threshold. Statistical thresholds are thereby less affected by corpus size than term-occurrence-based thresholds [13]. Yetisgen-Yildiz and Pratt compared several ranking measures and found that Linking Term Count with Average Minimum Weight as a tie breaker (LTC-AMW) performed the best [36]. Arrowsmith incorporates a logistic regression model that estimates the probability for each linking term to be relevant. The linking terms are ranked based on their predicted relevance. BITOLA computes frequency and confidence for restricting and ranking terms. SemBT uses the number of articles a term occurs in for term ranking.

January 2021

3.8. Results display

The most common output of LBD systems is a ranked list of linking and target terms. As this method does not provide a sufficient explanation of the discoveries, many systems display additional information [13]. Arrowsmith and SemBT present the user the articles that were involved in linking two terms together. The user may read the articles and decide for herself, if the relationship found is plausible. BITOLA forwards the user to a PubMed search for the terms involved in the linkage. Since the user cannot determine which of the articles presented by PubMed were responsible for linking the terms together, this method is much less useful and transparent than the approach of Arrowsmith and SemBT.

3.9. Evaluation

The evaluation of LBD systems is intrinsically difficult, since their purpose is to uncover relationships that have been previously unknown. However, Henry and McInnes describe three evaluation methods that have become standard for system evaluation [13].

3.9.1. Discovery replication

Discovery replication works by replicating one or more discoveries previously made, for example, using other LBD systems. While not explicitly mentioned in [13], there seems to be no good reason for why discoveries made without the use of LBD systems should not be replicated for system evaluation as well. Considering COVID-19 research, previously proposed drugs for the treatment of COVID-19, such as remdesivir and chloroquine, may be used for evaluating a systems performance, especially regarding COVID-19 drug repurposing. The most commonly replicated discovery is the relationship found between Raynaud's disease and fish oil by Swanson in 1986 [28]. For discovery replication, only literature published before the to-be replicated discovery is used. For example, to replicate Swanson's Raynaud's disease-fish oil discovery, only literature published before 1986, the year Swanson's paper was published, may be used. Based on the pre-discovery literature, discoveries are generated by the LBD system. Discovery replication may be used for evaluating open, as well as closed discovery. For evaluating open discovery, the chosen discovery is considered successfully replicated if the target terms contain the to-be replicated discovery. For the evaluation of closed discovery, the linking terms including the discovery of interest constitutes a successful replication. The presence of the replicated discovery in the linking or target terms is not enough to conclude, that the discovery would also be found by a researcher, since the output may consist of too many terms for the researcher to manually go through. Thus, reporting the rank of the discovery in the linking or target term list is crucial for evaluating a systems performance.

3.9.2. New discovery proposal and empirical evaluation

A major limitation to discovery replication is, that it does not evaluate a systems capability to actually make new discoveries. New discovery proposal, on the contrary, shows a system is able to generate practical new knowledge. Therefore, discovery replication is often combined with new discovery proposal to evaluate a systems performance. However, in order to be sufficient, new discovery proposal has to be validated by expert vetting or empirical evaluation. Expert vetting may consist of evaluation by an expert or

a publication in the respective research area. For drug repurposing, this concerns the biomedical domain. Expert vetting allows for the elimination of obvious, uninteresting or incorrect hypotheses. Promising hypotheses are kept and empirically evaluated via laboratory testing. For drug repurposing, promising candidates are tested in clinical trials in terms of efficacy and safety concerning the new indication.

3.9.3. Time slicing

Another drawback to discovery replication is, that it only shows a systems capability to generate a single discovery and is thus prone to overfitting. Time slicing alleviates this problem by showing a systems ability to generalize and produce many new discoveries. For that, a cutoff date is specified, that divides the literature into two distinct sets of articles published before and after the cutoff date. Articles published before the cutoff date are used to generate new discoveries, while articles published after the cutoff date are used to evaluate the generated discoveries. The perfect validation dataset would be a list of all new real world knowledge discovered after the cutoff date. However, that being impossible to attain, the validation dataset is comprised of relationships present in the test set and absent from the training set. These relationships represent new discoveries, but raise the same question as the system design itself, of what constitutes a relationship. For the sake of simplicity and consistency, the same method as used by the system itself could be used to determine what counts as a new discovery, but other methods may be used for that just as well. Precision and recall, as well as other derived statistical measures are used to quantify time slicing evaluation. An overview of popular methods is given in [13].

4. LBD systems

Based on the methods described previously and other LBD techniques, several LBD systems have been developed in the past. We chose three LBD systems for further evaluation, Arrowsmith, BITOLA and SemBT. An overview of other LBD systems is provided in [34]. We briefly explain the functionality of the systems and demonstrate open and closed discovery for each system, based on example experiments. Additionally, we provide a conclusion for each system, that outlines its major limitations.

4.1. Arrowsmith

Based on Swanson's manual LBD discoveries of the connections between Raynaud's disease and fish oil [28], and between migraine and magnesium [29], Swanson and Smalheiser developed the Arrowsmith LBD system¹⁰ [30]. Arrowsmith is the very first semi-automatic LBD system and uses a co-occurrence model in order to constitute relationships between terms appearing together in the literature. To overcome the limitations of co-occurrence based models, Smalheiser later improved the system by integrating biomedical knowledge resources including Medical Subject Headings (MeSH¹¹) [31] and the UMLS [35]. MeSH is a hierarchically-organized vocabulary controlled by the

¹⁰http://arrowsmith.psych.uic.edu/cgi-bin/arrowsmith_uic/start.cgi

¹¹<https://www.nlm.nih.gov/mesh/meshhome.html>

National Library of Medicine (NLM). It includes subject headings appearing in MEDLINE/PubMed, the NLM catalog and other NLM databases and is used for indexing, cataloging and searching of biomedical information.

Arrowsmith allows for open and closed discovery, which are referred to as one- and two-node searches. One-node searches may also be performed with two-node searches. In order to perform a one-node search with a two-node search, the A literature is limited to a small set of articles representing a specific problem, while the C literature may consist of a broad spectrum of articles, for example, all articles belonging to a specific MeSH category [26]. Arrowsmith allows the user to input PubMed queries in order to define the A and C literature. For that purpose, a simplified version of the PubMed query box was integrated into the system.

4.1.1. Two-node search

To perform a two-node search, the user is asked to input two separate PubMed queries to define the A and C literature. We used the two-node search to investigate linking concepts that may explain the mechanisms of action involved in remdesivir treating COVID-19. To define the A literature, we used the start term *coronavirus disease 2019*. This query yielded 25,000 articles dealing with COVID-19, which is the maximum number of articles considered for the A literature. For the C literature, we chose *remdesivir* as target term. That resulted in 526 articles about remdesivir. Theoretically, up to 25,000 articles could be considered by the system for the C literature, so the total number of articles considered for a two-node search adds up to 50,000 articles [26]. Thereby, always the latest 50,000 articles are considered. Articles present in both literature sets are removed, so only indirect relations between the literature sets via linking concepts are captured. However, the removed articles are kept and the user can view them if she wants. For *coronavirus disease 2019* and *remdesivir*, this affected 457 articles, leaving 68 articles that deal with *remdesivir* but not *coronavirus disease 2019*. Next, the system searches for words and two- and three-word phrases that occur in article titles in both the A and C literature. The resulting linking terms are processed and ranked according to the predicted probability that they will be relevant to the user. Torvik and Smalheiser integrated this feature into Arrowsmith to solve the problem of predicting which of the hundreds to thousands of linking terms returned for a single query are most likely to be relevant to the user [35]. They developed a logistic regression model that estimates the probability for each linking term to be relevant. Based on their predicted relevance, the linking terms can be ranked. Furthermore, the model allows to estimate the total number of relevant linking terms for a given two-node search. For *coronavirus disease 2019* and *remdesivir*, 498 linking terms were generated, 89 of which were predicted to be relevant. To limit the linking terms to concepts that may explain the mechanism of action involved in remdesivir treating COVID-19, we restricted the linking terms to the semantic types *Anatomy*, *Chemicals & Drugs*, *Genes & Molecular Sequences*, and *Gene & Protein Names*, and *Physiology*. That narrowed down the list to 96 linking terms. 13 of these terms were predicted to be relevant and are shown in table 4.1.1.

Arrowsmith allows the user to inspect the articles connected through a linking term by selecting a term of interest. Multiple linking terms may be selected as well. Of the 13 terms, only *rna dependent rna*, *dependent rna polymerase* and *dependent rna* explain the physiologic link between COVID-19 and remdesivir. It should be noted that these terms are synonyms for RNA-dependent RNA polymerase and thus refer to the same concept.

Rank	Probability	B-term
1	0.99	ritonavir
2	0.99	respiratory syndrome coronavirus
3	0.99	lopinavir ritonavir
4	0.98	ebola
5	0.97	cov
6	0.96	rna dependent rna
7	0.94	dependent rna polymerase
8	0.94	dependent rna
9	0.74	antiviral strategy
10	0.74	chloroquine
11	0.64	provide insight
12	0.59	sars
13	0.48	interferon beta

Table 2. 13 linking terms predicted to be relevant by Arrowsmith in closed discovery to explain the relationship between *coronavirus disease 2019* and *remdesivir*. The linking terms were limited to the semantic types *Anatomy, Chemicals & Drugs, Genes & Molecular Sequences, and Gene & Protein Names, and Physiology*.

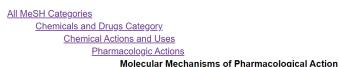


Figure 4. Position of the *Molecular Mechanisms of Pharmacological Action* category in the MeSH hierarchy (<https://www.ncbi.nlm.nih.gov/mesh/68045504>).

Remdesivir inhibits the RNA-dependent RNA polymerase of MERS [12], various flaviviruses [18], Ebola virus [33], and human endemic and zoonotic deltacoronaviruses [7]. Since SARS-CoV-2 also relies on an RNA-dependent RNA polymerase for the catalyzation of the RNA replication process [20], this could provide an explanation of the mechanisms of action involved in remdesivir treating COVID-19.

4.1.2. One-node search

To perform a one-node search, the user inputs a single PubMed query to define the A literature. We used the start term *coronavirus disease 2019* in order to limit the A literature to articles dealing with COVID-19. Next, the user is asked to choose a MeSH category to narrow down the C literature to search for target terms. Since the objective is to find existing drugs that may be repurposed for COVID-19 treatment, the target terms should be drugs. Therefore, we chose the MeSH category *Molecular Mechanisms of Pharmacological Action*, which includes twenty classes of drugs. The position of this category in the MeSH hierarchy is shown in figure 4.

Once the user chose a category, the system performs a number of two-node searches between the defined A literature and all subcategories of the defined MeSH category (C literature). In case of the *Molecular Mechanisms of Pharmacological Action* category, all drug subcategories are searched. For each subcategory, a number of metrics is calculated that quantify the result.

- nC = number of articles in the C literature

Rank	Job ID	C-query	nC	nAC	nTot	nR	pR
1	1901117	Peptidomimetics [mh]	1160	3	2621	481	0.184
2	1901118	Enzyme Inhibitors [mh]	50000	381	28846	5081	0.176
3	1901111	Angiotensin Receptor Antagonists [mh]	17334	141	13995	2165	0.155
4	1901110	Fibrin Modulating Agents [mh]	38061	33	22974	3385	0.147
5	1901118	Radiopharmaceuticals [mh]	50000	7	22273	3184	0.143
6	190111	Alkylating Agents [mh]	17397	2	13492	1793	0.133
7	1901112	HIV Fusion Inhibitors [mh]	1213	0	2931	384	0.131
8	1901114	Antimetabolites [mh]	50000	28	28015	3561	0.127
9	1901115	Neurotransmitter Agents [mh]	50000	16	27309	3355	0.123
10	1901113	Membrane Transport Modulators [mh]	50000	5	24221	2870	0.118
11	1901111	Heparin Antagonists [mh]	1288	0	2651	270	0.102
12	1901114	Mitosis Modulators [mh]	15530	4	11378	1122	0.099
13	1901117	Enzyme Activators [mh]	2610	0	4604	453	0.098
14	1901112	Antacids [mh]	6292	0	6840	657	0.096
15	1901116	Nitric Oxide Donors [mh]	6946	0	6697	591	0.088
16	1901115	Antioxidants [mh]	50000	20	19749	1673	0.085
17	1901113	Antifoaming Agents [mh]	192	0	793	59	0.074
18	1901119	Enzyme Reactivators [mh]	1749	0	2742	191	0.070
19	1901119	Sequestering Agents [mh]	31538	2	16122	1044	0.065
20	1901116	Cerumenolytic Agents [mh]	27	0	132	2	0.015

Table 3. Result of the Arrowsmith open discovery subcategory search for the MeSH category *Molecular Mechanisms of Pharmacological Action* for the start term *coronavirus disease 2019*. Clicking on the Job IDs shows the linking terms that have been generated for the respective subcategory.

- nAC = number of articles in both the A and C literature
- nTot = total number of linking terms in the subcategory search
- nR = number of linking terms predicted to be relevant
- pR = percentage of linking terms predicted to be relevant (high pR values indicate that the A and C literature share a lot of implicit information. $pR < 0.1$ is near chance level whereas $pR > 0.3$ is a relatively high value).

Table 4.1.2 shows the metrics computed for each subcategory. The user may click each of the Job IDs to investigate the linking terms that have been generated with the respective subcategory. Peptidomimetics scored the highest pR and were ranked first.

Clicking on a Job ID opens a similar interface as for the two-node search. The user is presented the generated linking terms between the start term and selected target term. As with the two-node search, the user may select one or more linking terms and inspect the articles containing the start and linking term, and the linking and target term. Unfortunately, unlike as with the two-node search, right now the linking terms can not be restricted using semantic type filtering. Clicking on the respective button results in an internal server error (last try 13 November 2020 16:15 CEST).

For *coronavirus disease 2019* and *peptidomimetics*, 2,621 linking terms were generated, 481 of which were predicted to be relevant. 445 articles appeared in both literature

Rank	Probability	B-term
1	0.99	respiratory syndrome coronavirus
2	0.99	molecular dynamic simulation
3	0.99	molecular docking
4	0.99	syndrome coronavirus
5	0.99	3c protease
6	0.99	dynamic simulation
7	0.99	docking study
8	0.99	docking molecular
9	0.99	furin
10	0.99	proteasome inhibitor

Table 4. First ten linking terms predicted by Arrowsmith using open discovery to be relevant to explain the relationship between *coronavirus disease 2019* and *peptidomimetics*.

sets and were not included in the search for linking terms. Table 4.1.2 shows the first ten linking terms predicted to be relevant by Arrowsmith.

Linking terms that could stimulate further research include *3c protease*, *furin* and *proteasome inhibitor*. For example, ten articles investigate the effects of 3C-like protease inhibition on SARS-CoV-2 and related coronaviruses (e.g. [9]). Four studies research peptidomimetics as 3C-like protease inhibitors [4,38,27,10].

4.1.3. Discussion

The results presented above demonstrate Arrowsmith's potential regarding the search for COVID-19 treatments. However, despite these encouraging results, the obvious limitations to Arrowsmith resulting from the underlying co-occurrence model cannot be denied. Although the system was improved with the integration of MeSH categories and UMLS semantic type filtering, the system still outputs a lot of irrelevant linking terms. This is due to the fact, that co-occurrence based models do not make use of known semantic knowledge in biomedical text. The problem of finding relevant linking terms worth researching is aggravated by the fact, that semantic type filtering for linking terms generated by one-node searches does not work at the moment.

4.2. BITOLA

BITOLA¹² is a LBD system developed by Hristovski et al. [15], which supports both open and closed discovery. It uses association rule mining, a variant of the co-occurrence model [13], to identify relationships. The system uses medical subject headings for indexing MEDLINE and human genes from HUGO. Like Arrowsmith, BITOLA searches MEDLINE articles for linking concepts. Hristovski et al. refer to A, B, and C as X, Y, and Z. BITOLA computes frequency and confidence for restricting and ranking terms. The retrieved terms may be filtered using MeSH semantic groups and types, and the computed frequency and confidence.

The use of medical subject headings turned out to be a limiting factor, at least concerning the search for COVID-19 treatments. For the sake of comparability, we intended

¹²<https://ibmi.mf.uni-lj.si/bitola>

Table 5. Part I - First ten linking terms generated by BITOLA using closed discovery for the start term *Severe Acute Respiratory syndrome* and the target term *Chloroquine*. The linking terms were limited to the semantic type *Enzyme*.

Concept Name	Semantic Type	FreqXY	ConfXY (%)	FreqYZ
Lactate Dehydrogenase	Enzyme	19	0.768	42
Cysteine Protease	Enzyme	8	0.323	31
Endopeptidases	Enzyme	4	0.162	41
Cathepsins	Enzyme	1	0.040	43
Alanine Transaminase	Enzyme	12	0.485	11
Aspartate Transaminase	Enzyme	7	0.283	20
RNA-Directed RNA Polymerase	Enzyme	7	0.283	1
paired basic amino acid cleaving enzyme	Enzyme	1	0.040	3
Peptide Hydrolases	Enzyme	2	0.081	46
ALANINE AMINOPEPTIDASE	Enzyme	5	0.202	1

to investigate the relationship between COVID-19 and remdesivir, as we did with Arrowsmith. However, although *COVID-19*¹³ and *remdesivir*¹⁴ both exist as medical subject headings, BITOLA does not allow to use them as start or target terms. This might be due to the fact that both *COVID-19* and *remdesivir* are currently classified as MeSH Supplementary Concept Data, which may not be recognized by BITOLA. *Severe Acute Respiratory Syndrome*¹⁵, for example, classified as MeSH Descriptor Data, is recognized by BITOLA. Same goes for *Chloroquine*¹⁶.

4.2.1. Closed discovery

Therefore, we used *Severe Acute Respiratory Syndrome* and *Chloroquine* to demonstrate the use of BITOLA for closed discovery. *Severe Acute Respiratory Syndrome* and *Chloroquine* appeared together in two MEDLINE articles. In total, 1,676 linking terms were generated. We restricted the linking terms to the semantic type *Enzyme*, as we were interested in an potential inhibitory effect of chloroquine on enzymes involved in the replication of SARS. This narrowed down the list to 32 linking terms. Table 5 shows the ten linking terms ranked first.

Clicking on FreqXY or FreqYZ performs a PubMed query consisting of the respective X and Y or Y and Z terms. Unfortunately, the user cannot determine which of the articles returned by the PubMed query were involved in linking the terms together. For example, clicking on FreqXY of *Lactate Dehydrogenase* leads to a PubMed search that returns 246 articles¹⁷ (30 August 2020). But it cannot be determined, which of these articles are part of the 19 articles that linked *Severe Acute Respiratory Syndrome* to *Lactate Dehydrogenase*. This limits the tools usability for further research, especially compared to Arrowsmith, which presents the user exactly the articles, that were involved in linking two terms together.

¹³<https://meshb.nlm.nih.gov/record/ui?ui=C000657245>

¹⁴<https://meshb.nlm.nih.gov/record/ui?ui=C000606551>

¹⁵<https://meshb.nlm.nih.gov/record/ui?ui=D045169>

¹⁶<https://meshb.nlm.nih.gov/record/ui?ui=D002738>

¹⁷<https://pubmed.ncbi.nlm.nih.gov/?cmd=Search&term=Severe%20Acute%20Respiratory%20Syndrome%5BMH%20NM%5D%20AND%20Lactate%20Dehydrogenase%5BMH%20NM%5D&dispmx=50>

January 2021

Table 6. Part II - First ten linking terms generated by BITOLA using closed discovery for the start term *Severe Acute Respiratory syndrome* and the target term *Chloroquine*. The linking terms were limited to the semantic type *Enzyme*.

Concept Name	ConfYZ(%)	FreqXY*FreqYZ	ConfXY*ConfYZ
Lactate Dehydrogenase	0.137	798	0.105
Cysteine Protease	0.293	248	0.095
Endopeptidases	0.195	164	0.032
Cathepsins	0.747	43	0.030
Alanine Transaminase	0.061	132	0.030
Aspartate Transaminase	0.101	140	0.028
RNA-Directed RNA Polymerase	0.069	7	0.020
paired basic amino acid cleaving enzyme	0.464	3	0.019
Peptide Hydrolases	0.201	92	0.016
ALANINE AMINOPEPTIDASE	0.070	5	0.014

Concept Name	Semantic Type	Freq	Conf(%)
Lactate Dehydrogenase	Enzyme	19	0.768
ACE2 enzyme	Enzyme	12	0.485
Carboxypeptidase	Enzyme	12	0.485
Alanine Transaminase	Enzyme	12	0.485
Cysteine Protease	Enzyme	8	0.323
Creatine Kinase	Enzyme	8	0.323
Aspartate Transaminase	Enzyme	7	0.283
RNA-Directed RNA Polymerase	Enzyme	7	0.283
3C-like proteinase, Coronavirus	Enzyme	6	0.243
ALANINE AMINOPEPTIDASE	Enzyme	5	0.202

Table 7. First ten linking terms generated by BITOLA using open discovery for the start term *Severe Acute Respiratory Syndrome*. The linking terms were limited to the semantic type *Enzyme*.

4.2.2. Open discovery

To perform open discovery with BITOLA, the user is asked to enter a medical subject heading as start term. Because *COVID-19* is not recognized as a medical subject heading by BITOLA, we used *Severe Acute Respiratory Syndrome* instead. Without further restriction, BITOLA generated 2,848 linking terms. Since this is way too much for manual evaluation, we limited the linking terms to the semantic type *Enzyme*. This reduced the number of linking terms to 52, the first ten of which are shown in table 4.2.2. Next, the user has to select the linking terms she wants to use to search for target terms. Unfortunately, there is no option to select/deselect all generated linking terms at once. They have to be selected/deselected one by one. We decided to investigate *Lactate Dehydrogenase*, which was ranked first.

Without further restriction, the search generated 22,069 target terms. Therefore, we limited the target terms to the semantic type *Pharmacologic Substances*, which includes *Enzyme Inhibitors*. That shrank the linking terms down to 2,538. Unfortunately, there is no option to further restrict the target terms, for example, to exclusively include *Enzyme Inhibitors*. Table 8 shows the first ten target terms generated for *Lactate Dehydrogenase*.

Table 8. First ten target terms generated by BITOLA using open discovery for the start term *Severe Acute Respiratory Syndrome* and the linking term *Lactate Dehydrogenase*. The target terms were restricted to the semantic type *Pharmacologic Substance*.

Concept Name	Rank Freq	Rank Conf	Count	Ys	Freq	Conf	"Discovery?"
Lactate	26068	3,4285	1	1372	4.463	YES	
Adenosine Triphosphate	20368	2,6789	1	1072	3.487	YES	
Lactic acid	9481	1,247	1	499	1.623	YES	
Enzyme Inhibitors	8569	1,127	1	451	1.467	NO	
Amino Acids	7676	1,0096	1	404	1.314	NO	
Antioxidants	7296	,9596	1	384	1.249	NO	
Superoxide Dismutase	7011	,9221	1	369	1.200	YES	
Recombinant Insulin	6479	,8521	1	341	1.109	NO	
Amylases	6213	,8172	1	327	1.064	YES	
Hydrogen Peroxide	5662	,7447	1	298	0.969	YES	

4.2.3. Discussion

The example of COVID-19 as search term has shown, that the use of medical subject headings can limit the usability of BITOLA, since COVID-19 was not recognized by the system. Another drawback to BITOLA is, that terms may only be restricted using a limited subset of MeSH semantic groups and types. The system does not allow for further restriction using lower levels of the MeSH hierarchy. This is demonstrated by the open discovery search for *Enzyme Inhibitors*, which, despite the restrictions put in place, returned 451 articles dealing with *Lactate Dehydrogenase* and *Enzyme Inhibitors*. An option to further limit the *Enzyme Inhibitors* to agents specifically targeting *Lactate Dehydrogenase* would have been useful. What's also a limiting factor to the usability of BITOLA, is the fact, that the user is not presented the articles that were involved in term-linking. Instead, the user is referred to a general PubMed search for the terms involved, which impairs the systems transparency. The biggest limitation to BITOLA remains its underlying co-occurrence model. Without the use of semantic knowledge, the system generates to many unrelated terms. To increase the quality of the generated terms, Hristovski et al. proposed the use of discovery patterns, which were discussed in section 3.6. Although this is a promising approach, it requires external semantic parsers like SemRep and BioMedLEE, to extract semantic knowledge from the corpus.

4.3. SemBT

To address some of BITOLAs issues, Hristovski developed SemBT¹⁸ (Semantic BITOLA), the semantic version of BITOLA, which takes advantage of semantic knowledge extracted from biomedical text with SemRep. SemBT supports both open and closed discovery. Search queries cannot be entered in natural language, but instead have to be formulated as so called questions, which may consist of subject, relation, and object. These refer to the different components of SemRep semantic predications (see section 3.2). At least one of the components must be specified, but two, or even all three components may be specified as well. The question is forwarded to Lucene, which means

¹⁸<http://semtb.mf.uni-lj.si/>

that full Lucene query syntax is allowed¹⁹. Like BITOLA, SemBT currently neither recognizes COVID-19 nor related terms, which limits the tools usability for searching COVID-19 treatments.

1. Chloroquine: Simple question with only one component specified. The concept *Chloroquine* may be either the subject or the object. The question will return any biomedical concepts related to *Chloroquine*.
2. Chloroquine TREATS: More specific question, where both a concept and a relation are specified. The concept *Chloroquine* may be either the subject or the object, regardless of whether it is placed before or after the relation TREATS. The question will return any biomedical concepts that are related to *Chloroquine* via the TREATS relation.
3. Chloroquine TREATS Malaria: Concrete question where all three components are specified. Both concepts may be either subject or object. The question will return any semantic relations that match the specified criteria.

The generated subjects and objects, and the relations may be filtered using semantic types²⁰ and relations²¹. The semantic types have to be abbreviated. The subject and object, and their semantic type, may be referred to explicitly using qualifiers. When it should not be distinguished between subject and object, the `arg` qualifiers may be used. The relation may also be referred to explicitly.

- `sub_name`: subject name
- `sub_semtype`: subject semantic type abbreviation
- `obj_name`: object name
- `obj_semtype`: object semantic type abbreviation
- `arg_name`: subject or object name
- `arg_semtype`: subject or object semantic type abbreviation
- `relation`: relation name

4 shows a fully qualified question making use of the qualifiers.

4. `sub_name:Chloroquine sub_semtype:phsu relation:TREATS obj_name:Malaria obj_semtype:dsyn`: Fully qualified question. `phsu` refers to the abbreviated semantic type *Pharmacologic Substance*, `dsyn` to *Disease or Syndrome*.

4.3.1. Closed discovery

Since SemBT uses SemRep for semantic predication extraction, the discovery patterns described in section 3.6 may be applied for exploration. To demonstrate the use of SemBT for closed discovery, we chose the *May Inhibit* discovery pattern described in section 3.6.2. Because SemBT does not allow for COVID-19 or related terms as arguments, we used *Chloroquine* and *Malaria* instead. For the question defining the XY

¹⁹http://semtb.mf.uni-lj.si/user_guide/qa_user_guide.html

²⁰http://semtb.mf.uni-lj.si/user_guide/SemBT_semantic_types.html

²¹http://semtb.mf.uni-lj.si/user_guide/SemBT_relation_types_and_instances_counts.html

Rank	Value	Count XY	Count YZ
1	TNF	4	4
2	Tumor Necrosis Factor-alpha	2	4
3	Antibodies	1	4
4	CD4	1	2
5	TNF gene	1	2
6	CD4 gene	1	2
7	cytokine	1	2
8	Proteins	1	2
9	Peptide Hydrolases	2	1
10	NOS2	1	1

Table 9. First ten linking terms generated by SemBT using closed discovery for the start term *Chloroquine* (semantic type *Organic Chemical*) and target term *Malaria* (semantic type *Disease or Syndrome*). The linking terms were limited to the semantic types *Amino Acid*, *Peptide*, or *Protein*, and *Gene or Genome*. The XY literature was restricted to INHIBITS relations, the YZ literature to CAUSES relations.

literature, we qualified *Chloroquine* as subject and INHIBITS as relation. As semantic type for *Chloroquine* we used *Organic Chemical* (orch), which includes drugs such as *Chloroquine*. The objects we limited to the semantic types *Amino Acid*, *Peptide*, or *Protein* (aapp), and *Gene or Genome* (gngm). For the question that defines the YZ literature, we set *Malaria* as object with semantic type *Disease or Syndrome* (dsyn). We specified CAUSES as relation and limited the subjects to *Amino Acids*, *Peptides*, or *Proteins* and *Genes or Genoms*. 1 shows the fully qualified question specifying the XY literature, 2 the one defining the YZ literature.

1. sub_name:Chloroquine sub_semtype:orch relation:INHIBITS
obj_semtype:(aapp OR gngm)
2. sub_semtype:(aapp OR gngm) relation:CAUSES obj_name:malaria
obj_semtype:dsyn

SemBT found 215 XY relations, 112 YZ relations, and 24 common Ys. The first ten Y terms are shown in table 4.3.1.

While our limited knowledge in biomedicine does not allow us to assess the results properly, the generated linking terms appear to be relevant for connecting *Chloroquine* to *Malaria*. Though it should be noted that tumor necrosis factor (TNF) and the CD4 gene are included redundantly in the list. The user may click any of the generated linking terms to inspect the articles that were involved in linking the start term to the target term. Other than BITOLA, SemBT presents the user the articles responsible for the linkage, instead of referring to a general PubMed search for the terms involved.

4.3.2. Open discovery

With SemBT, open discovery has to be performed in a different way than with BITOLA. While BITOLA allows the user to search for linking terms related to the start term, and subsequently to search for target terms related to one or more linking terms, SemBT requires the user to perform two separate searches. When searches performed with *Severe Acute Respiratory Syndrome* produced no meaningful results, we turned to *Chloroquine* and *Malaria* once again. To search for linking terms related to the start term *Malaria*, we

Subject	Sem Relation	Object	Frequency
cytokine	CAUSES	Malaria, Cerebral	12
cytokine	CAUSES	Malaria	10
Antibodies	CAUSES	Malaria	8
Tumor Necrosis Factor-alpha	CAUSES	Malaria	5
Tumor Necrosis Factor-alpha	CAUSES	Malaria	4
Intercellular adhesion molecule 1	CAUSES	Malaria, Cerebral	4
chemokine	CAUSES	Malaria, Cerebral	3
Genes	CAUSES	Malaria	3
Proteins	CAUSES	Malaria	3
TNF—TNF gene	CAUSES	Malaria	3

Table 10. First ten linking terms generated by SemBT using open discovery for the start term *Malaria* (semantic type *Disease or Syndrome*). The linking terms were limited to terms of the semantic types *Amino Acid*, *Peptide*, or *Protein*, and *Gene or Genome*, that are related to the start term through the CAUSES relation.

set *Malaria* as object with semantic type *Disease or Syndrome*. The subjects we limited to semantic types *Amino Acid*, *Peptide*, or *Protein*, and *Gene or Genome*. As relation we defined CAUSES. The full query is shown in 1.

```
1. sub_semtype:(aapp OR gngm) relation:CAUSES obj_name:Malaria
   obj_semtype:dsyn
```

The search generated 22 linking terms. Table 4.3.2 shows the ten linking terms ranked first.

We chose the linking term *Tumor Necrosis Factor-alpha* to search for target terms. To search for agents that may inhibit *Tumor Necrosis Factor-alpha*, we qualified it as object and set its semantic type to *Gene or Genome*. The subjects we limited to *Organic Chemicals*. As relation we specified INHIBITS. 1 shows the resulting question.

```
1. sub_semtype:orch relation:INHIBITS obj_name:Tumor Necrosis
   Factor-alpha obj_semtype:gngm
```

SemBT generated 239 target terms. As shown in table 4.3.2, *Chloroquine* showed up on the fourth place.

4.3.3. Discussion

While this admittedly is a constructed example, it nonetheless demonstrates SemBT's potential regarding the search for drug-disease-associations. However, the fact that COVID-19 and related terms are currently not allowed as arguments, limits the tools usability for searching COVID-19 treatments. Once this limitation is removed, we believe that SemBT could be a useful tool to search for COVID-19 treatments.

5. Conclusion

When it comes to LBD, we conclude that the use of semantic models and corresponding methods is the best approach for searching new COVID-19 treatment options. Other than

Subject	Sem Relation	Object	Frequency
Pentoxifylline	INHIBITS	TNF	32
Methotrexate	INHIBITS	TNF	9
Curcumin	INHIBITS	TNF	7
Chloroquine	INHIBITS	TNF	6
Ketamine	INHIBITS	TNF	5
vesnarinone	INHIBITS	TNF	4
Rolipram	INHIBITS	TNF	4
Aspirin	INHIBITS	TNF	4
triptolide	INHIBITS	TNF	4
Ethanol	INHIBITS	TNF	4

Table 11. First ten target terms generated by SemBT using open discovery for the linking term *Tumor Necrosis Factor-alpha* (here abbreviated TNF, semantic type *Gene or Genome*). The target terms were limited to terms of the semantic type *Organic Chemicals*, that are related to the linking term through the INHIBITS relation.

co-occurrence models, semantic models make use of semantic knowledge derived from research literature with semantic parsers. The use of semantic models becomes especially appealing in the biomedical domain, which provides several knowledge resources like the UMLS and MeSH, as well as semantic parsers for knowledge extraction, such as SemRep and BioMedLEE. Semantic models become even more useful with the use of discovery patterns. We compared three existing LBD systems, concerning their suitability for searching novel COVID-19 treatments, Arrowsmith, BITOLA, and SemBT. Although Arrowsmith is based on a co-occurrence model, we believe it is the best choice at the moment. This is due to the fact, that both BITOLA and SemBT currently do not recognize COVID-19 or related terms, which makes them basically useless for COVID-19 related searches. This limitation is probably caused by MeSH, which both BITOLA and SemBT use for restricting the allowed search terms. COVID-19 is currently classified as MeSH Supplementary Concept Data, which might not be recognized by BITOLA and SemBT. Once COVID-19 is classified as MeSH Descriptor Data, this limitation may disappear. If the restriction is removed, SemBT is the best choice in our opinion. SemBT uses a semantic model that incorporates SemRep for semantic knowledge extraction. This reduces the number of irrelevant terms generated by the system. SemBT's potential has been shown using the example of chloroquine and malaria. Both in open and closed discovery, the system identified relevant mechanisms of action involved in chloroquine treating malaria. The development of a new LBD systems seems unnecessary. The existing systems viability has been proven by new discoveries made and validated in the past, and the systems make good use of the knowledge resources available in the biomedical domain. Also, the development of a new LBD system would take a considerable period of time, which is sparse in the middle of a progressing pandemic. Future work should instead be focusing on searching new treatment options for COVID-19 using existing systems.

References

- [1] C. P. Adams and V. V. Brantner. Estimating the cost of new drug development: is it really 802 million dollars? *Health Aff (Millwood)*, 25(2):420–428, 2006. [DOI:10.1377/hlthaff.25.2.420] [PubMed:16522582].
- [2] M. L. Agostini, E. L. Andres, A. C. Sims, R. L. Graham, T. P. Sheahan, X. Lu, E. C. Smith, J. B. Case, J. Y. Feng, R. Jordan, A. S. Ray, T. Cihlar, D. Siegel, R. L. Macknam, M. O. Clarke, R. S. Baric, and M. R. Denison. Coronavirus Susceptibility to the Antiviral Remdesivir (GS-5734) Is Mediated by the Viral Polymerase and the Proofreading Exoribonuclease. *mBio*, 9(2), 03 2018. [PubMed Central:PMC5844999] [DOI:10.1128/mBio.00221-18] [PubMed:12966103].
- [3] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindfleisch. Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annu Symp Proc*, pages 6–10, Oct 2007. [PubMed Central:PMC2655783] [PubMed:11080015].
- [4] M. J. Ang, Q. Y. Lau, F. M. Ng, S. W. Then, A. Poulsen, Y. K. Cheong, Z. X. Ngho, Y. W. Tan, J. Peng, T. H. Keller, J. Hill, J. J. Chu, and C. S. Chia. Peptidomimetic ethyl propenoate covalent inhibitors of the enterovirus 71 3C protease: a P2-P4 study. *J Enzyme Inhib Med Chem*, 31(2):332–339, 2016. [DOI:10.3109/14756366.2015.1018245] [PubMed:25792507].
- [5] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32:D267–270, Jan 2004. [PubMed Central:PMC308795] [DOI:10.1093/nar/gkh061] [PubMed:10842744].
- [6] M. S. Boguski, K. D. Mandl, and V. P. Sukhatme. Drug discovery. Repurposing with a difference. *Science*, 324(5933):1394–1395, Jun 2009. [DOI:10.1126/science.1169920] [PubMed:19520944].
- [7] A. J. Brown, J. J. Won, R. L. Graham, K. H. Dinnon, A. C. Sims, J. Y. Feng, T. Cihlar, M. R. Denison, R. S. Baric, and T. P. Sheahan. Broad spectrum antiviral remdesivir inhibits human endemic and zoonotic deltacoronaviruses with a highly divergent RNA dependent RNA polymerase. *Antiviral Res.*, 169:104541, 09 2019. [PubMed Central:PMC6699884] [DOI:10.1016/j.antiviral.2019.104541] [PubMed:26733065].
- [8] P. Bruza, R. Cole, D. Song, and Z. Bari. Towards operational abduction from a cognitive perspective. *Logic Journal of the IGPL*, 14(2):161–177, 2006.
- [9] Y. W. Chen, C. B. Yiu, and K. Y. Wong. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CL pro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Res*, 9:129, 2020. [PubMed Central:PMC7062204.2] [DOI:10.12688/f1000research.22457.2] [PubMed:32173287].
- [10] C. P. Chuck, C. Chen, Z. Ke, D. C. Wan, H. F. Chow, and K. B. Wong. Design, synthesis and crystallographic analysis of nitrile-based broad-spectrum peptidomimetic inhibitors for coronavirus 3C-like proteases. *Eur J Med Chem*, 59:1–6, Jan 2013. [PubMed Central:PMC7115530] [DOI:10.1016/j.ejmech.2012.10.053] [PubMed:16219322].
- [11] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. The price of innovation: new estimates of drug development costs. *J Health Econ*, 22(2):151–185, Mar 2003. [DOI:10.1016/S0167-6296(02)00126-1] [PubMed:12606142].
- [12] C. J. Gordon, E. P. Tchesnokov, J. Y. Feng, D. P. Porter, and M. Götte. The antiviral compound remdesivir potently inhibits RNA-dependent RNA polymerase from Middle East respiratory syndrome coronavirus. *J. Biol. Chem.*, 295(15):4773–4779, 04 2020. [PubMed Central:PMC7152756] [DOI:10.1074/jbc.AC120.013056] [PubMed:31924756].
- [13] Sam Henry and Bridget T. McInnes. Literature based discovery: Models, methods, and trends. *Journal of Biomedical Informatics*, 74:20 – 32, 2017.
- [14] D. Hristovski, C. Friedman, T. C. Rindfleisch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. *AMIA Annu Symp Proc*, pages 349–353, 2006. [PubMed Central:PMC1839258] [PubMed:11080015].
- [15] D. Hristovski, B. Peterlin, J. A. Mitchell, and S. M. Humphrey. Improving literature based discovery support by genetic knowledge integration. *Stud Health Technol Inform*, 95:68–73, 2003. [PubMed:14663965].
- [16] X. Hu, X. Zhang, I. Yoo, and Y. Zhang. A semantic approach for mining hidden links from complementary and non-interactive biomedical literature. volume 2006, pages 200–209, 2006.
- [17] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosembat, and T. C. Rindfleisch. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, 28(23):3158–3160, Dec 2012. [PubMed Central:PMC3509487] [DOI:10.1093/bioinformatics/bts591] [PubMed:14728234].

- [18] E. Konkolova, M. Dejmeek, H. Hřebabecký, M. Šála, J. Böserle, R. Nencka, and E. Boura. Remdesivir triphosphate can efficiently inhibit the RNA-dependent RNA polymerase from various flaviviruses. *Antiviral Res.*, 182:104899, Aug 2020. [PubMed Central:PMC7403104] [DOI:10.1016/j.antiviral.2020.104899] [PubMed:32763313].
- [19] Y. Lussier, T. Borlawsky, D. Rappaport, Y. Liu, and C. Friedman. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*, pages 64–75, 2006. [PubMed Central:PMC2906243] [PubMed:10802651].
- [20] D. L. McKee, A. Sternberg, U. Stange, S. Laufer, and C. Naujokat. Candidate drugs against SARS-CoV-2 and COVID-19. *Pharmacol Res*, 157:104859, 07 2020. [PubMed Central:PMC7189851] [DOI:10.1016/j.phrs.2020.104859] [PubMed:32360480].
- [21] Wanda Pratt and Meliha Yetisgen-Yildiz. Litlinker: Capturing connections across the biomedical literature. In *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP '03*, page 105–112, New York, NY, USA, 2003. Association for Computing Machinery.
- [22] J. Preiss, M. Stevenson, and R. Gaizauskas. Exploring relation types for literature-based discovery. *J Am Med Inform Assoc*, 22(5):987–992, Sep 2015. [PubMed Central:PMC4986660] [DOI:10.1093/jamia/ocv002] [PubMed:19124086].
- [23] Judita Preiss. Seeking informativeness in literature based discovery. In *Proceedings of BioNLP 2014*, pages 112–117, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
- [24] T. C. Rindfleisch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J Biomed Inform*, 36(6):462–477, Dec 2003. [DOI:10.1016/j.jbi.2003.11.003] [PubMed:14759819].
- [25] C. Scavone, S. Brusco, M. Bertini, L. Sportiello, C. Rafaniello, A. Zoccoli, L. Berrino, G. Racagni, F. Rossi, and A. Capuano. Current pharmacological treatments for COVID-19: What's next? *Br. J. Pharmacol.*, Apr 2020. [PubMed Central:PMC7264618] [DOI:10.1111/bph.15072] [PubMed:32214286].
- [26] N. R. Smalheiser, V. I. Torvik, and W. Zhou. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Comput Methods Programs Biomed*, 94(2):190–197, May 2009. [PubMed Central:PMC2693227] [DOI:10.1016/j.cmpb.2008.12.006] [PubMed:16507357].
- [27] S. E. St John, M. D. Therkelsen, P. R. Nyalapatla, H. L. Osswald, A. K. Ghosh, and A. D. Mesecar. X-ray structure and inhibition of the feline infectious peritonitis virus 3C-like protease: Structural implications for drug design. *Bioorg. Med. Chem. Lett.*, 25(22):5072–5077, Nov 2015. [PubMed Central:PMC5896745] [DOI:10.1016/j.bmcl.2015.10.023] [PubMed:9527924].
- [28] D. R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.*, 30(1):7–18, 1986. [DOI:10.1353/pbm.1986.0087] [PubMed:3797213].
- [29] D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspect. Biol. Med.*, 31(4):526–557, 1988. [DOI:10.1353/pbm.1988.0009] [PubMed:3075738].
- [30] Don R. Swanson and Neil R. Smalheiser. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183 – 203, 1997. Scientific Discovery.
- [31] Don R. Swanson, Neil R. Smalheiser, and Vette I. Torvik. Ranking indirect connections in literature-based discovery: The role of medical subject headings: Research articles. *J. Am. Soc. Inf. Sci. Technol.*, 57(11):1427–1439, September 2006.
- [32] Michael Symonds, Peter Bruza, and Laurianne Sitbon. The efficiency of corpus-based distributional models for literature-based discovery on large data sets. In *Proceedings of the Second Australasian Web Conference - Volume 155, AWC '14*, page 49–57, AUS, 2014. Australian Computer Society, Inc.
- [33] E. P. Tchesnokov, J. Y. Feng, D. P. Porter, and M. Götte. Mechanism of Inhibition of Ebola Virus RNA-Dependent RNA Polymerase by Remdesivir. *Viruses*, 11(4), 04 2019. [PubMed Central:PMC6520719] [DOI:10.3390/v11040326] [PubMed:24590073].
- [34] Menasha Thilakarante, Katrina Falkner, and Thushari Atapattu. A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Comput. Surv.*, 52(6), December 2019.
- [35] V. I. Torvik and N. R. Smalheiser. A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics*, 23(13):1658–1665, Jul 2007. [DOI:10.1093/bioinformatics/btm161] [PubMed:17463015].
- [36] M. Yetisgen-Yildiz and W. Pratt. A new evaluation methodology for literature-based discovery systems. *J Biomed Inform*, 42(4):633–643, Aug 2009. [DOI:10.1016/j.jbi.2008.12.001] [PubMed:19124086].
- [37] Meliha Yetisgen-Yildiz. Litlinker : A system for searching potential discoveries in biomedical literature.

January 2021

2006.

- [38] Y. Zhai, Y. Ma, F. Ma, Q. Nie, X. Ren, Y. Wang, L. Shang, and Z. Yin. Structure-activity relationship study of peptidomimetic aldehydes as enterovirus 71 3C protease inhibitors. *Eur J Med Chem*, 124:559–573, Nov 2016. [DOI:10.1016/j.ejmech.2016.08.064] [PubMed:27614190].

Exploring SPA Semantic Computing Method for Heating Network Leaks Monitoring

Ilia TRIAPITCIN^a and Ajantha DAHANAYAKE^a

^a *Lappeenranta-Lahti University of Technology Lappeenranta,
Finland, Yliopistonkatu 34
ilia.triapitcin@student.lut.fi,
ajantha.dahanayake@lut.fi*

Abstract. Modern district heating (DH) systems are complex engineering structures that play an essential role in large city infrastructures. DH networks have many sensors, nodes, and methods for monitoring the status of the DH network. Sensing, processing, analytical actuation (SPA) of incoming information handled by the SPA semantic Computing method can be applied to similar problems. The SPA Semantic Computing method searches for correlations between the sets of incoming data and to identify the correct scenario to respond to events. This article explores the integration of SPA functions to analyze multivariate sensing data, including data from multivariable sensors and infrared images, for creating a monitoring system for DH networks. The focus is to assess whether the SPA approach is a suitable candidate to use to monitor the emergency events of the DH network. Specific target data for the assessment are [1] multi-parameter DH network sensor data, such as water temperature, sweat rate, energy delivered, etc., and [2] infrared image data from a camera mounted on the unmanned aerial vehicle (UAV) for monitoring the location of the underground DH network leaks. A multivariate computational model, a mathematical model of meaning (MMM), and a spatial image filtering method are proposed for integrating SPA semantic computing for emergency leak detection in DH networks.

Keywords. district heating, geographical information system, sensor data, differential computing, semantic computing.

1. Introduction

District heating (DH) or remote heating is a method for distributing heat generated in a heat center (HC) through insulated pipes for individual and commercial heating needs such as space heating and water heating. Unfortunately, water leaks from the district heating network are disastrous during Nordic winters. Therefore, finding leaks as soon as possible is necessary to reduce any unwanted accidents.

Modern DH networks are complex engineering structures monitored by hundreds of sensors in real-time. In addition, there are many different analysis methods for localizing accidents in the DH system [1–3]. Therefore, many highly qualified staffs are engaged to analyze all the incoming information in the shortest possible time. This study explores the possibility of introducing sensing, processing, analyzing, and actuation (SPA) semantic computing method [4] to automate the DH network states monitoring and leak detection. As a result, it helps reduce the number of company staff needed for constant manual monitoring and, in turn, increases the response time during an emergency.

This article presents how the industry can use the SPA semantic computing method can be used to analyze the DH networks' multivariate sensor data and thermal images around the surface of heating water pipelines for monitoring the system. This research suggests that integrating multimedia data and real-time sensor data analysis improves DH network analysis, facilitating emergency activity monitoring. Specific target data of the DH network are:

1. DH network sensor data, such as the incoming and outgoing water temperature, the movement of the incoming and outgoing flow, date of DH network construction, sensors' alarm data, water pressure inside tubes, and
2. the image data in the infrared range was obtained using a camera installed on the unmanned aerial vehicle (UAV).

The objectives of this study are to explore whether it is possible:

1. to analyze and generalize the spatial dynamics of the thermal data of the soil surface in the pipe-laying zone of the DH network, and
2. to integrate the semantic sequencing function and the spatial dynamics of the DH network data to assess the physical state of the DH network.

The research approach bases on the design science framework [4]. This research follows the design science research guidelines developed by Hevner et al. [4] for solving a relevant problem of the research field by creating an artifact to solve the problem while using the existing body of knowledge to arrive at an innovative solution. Finally, the artifact is validated for its relevance for the application domain. It extends the current scientific knowledge base with the new knowledge formulated for problem-solving in the environment of the research field. This process is illustrated in figure 1.

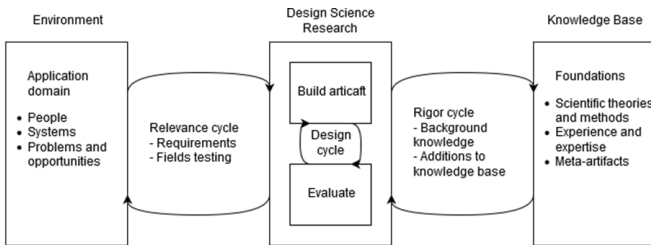


Figure 1. Design Science Research development life cycle.

The knowledge used from the body of knowledge is the SPA semantic computing concepts and method and the image processing and filtering methods. This knowledge foundation is used to explore an innovative solution for the DH network leak monitoring approach.

This article presents an overview of the five-dimensional monitoring and data analysis system for DH networks, and an overview of the system concept is provided and presented in Section 2. The method of the proposed detection, processing, and actuation functions in real-time is described in Section 3 while evaluating the proposed functions in Section 4. Finally, section 5 gives the conclusions together with further research directions.

2. Overview of the Theoretical Foundation

In Y. Kiyoki et al. [5], the "5D World Map System" framework is a multivisualized and flexible information retrieval system used for environmental analysis and semantic computation. This framework defines the system as made up of five dimensions. The first dimension is the temporal dimension, then the second, third and fourth are spatial dimensions. The last one is the semantic dimension reflecting a large-scale and multi-dimensional semantic space based on the associative semantic computing system. The semantic dimension stores data from resources to correlate temporal and spatial areas. It implements the 5D world map to dynamically build Spatio-temporal and multiple semantic representations for diverse media resources.

The parameters of the DH network are treated as the fifth dimension in this article. This fifth dimension is depicted as a multi-dimensional space based on a mathematical model of meaning (MMM). MMM uses a semantic associative search approach for defining the concepts of "semantics" and "impressions" based on the "context" of infrared image tools. Figure 1 depicts the systems architecture for monitoring the condition of DH networks based on the design described in Y. Kiyoki et al. [5][6].

The DH network condition monitoring system is explored based on the "Sensing, Processing and Analytical Actuation" concept (SPA) [5][7][8]. "SPA" is efficient and beneficial in defining environmental phenomena in real space for real data tools. They map them to cyber-physical space to build analytical and semantic computation and visualize the analytically computed findings to the real space to convey environmental phenomena with causalities. In this study, the SPA definition is extended to semantic computation (see Figure 2). This system tries to reproduce human steps to handle DH systems' monitoring conditions and automate the monitoring process.

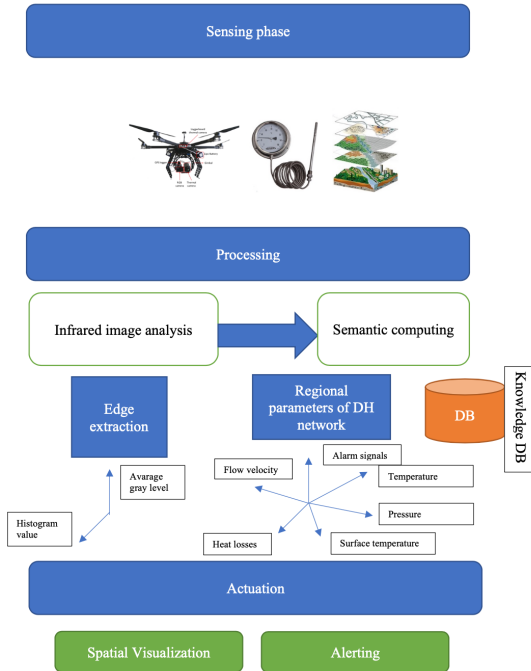


Figure 2. Concept of the system.

3. District Heating Measurements Analysis

Semantic computing is based on semantic in terms of context, meaning, or intention. "SPA" is a foundational framework for implementing an environmental system of three essential functions, "Sensing, Processing and Analytical Actuation," to develop Cyber-Physical integration and industrial engineering information systems. SPA helps define industrial phenomena in a physical space as real data tools, mapping them to cyber-space to build analytical and semantic computation, and visualize the analytically computed findings to the real space to express industrial phenomena with causalities [5]. Collected sensing data make the analysis space for finding correlations between different values. There are two other analysis spaces created in this research. The first space contains DH measurements from digital sensors inside the DH network. The second space is from digital thermal images collected by UAV. Sensing data and analysis spaces are explained below.

3.1. Multi-Parameter Sensing Data

The DH network's stability is controlled by many parameters and is critical for the stability of the infrastructure. The multi-dimensional space for expressing the DH system's parameters for monitoring the DH network is shown in Figure 3 [9,10].

- *Air temperature axis:* shows outside air temperature.
- *Sensor's identification axis:* is a unique number of a sensor in the DH network.
- *Water temperature axis:* shows the temperature of the water inside the DH network.
- *Alarm indicator:* shows emergency in the DH network branch
- *The pressure inside tubes:* shows water pressure inside the DH network pipe
- *Flow velocity:* parameters of the pressure loss of the water
- *Customer needs:* the amount of heat consumed by customers
- *Building date:* date is when the DH network or branch was built or renovated

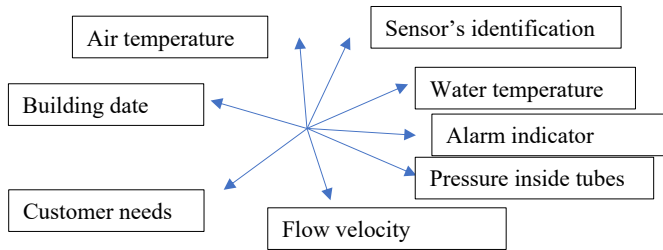


Figure 3. The multi-dimensional DH quality parameters.

The correlation between outside air temperature and input/output water temperature is presented in Figure 4. Also, the green line is the correlation between outside and inside temperature.

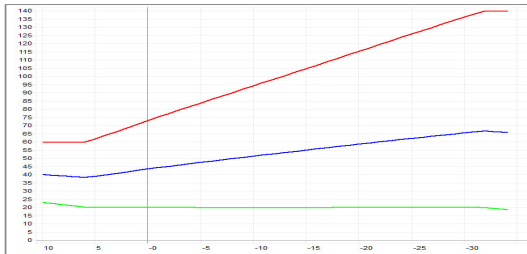


Figure 4. The temperature graph show correlation between output water temperature (ordinate) and air temperature (abscissa): the red line is the input temperature of water; the blue line is output temperature of water; the green line is air temperature inside a customer building;[11].

The list of parameters is created at the design and modeling stage before constructing the industrial facility. However, in reality, the DH network is monitored by a lot of parameters. Therefore, a data vector is created of multiple DH parameters. The data vector contains a value of air temperature outside, inside customer's building, sensor's unique identification, alarm indicator, the water temperature on input and output of a pipe, the pressure inside a pipe, the velocity of the flow, total customer needs in power, building or renovation date. The data vector structure is presented in Table 1. These data are collected in real-time and stored in a database.

Table 1. Data vector of DH parameters

Sensor Id	Temperature, °C	Pressure, MPa ($10^6 \frac{N}{m^2}$)	...	Velocity, $\frac{m^3}{h}$	Needs, <i>GWt</i>
V_{i1}	V_{i1}	V_{j1}	...	V_{k1}	V_{m1}
V_{i1}	V_{i2}	V_{j2}	...	V_{k2}	V_{m2}
...
V_{i1}	V_{ik}	V_{jk}	...	V_{kn}	V_{mn}

3.2. Spatial District Heating Network Data

Information about spatial DH networks is stored in the geographical information system. This data includes the spatial position of objects and all physical and technical parameters of the entities. The DH network example is shown in Figure 5.

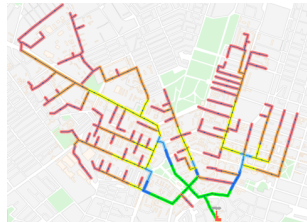


Figure 5. Spatial presentation of the DH network [11].

3.3. Infrared Images

The second space is for sensing images. UAV drone collects thermal images of the surface above the DH network. The collected images are handled during the preprocessing phase. This phase aims to find thermal anomalies on the soil surface associated with the DH network heat losses and water leaks.

3.3.1. Image Preprocessing

The original thermogram contains the absolute temperature values encoded in RGB palettes. It means that color code #00000 from the RGB palette corresponds to 0 °K or -273.16 °C. In reality, the thermogram of the Earth's surface in a settlement contains a temperature difference in the range of several tens of degrees. The minimum temperature value should be set arbitrarily to improve the visualization of gradients. The criterion for

choosing a color is the distance from the opposite value. The optimal choice is blue (color code #0000FF in the RGB palette) for the coldest area and red for the hottest site (color code #FF0000 in the RGB palette). The original thermogram as a result of applying the filter is shown in Figure 6.

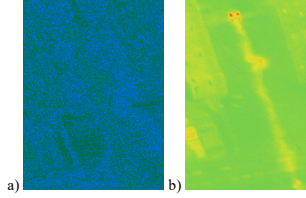


Figure 6. Selection of warm and cold areas: a) the original image; b) with the use of a filter; [11].

The feature extraction process from thermal images is based on the contrast enhancement of the image method proposed by Chiwu Bu et al. [12]. Gray level changes are used to describe the picture corresponding to the derivative of the picture's two-dimensional function. The highest difference in gray image value is the local peak value of the first derivative and the zero point of the two order derivative. Contrast enhancement is performed on an infrared picture using the Top and Bottom-Hat transforms of Chiwu Bu et al. [12]:

$$HAT(A) = A - (A \circ B) \quad (1)$$

$$HAT'(A) = (A \cdot B) - A \quad (2)$$

$$f \circ g = (f \ominus g) \oplus g \quad (3)$$

$$f \cdot g = -[(-f) \circ (-g)] \quad (4)$$

$$HAT(A) = A - (A \circ B) = A - (A \ominus B) \oplus g \quad (5)$$

$$HAT'(A) = (A \cdot B) - A = -A - [(-A) \circ B] = HAT(-A) \quad (6)$$

Where:

- A is the infrared image;
- B is the thermal anomaly element on the infrared image;
- f is the gray value function of pixels x and y ;
- g is the filter function of pixels x and y ;
- $f \circ g$ is the opening operation of the gray value;
- $f \cdot g$ is the closing operation of the gray value;
- $HAT(A)$ is the Top-Hat transformation;
- $HAT'(A)$ is the Bottom-Hat transformation;

The Top-Hat transform can identify relatively bright things against a dark backdrop, whereas the Bottom-Hat transform can identify somewhat unclear things against a light

background. As a result, pixel locations of hot areas in an infrared picture may be recognized using the Top-Hat, and Bottom-Hat transforms [11]. This research looks into this method to localize anomalies and then calculate spatial coordinates of the thermal anomalies. Processing results are shown in Figure 7. Because edge detection is sensitive to picture noise, the first step is to eliminate the picture's noise with a Gaussian filter. Chiwu Bu et al. [11] use the 5x5 parameter for the Gaussian filter. This research is checked by using 3x3, 5x5, and 8x8 parameters for the Gaussian filter. The optimal value between noise and data loss is the 5x5 parameter. The second step after blurring is to apply the watershed algorithm [13].

The watershed algorithm is a type of picture segmentation method based on morphology theory. Grayscale levels are presented in Figure 7a. The picture following the Top-Hat transform is displayed in Figure 7b, in which brighter items in the black backdrop falling can be measured, and the size can be approximated using the average radius of the measured objects in the picture. The picture following the Bottom-Hat transform is shown in Figure 7c. The merge of Top-Hat and Bottom-Hat creates full anomaly size for opening operation. The result of the merge is presented in Figure 7d. In the end, it requires applying a grey value opening operation to create a binary bitmap for the subsequent edge detection [14].

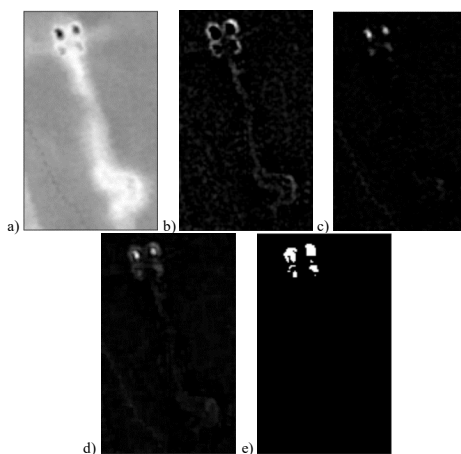


Figure 7. Processing result: a) grayscale change; b) Top-Hat transform; c) Black-Hat transform; c) gray value opening operation; d) Top-Hat transform + Black-Hat transform; e) gray value opening operation; [11].

3.3.2. Object Edge Detection

The characteristic knowledge of the deficient area is obtained after segmentation of the Watershed. It is then required to classify candidates' edges using a popular edge detection algorithm- the Canny Edge Detection [12], as seen in Figure 8b. As a consequence of the processing step, a significant amount of worthless and untrue boundary data can be minimized, increasing the leakage extraction function effect: Figures 8a and 8b show the

approach proposed in this explorative research for filtering relevant data. Figures 8c and Figure 8d show the classical method of filtering.

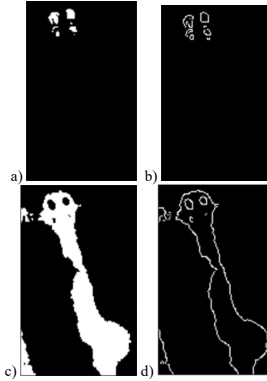


Figure 8. Result of edge detection: a) thermal image with gray value opening operation; b) edge detection by Canny operator in the thermal image with gray value opening operation; c) classic approach of opening operation with basic thresholding; d) edge detection by Canny operator in the thermal image with basic thresholding; [11].

3.3.3. Leakage Position Extraction

As a result of this phase, the thermal data is represented as spatial objects. The detected object on the thermal image is assigned to one of three categories: water leakage, heat leakage, or unrelated. The water leakage object means that the DH network has an emergency. The heat leakage implies that part of the DH network has a thermal insulation issue. This issue is not critical and can be postponed. A not related type of problem means that other reasons cause the thermal anomaly. It is worth noting that some of the intensely glowing objects can also be detected as leaks. As a result of this step, the thermal data is presented in the form of spatial coordinates. Next, it is necessary to calculate the expected place of the leak to carry out a semantic search in the future. It is used the algorithm for finding the center of the blob (centroid). The centroid of the shape is the arithmetic mean of all points in the shape of the figure has n different points with coordinates $x_1 \dots x_n$ and $y_1 \dots y_n$, then the centroid $c(x, y)$ is given as

$$c_x = \frac{1}{n} \sum_{i=1}^n x_i, c_y = \frac{1}{n} \sum_{i=1}^n y_i \quad (7)$$

Each form in image processing and computer vision is built up of pixels, and the centroid is simply the weighted average of all the pixels that comprise the shape. After the centroid is calculated, most of the incorrectly identified areas are removed. However, some large, unrelated objects can still be preserved as heat anomalies. Spatial Data on the DH network is used for checking the anomaly's intersection with the network to remove heat anomalies not related to the DH network or permissible heat losses at

engineering facilities. An algorithm (**Algorithm 1**) for eliminating thermal noise from a heat map for an area of underground DH networks is described below.

Algorithm 1: Filtering relevant data (X, C)

Input: Dataset $X \in \mathbb{R}^{d \times n}$,
Dataset $C \in \mathbb{R}^{d \times n}$

Output: $X^* \in \mathbb{R}^{d \times n}$

$S_i(x, y) \leftarrow$ Find shapes on a picture;

REPEAT i TIMES:
 FOR OBJECT $S_i \in X$:
 REPEAT j TIMES:
 FOR OBJECT $L_j \in C$:
 $K(S_i, L_j) \leftarrow$ Find intersection of the centroid with DHN;
 $x_i^* \leftarrow$ Remove S_i where $K(S_i, L_j) = \emptyset$, $x_i^* \in X^*$;

4. Semantic Space for Calculation

The research work of S. Sasaki et al. [6] use semantic range as simple fuzzy logic in semantic computation to define a range of semantics according to each standard. Each element of the DH network is assessed according to the range of parameter values [15]. Those values are interpreted as an emergency, problematic and good. For example, if there is a hole in the pipeline in a network section, the DH network goes into an emergency state. The DH network becomes a problematic state when the DH network doesn't have water leakages but presents thermal insulation problems of heat leaks. If a section or the entire DH network sensor values have acceptable values, such a section or DH network has a good condition. The number of states can be expanded depending on the events needing to be monitored. For the DH network state evaluation method, these influences are defined as the semantics of the DH network state.

The most important feature of the method is that the system provides an interpreter [1] to calculate the quality level and convert influences/meanings into a sentence or a set of words that even non-specialists or ordinary people can understand. Secondly, [2] implement semantic calculations to target the values of multi-parameter sensing. Thus, the semantic space for calculating relationships or correlations between each parameter's values and the semantics is presented in Figure 9.

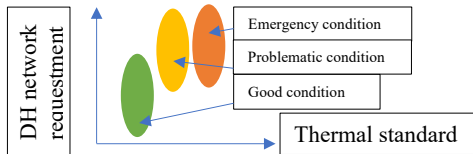


Figure 9. Semantic space for the DH measurement analysis.

Sensing Measurements Range Calculation and Priority Factor

In reality, sensing measurement semantics have a higher priority than infrared image semantics. In the absence of emergency readings from the sensors, the system cannot go into an emergency state even if thermal anomalies are present in the infrared images. Therefore, a priority factor k is introduced to add weight to the calculation. S. Sasaki et al. [6] apply simple Fuzzy logic to defining the range of semantics according to each standard called Semantic Range (SR). This research also uses the same solution. The minimum value of a parameter in measurement is defined as $SR-min$, the maximum value is defined as $SR-max$, and the $SR-mean$ is defined as a mean value. For each parameter value, the measured values' total point in the fuzzy-set interval is determined as the sum of each weight [6]. Weight uses 0 value to excluded un-used measurement and 1 to include measurement in the semantic calculation.

Table 2. Definition of fuzzy-set interval and the membership function

Definition	Weight
$SR-min$	0
$SR-mean$	1
$SR-max$	0

5. Conclusion

This paper describes how the industry can use semantic computing SPA to identify water leaks using multi-parameter sensors and thermal images of the ground surface near DH pipelines to monitor emergencies in the system. The paper explores the integration of SPA functions to analyze multi-dimensional data to create a monitoring system for DH networks. It is assessed that the SPA approach is a suitable candidate to be used for monitoring emergency events in the DH network. It is found that during the research process, several dozen conditions are applied to verify the problem identification conditions. In this case, it is possible to use the knowledge of the correct condition as a condition for verification. It would be good to separate the knowledge of the correct condition and observe the condition. Thus, it would give fast state monitoring without changing the correct state knowledge. The main feature of the proposed approach is the integration of sensor data semantics and thermal anomaly search from infrared image databases to identify the leak location. This research aims to implement a search for the coordinates of the problem based on sensor and thermal image data and create a list of problems based on the search results.

References

- [1] Lah AAA, Dziyauddin RA, Md Yusoff N. Localization Techniques For Water Pipeline Leakages: A Review. In: 2018 2nd International Conference on Telematics and Future Generation Networks (TAFGEN) [Internet]. IEEE; 2018. p. 49–54. Available from: <https://ieeexplore.ieee.org/document/8580467/>
- [2] Zhou S, O'Neill Z, O'Neill C. A review of leakage detection methods for district heating networks. Appl Therm Eng [Internet]. 2018 Jun;137:567–74. Available from:

<https://linkinghub.elsevier.com/retrieve/pii/S1359431118301169>

- [3] Chan TK, Chin CS, Zhong X. Review of Current Technologies and Proposed Intelligent Methodologies for Water Distributed Network Leakage Detection. *IEEE Access* [Internet]. 2018;6:78846–67. Available from: <https://ieeexplore.ieee.org/document/8565861/>
- [4] Hevner, March, Park, Ram. Design Science in Information Systems Research. *MIS Q* [Internet]. 2004;28(1):75. Available from: <https://www.jstor.org/stable/10.2307/25148625>
- [5] Kiyoki Y, Chen X, Veesommai C, Rachmawan IE, Chawakitchareon P. A SPA-Based Semantic Computing System for Global & Environmental Analysis and Visualization with “5-Dimensional World-Map”:“Towards Environmental Artificial Intelligence.” *Inf Model Knowl Bases XXXI*. 2020;321:285.
- [6] Sasaki S, Kiyoki Y. Analytical Visualization Function of 5D World Map System for Multi-Dimensional Sensing Data. *Inf Model Knowl Bases*. 2018;29:71–89.
- [7] Rungsupa S, Chawakitchareon P, Hansuebsai A, Sasaki S, Kiyoki Y. Photographic assessment of coral stress: Effect of low salinity to *Acropora* sp. *Goniopora* sp. and *Pavona* sp. at Sichang Island, Thailand. *Inf Model Knowl Bases XXIX*. 2018;301:137–48.
- [8] Kiyoki Y, Chen X, Veesommai C, Sasaki S, Uraki A, Koopipat C, Chawakitchareon P, Hansuebsai A. An Environmental-Semantic Computing System for Coral-Analysis in Water-Quality and Multi-Spectral Image Spaces with” Multi-Dimensional World Map. *Inf Model Knowl Bases XXIX*. 2018;301:52–70.
- [9] Nemchenko VI, Zheltukhin AA. The System Analysis Of Regulation Of Thermal Loading And Increase Of Efficiency Heat Supply Of Micro-area Of Samara. *Constr Archit* [Internet]. 2010;(7):172–9. Available from: <https://cyberleninka.ru/article/n/sistemnyy-analiz-regulirovaniya-teplovoy-nagruzki-i-povysheniye-effektivnosti-teplosnabzheniya-mikrorayona-g-samary>
- [10] Kudinov I V., Kolesnikov S V., Eremin A V., Branfileva AN. Computer models of complex multiloop branched pipeline systems. *Therm Eng* [Internet]. 2013 Nov 11;60(11):835–40. Available from: <http://link.springer.com/10.1134/S0040601513080053>
- [11] Politerm, OOO [Internet]. 2021 [cited 2021 Jul 25]. Available from: <https://www.politerm.com/>
- [12] Bu C, Sun Z, Tang Q, Liu Y, Mei C. Thermography Sequence Processing and Defect Edge Identification of TBC Structure Debonding Defects Detection Using Long-Pulsed Infrared Wave Non-Destructive Testing Technology. *Russ J Nondestruct Test* [Internet]. 2019 Jan 26;55(1):80–7. Available from: <http://link.springer.com/10.1134/S1061830919010030>
- [13] Moga AN, Gabbouj M. Parallel image component labelling with watershed transformation. *IEEE Trans Pattern Anal Mach Intell* [Internet]. 1997 May;19(5):441–50. Available from: <http://ieeexplore.ieee.org/document/589204/>
- [14] Peters JF. Foundations of Computer Vision: Computational Geometry, Visual Image Structures and Object Shape Detection. *Foundations of Computer Vision*. Cham: Springer International Publishing AG; 2017. (Intelligent systems reference library; vol. 124).
- [15] Kiyoki Y, Kitagawa T, Hayama T. A metadata system for semantic image search by a mathematical model of meaning. *ACM SIGMOD Rec* [Internet]. 1994 Dec;23(4):34–41. Available from: <https://dl.acm.org/doi/10.1145/190627.190639>

Human digital twins in acquiring information about human mental processes for cognitive mimetics

Pertti Saariluoma, Antero Karvonen & Lotta Sorsamäki

University of Jyväskylä, Finland

Technical Research Centre of Finland Ltd. (VTT), Finland

Abstract: Modern information technology makes it possible to redesign the ways people work. In the future, machines can carry out intelligence-requiring tasks, which previously were done by people. It is thus good to develop methodologies for designing intelligent systems. An example of such methods is cognitive mimetics, i.e. imitating human information processing. Today, machines cannot by themselves navigate in archipelagos. However, the fact that people can take care of ship steering and navigation means that there is an information process, which makes it possible to navigate ships. This information process takes place inside the minds of navigating people. If we are able to explicate the information processing in the navigator's mind, the knowledge of it can be used in designing intelligent machines.

Replicating physical objects and industrial processes by means of digital computers is called digital twinning. Digital twins (DTs), which are digital replicas of physical systems and processes, have recently become tools for working with complex industrial processes. A crucial question for DTs is should human actions be added to them? As the answer is positive, such models of human information processing can be called human digital twins (HDTs).

The knowledge of human tacit and explicit information processes can be represented by human digital twins. Models can be used in the search for a deeper understanding of human intelligent information processes. Human digital twins can thus be used as methodological tools in cognitive mimetics. In our present study, we modeled paper machine operators' thinking. Specifically, we developed an ideal-exception-correction (IEC) model for paper operators' control logic. The model illustrates how research and HDT-modeling can be used for explicating the subconscious or tacit information processing of people.

Introduction: developing intelligent technologies

The emergence of intelligent technologies is becoming a sign of our time. Modern information technology makes it possible to redesign the ways of human work and carry out tasks people that used to do. Typical examples of innovative solutions can be found – for example, in transportation, in healthcare, as well as in industrial, legal, and administrative information processing (Ford 2015, Fukuda 2020, Tegmark 2017). People can be freed from many routine tasks, if only it is possible to innovate, design, and develop solutions taking care of human functional roles. The emergence of intelligent societies and artificial intelligence (AI) technologies is a central challenge of our times. Therefore, it is essential to develop methodological practices to enable designers to create working artefacts taking care of human roles in practical tasks.

Machines have always been designed, developed, and manufactured to enable people to reach their action goals (Bernal 1969; SaariLuoma, Cañas and Leikas 2016). People have their needs, which motivate their actions. In the beginning, there were only simple tools, such as hand axes or spears. However, by means of elementary technologies, it was possible to construct huge buildings, from palaces to pyramids (Bernal 1969). After the development of precision work and steam energy, it was possible to begin industrialization, which used mechanical machines and later electromechanical systems. Finally, electromechanical technologies led to information technology with the rise of computing as the latest game-changer (Minsky 1967, Turing 1936-7, Waldrop 2001).

Industrialism changed the way people worked and lived. There is no doubt that information technology has already changed and will again change how people satisfy their needs and carry out work processes. One essential property of computational thinking is the possibility to create intelligent technologies, which can carry out intelligence-demanding tasks (Engelbrecht 2007, Russell and Norvig 1995, Turing 1950). Traditionally, people have had to operate machines as there are numerous control decisions to be made, which could not be done by machines. Automatization, AI, autonomic technologies, and machine intelligence will free people from many present jobs, as they already have done. However, before it is possible to replace people in intelligence-demanding tasks with machines, it will be necessary to innovate, design, develop, and manufacture technological solutions capable of taking care of these tasks.

It is important to think about the foundations of how to design intelligent technological solutions. Here, we will begin with a simple idea. If people today are able to carry out some intelligence-demanding tasks, this knowledge could be used to solve the problems of designing intelligent technologies for various contexts. If people can navigate ships through complex archipelagos, it is clear that there exists an information process that can carry out such a task. Logically, designers should be able to use the knowledge of the present information process to develop intelligent machines to take care of the same task.

In cognitive research, the idea that the same information process can be carried out by different physical entities is called “multiple realizability” (Bickle 2020). For example, people are able to do mental calculation; however, pocket calculators can also effectively realize the processes of calculation. The focus or research should not thus be on the physical entities realizing information processes, but the information processes themselves. Information processes provide an independent level of conceptualization. The problem transforms into how designers can best use the knowledge of human information processes in creating intelligent technologies.

Content-based thinking

It is possible to realize information processes in different physical systems from humans and animals to computers. The phenomenon has been termed “multiple realizability” (Bickle 2020). The use of multiple realizability in designing intelligent systems is possible, if one knows how people process information in some particular tasks. A necessary presupposition for constructing a pocket calculator is to know how people process arithmetic information. The goal of our work is to understand how paper machine operators process information in their work. Our focus is on the contents of processed information and for this reason our approach has been called content-based analysis of human information processing (Saariluoma 1995, 1997, Saariluoma, Cañas and Leikas 2016).

The roots of content-based thinking can be found in the history of cognitive science. Content-based thinking begins with Turing’s (1936-7, 1950) modeling of the mathematical mind and his idea that machines can process information like human beings. Newell and Simon (1972) developed Turing’s thinking. They assumed that people are information processing systems and, unlike Turing (1950), they began to study empirically how the human mind operates (Newell and Simon 1972). They modeled human information processes computationally and initiated a wide research on the role of capacity in human information processing (Anderson 1993, recently overviewed in the collection by Polk and Seiffert 2002).

However, the early tradition gave much more weight to limited capacity than to mental contents, because the analysis of limited capacity as an explanatory ground was highly successful in working with problems of human technology interaction (Anderson 1993, Broadbent 1958, Miller 1956, Saariluoma 1997). Furthermore, the figure of Shannon (1948) and his information theory influenced this as well (Aspray 1985, Waldrop 2001). Shannon’s information theory was essentially capacity-oriented, as he deemed the contents of messages irrelevant (to the engineering problem) (Shannon 1948).

Newell and Simon (1972), as well as many other researchers, saw the problem of mental content but paid much less attention to it than the problems of the limited capacity of human information processing. The focusing on mental capacity instead of mental contents has not been good for developing intelligent technologies, because the essence of human information processing is in its capacity to analyze, process and create new mental contents (Saariluoma 1997).

Some steps towards the analysis of mental contents can be found in concepts of cognitive simulation models, such as production systems (Anderson 1976, 1993, Newell and Simon 1972), theories of mental models (Johnson-Laird 1983, 2008), and semantic networks (Anderson 1976, Collins and Quillian 1969). From the present point of view, the use of these concepts has rather turned on the human limited capacity to have these entities in working memory and mind than on their information contents. On the other hand, the ground concepts created over the past four decades may in the future be effectively used in analyzing how people process mental contents.

Content-based analysis of information processing requires that the researchers are able to learn to understand the actual content of people’s minds. This fact entails some prerequisites of the research process. Content-based analysis is different from those approaches that analyze human action on the external level. Content-based thinking focuses on the internal properties of information processes (Newell and Simon 1972, Saariluoma 1995). We do not deny the importance of the action level studies. These studies provide one element of understanding. Our purpose is, however, to point out the importance of internal information processes, which make it possible for people to guide and control, in this case study, paper machines. Ultimately, the explanation and the reason for particular actions is in the mental contents (conscious or unconscious) of actual human operators.

Content-based analysis of human information processing has several phases. Firstly, it is important to collect raw data on information in the minds of operators in different situations. This can be done with a

number of qualitative methods used to explicate mental contents (Ericsson and Simon 1984, Patton 1990). Typical examples of data collection methods are protocol analysis (Ericsson and Simon 1984), observation, discourse analysis, narrative methods, documentary analysis interviews, focus groups, and even qualitative tests (Patton 1990). The main goal of data collection is to get as good an idea as possible about how operators, or people in general, mentally represent their work situation and respective actions.

Explicating mental contents, i.e. information contents of mental representations is the core activity of content-based analysis (Saariluoma 1997, 2001). The raw data must be turned into explicit descriptions of mental contents as this is the way one can use mental contents as explanatory ground for human actions. For example, if paper machine operators represent in their mind that the process runs too fast, they reduce the machine speed following their mental models for the situation (Saariluoma, Nevala and Karvinen 2006). If they misinterpret and misrepresent the same situation, it is possible that their actions will lead to operational failures. Thus, the mental contents of the operators make it understandable why they choose a non-optimal manner of action. To be able to analyze and to explain such situations, researchers have to have a clear idea of the mental contents of operators in the particular situations.

An understanding of the action-relevant mental contents enables a researcher to study the digitalization of intelligent information processes. As people are able to carry out the intelligence-demanding task, knowledge of how people process information is vital for developing technical artefacts, which can carry out the same tasks.

Mimetics in designing intelligent technologies

Existing intelligent information processes can be used to develop new information processes in new physical entities. The process of developing new information processes on the ground of old ones can be called mimicking. Mimetic design is one important method of design today.

Mimicking nature is a well-founded branch in technology design (Vincent et al. 2006). Mostly its focus has been on creating physical entities by taking inspiration from the solutions found in nature. This kind of design thinking is called biomimetics or biomimicry. The main idea of biomimetics is to solve complex technical problems by imitating the solutions that nature has developed for similar problems. Many physical artefacts from clothes, spades, and airplanes to Velcro tape, have their origins in biomimetic thinking. However, in developing intelligent technologies, a new model of mimicking is required. Instead of mimicking biological structures, it is good to focus on human information processes (focusing on contents) and imitate them in developing intelligent systems. Mimetics based on an analysis of human information processes in developing intelligent technological solutions can be called "cognitive mimetics."

The basic shape of any mimetic design process has a source S, a target T, and a mapping (m) relation between them. Broadly speaking, for cognitive mimetics the source is the human mind and the target is a computer system.

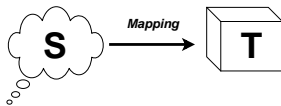


Figure 1: Mimetic mapping. S: Human mind, T: Simulation model.

Both the source and the target include here aspects such as context, environment, task, and task-requirements. What the designer is looking for in mimetic design is, first, what makes the source (s) an effective solution in its context, and second, how to map (m) or transform that into a technological form (t). Thus, the process can be understood as “an abstraction to concretization loop,” where ideas are abstracted from the source and transformed into designs and concrete results. Of course, no design process is linear, and the reality is an iterative dance between two ultimate poles of source and target with many sub-loops in between. One aspect is the appropriate abstraction(s) of the source. A bird achieves flight (source) in the air (source environment) exploiting aerodynamic laws (abstraction of the coupling), for example. It achieves propulsion, lift and can control itself (abstract task requirements). These are among the right abstract ingredients for *human* flight, and yet, da Vinci’s idea of copying the wing-flapping of birds was not the best path. Thus, we see an example of multiple realizability by mimetic means – one does not get very far by direct copying, because the context of the target constrains possibilities in its own way (and indeed is constructed around those constraints). The same is true for computers and intelligence. Computers have their own way of working and mimetic solutions must be built around those constraints. Of course, nothing stops an inventive person from redesigning computers’ operational principles through mimetics, but that is not our focus here. The crucial point for now is that the noetic resources, to borrow a term from Floridi (2013), used by the human in completing intelligence-requiring tasks in a context, constitute the source for cognitive mimetics. Among those aspects, we highlight the role of mental contents and their transformations as the key.

Human digital twins in explicating tacit knowledge

A difficulty with analyzing human information processing is that most mental contents are subconscious or tacit (de Groot 1965). People are not aware of what they should do. Therefore, it is necessary to investigate the tacit levels of information in human mind in order to get a reliable and valid picture of how people process information during intelligence-requiring tasks.

One way to obtain a deeper understanding is to get a clear method of explicating mental contents, and a good method for presenting how information processing works in the minds of people. In the research on human information processing and thinking, it has been thought that computer simulations might provide a means to represent what happens in human minds when, for example, they solve difficult problems (Anderson 1993, Newell and Simon 1972).

Originally, the idea of imitating human thinking by means of computer programs was presented by Turing (1936–1937, 1950). Actually, the basic model of computational algorithms has been the Turing Machine, which is a model of how mathematicians think (Petzold 2008 Turing 1936-7). After Turing, Newell and Simon (1972), among many other researchers, began to develop the idea that people are some kind of computers (Boden 2016, Newell and Simon 1972).

Recent work on mimicking physical objects and industrial processes by means of digital computers has been called digital twinning. Digital twins (DTs), which are digital replicas of technical systems and processes, have recently become a tool for working with complex industrial processes (Barricelli, Casiraghi and Fogli 2019). They model components and functionalities of the systems and, therefore, support design,

operational planning, and maintenance. A crucial question for DTs is how should human actions be analyzed? When are they relevant? How should human digital twins be modeled? From which level of abstraction should the human be analyzed? These questions and possibilities are a natural follow-up from the original DT prospect and have been identified in parallel (Saariluoma, Cañas and Karvonen 2020) by at least Kaivo-oja et al. (2020) and Hafez (2019). As mimetic design shows, there are legions of different levels of abstraction' (LoAs) (Floridi 2013) one can take on the human. What is a rational and pragmatic level of abstraction is dictated by the purpose of the human digital twin (HDT). If the HDT is about physical ergonomics of a workplace, one is likely to look at anatomical models. If, on the other hand, it is about human action with respect to technology, a good LoA could be the cognitive perspective. If the model is used to analyze fuel combustion processes in a turbine, user knowledge is of less relevance than in the case of thinking how the users control artefacts. As digital twins typically take a wide perspective into a technical system, it becomes likely that human action should often be included in the twins. For this reason, it is vital to discuss different approaches to modeling human actions as parts of digital twins.

Human digital twins (HDTs) are models of human actions when interacting with technologies. They can focus on narrow issues such as usability and even user experience, but there are no obstacles to modeling technical artefacts as parts of human work and human life. The issues of designing human machine collaboration processes can benefit from digital twinning, but they can also be used, for example, to explicate expertise and other types of tacit knowledge. This in turn can be used for artificial intelligence (as in cognitive mimetics) and many other purposes, like learning and developing organizational knowledge. In traditional engineering, such as building cyber physical systems, interacting with human actions can be seen as functional machine elements. This means that a number of machine functions are supposed to wait for input from users. Thus, car drivers have a set of controls they can use to get their car to behave as they want. In such examples, people are seen as a kind of input machine. However, it is also possible to study how it is possible for people to use technical systems. Usability, UX (user experience), and life-based design are typical examples of looking at human–technology interaction through the concepts of cognitive and human research (Saariluoma, Cañas and Leikas 2016).

Human digital twins are likely to present the next frontier of digital twinning. We have here only scratched the surface of their possibilities. We will next outline through a case study one approach for human digital twinning, oriented around the mimetic perspective. One way to understand our approach is to see it as a first layer of intelligence on a digital twin of a factory process.

Our domain – paper machine operators' information processes

Paper machines produce paper fast and in large quantities. They transform pulp suspensions, containing even more than 99% of water, to paper webs in seconds. The web is dried to 1.5–10-meter-wide paper sheets with less than .02 mm tolerance at a speed of over 80–90 km/h. Obviously, paper machines require high-precision engineering, and they are no less complex than big airplanes. Papermaking has a number of human-driven process parts, which may eventually be replaced by intelligent technology solutions. Interestingly, paper machines have a kind of mimetic origin: "The Fourdrinier machine represented a straightforward mechanization of what was formerly done by hand." (Särkkä, Gutiérrez-Poch, and Kuhlberg, 2018). Automation, more broadly, is the continuation of this same trend but it begins to encompass the further mimetics of the information processes, rather than manual labor of paper-machine operators. The first proportional-integral-derivative (PID) controller, for example, was a technical solution directly based on the mental processes of a steersman in ships (Bennett 1984, Minorsky 1922,). Our work here can be placed into this continuum by applying cognitive perspectives to the problems of creating digital intelligence for paper industry processes.

The studied environment was a pilot scale paper machine designed for research purposes. This means that in contrast to industrial-sized paper machines, paper grade changes and optimization of running parameters take place far more frequently (even several times within an hour). Interviews with the operators showed that this

means that the process is run more manually than normal paper machines and therefore it requires more human thought and action. On the other hand, pilot runs are, in general, not so different from the operation of normal paper machines. In both cases, there are targets for the paper specifications and the operator's task is to align the process and the raw materials with those goals. However, in the pilot environment there may be fewer variables to consider and adjust than in normal paper machines. Here, the focus is typically on optimal conditions against a few – and even, in some cases, only one – variables and there is flexibility with the rest. In industrial paper machines, the product must be as good as possible, and conditions of the run are secondary and may be suboptimal.

In the studied pilot paper machine environment, the operator works in a control room that is separated from the paper machine by a windowed wall. The process is operated and controlled through four medium-sized computer screens. The operator obtains information from the process by four principal means:

1. Graphical and numerical information from the Valmet DNA Distributed Control System (DCS) and the Trimble Wedge data analytics system. The Valmet DNA is a user interface that gives the operator the control of all processes. Through the DNA, the operator sees the prevailing process conditions (flows, levels of tanks, consistencies, pressures, etc.) and the operation status of the main process equipment (opening degree of valves, running of pumps, etc.), and is able to adjust/control the process. The Wedge data system, on the other hand, contains over 600 online measurement points from the pilot paper machine. With Wedge, the operator may follow the changes in the most important process parameters online as well as check the historical trends of the parameters.
2. Visual information from several cameras placed around the paper machine through a screen. The operator may also observe some parts of the process directly through the windowed wall.
3. Audio information (radio communication) from operators (2–3) in the field. The communication concerns mainly the tasks and task-status communication, and information transmission.
4. Audio information from the paper machine system. This is a largely tacit dimension where sounds, especially expected or anomalous, are used as an additional source of information. For example, the sound of a pump changes when the level of the tank it is pumping is below the level detectable by the level controller. This can be used as information to stop the pumping.

In the studied pilot paper machine, four operators run the process. Three of them operate in the field, i.e. in the immediate vicinity of the machine. Their tasks include manual procedures like sampling, opening valves, feeding the tank with fiber, etc. They communicate with the fourth operator via radio phones. The fourth operator operates in the control room and has the main responsibility. The operator controls and adjusts the process with the DNA system. He performs a set of actions during the start-ups and shutdowns of the machine, which is fairly routine. However, adjusting the process during the actual trial point to achieve the set targets for the end product or running parameters is not so straightforward. Achieving the specifications for the end product or a sufficient runnability of the machine with a totally new raw material requires thorough knowledge of how the machine works, the capability to exploit know-how and theory from earlier pilot runs, and also fast problem-solving skills. Even though every pilot trial point is unique, the operator utilizes the experience from earlier trials.

The research material was collected by video recording. The camera was positioned "over the shoulder" of the operator and mainly captured the screens that showed the process control system(s). The main idea was to capture sound, namely the think-aloud protocols of the operators. The operators ($n = 2$) were reminded to say what was on their minds and what they were thinking or watching. All in all, we captured approx. 7 hours of material, which was then transcribed into text and analyzed into episodes and phases. This analysis is still ongoing.

Example from protocol

We collected verbal think-aloud protocols on operators during work (Ericsson & Simon 1984). The main goal was to collect information about the basic structure of operators' thoughts in order to model them. The following presents an excerpt from protocol materials to demonstrate the highest-level structure of operators' actions.

The Fluctuating Flow of the Headbox Feed Pump Episode:

K1: [0:20:12] What are you looking at?

M1: [0:20:14] The speed of the middle ply alters quite a bit, sometimes it is 25 and then 34. [pause 5 s]

M3: [from the radio] [0:20:26] What do you mean by speed?

M1: [responds] [0:20:31] I mean volume flow [pause 5 s]

M3: [from the radio] [0:20:41] (inaudible) a lot of fluctuation (inaudible) by eyesight. [pause 11 s]

M1: [responds] [0:20:54] 26–37, fluctuates between there.

M3: [from the radio] [0:21:01] Is air removal on? [pause 7 s]

M1: [responds] [0:21:10] No, I'm putting it on. [thinks aloud] [0:21:12] So the air removal had been left off [pause 5 s] [talks to radio] [0:21:23] it's on. [pause 12 s]

M3: [from the radio] [0:21:38] (inaudible)

M1: [responds] [0:21:48] Yeah, let's see where it settles. [pause 10 s] [talks to radio] [0:22:02] I'm going to decrease the flow of the headbox feed pump [pause 14 s]

K1: [0:22:21] What are you thinking now?

M1: [thinks aloud] [0:22:26] That [I will decrease] the headbox feed pump, so that the flow doesn't go too high. [0:22:37] Now it is pretty good, fluctuates between 32 and 34, I will try to adjust the (inaudible) [pause 10 s] [0:22:55] That's why it was fluctuating, the air removal was not on. [pause 12 s]

In a nutshell, the ideal state was violated by the fluctuating flow of the headbox feed pump. This resulted in a quick problem-solving episode, where the field operator suggested that it might be because the air removal was not on. Turning it on corrected the fluctuation.

The follow-up interview showed that this cause was "a classical one," accounting for 95% of cases in their estimation. The second "go-to" reason would have been blockage in the headbox hoses. The operators could list about 10 reasons for the event off the top of their heads.

Example 2:

M1: [thinks aloud] [0:43:33] Now the level in tank A is a bit high. I'm putting it on manual and reducing the level control, and then putting it back on automatic. [pause 9 s] [0:43:55] Also I will decrease the chemical

dosing (so that the density won't go too low.) [0:44:05] Also, I will manually put the dilution water valve to zero (-) [silent talking] to automatic. [0:44:19] Tank A shouldn't overflow, you can follow that from the camera. [0:44:22] If it overflows, you lose fiber. [0:44:29] Now the level is OK, density also pretty close [pause 7 s] [0:44:42] It's coming down pretty fast [the density], I'm going to turn off the chemical pump so it doesn't go too low. [pause 8 s] [0:44:58] Checking the density in tank B, the level in tank A is good, it is in the set point. [0:45:21] Flows are good, looks like the density in tank A is altering. [pause 11 s] [0:45:40] (-) I'm looking to see if I still need to do something about it. [0:45:47] No need, its going fine, it is 520 and (-) going down (-) and the (chemical) pump is off, so we are getting close to the trial point. [pause 11 s] [0:46:09] The density in tank A has set to 520. [0:46:16] I have to monitor if it still starts to change (-) [pause 5 s] [0:46:25] Still going down. [0:46:28] Then I must remember to decrease the flow of the headbox feed pump to 32, it is still too high [pause 8 s] [0:46:46] Density is good. [0:46:50] I'm going to decrease the flow of the headbox feed pump a little, just so that I reach the targeted flow. [pause 7 s] [talks to radio] [0:47:04] Densities are pretty much there. I'll still decrease the flow of the headbox feed pump, so we get the flow right, to 32.

Figure 2: Two examples of operator activity protocols.

In both examples we can see the same structure. Operators recognize that something is not as it ideally should be. Therefore, they look for methods to correct the state of affairs. Interestingly, the contents of their thinking are not precisely in the ongoing situations, but they foresee how things will go wrong if they don't make their correcting operations. Thus, they anticipate possible future events and actions based on past and present information. This kind of anticipation has long been known to be typical for human thinking (Selz 1913).

The operators' thought of processes on a higher level apparently take the form of ideal-exception-correction. They see how things are straying from the path they should be on and consequently they understand that they have to do something to prevent things from reaching that state. They compare their information on the present state of the process with the idea to find out means to reach an ideal state after correcting operations.

Based on the protocols we can see the basic logic of the operators' 'thinking. They have an ideal state in their minds. What the ideal state is depends on a number of issues, such as the quality of paper they are now producing, the raw materials they have at their disposal, and the state of the production process. Operators encode the present state of the technical process and register exceptions from the ideal. They can register deviations from the ideal state by comparing the present state with the expected ideal. Here it is important that operators can predict possible critical situations based on the indicators they have about the process in its present state and their own knowledge of the paper machine behavior. Finally, operators have in their minds a list of possible corrective actions, which they apply to bring the process to the ideal state.

Modeling mind IEC_0.81

In order to get concrete clarity on the use of a digital twin in acquisition of knowledge on mental contents, we developed a small model for operator information processing called IEC_0.81. The model is a very simple process model, but it contributes as it can be used to develop the use of human digital twins in acquiring information about human information processes in controlling paper machines.

IEC_081 assumes, based on collected empirical knowledge, that paper machine operators' thinking has an IDEAL-EXCEPTION-CORRECTION (IEC) loop. This means that operators observe the behavior of the machine process by means of measurement instruments and visual contact in the control room. Information is in addition passed "from the field" by other operators who work near the paper machine. When they observe an unexpected state of the process (or rather a deviation from the ideal), they make respective corrective actions following their models for corrective actions.

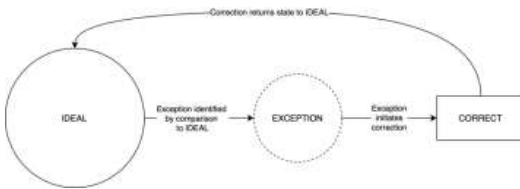


Figure 3: Ideal-exception-correction model.

The model is based on the idea that all human involvement points (HIPs) can be defined. Such a point is a place in a process in which people become involved in the process. The point entails a set of observation values (OVs) and a set of possible actions (PAs). Since the machines are closed and defined systems, they have for each HIP a limited set of OVs and PAs. All the possible human actions of involvement in the ongoing machine process can be thus defined in terms of HIPs, OVs, and PAs.

In the case of analogue controls, which are in principle continuous, one can digitalize the operations. For example, a rudder that has an infinite number of positions can be divided into a finite number of possible states by means of assuming it is digitalized. As another example, in recordings the voice varies continuously, but it can be represented with sufficient accuracy in digital recordings. Thus, the basic analysis of a finite number of possible HIPs can be kept.

As said, IEC_081 is very simple, but it can still give us an idea about the role of digital twin models in information collection. The model gives an interpretation for one possible solution to the problem of how human information processes and their content operate in controlling paper machines. The problem is that in its present elementary state, IEC_081 does not present the possible HIPs with sufficient accuracy. However, it can point out the open points in the control process and thus direct further collection of information. One can go to the operators and collect knowledge required in the description of HIPs. This in turn is what we have called the "mental contents."

IEC_081 does not yet have a detailed description of ideal processes, and ideal states, or corrective actions. It does not yet have detailed descriptions of detailed operator actions. Nevertheless, the model can be

developed further by studying how operators carry out their actions in different situations. Thus, the model can very effectively aid in directing information collection on operators' mental contents.

The model also enables researchers to test the logic of their interpretation of data. If simulations work, this suggests that the interpretations do not have a problem in their formal structures. If simulations do not work, it means that the interpretations must be reanalyzed. Internally contradictory models cannot be possible and thus simulation makes it possible to perform a self-corrective analysis and interpretation of data on mental contents.

Discussion: Computational thinking in mimetics

Intelligent technologies are constructed conceptually on Turing machines (1936–1937). The Turing machine is different from biomimetical artefacts as they are based on analogy between human and machine information processes. Turing machines do not imitate biological structures, but human thinking. Turing's machine was originally a machine version of how mathematicians process information (Petzold 2008, Turing 1936–1937). As Turing machines need not limit themselves to mathematical information – for example, the symbols in them could be Chinese letters – they can also be used as models of the mind as an information processing system. Often intelligent technologies, whether theoretical or practical, have their origins in imitating how people process information. Cognitive information processing models describe how people process information (Lindsay & Norman 1977).

To be able to imitate human information processes, designers need to know what these processes are like. Since vast parts of human information processes are subconscious or tacit, it is not easy to gain a solid understanding of what happens in the minds of people when they carry out some intelligence-requiring task. It is necessary to explicate the tacit information in order to acquire a sufficient understanding of how people are able to carry out these processes.

One method of explication is to construct digital twins, which are computational models of source processes. Here, the source processes would be human information processes and thinking. In basic research, much work has been done to computationally model the human information processing system (Anderson 1993, Newell and Simon 1972). Thus, simulative cognitive psychology can provide many tools for explicating human information processing and thus give ideas for developing intelligent information systems.

Interestingly, we found that modeling can guide empirical research. IEC_0.81 is a very primitive model for operators' information processing. It basically analyzes how operators detect exceptions and how they revise the processes into the ideal course of affairs. In the present conceptual version of IEC, it tests constantly the state of the machine process in HIPs or human interaction points. They are states in which operators have controls to regulate process. These points are places in which people become involved in the paper process and change them in some way. The ways they can do something is defined by machines.

In the model, all corrective actions are under one function called CORRECT. In order to move forward it is essential to study the mental contents that enable people to do various types of operations typical for CORRECT. Thus, the design of CORRECT functions enables modelers to focus their data collection on definable points in operators' information processing. This means that modeling proceeds through steps, in which researchers collect information about operators' mental contents. For example, the present version of IEC_0.81 has an operator called "control," but it does not yet analyze the methods of controlling and the connections between states of paper machine and respective controls. This analysis and modeling can be done in later versions. They re-evaluate what has been modeled so far to be able to go further.

In data collection, researchers can use the same methodological thinking as above. Only the topic can specifically be focused on the relevant points in operators' information processing. The collection of information can again be based on qualitative methodologies typical to modern psychology (Ericsson and Simon 1984, Patton 1990). The goal of the research process is to explicate tacit and explicit information contents in the minds of operating people.

The process outcome of designing intelligent information processes has an iterative structure. Empirical research of human information processes and especially their information contents are used to model what happens in human minds during some specific information process. The model can be used to direct information collection, and mimetically in generating technical systems that can carry out the tasks.

References

- Anderson, J. R. (1976). *Language, Memory and Thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. R. (1993). *Rules of the Mind*. Hillsdale, NJ: Erlbaum.
- Aspray, W. F. (1985). The scientific conceptualization of information: A survey. *Annals of the History of Computing*, 7(2), 117-140.
- Barricelli, B. R., Casiraghi, E., & Fogli, D. (2019). A survey on digital twin: Definitions, characteristics, applications, and design implications. *IEEE Access*, 7, 167653-167671.
- Bennett, S. (1984). Nicolas Minorsky and the automatic steering of ships. *Control Systems Magazine*, 10-15.
- Bernal, J. D. (1969). *Science in History I–VI*. Harmondsworth: Penguin.
- Bickle, J. (2020). Multiple realizability. In: E. Zalta (Ed.), *Stanford Encyclopaedia of Philosophy* (Summer 2020).
- Broadbent, D. (1958). *Perception and Communication*. London: Pergamon Press.
- Collins, A. & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, 8, 240-248.
- de Groot, A. (1965). *Thought and Choice in Chess*. The Hague: Mouton.
- Engelbrecht, A. (2007). *Computational Intelligence*. Chichester: Wiley.
- Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis*. Cambridge, MA: MIT Press.
- Floridi, L. (2013). What is a philosophical question? *Metaphilosophy*, 44(3), 195-221.
- Ford, M. (2015). *Rise of the Robots*. New York: Basic Books.
- Fukuda, K. (2020). Science, technology and innovation ecosystem transformation towards society 5.0. *International Journal of Production Economics*, 220, 3-14.
- Hafez, W. (2019, September). Human digital twin: Enabling human-multi smart machines collaboration. In Proceedings of SAI Intelligent Systems Conference (pp. 981-993). Springer, Cham.
- Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.
- Johnson-Laird, P. (2008). *How We Reason*. Oxford: Oxford University Press.

- Kaivo-oja, J., Knudsen, M. S., Lauraeus, T., & Kuusi, O. (2020). Future knowledge management challenges: Digital twins approach and synergy measurements. *Management*, 8(2), 99-109.
- Lindsay, P., & Norman, D. (1977). *Human Information Processing*. New York: Academic Press.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Minorsky, N. (1922). Directional stability of automatically steered bodies. *Journal of the American Society of Naval Engineers*, 280-309.
- Minsky, M. L. (1967). *Computation: Finite and Infinite Machines*. Englewood Cliffs, NJ: Prentice-Hall.
- Newell, A. & Simon, H. A. (1972). *Human Problem Solving*. Englewood Cliffs, NJ: Prentice-Hall.
- Patton, M. (1990). *Qualitative Evaluation and Research Methods*. London: Sage.
- Petzold, C. (2008). *The Annotated Turing*. Indianapolis: Wiley.
- Polk, T., & Seiffert, C. (2002). *Cognitive Modelling*. Cambridge, Mass.: MIT-press.
- Russell, S., & Norvig, P. (1995). *Artificial Intelligence*. Upper saddle river: Prentice Hall.
- Saariluoma, P. (1995). *Chess Players' Thinking*. London: Routledge.
- Saariluoma, P. (1997). *Foundational Analysis: Presuppositions in Experimental Psychology*. London: Routledge.
- Saariluoma, P. (2001). Chess and content-oriented psychology of thinking. *Psicologica*, 22, 143–164.
- Saariluoma, P., Cañas, J., & Karvonen, A. (2020, August). Human digital twins and cognitive mimetic. In *International Conference on Human Interaction and Emerging Technologies*(pp. 97-102). Springer, Cham.
- Saariluoma, P., Cañas, J. & Leikas, J. (2016). *Designing for Life*. London: Macmillan.
- Saariluoma, P., Nevala, K., & Karvinen, M. (2006). Content-based analysis of modes in design engineering. In *Design Computing and Cognition '06* (pp. 325-344). Springer, Dordrecht.
- Selz, O. (1913). *Ueber die Gesetze des geordneten Denkverlaufs* [On the laws of organized thinking]. Stuttgart: Spemann.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3), 379-423.
- Särkkä, T., Gutiérrez-Poch, M., & Kuhlberg, M. (Eds.). (2018). *Technological Transformation in the Global Pulp and Paper Industry 1800–2018: Comparative Perspectives* (Vol. 23). Springer.
- Tegmark, M. (2017). *Life 3.0*. Harmondsworth: Penguin Books.
- Turing, A. M. (1936–1937). On computable numbers, with an application to the entscheidungs problem. *Proceedings of the London Mathematical Society*, 42, 230–65.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, LIX, 433-460.

Vincent, J. F., Bogatyreva, O. A., Bogatyrev, N. R., Bowyer, A., & Pahl, A. K. (2006). Biomimetics: its practice and theory. *Journal of the Royal Society Interface*, 3(9), 471-482.

Waldrop, M. M. (2001). *The Dream Machine: JCR Licklider and the Revolution that Made Computing Personal*. Viking Penguin.

Improving Maximum Entropy Model by GIS

Junichi KAWAMI^a and Takao MIURA^a

^a*Dept. of Advanced Sciences, HOSEI University
Kajinocho 3-7-2, Koganei, Tokyo, Japan*

Email: junichi.kawami.8x@stu.hosei.ac.jp miurat@k.hosei.ac.jp

Abstract. Maximum Entropy Model (MEM)[1][4] estimates probability distribution functions, by which current state of knowledge is described in the context of prior data. Here we examine Generalized Iterative Scaling (GIS)[1] algorithm to determine optimum feature weights with feature selection during learning. Maximum Entropy principle[1] provides us with all the characteristics of the data given in advance and we could expect robust distribution against outlier. However it takes much time until convergence because the computation depends heavily on the number of classes. We introduce a novel approach random sampling of Monte Carlo method into GIS for improved computation.

Keywords. Natural Language Processing, Multiple Classification, Maximum Entropy Model, Monte Carlo method, Sampling

1. Introduction

Recently wide-spread internet allows us to analyse and extract what we could have and how we could do from the view point of both quantity and quality. Most of data are written in text and we should examine them with natural language processing (NLP). For example, very often we classify document d into one of given classes $C = \{c_1, \dots, c_m\}$, the problem is called *document classification*. Let D be a set of documents over words W and a document d in D , we consider d as a vector $[t_1, t_2, \dots, t_n]$ over words $W = \{w_1, w_2, \dots, w_n\}$, where t_j means frequency of w_j appeared in d . Note d is, in fact, a vector over W not a list. There have been many approaches proposed, but Maximum Entropy Model (MEM) works very well in the classification problem. As well know, we face to data sparseness problems in NLP. MEM helps us to extract characteristic context by entropy and to make inferences on the basis of partial information.

Each word may carries several meanings. It is hard to identify interest words suitable for the current context of documents. N-grams or collocations mean a set of words to carry single semantics as a whole. We can separate them onto word to obtain the semantics. All these aspects cause hard tasks to solve classification problems correctly.

One of problems over MEM comes from how to estimate probabilities, GIS is one of algorithms for estimating the parameters of MEM. It helps us to compute these parameters empirically and approximately. However it takes much times until convergence because the computation depends heavily on the number of classes.

In this work, we propose a novel approach Generalized Iterative Scaling (GIS) based on random sampling improves computation. In fact, the marginal probability causes the heavy computation of the probability Summarization to all the classes, as integral calculation by random sampling. Compared to GIS which computes marginal probability with all classes during learning, it becomes faster to obtain approximation.

Our results contribute to NPL research focusing on the following points; (1) Our approach can improve efficiency of GIS algorithm with the help of sampling techniques and (2) We propose a sophisticated technique to introduce feature selection. By examining a collection of learning data, we *mine* effective functions in terms of association rules so that we improve classification dramatically and that we can complete feature selection automatically.

The rest of the paper is organized as follows. In section 2 we describe fundamental roles of document vectors, vector space model and document classification. In section 3, we describe Maximum Entropy Model and the model generation. Section 4 contains how to apply random sampling to the model calculation, and our sample generation in section 5. Section 6 contains some experimental results to see the effectiveness and we conclude this investigation in section 7.

2. Maximum Entropy Model

Conceptually MEM approach helps us to model all that is known in advance and few about what is unknown. In other words, we like to obtain probability model satisfying a set of constraints which represent "evidence" and choose the *most uniform* distribution otherwise because the distribution carries the maximum entropy or the minimum commitment. One way to represent evidence is to encode characteristic facts as *features*. Any kind of contextual feature can be used in the model, and experimenters generally need to focus their efforts on deciding what features to use. The representation of the evidence discussed below, then determines the form p .

Given an input vector \vec{x} of a document over words, we like to classify the \vec{x} , i.e., to estimate a class c to which \vec{x} belongs. To build a classifier from the viewpoint of MEM, this means we like to estimate a class c' of the maximum probability $p(c'|\vec{x})$, i.e., $c' = \arg \max_c p(c|\vec{x})$.

2.1. Modeling by Maximum Entropy

In MEM, we assume a set of *features* f in advance to a word w in W and a class c in C which we intend to mention w is characteristic to c , i.e., w is a *feature word* of c . Given w , a feature is a function $f_w(\vec{x}, y)$ where \vec{x} means a document vector and y a class:

$$f_{w,c}(\vec{x}, y) : W \times C \rightarrow \{0, 1\} \quad (1)$$

$$f_{w,c}(\vec{x}, y) = 1 \text{ if } w \in \vec{x} \text{ and } y = c, 0 \text{ otherwise} \quad (2)$$

Since features show characteristic aspects of our documents of interests, it could also be useful to describe classification. Here we assume each document may belong to one class as well as a word. To have conditional probability distribution $p(c|\vec{x})$

given a class c and an input \vec{x} . We also assume q_w if we give a constraint " a document d containing a word w belongs to a class c " in terms of expects over features and their distribution. Then let q_w be a distribution of relative frequency over $W \times C$, and can be seen as a probability function of (\vec{x}, y) as a constraint defined as :

$$E_p[f_w] = \sum_{\vec{x}, y} q_w(\vec{x}, y) f_w(\vec{x}, y) = \frac{1}{N} \sum_{\vec{x}, y} f_w(\vec{x}, y) \quad (3)$$

Note N means the size of domains $W \times C$. Then we give an expect of the distribution p as our constraints of w, \dots :

$$E_p[f_w] = \sum_{\vec{x}, y} p_w(\vec{x}, y) f_w(\vec{x}, y) \quad (4)$$

The constraints wrt (w, c) can be described as:

$$E_p[f_w] = E_q[f_w] \quad (5)$$

$$\sum_{\vec{x}, y} p_w(\vec{x}, y) = 1 \quad (6)$$

Since the objective is to maximize entropy $H(p) = \sum p(\vec{x}, c) \log(1/p(\vec{x}, c))$ subject to the constraints above. To estimate the distribution p , we apply *Lagrange Multipliers* to our model by maximizing $L(p)$

$$L(p) = H(p) + \sum_w \lambda_w (E_p[f_w] - E_q[f_w]) + \lambda_0 \left(\sum_{(\vec{x}, y)} (p(\vec{x}, y) - 1) \right) \quad (7)$$

Then we have the solution below:

$$\begin{aligned} p(\vec{x}, y) &= \exp\left\{ \sum_{w \in W} \lambda_w f_w(\vec{x}, y) \right\} / Z \\ Z &= \exp\{1 - \lambda_0\} \\ &= \sum_{\vec{x}, y} \exp\left\{ \sum_{w \in W} \lambda_w f_w(\vec{x}, y) \right\} \end{aligned} \quad (8)$$

Note $p(y|\vec{x}) = p(\vec{x}, y) / p(\vec{x}) = p(\vec{x}, y) / \sum_y p(\vec{x}, y)$.

Ratnaparkhi has examined several models : it is always possible to get such p in a unique manner and discussed how to do that. Once we obtain MEM, we could classify documents. Clearly the results depends heavily on both the selection of features and the parameters λ_w . There have been several algorithms such as *Generalized Iterative Scaling* (GIS) and *Improved Iterative Scaling* (IIS) proposed so far. However, all of them take much time to obtain the values.

2.2. Training parameters using GIS

To obtain model-parameters λ s of MEM probability function, there have been proposed two useful algorithms, Generalized Iterative Scaling(GIS) and Improved Iterative Scaling(IIS). Both algorithms work in an iterative scaling manner based on a gradient method. The parameters shows how important role the feature plays to classification task.

Let us illustrate how GIS works in a case of single classification in a Table 1:

Table 1. An outline of GIS algorithm

1. Assume all the feature f_1, \dots, f_K are given in advance.
And also assume q an initial distribution.
2. Let C and f_{K+1} be an auxiliary constant and a feature. $C = \max_{\vec{x}, y} \sum_{j=1}^K f_j(\vec{x}, y)$
3. Set $\omega_i^0 = 0.0, i = 1, \dots, K + 1$
4. improve ω_i^k as follows where N is the size of traing data:
$$\omega_i^{k+1} = \omega_i^k + \log \frac{1}{C} \frac{E_q[f_j]}{E_p[f_j]}$$
5. Repeat 4 until convergence.

We like to obtain our goal, a probability density function $p(\vec{x}|y; \omega)$. To do that, we have to estimate parameters ω . Note in 2 we define a constraint C and a feature f_{K+1} additionally to simplify the algorithm. Step 4 describes our constraints $E_q[f_j]$ in terms of the features:

$$E_p[f_j] = \sum_{\vec{x}, y} q_j(\vec{x}, y) f_j(\vec{x}, y) = \frac{1}{N} \sum_{\vec{x}, y} f_j(\vec{x}, y) \quad (9)$$

Also $E_p[f_j]$ in step3 describes our constraint of probability distribution p in terms of the features.

$$E_p[f_j] = \sum_{i=1}^N \sum_{y \in \mathcal{Y}(x_i)} p_j(\vec{x}, y) f_j(\vec{x}, y) \quad (10)$$

Similarly we repeat the whole process to improve ω_i values until we get to $E_p[f_j] = E_q[f_j]$. Then we eventually obtain our model p . During GIS processes, it is impossible to avoid heavy computation. In fact, once we obtain $E_q[f_i]$ for initialization, we approximate $E_p[f_i] \propto (|P(D)| \times |C|)$ times for each feature f_j .

3. Multiple classification and Feature Selection

By a word "multiple classification", we mean that an object belongs to a class softly. That is, we assume it belongs to a single class but we don't know explicitly, and we could have some knowledge with possibility by means of distribution over classes. We discuss multiple classification with MEM approach. To do that, we explore how to extract feature functions based on frequent patterns appeared in training data. We define our patterns as features, by which we can consider a set of words automatically as single feature so that we can construct MEM for multiple classification.

GIS helps us to compute parameters empirically and approximately such as feature weights. However it takes much time until convergence because the computation depends

heavily on the number of classes. Multiple classification allows us to label a document multiple class, GIS plays critical role on classification task. In this investigation, we discuss how to apply MEM to multiple classification. Here let us discuss how to extend MEM, especially feature functions, and GIS algorithm.

3.1. Feature function for multiple classification

By a word "multiple classification", we mean that an object belongs to a class softly. That is, we assume it belongs to a single class but we don't know explicitly, and we could have some knowledge with possibility by means of distribution over classes.

Since features show characteristic aspects of our documents of interests, it could also be useful to describe multiple classification. For example, a sentence *scientists who study viruses say they don't know what a pandemic strain would look like* could belong to class "health". Similarly a word *pandemic* is characteristic to the class. However if class set C contain "economy", only *pandemic* could not assign a document to 'health' or "economy". In the sentence, we could get information that *pandemic* and *virus* could be characteristic to "health". In other words, for classification, characteristics to a class are not a word, but they are set of words at same time in a document, and the discovery of interesting associations and correlations between a set of words and classes helps us to assign documents to classes.

The association rule is an implication of the form $U \Rightarrow c$ where U is a set of words and c is a class. The rule $U \Rightarrow c$ holds in the document set D with *support*, where *support* is the percentage of documents in D that contain U and c . The rule $U \Rightarrow c$ has *confidence* in the documents set D , where *confidence* is the percentage of document in D containing U that also contain c . This is taken to be the conditional probability, $p(c|U)$. That is,

$$support_{U,c} = \frac{\text{Frequency of documents containing } U \text{ and } c}{|D|} \quad (11)$$

$$confidence_{U,c} = \frac{\text{Frequency of documents containing } U \text{ and } c}{\text{Frequency of document containing } U} = p(c|U) \quad (12)$$

a rule that satisfies both a minimum support threshold ($minsup_{U,c}$) and minimum confidence threshold ($minconf_{U,c}$) helps us to solve classification problems. The occurrence frequency of a set of words is the number of documents that contains the set of words. If the support of a set of words U satisfies a prespecified *minsup*, then U is a frequent set of words. We describe how to be corresponded rules to feature functions.

Given a document \vec{x} and a class y over C , we extend the definition of a feature function $f(\vec{x}, y)$ as follows:

$$f : P(W) \times C \rightarrow [0, 1]$$

$$f_{U,c}(\vec{x}, y) = p_{U,c} \text{ if } U \subseteq \vec{x} \text{ and } y = c, 0 \text{ otherwise}$$

$$p_{U,c} = p(c|U) = confidence_{U,c}$$

The constraint wrt (\vec{x}, y) can be described as:

$$\sum_{y \in C} f_{U,c}(\vec{x}, y) = 1$$

3.2. Feature Selection

Let us describe how to examine feature functions from learning data. To select feature function of MEM, we propose feature selection to use a *minsup* threshold to ensure the generation of a set of frequencies a set of words and a *minconf* threshold to ensure a set of correlations of a set of words. The discovery of interesting associations and correlations between a set of words and classes helps us to assign documents to classes.

Let us illustrate how the feature selection works in a case of multiple classification, when *minsup*=0.3, *minconf*=0.4.

Table 2. DB

Document	authority	virus	impact	pandemic	Class
x_1	1	1	0	1	economy
x_2	0	0	1	1	economy
x_3	0	1	0	1	health
x_4	0	1	0	0	health

Table 3. Frequency a set of words and classes in DB

A set of words	Frequency	economy	health
{authority}	1	1	0
{virus}	3	1	2
{impact}	1	1	0
{pandemic}	3	2	1
{authority, virus}	1	1	0
{authority, pandemic}	1	1	0
{virus, pandemic}	2	1	1
{authority, virus, pandemic}	1	1	0

Table 4. Confidence

A set of words	frequency	confidence	
		economy	health
{virus}	3	0.33	0.66
{pandemic}	3	0.66	0.33
{virus, pandemic}	2	0.5	0.5

Table 2 shows a DB, Table 3 shows frequencies of a set of words and classes at the same time in the DB and Table 4 shows *confidence* which equals to $p_{U,c}$ of each a set of words exceeding *minsup*=0.3 and classes.

For all that have the same set of words, we select feature functions based on a set of words and classes which have confidence exceeding *minconf*=0.4. We show selected feature below:

$$\begin{aligned}
 f_{\{virus\},health}(\vec{x},y) &= 1.0 & f_{\{pandemic\},economy}(\vec{x},y) &= 1.0 \\
 f_{\{virus,pandemic\},economy}(\vec{x},y) &= 0.5 & f_{\{virus,pandemic\},health}(\vec{x},y) &= 0.5
 \end{aligned}$$

In the selected feature functions, U helps us to assign documents to class c to follow rules ($U \Rightarrow c$).

3.3. GIS for multiple Classification

Here we have to discuss how to extend MEM, especially GIS algorithm. GIS provides us with the parameters λ_w to feature functions $f(\vec{x},y)$ to estimate $p(c|\vec{x})$. Note it takes heavy computation.

However, we see GIS is not efficient because of Z , a normalization term. In fact, Z is nothing but marginalization of each class so that we need the integral values. When we estimate $p(y)$ for each class y in such a way that $Z(\vec{x}) = \int_y p(\vec{x}, y) dy$:

$$\begin{aligned}
 Z(\vec{x}) &= \sum_{y \in P(Y)} \exp\{\phi(\vec{x}, y)\} \\
 &= \sum_{y \in P(Y)} \frac{\exp\{\phi(\vec{x}, y)\}}{p(y)} p(y) \\
 &\approx \frac{1}{M} \sum_{k=1}^M \frac{\exp\{\phi(\vec{x}, y_k)\}}{p(y_k)} \quad y_k \sim p(y)
 \end{aligned} \tag{13}$$

As input (\vec{x}, y) , we give training data (with class y). In this investigation, using random sampling, we approximate the computation appropriately and efficiently. That is, given a set of multi-class distribution Y , we take samples y_1, y_2, \dots, y_M where $y \sim p(Y)$. As $p(Y)$, we assume our base probability as follows.

$$p(y_i) = \frac{\text{FrequencyDistribution} + 1}{|D| + |Y|} \tag{14}$$

To generate samples, we generate $0.0 \leq u \leq 1.0$ through uniform distribution and obtain v such that $\sum_{i=1}^v p(y_i) \leq u < \sum_{i=1}^{v+1} p(y_i)$. Since we assume all the classes are independent with each other, we take M times.

4. EXPERIMENTS

We show our experimental results to see how well the proposed approach works. We discuss our results for multiple classification using MEM. As the baseline, we compare normal GIS with our approach to focus on accuracy of classification, learning time, time which computes normalization term, and a rate of time to calculate denominator in learning.

4.1. Preliminaries

UCI KDD Archive contains datasets[6], as such Reuters-21578 Text Categorization Collection, for document classification. The dataset is composed of text and category labeled topic, people, place, and orgs. We examine the corpus containing documents labeled topic expect class of *earn* and *acq* in documents. Table 5 shows details of learning data and test data.

In feature selection, we select feature functions based on a set of words and classes satisfying $minsup = 0.15$, $minconf = 0.05$. Table 6 shows detail of features, Table 7 shows the number of feature functions for each of classes, *confidence* is not equal to zero, and Table 8 shows the number of documents of each class in both of learning data and test data documents topic as one piece of data in this experiment. We examine MEM by GIS based on random sampling and normal GIS. The number of update, feature function, learning data and test data of GIS based on random sampling has the same as the baseline.

Table 5. Details of corpus

	Learning data	Test data
File	reut2-000	
	reut2-001	
	reut2-002	reut2-014
	reut2-003	reut2-015
	reut2-004	reut2-016
	reut2-005	
Using Tag	Text : <BODY> Class:<TOPIC>	Text : <BODY> Class:<TOPIC>
The number of documents	800	336
The number of kind of class which appeared in data.	46	46

Table 6. Details of features

	Baseline	Proposed GIS
The number of feature functions	191	191
	(Including correction function)	(Including correction function)
The number of kind of extracted class	18	18
The number of kind of sets of words	32	32
MinSupport	0.15	0.15
MinConfidence	0.05	0.05

4.2. Results

In Tables 10, 11, 12, and 13, let us illustrate our results of accuracy, recall, precision and F-measure in both baseline and GIS based on random sampling as the number of sample is 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Table 9 shows results of learning time, time which computes normalization term, and a rate of time to calculate denominator in learning.

4.3. Discussion

4.3.1. Classification

Let us discuss what our results of classification means. In the case that the number of sample is 10 and 20, there are difference of results to accuracy, recall, precision and F-measure compared to the baseline with more nine kinds of classes shown in Tables 10,11,12,13. In the case that the number of sample is 30, we got the result of 8.3% and 0.3% difference to recall and F-measure in the classes of "cpi" and "sugar" but the same results to accuracy and precision. In the case that the number of sample is 40, we got result of 8.3% and 0.3% difference to accuracy, precision, recall and F-measure in classes of "trade" and "coffee". In the case that the number of sample is more 50, we got result of the same baseline .

Table 7. Class containing feature functions

Class	The number of feature functions
trade	29
crude	29
coffee	22
money-fx	17
ship	16
money-supply	15
interest	15
gnp	9
sugar	9
cpi	8
reserves	6
rubber	4
bop	3
gold	2
ipi	2
jobs	2
grain	1
wpi	1

Table 8. In learning data and test data the number of document to each of class

Class	The number of document	
	Learning data	Test data
alum	11	5
bop	12	4
carcass	1	0
cocoa	10	5
coffee	44	9
copper	12	9
cotton	7	8
cpi	18	8
cpu	2	0
crude	122	22
dfr	0	2
fuel	0	2
gas	2	1
gnp	24	6
gold	34	11
grain	17	9
groundnut	0	1
heat	4	5
housing	9	0
instal-debt	2	1
interest	42	28
inventories	1	0
ipi	11	7
iron-steel	7	4
jobs	16	7
lead	1	4
lei	5	2
livestock	7	4
lumber	3	0
meal-feed	2	1
money-fx	53	31
money-supply	35	14
nat-gas	9	5
nickel	1	1
oilseed	4	1
orange	10	6
pet-chem	4	1
potato	1	3
reserves	15	2
retail	9	2
rice	0	1
rubber	16	4
ship	60	8
strategic-metal	5	3
sugar	39	19
tea	0	1
tin	5	7
trade	88	48
veg-oil	7	0
wpi	8	6
yen	0	6
zinc	5	2

On the other hand, in case that the number of samples is less than 40, Table 13 shows parameter robustly update with GIS based on Monte Carlo method, so that the more less the number of samples decrease, the more the result is difficult. Our experiments show that decreasing the number of sample causes difference to result of baseline and GIS based on random sampling.

4.3.2. Learning Time

Let us discuss what our results of leaning time means. Table 9 shows the results of learning time and a rate of time to calculate denominator in learning. To all because baseline the computation of the probability Summarization to 46 kind of classes. In the case that samples are less than 40, each of time is less than learning time of the baseline, and is more than learning time of baseline in the case that samples are more than 50. On the

Table 9. Time which compute GIS

	BaseLine	The number of sample									
		10	20	30	40	50	60	70	80	90	100
Learning time(ms)	309471697	76025268	143312229	211311304	279246943	348131759	416124096	485274302	556640097	621945915	688576573
Time which compute Z (ms)	302529497	68669500	136319748	204458945	272371393	341020496	409203682	478347500	549176299	614543047	681476305
A rate of time to compute Z	0.9775	0.9029	0.9512	0.9675	0.9753	0.9795	0.9833	0.9857	0.9865	0.988	0.9896

Table 10. Accuracy

Class	BaseLine	The number of sample									
		10	20	30	40	50	60	70	80	90	100
heat	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nat-gas	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
hop	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
meal-feed	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Fuel	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
pet-chem	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
cotton	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ship	0.909	0.926	0.914	0.909	0.909	0.909	0.909	0.909	0.909	0.909	0.909
retail	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
crude	0.955	0.951	0.959	0.955	0.959	0.955	0.955	0.959	0.955	0.955	0.955
money-fx	0.881	0.877	0.881	0.881	0.881	0.881	0.881	0.881	0.881	0.881	0.881
gold	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955	0.955
tea	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
let	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
interest	0.864	0.864	0.864	0.864	0.864	0.864	0.864	0.864	0.864	0.864	0.864
potato	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
rubber	0.984	0.984	0.984	0.984	0.984	0.984	0.984	0.984	0.984	0.984	0.984
livestock	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
tin	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gas	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
copper	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
gram	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963
zinc	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
cpi	0.955	0.955	0.955	0.951	0.955	0.955	0.955	0.955	0.955	0.955	0.955
strategic-metal	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
groundnut	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
nickel	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
ipi	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971	0.971
yen	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
jobs	0.963	0.967	0.967	0.963	0.963	0.963	0.963	0.963	0.963	0.963	0.963
gdp	0.942	0.947	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942
reserves	0.996	0.992	0.992	0.996	0.996	0.996	0.996	0.996	0.996	0.996	0.996
oilseed	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
dlt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
rice	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
cocoa	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
alum	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
lead	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
orange	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
wpt	0.975	0.971	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975
trade	0.881	0.881	0.877	0.881	0.872	0.881	0.881	0.877	0.881	0.881	0.881
coffee	0.909	0.914	0.905	0.909	0.905	0.909	0.909	0.909	0.909	0.909	0.909
money-supply	0.942	0.938	0.938	0.942	0.942	0.942	0.942	0.942	0.942	0.942	0.942
iron-steel	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
metal-debt	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
sugar	0.868	0.885	0.872	0.868	0.868	0.868	0.868	0.868	0.868	0.868	0.868
MicroAve	0.975	0.976	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975	0.975

other hand, it takes time to sample class by the inverse function method. We consider, if the number of samples equals to the number of classes while baseline appeared, learning time by GIS based on random sampling is more than learning time of baseline.

In the sampling, GIS based on random sampling without rejection, adopt all classes generated by sampling. Thus overhead hardly happens and improve GIS. Finally, the proposed GIS expects that the we expand dimension in probability distribution, the learning time decreases learning time.

5. CONCLUSION

In this work, we have proposed an approach of improved GIS based on random sampling by which the marginal probability causes the computation of the probability Summariza-tion to all the classes.

Table 11. Recall

Class	BaseLine	The number of sample									
		10	20	30	40	50	60	70	80	90	100
heat	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nat-gas	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bop	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
meal-feed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fuel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pet-chem	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cotton	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ship	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
retail	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
crude	0.909	0.909	0.909	0.909	0.909	0.909	0.909	0.909	0.909	0.909	0.909
money-fx	0.419	0.419	0.419	0.419	0.419	0.419	0.419	0.419	0.419	0.419	0.419
gold	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lei	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
interest	0.607	0.607	0.607	0.607	0.607	0.607	0.607	0.607	0.607	0.607	0.607
potato	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rubber	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
livestock	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tin	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gas	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
copper	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gram	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
zinc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cpi	0.250	0.375	0.375	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
strategic-metal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
groundnut	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nickel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ipi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
yen	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
jobs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
gdp	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167	0.167
reserves	0.500	0.000	0.000	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
oilseed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dfr	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rice	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cocoa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
alum	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lead	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
orange	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wpi	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
trade	0.771	0.833	0.771	0.771	0.750	0.771	0.771	0.771	0.771	0.771	0.771
coffee	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
money-supply	0.786	0.786	0.786	0.786	0.786	0.786	0.786	0.786	0.786	0.786	0.786
iron-steel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
instal-debt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
sugar	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053	0.053
MicroAve	0.436	0.449	0.436	0.436	0.432	0.436	0.436	0.436	0.436	0.436	0.436

Our experimental results showed that improved MEM by GIS takes advantages to the traditional GIS: we got 68% learning time more compared to the baseline, keeping the same results to accuracy and precision, and 8.3% and 0.3% difference to recall in two kind of class.

We expect to apply our approach to other kinds of languages in classification problem.

References

- [1] Adwait Ratnaparkhi. "A Simple Introduction to Maximum Entropy Models for Natural Language Processing." 1997
- [2] J.N.DARROCH and D.RATCLIFF "GENERALIZED ITERATIVE SCALING FOR LOG-LINEAR MODELS." 1972
- [3] Bing Liu Wynne Hsu and Yiming Ma "Integraing Classification and Association Rule Mining." 1998
- [4] Ambedkar Dukkipati, Abhay Kumar Yadav and M. Narasimha Murty "Maximum entoropy model based Classification with Feature Selection." 1972
- [5] <https://sites.google.com/site/kevinbouge/stopwords-lists>
- [6] <https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html>

Table 12. Precision

Class	BaseLine	The number of sample									
		10	20	30	40	50	60	70	80	90	100
heat	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nat-gas	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bop	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
meal-feed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fuel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pet-chem	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cotton	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ship	0.111	0.143	0.118	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111
retail	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
crude	0.690	0.667	0.714	0.690	0.714	0.690	0.690	0.714	0.690	0.690	0.690
money-fx	0.542	0.520	0.542	0.542	0.542	0.542	0.542	0.542	0.542	0.542	0.542
gold	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
interest	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
potato	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rubber	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
livestock	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tin	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gas	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
copper	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gram	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zinc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cpi	0.286	0.333	0.333	0.250	0.286	0.286	0.286	0.286	0.286	0.286	0.286
strategic-metal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
groundnut	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nickel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ipi	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
yen	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
jobs	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
gdp	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100	0.100
reserves	1.000	NaN	NaN	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
oilseed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dfr	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rice	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cocoa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
alum	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lead	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
orange	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wpi	NaN	0.000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
trade	0.673	0.656	0.661	0.673	0.655	0.673	0.673	0.673	0.673	0.673	0.673
coffee	0.067	0.071	0.063	0.067	0.063	0.067	0.067	0.067	0.067	0.067	0.067
money-supply	0.500	0.478	0.478	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
iron-steel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
instal-debt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
sugar	0.067	0.091	0.071	0.071	0.067	0.067	0.067	0.067	0.067	0.067	0.067
MicroAve	0.436	0.449	0.436	0.436	0.432	0.436	0.436	0.436	0.436	0.436	0.436

Table 13. F-measure

Class	BaseLine	The number of sample									
		10	20	30	40	50	60	70	80	90	100
heat	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nat-gas	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
bop	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
meal-feed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
fuel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
pet-chem	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cotton	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ship	0.154	0.182	0.160	0.154	0.154	0.154	0.154	0.154	0.154	0.154	0.154
retail	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
crude	0.784	0.769	0.800	0.784	0.800	0.784	0.784	0.800	0.784	0.784	0.784
money-fx	0.473	0.464	0.473	0.473	0.473	0.473	0.473	0.473	0.473	0.473	0.473
gold	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lea	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
interest	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
potato	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rubber	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
livestock	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
tin	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gas	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
copper	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gram	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
zinc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cpi	0.267	0.353	0.353	0.250	0.267	0.267	0.267	0.267	0.267	0.267	0.267
strategic-metal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
groundnut	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
nickel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ipi	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
yen	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
jobs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
gdp	0.125	0.133	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
reserves	0.667	NaN	NaN	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
oilseed	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
dfr	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rice	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cocoa	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
alum	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
lead	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
orange	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
wpi	NaN	0.718	0.734	0.712	0.718	0.699	0.718	0.712	0.718	0.718	0.718
trade	0.083	0.087	0.080	0.083	0.080	0.083	0.083	0.083	0.083	0.083	0.083
coffee	0.061	0.095	0.095	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061
money-supply	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
iron-steel	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
instal-debt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
sugar	0.059	0.067	0.061	0.061	0.059	0.059	0.059	0.059	0.059	0.059	0.059
MicroAve	0.436	0.449	0.436	0.436	0.432	0.436	0.436	0.436	0.436	0.436	0.436

Quality characteristics for user-generated content

Jiri MUSTO^{a,1} and Ajantha DAHANAYAKE^a

^a*Lappeenranta-Lahti University of Technology LUT, Lappeenranta, Finland*

Abstract. Today there are vast amounts of data collected from the internet. The general public generates most data with the use of social networks. There is a need to have a comprehensive approach to characterize the quality of such user-generated data collection from the internet. The data quality characteristics accepted among database and computer science communities have definitions that are not domain-specific. Therefore, there is no clear understanding of the data quality characteristics specific to user-generated content. In this research, different user-generated content platforms are examined against the general data quality characteristics to determine which quality characteristics are essential for user-generated content. The research contributes to a list of definitions of those data quality characteristics specific to user-generated content. These definitions help identify quality characteristics useful for user-generated content platforms and their implementations. The quality of the content of Atlas of Living Australia, Twitter, YouTube, Wikipedia, and WalkingPaths is evaluated to assess the essence of the quality characteristics defined in this research.

Keywords. data collection, data quality, information quality, quality characteristics, user-generated content

1. Introduction

Content generation involving the general public is a lucrative practice today. Such user-generated content (UGC) instigates heated discussions concerning the quality of the collected data. UGC platforms, such as social media platforms, have over three billion users worldwide, and users are averaging over two hours daily on these platforms. According to an article [1], over a billion stories are created daily on Facebook.

UGC is primarily unstructured content gathered and used for a variety of purposes. Social media platforms such as Facebook, Twitter, and Instagram, crowdsourcing platforms such as Wikipedia and OpenStreetMap, and citizen science platforms such as eBird and iNaturalist, are examples of UGC gathering platforms. UGC has been demonstrated to be used for investigating customer feedback [2,3], monitoring catastrophic environmental effects [4], tracking visitors in protected areas [5], flood research [6], emergency reporting [7], future prediction [8], service quality analysis [9], managing online encyclopedia [10], and targeting advertisements and recommendations for potential customers [11,12].

Social media platforms are designed for connecting users and sharing content within the community. Most users use social media platforms to interact with others and seek

¹ Corresponding Author, Jiri Musto, School of Engineering Science, LUT University, Yliopistonkatu 34, 53850 Lappeenranta, Finland; E-mail: jiri.musto@lut.fi.

information about events, businesses, deals, and products [13]. The content shared on these platforms is mainly subjective. However, social networks have been increasingly used as sources for news among the younger generation, which are easily influenced by good or fake news [14]. Without social networks, such users may lack any knowledge of the surrounding world [15]. Furthermore, the younger generation may not read actual news, and as a result, social media has become their predominant world events and news channel.

Content generated by users is said to pertain to cases of unverified, misleading, or erroneous information that diminishes the credibility and lowers the quality of data [16–18]. Because of this, low data quality is one of the significant concerns in UGC [19] that can lead to poor decisions [20,21], or in rare cases, generate errors that eventually crash the underlying platforms [22].

Researchers and organizations have defined data quality as a collection of dimensions or characteristics [23–25]. This definition has been widely adopted and accepted [26,27]. There are over 40 different data quality characteristics, but many overlap with each other [23,25]. Quality characteristics frequently have a different definition depending on the domain; precision in healthcare has a different definition than precision in geographic information. Consequently, there is no clear consensus and agreement on what characteristics fulfill the data quality in each context and use case [24,26,28–30].

Data quality is essential because a massive amount of content can lead to wrong conclusions if the quality is compromised [31,32]. Some platforms suffer from the abuse of “quantity over quality.” One extreme example of such abuse is review bombing, where a group of people collectively gang up on one person or product [33,34]. Review bombing is a significant problem in online shops and reviews sites [35,36].

In order to overcome the ambiguity of UGC’s data quality, this research examines the following research question:

What are the quality characteristics of user-generated content?

Researchers, organizations, and communities have promoted a plethora of formulations of data quality characteristics. This research aims to establish concise formulations of data quality characteristics for UGC by applying formulations found in [23–25,37] as the base. The works are selected based on their citation count and wide usage among researchers. In addition, this research aims to provide a solution for improving the data and information quality in a citizen science platform by integrating quality characteristics into the design of a platform that collects walking path observations

Because of the influence of UGC in modern businesses [38,39], the data quality of UGC is highly contested. Therefore, this research investigates the formulation of quality characteristics of UGC based on available literature. Formal formulations are based on existing formal definitions when applicable to UGC. When hardly any formal definitions exist, the definitions are formulated based on the context and use cases.

The main contributions of this research are:

- Giving exposure to the current status of data quality in UGC platforms
- Formalization of a comprehensive but not exhaustive list of quality characteristics for the domain of UGC
- A comprehensive list of quality characteristics to choose from during the design and implementation of future UGC platforms with substantially improved data quality in the generated content.

2. Background

2.1. Data quality research

Data quality is a widely discussed topic in computer science and database technology. The systematic analysis of keyword-based article searches in scientific databases given in Table 1 accounts for the present (2020) status of data quality research.

The number of articles drastically reduces when the term “data quality” is combined with a keyword. It demonstrates that the actual research on data quality is a fraction of the many articles that mention “data quality” as a loud and popular buzzword.

Table 1. Results of keyword-based article search in scientific databases

Search terms	Scopus	IEEE	Springer	ACM
“data quality”	95069	20933	50586	4892
AND “citizen science”	1143	38	393	99
AND “big data”	5547	1466	3726	721
AND “remote sens*”	8 715	2497	3672	2
AND “crowdsourc*”	2796	311	1001	0
AND “user generated”	705	30	574	186
AND “social media”	22327	150	2262	520
“data quality defin*”	20	42	59	0
“data quality model”	407	123	193	39
“data quality dimension”	1154	62	455	49
“data quality characteristic”	40	13	86	2
“data quality framework”	319	56	109	12

Some widely cited data quality research works belong to the 1990s [20,25,40], and new research works and standards extend them [23,24,41,42]. Researchers and standards define data quality as:

- Multidimensional, divided into characteristics
- Contextual
- Characteristics’ importance is subjective
- Quality is measured through the characteristics.

[25] generalizes the data quality characteristics under four categories: intrinsic, contextual, accessibility, and representational characteristics. ISO standard [24] categorizes data quality characteristics into inherent, inherent and system dependent, and system dependent categories.

Different assessment processes and frameworks have proposed specific steps and metrics to evaluate quality and improvement ideas when quality is low [30]. An extensive survey of existing data quality frameworks is provided in [43]. However, there is a lack of actual assessment or evaluation methodology [27,44]. Some frameworks have implemented data quality evaluation for one specific use-case, such as social media, but the final test only consists of one characteristic [30].

2.2. User-generated Content

Data quality in UGC has been explored since social networking, and social media platforms took off during the 21st century. As data quality is contextual, definitions for each characteristic in UGC can be different from other domains. Moreover, even within the UGC domain, there are different definitions for the same characteristics [45–47].

Data quality in UGC is crucial as regular citizens generate the content. The quality of data in UGC is often questioned as users are not experts. As a result, UGC is more vulnerable to low-quality data compared to other domains [48]. For this reason, some projects use specific tools, like sensors, for data collection to make data more reliable compared to just human-computer interaction [49]. Several methods for improving UGC have been proposed, such as participant selection [50], task allocation [51], and reputation models [52].

3. Data quality characteristics

ISO quality characteristics [24] are used as the starting point to develop a list of UGC data quality characteristics. These characteristics are presented in Table 2.

Table 2. List of initial data quality characteristics

ISO Data quality characteristics [24]	
accessibility	availability
completeness	compliance
consistency	confidentiality
credibility	currentness
efficiency	portability
precision	recoverability
semantic accuracy	syntactic accuracy
traceability	understandability

From the ISO characteristics, *accessibility*, *availability*, *efficiency*, *portability*, and *recoverability* are discarded as they are related to the underlying system and not data itself. The list in Table 2 is further extended to accommodate the UGC domain’s data quality characteristics with contributions from domains of general data quality, social media, and big data. These additional characteristics are presented in Table 3.

Table 3. Data quality characteristic from other domains

Extended data quality [23,40,53]	Social media [5,54]	Big data [27,55,56]
objectivity	privacy	relevance
provenance	usability	value
timeliness		volume

To formulate practical definitions for specific characteristics, it is essential to be clear with the general understanding of the term, limiting misinterpretation. Therefore, the formal data quality definitions of the characteristics listed are formulated using existing literature.

Accuracy: Closeness between data values v and v_0 , where v_0 is the correct representation of what the data value v aims to represent. Based on syntactic and semantic accuracy [23].

Syntactic accuracy: Closeness of words in the text to a reference vocabulary. K is the number of words, w_i is a word in the text, and V is the vocabulary used in the text (1)[37].

$$\text{syntactic acc} = \frac{\sum K_{\text{closeness}}(w_i, V)}{K} \quad (1)$$

Semantic accuracy: How correctly the meaning of values represents real-world facts. An object identification problem where α and β are a pair of tuples to be matched, M is the set that contains a record of similar existing pair, U is the set that represents nonmatch and \underline{x} is a random vector of n number of attributes, and $p()$ is the probability of matching (2)[23,57].

$$\langle \alpha, \beta \rangle \in \begin{cases} M & \text{if } p(M|\underline{x}) \geq p(U|\underline{x}) \\ U & \text{otherwise} \end{cases} \quad (2)$$

Completeness: Completeness of a tuple with respect to the values of all its fields where T_v is the number of null values in a tuple and N_v is the total number of values in a tuple (3)[23,58].

$$\text{completeness} = 1 - \frac{T_v}{N_v} \quad (3)$$

Consistency: Violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. g is the data value, and N is the number of rules for g (4)[59].

$$r_n(g) = \begin{cases} 0, & \text{if } g \text{ fulfills rule } r_n \\ 1 & \text{else} \end{cases}, \text{cons}(g) = 1 - \frac{\sum_{n=1}^N r_n(g)}{N} \quad (4)$$

Credibility: How data are accepted or regarded as true, real, and credible, where $dist$ is the distance between the sensor s and entity e , and d_{max} is the maximum distance acceptable (5)[60].

$$\text{credibility} = \begin{cases} 1 - \frac{dist}{d_{max}} & \text{if } d(s, e) < d_{max} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Objectivity: Data is unbiased and impartial, where E is evidence, H is a hypothesis (assumed value), and $p()$ denotes the probability (6)[61].

$$w(E, H, H') = \log \frac{p(E|H)}{p(E|H')} \quad (6)$$

Precision: Precision refers to the amount of detail that can be discerned in space, time, or theme. Using Levenshtein edit distance where a and b are the given values, i and j are the indexes (7)[57,62].

$$\text{lev}_{a,b}(i, j) = \begin{cases} \max(i, j) & \text{if } \min(i, j) = 0, \text{ otherwise} \\ \min \begin{cases} \text{lev}_{a,b}(i-1, j) + 1 \\ \text{lev}_{a,b}(i, j-1) + 1 \\ \text{lev}_{a,b}(i-1, j-1) + 1_{(a_i \neq b_j)} \end{cases} & \end{cases} \quad (7)$$

Volume: Appropriate amount of data: the extent to which the quantity or volume of available data is appropriate. Sample size formula where z is z-score, e is the margin of error, p is standard deviation, and N is population size (8)[63].

$$\text{sample size} = N * \frac{\frac{z^2 * p * (1-p)}{e^2}}{\left[N - 1 + \frac{z^2 * p * (1-p)}{e^2} \right]} \quad (8)$$

Compliance: Defining and evaluating the compliance between data and schemas measure of relationship (similarity, relatedness, distance, etc.) between two entities.

Where a and b are values of elements in minimum distance and \bar{a} and \bar{b} are means of all elements (9)[64].

$$compliance (degree of variance) = \frac{\Sigma(a-\bar{a})(b-\bar{b})}{\sqrt{\Sigma(a-\bar{a})^2 \Sigma(b-\bar{b})^2}} \quad (9)$$

Currentness: Currency concerns how promptly data are updated with respect to changes occurring in the real world (10)[23].

$$currentness = Age + (DeliveryTime - InputTime) \quad (10)$$

Timeliness: Data is sufficiently up to date for the task at hand. *Volatility* is the defined length of how long data remains valid (11)[23].

$$timeliness = \max\{0, 1 - \frac{currentness}{volatility}\} \quad (11)$$

Privacy: Data is hidden or concealed from others. S is the sensitivity of a data item, and V is the visibility in a given context, and R is relatedness. a , b and c are real numbers (12)[65].

$$PrivacyRisk_{(i,j)} = \frac{s^a \times v^b_{(i,j)}}{r^c_{(i,j)}} \quad (12)$$

Relevance: The extent to which data are applicable and helpful for the task at hand. n is the number of words in a sentence, m is the number of characters in a word, and *WordSimilarity* is the similarity between two words between 0 and 1 (13)[66].

$$SentenceSimilarity(Q, Q') = \frac{1}{n} \sum_{1 \leq j \leq n} (\max_{1 \leq i \leq m} WordSimilarity(w_j, w'_i)) \quad (13)$$

Usability: A collection of other characteristics characterized by usability aspects, verifiability, imperfection, and integration (14)[67].

$$usability = avg(accuracy + credibility + completeness + currentness + relevance + granularity + accessibility) \quad (14)$$

Value: The extent to which data are beneficial and provide advantages from their use (15)[68]

$$DataValue(t) \geq (GatherCost + MaintainCost + AccessCost)/GB/yr * RetentionPeriod \quad (15)$$

Confidentiality: Data is available to authorized persons when and where needed (especially in the medical field). W_c is the weight of confidentiality for a subsystem, x_i is a dependency score for a subsystem, and n is the number of subsystems in an information security system (16)[69].

$$confidentiality = \frac{\sum_{i=1}^n W_{c_i} \times x_{s_i}}{\sum_{i=1}^n W_{c_i}} \quad (16)$$

Granularity: Granularity concerns the ability to represent and operate on different levels of detail in data, information, and knowledge located at their appropriate level. Shannon entropy in terms of Hartley entropy for partition granularity (17)[70].

$$granularity = \log|U| - \sum_{i=1}^n \frac{|X_i|}{|U|} \log(|X_i|), U \text{ is a universal set and set } X \subseteq U \quad (17)$$

Traceability: The extent to which data are well documented, verifiable, and easily attributed to a source. R is a source, Ω is a set of R , $E(\Omega)$ is a measure of uncertainty, and λ is the number of reports (18)[71].

$$Network \text{ traceability entropy } (NTE), E^\lambda = \sum_{\Omega:|\Omega|=\lambda} E(\Omega) / \binom{|R|}{\lambda} \quad (18)$$

Provenance: Provenance of a resource is a record of metadata containing descriptions of the entities and activities involved in producing and delivering or otherwise influencing a given object. Q is a query, I is an instance, and t is a tuple in U (19)[72].

$$\begin{aligned} \text{whyProvenance}(Q, I, t) &= \{J \in I \mid t \in Q(J)\} \\ \text{whereProvenance}(\{u\}, I, t) &= \begin{cases} (A : \emptyset)_{A \in U}, & \text{if } t = u \\ \perp, & \text{otherwise} \end{cases} \end{aligned} \quad (19)$$

$$\text{howProvenance}(Q, I, t) = Q^{K_{\text{How}}}(I_{\text{How}})t$$

Understandability: The ease with which data can be comprehended without ambiguity and be used by a human information consumer (20)[73].

$$\begin{aligned} \text{understand.} &= -0.33 * \text{Abstraction} + 0.33 * \text{Encapsulation} + 0.33 * \\ &\text{Coupling} + 0.33 * \text{Cohesion} - 0.33 * \text{Polymorphism} + 0.33 * \\ &\text{Complexity} - 0.33 * \text{DesignSize} \end{aligned} \quad (20)$$

Readability: Reading easiness, the ease of understanding written text using Gunning-Fox index (21)[60,74].

$$\text{readability} = 0.4 * \left[\left(\frac{\text{words}}{\text{sentence}} \right) + 100 * \left(\frac{\text{complexwords}}{\text{words}} \right) \right] \quad (21)$$

4. Case studies: user-generated content creating platforms

Citizen science platforms are famous for using public submitted context-specific content. There are over 1000 citizen science platforms (<https://scistarter.org>), with content related to wildlife, environment, and city management. [75] gives a detailed overview of close to 100 citizen science platform evaluations.

Atlas of Living Australia (ALA) (<https://ala.org.au/>) is an Australian citizen science platform for plant and wildlife monitoring. ALA has integrated another citizen science platform called *iNaturalist* (<https://inaturalist.ala.org.au/>), allowing data from iNaturalist to be sent to ALA. Features of ALA can be generalized because most citizen science platforms operate using similar functionalities. In citizen science platforms, citizens send reports with a varying number of fields that often include multimedia. Citizens may give a username when submitting reports, and reports can be updated later. Reports can have automated tests for quality and be voted by the community. Some issues specific to citizen science platforms, such as content submitted by regular citizens making credibility questionable, personal details of users are sometimes shown to the public, and some reports stay incomplete.

Twitter (<https://twitter.com/>) is a social media platform where users share short texts and images called tweets. Users can comment, like or reshare other people's tweets. These actions provide context on how well tweets are received. On Twitter, tweets and accounts can be made private. In addition, users have a number of followers and followed, and tweets can have hashtags that work like keywords. Most content comes from individuals without any source material. Thus it is challenging to define credible information. Tweets are occasionally in another language or nonsensical, and some people make fake accounts pretending to be someone else.

Worldometer (<https://www.worldometers.info/>) is a crowdsourcing platform that collects and aggregates information from multiple sources. The sources vary from news articles and healthcare-operated sites to third-party organizations. Worldometer is widely referenced as a reliable real-time information provider during the Covid-19 pandemic. In Worldometer, information is primarily numbers and based on a source. Worldometer is continuously updated and considered to be reliable based on the sources it uses. The information is presented in text, graphs, and tables. However, some information requires users to contribute, leading to incompleteness. The credibility of information must be

checked before sharing it with the public, and inaccurate information from users requires further administrator reviews.

Wikipedia (<https://www.wikipedia.org/>) is an online encyclopedia where registered users create and modify content, and more reputable volunteers act as moderators. Wikipedia requires a source before it accepts content as valid information. In addition, Wikipedia has a specific style that articles must follow. Because community updates and moderates Wikipedia, it is updated fast in the native language compared to translations. Most information is written clearly and understandably.

Nevertheless, there are cases when information is not correct in Wikipedia or correct information is not accepted because of the source. Sometimes, the source material's credibility can be questionable, and volunteer administrators' opinions may be reflected in the accepted content. Few articles are left incomplete because of the lack of contributions.

YouTube (<https://www.youtube.com/>) is a video-sharing platform owned by Google. Anyone can view public videos, but only registered users can upload new videos. Videos are not allowed to infringe any copyright laws, and the content must not be harmful or hateful. YouTube has similar characteristics to Twitter, such as videos have a number of views, and they can be liked/disliked and commented on. As regular citizens make most videos, the information may not be credible, and there is no guarantee of objectivity. It is challenging to validate the official channels from other forms of propaganda, and some users purposefully report videos they do not like.

Each of the introduced platforms has different use-cases and contexts. Content in Wikipedia and Worldometer are meant for public consumption, but their context is different. Content in citizen science platforms is used for research and context changes from one platform to another. Twitter and YouTube are used for connecting with others and sharing subjective content. So Twitter and YouTube have the same context, but the provided content is vastly different. With this in mind, the public uses all introduced platforms. Thus the platforms are expected to have some level of quality in the content.

Table 4 presents the mapping of data quality characteristics listed in Section 3 to the described UGC platforms. Characteristics are examined from the platform's context (credibility relates to the user's credibility). The data quality of UGC is governed by the quality of the content requested from the user. The context defines the limits and requirements for the data quality that the content needs to fulfill. Some characteristics require a specific use-case for the content, such as relevance and value. Each characteristic is given a value as follows:

- 1: The platform takes into consideration by requiring specific content.
- 0: The platform does not take into consideration. The user can submit content without any limitations.
- ?: Unclear if the system considers that characteristic or not.
- +/-: Situation dependent and only applicable to specific use cases.

Table 4 shows that Twitter and YouTube care less about information correctness than the other UGC platforms. Twitter has no regard for completeness, but ALA, Wikipedia, and Worldometer have minimum requirements for submissions. In addition, there are situations when a data quality characteristic needs a degree of variation. In ALA, timeliness is sometimes essential in situations where the information must be from specific periods. When extracting data from the UGC platform, it is beneficial to know the quality of extracted data. When using Twitter and YouTube data, objectivity must be evaluated separately because the platforms place no importance on objectivity.

Table 4. Data quality characteristic mapping to platforms that curate UGC

Data quality characteristics	ALA	Twitter	Worldometer	Wikipedia	YouTube	Explanations of the characteristics in terms of information gathered by the platform
Syntactic accuracy	1	0	1	0	0	User submits information in the syntax expected by the system
Semantic accuracy	1	0	1	1	0	User submits information that follows semantic rules set by the system
Completeness	1	0	1	1	0	The system expects the user to submit a minimum amount of information
Consistency	0	0	0	0	0	Information is consistent in comparison to multiple users input
Credibility	1	1	1	1	1	User's credibility
Objectivity	1	0	1	1	0	User submits objective information
Precision	1	0	1	+/-	0	Information is detailed
Volume	1	1	1	1	1	Similar information from different sources
Compliance	?	?	?	?	?	Information is compliant with a standard
Currentness	1	1	1	1	1	Information is current
Timeliness	+/-	0	0	0	0	Information is from the correct time
Privacy	1	1	0	0	1	Personal information is not displayed
Relevance	1	1	1	1	1	User submits relevant information to the topic
Usability	1	+/-	1	1	+/-	Information is usable by others
Value	1	+/-	1	1	+/-	Information has value for others
Confidentiality	0	0	0	0	0	Sensitive information is inaccessible
Granularity	+/-	0	0	0	0	Information is split into specific parts
Traceability	1	1	1	1	1	Information origins are known
Provenance	0	0	0	0	0	Changes to information are known
Understandability (or readability)	1	1	1	1	1	Information is understandable (or readable)

Based on the above analysis and observations, the quality characteristics specific to UGC can be formulated as follows:

Traceability: How well the content is attributed to a specific source and time.

Twitter and YouTube record the user and time when content is created. In Worldometer and Wikipedia, the content has a specific source. Wikipedia tracks the user who has added or edited content. Similarly, citizen science platforms track the time created, the place where the content relates, and who submits it.

Credibility: How credible the content is based on who is giving the content.

In social media, credibility is subjective even when official channels of credible organizations or people are the creators. Credibility can be based on three factors: number of likes or followers, community opinion based on the comments, and user verification. For Wikipedia and Worldometer, credibility is based on the source material and in citizen science, credibility is based on community opinion and administration.

Currentness: How promptly content is updated with respect to changes occurring in the real world.

Twitter is designed for content to be created and shared as soon as possible. On YouTube, most content creators want to create content based on current hot topics. Wikipedia's purpose is to have current facts. Citizen science platforms' purpose is to get current information. Finally, Worldometer is continuously updating its content.

Relevance: How relevant the given content is to the platform context.

Worldometer, Wikipedia, and citizen science all have a specific purpose, and all three expect to get relevant content from users. YouTube and Twitter have opinion-based content, and the content always relates to some topics making it arguably relevant.

Accuracy: Accuracy is the closeness of given content to the expected content. Based on syntactic and semantic accuracy.

Syntactic accuracy: Closeness of the content syntax that the user gives depending on the platform context.

Twitter, Wikipedia, and YouTube all accept various types making information always syntactically accurate. Only Worldometer and citizen science limit what a user can give to ensure syntactic accuracy.

Semantic accuracy: How correctly the information within the content matches the real-world facts.

Twitter and YouTube are not interested in semantic accuracy. Worldometer and Wikipedia require sources to check semantic accuracy, and in citizen science, there are limits to what content can be given to have some semantic accuracy.

Completeness: How complete content is and not missing important information depending on the platform context.

Social media operates on more opinion-based content, and there is no minimum requirement of what needs to be given. In Wikipedia, short or incomplete information is marked by the platform automatically. Citizen science and Worldometer expect specific information at a minimum before any information can be sent.

Usability: How usable the content is based on the platform context. It is affected by accuracy, completeness, and credibility.

On Twitter and YouTube, content created by official channels of organizations is meant to be used by the public. Wikipedia and Worldometer are meant to be used by everyone, and unusable content is quickly removed. On the other hand, citizen science projects are meant for research.

Value: How useful the content is and provides advantages from its use.

Citizen science content is meant for research purposes that will lead to some value. Worldometer and Wikipedia are meant to be information sources making their content valuable. Twitter and YouTube provide value when combining a massive amount of content. However, individually, tweets and videos do not provide much value.

Understandability (and readability): How easily the information from the content can be comprehended without ambiguity by a human consumer within the platform context (and how easy written text is to read and comprehend).

Wikipedia is meant for the public, and many complex things are explained so that a novice can comprehend. Worldometer provides information in various formats making their content understandable. Citizen science often has maps and graphs to increase understandability. Only social media content can be challenging to understand, but more understandable content will be more popular and promoted.

Objectivity: How unbiased and impartial the content and its information are.

Twitter and YouTube are meant for opinion sharing making objectiveness non-essential. Worldometer and Wikipedia require sources to ensure objectiveness. In citizen science, content is subjective but made more objective by using community opinion.

Privacy: How much of the user's personal information is concealed.

Worldometer and Wikipedia do not handle private information, and social media platforms allow users to hide their information. In citizen science, content usually includes a location, but users are not required to use their names.

Volume: The amount of similar information given by multiple users.

All case platforms want to have a high volume of information. Wikipedia and Worldometer commend having more than one source. When collecting opinions from social media, having multiple people with similar opinions is valuable for researchers. In citizen science, if no one else agrees on a report, it is quickly deemed untrustworthy.

Precision: How detailed the given content is in the platform context.

Precision is not considered on Twitter or YouTube. For citizen science, precision is considered whenever there is location-based information given. In Wikipedia, precision is situational, but in most cases, no precision is required. On the other hand, Worldometer does not want any ambiguity in its information; thus, precise information is expected.

The listed characteristics can be used to define the data or information quality characteristics in UGC. Information is the content received from users, while data is the content stored in the database [76]. Only *precision* is not applicable in the context of information.

5. Integration Of Quality Characteristics Into The Citizen Science Platform: WalkingPaths

A citizen science web platform called WalkingPaths integrates the essential data quality characteristics listed in Table 2 into its design [76]. The platform is developed using ReactJS for the frontend and NodeJS for the backend with a MongoDB database. Mongoose middleware is used to enforce syntax restrictions on data.

The platform collects walking path information from citizens in Finland. Citizens are asked to fill a simple form consisting of the path's location and condition, and they are given an option to send an image with the observation. The data is collected from March 2020 to August 2020, and the final data set consists of 108 observations.

When integrating quality characteristics into the design, it is necessary to decide where these characteristics should be implemented. Characteristics should be integrated into the data model as well as the user interface. The database may store information related to these characteristics, but the interface is responsible for checking and enforcing them. Characteristics can be integrated into the user interface by limiting or extracting specific information from the content provider's input. For instance, the address is complete if geolocation exists. Similarly, the characteristics can be added as constraints in the database. A more detailed description of the integration of quality characteristics is found in [76].

Figure 2 shows the database schema using a snowflake model [77] of the platform WalkingPaths. In the center is the fact table *WalkingPathObservation*, and it is connected to several dimension tables. A snowflake schema can be easily transformed into a relational data model. Several data quality characteristics are integrated into the model as separate attributes, and these are bolded and cursive. These include precision, accuracy (syntactic and semantic), completeness, volume, credibility, privacy, objectivity, and traceability. The characteristics can store relevant quality evaluations.

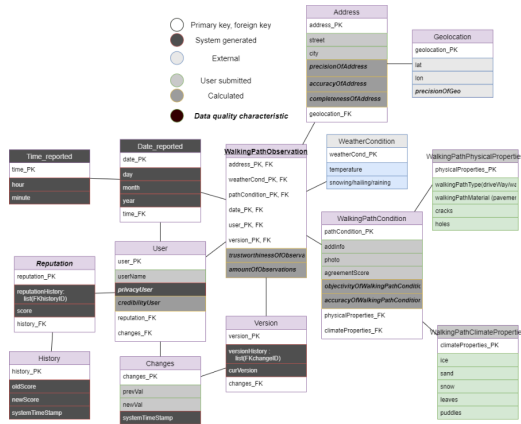


Figure 2. Snowflake schema for WalkingPaths

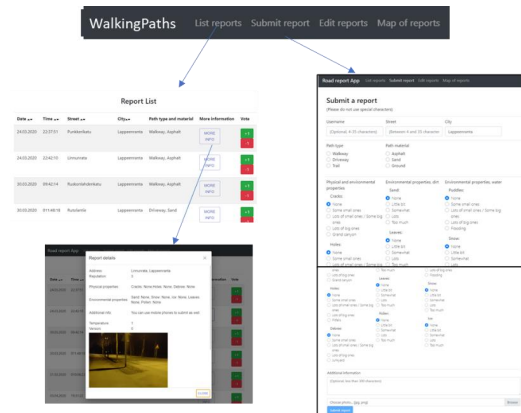


Figure 3. WalkingPaths list of observations and report submission

Figure 3 shows the transition using the navigation bar to listing observations and submitting new reports. The observation list only shows minimal details for each report, such as location and time. Users can open a *More information* pop-up window to reveal other information. Reports can be up-/downvoted, but as the platform does not require registration, some restrictions have been implemented in the voting mechanism to reduce misuse. Most choice boxes in the report window have predetermined values to guarantee

each report's completeness. Only two choice boxes do not have a value, but the report cannot be submitted before some value is given to both of them. The usage of choice boxes is an excellent method to increase the report's syntactic and semantic accuracy while enabling the content provider to know what to look for before submitting anything. Finally, additional information can be typed in the text box.

The report editing view is similar to submitting a report with the additional search box for finding existing reports. Map view presents a map where observations are shown as markers. More detailed figures are found in [76].

6. Case Study: Data Quality In User-generated Content Platforms

Data quality is evaluated by subjecting a data set from each platform to specific queries related to each quality characteristics presented in Section 4. The queries are performed using the data analytics platform RapidMiner (<https://rapidminer.com/>), a commercial software designed for data mining, analytics, and machine learning. Table 5 presents the general RapidMiner queries for each of the characteristics. The *value* characteristic for each data set is calculated based on other characteristics to simplify the definition.

RapidMiner query results are given as values between 0 and 1. Values indicate the percentage of correct data entities for each characteristic (conform to the given query). These resulting values are presented in Table 6. Not applicable (NA) results are deemed as zero because if something is not applicable, it does not exist. The number of entities in each data set is given in the headers of Table 6.

Table 5. General RapidMiner queries for DQ characteristics

Characteristic	General query	(Data mining) Technique
Syntactic accuracy	Data entities correspond to the expected syntax and format defined in the data set. This information is based on the headers and what data is expected, and in what format.	Text/content mining. Compare value syntax to expected (integer, string, date) and filter out incorrect values. Compare the number of correct values to the total number.
Semantic accuracy	Data is semantically correct compared to what is expected based on the headers	Value comparison. Headers define what data should be, for example, "date," "name," "country." Each value is checked to see if they are actual dates, countries, names.
Completeness	Each data set is checked for missing values for completeness.	Filter missing values and compare the amount to total (automated functionality)
Credibility	The credibility of the content provider giving the information.	Reputation model and calculation compared to the average score
Objectivity	Objectivity is based on how objective given information is. If multiple sources agree on the information, it is more likely to be objective.	Count how many entities from different sources/content providers have the same information and how many are only from singular content providers/sources.
Volume	For each data set, the volume is checked from similar data entities compared to all entities. The similarity is only based on a few attributes.	Count how many entities from different sources/content providers have relatable information based on selected attributes and how many are only from singular content providers/sources.
Currentness	Data has given a date/time. Compare that to the time data was extracted from the database	Content mining and comparison

Privacy	Privacy is measured based on the number of personal information stored with the data.	Filter out content providers whose possible real names are given and compare them to the total amount (text mining)
Relevancy	The relevance of the data to the given context regardless is the data correct or not.	Data comparison to given relevance factor such as the topic.
Usability	Usability is based on the context of usage for each data set	Content mining and comparison
Value	Value depends on the user. In this research, value = (Syntactic + Semantic + Credibility + Relevancy + Usability + Understandability) / 6	Calculation based on other characteristics
Traceability	Each data set provided attributes for time, location, and content provider that are checked for traceability.	Count how many entities have a valid time, location, and content provider/source compared to all entities
Understandability	Understandability is based on the content of information, in general, readability. Unreadable texts/characters and undefined acronyms reduce the understandability	Text mining of invalid words.

Table 6. Query result of data quality characteristics

Characteristic	WalkingPaths 108 entities	ALA 894 entities	Twitter 6012 entities	YouTube 750 entities	Wikipedia 19 797 entities	Worldometer 2996 entities
Syntactic accuracy	1.00	0.95	1.00	1.00	0.96	1.00
Semantic accuracy	0.96	0.96	0.93	NA	1.00	1.00
Completeness	1.00	0.72	0.89	0.99	0.95	1.00
Credibility	0.74	NA	0.32	0.82	0.32	NA
Objectivity	0.54	0.23	0.19	0.11	0.50	NA
Volume	0.36	0.42	0.61	0.69	NA	NA
Currentness	1.00	0.29	1.00	1.00	1.00	1.00
Privacy	1.00	0.86	0.67	1.00	1.00	1.00
Relevancy	1.00	1.00	1.00	1.00	1.00	1.00
Usability	1.00	0.79	NA	NA	0.85	1.00
Value	0.95	0.77	0.68	0.47	0.81	0.83
Traceability	1.00	0.76	0.66	0.67	0.67	1.00
Understandability	1.00	0.93	0.82	NA	0.72	1.00

Results show that chosen platforms do not support all quality characteristics, and Twitter and YouTube performed the worst out of all. These are social media platforms designed for opinion sharing and not for credible data collection and information sharing. It is necessary to consider integrating data quality characteristics into the design during its implementation to accommodate the maximum number of quality characteristics.

Overall, WalkingPaths scored similarly to Worldometer, aside from a few significant aspects. Semantic accuracy is less in WalkingPaths than in Worldometer because there are some misspellings in the addresses given in WalkingPaths. Semantic accuracy could be improved with easy addition to the user interface where a content provider is recommended the address during typing. However, if a similar platform is extended outside of one country, the list of cities and street names would inflate drastically. Other significant differences are credibility, objectivity, and volume that do

not apply to Worldometer. Compared to other platforms, WalkingPaths is better in objectivity and credibility and only loses in volume.

WalkingPaths achieved higher scores in everything except volume in comparison to ALA. ALA has been available for many years, so it is understandable for WalkingPaths to have a lower volume score. For completeness, currentness, and traceability, the most significant difference in scores is missing dates and times in ALA data, and a lot of data entities were before the year 2000. In some instances, time formatting changed. ALA data provided some information on the source of observations, but there are no methods to determine if the source is credible, making credibility unapplicable. While it can be argued that ALA performs worse because it collects different kinds of observations, the same techniques used in the development of WalkingPath can be utilized in any type of observation. The difference in observation types is negligible as both platforms' underlying principle stays the same.

7. Discussion

To improve the quality of information, the method of the collection must be improved. The improvement can be made by implementing checks or limits within the user interface to reduce misinformation drastically. In Worldometer, users can only give a limited amount and type of content through the user interface, thus ensuring that the information sent to the system is at least of decent quality. Social media platforms could use specific filters for information searches that are based on different criteria. Twitter already has hashtags implemented, but these are always user-defined. There could be some reserved hashtags that, when used, Twitter could enforce some quality control checks for the content shared while using the specific hashtag.

Another way to improve the collection is to remodel the user interface. Most users give content based on what is asked in UGC platforms. What is asked defines what is received, not just having checks or limits applied to the user interface but designing it to answer specific questions. Even if the input is not limited, most users will unconsciously avoid giving misinformation when answering questions.

Not all users may care about the quality. The content's quality could be evaluated by the application based on the selected quality characteristics. The results of these evaluations could be embedded, for example, as system-generated data to Twitter API. This way, regular users would not see these evaluation results, and they would only be visible in raw Twitter data. Another possibility would be to add an option for regular users to see these evaluation results, similar to the history of edits Facebook has implemented. It is not shown unless selected explicitly by the user.

A platform where quality characteristics of UGC are integrated into the user interface and data model is presented in [76]. The same platform is used in this research to evaluate the design against non-citizen science UGC platforms. The integration of quality characteristics brings advantages and disadvantages to the content provided.

Some of the advantages of implementing quality characteristics are:

- Receiving higher quality content from users
- Determine the quality of content
- Enables content filter for users (if necessary)
- Possible to show others the quality score of a given content (if necessary)

- The quality characteristics implementation can justify reusing data collected from the platform

Some disadvantages are:

- May limit what content users can share
- May limit the way content is shared and used
- May affect how data is stored

Designers and developers of UGC platforms should consider having some data and information quality control implementations. During the design phase, these decisions should be made to fully utilize appropriate methods and ensure the quality of the content shared through the platform. The disadvantages of such an approach, depending on how the characteristics are implemented. For example, when implementing checks for content completeness, it is possible to either require absolute completeness or allow incompleteness. If absolute completeness is required, users cannot submit any incomplete content. Thus, the content they share is limited. If incomplete content is allowed, the user may share this content and later edit it, or the system can mark the content as incomplete for others. It is possible to avoid the disadvantages through design decisions. Currently, Worldometer requires absolute completeness, while Wikipedia allows incomplete content.

The research presented has some limitations, such as:

- Only a limited number of platforms have been examined
- Only the data quality characteristics available in research works have been considered, but the list can be extended by integrating experiences from the practice.
- The definitions presented are only applicable to the UGC domain and are not designed to be used for other domains

8. Conclusion

Quality of content is an essential part of any platform that collects content from non-experts with varying levels of expertise and knowledge. Unfortunately, UGC platforms are considered untrustworthy because the quality of content is questionable [16–18].

It is necessary to understand what quality is to improve data quality. Data and information quality must be defined for each domain, and there are no existing definitions for UGC. This research provides an extensive but not exhaustive list of quality characteristics with definitions specifically tailored for UGC. The importance of quality characteristics depends on the platform, and different contexts for the platform will change what characteristics should be emphasized.

Considering and integrating quality characteristics during the design of a platform has been presented in [75,78]. The articles provide general guidelines on how the quality characteristics can be implemented in the design of a platform. A citizen science platform for collecting WalkingPaths information is created to experiment with the proposed methodology, and the quality of collected content is evaluated against existing citizen science platforms [76].

Results show that integrating quality characteristics into the design increases the overall quality of UGC platforms. Most characteristics can be easily integrated into the design without significant changes. This method can be used in any platform and even applied to an existing platform if necessary. The most important part is identifying which characteristics are essential in each platform, and this has to be done by considering the

context where the information will be used. The definitions of quality characteristics for UGC are helpful instruments for identifying essential characteristics for a UGC platform's content.

This research contributed to the formulation of specific quality characteristics definitions specifically for the UGC domain that collects content using social networks and web technology. The presented definitions are based on existing definitions of general data quality characteristics but modified for UGC usage. Quality characteristics depend on the context of the platform. Even within the same domain, different contexts for the platform will change what characteristics should be emphasized. This research contributes to building a cumulative tradition of building a sound set of UGC's quality characteristics.

References

- [1] Influencer Marketing Hub. 42 Essential Social Media Statistics for 2020 [Internet]. Influencer Marketing Hub. 2020 [cited 2020 Apr 28]. Available from: <https://influencermarketinghub.com/social-media-statistics-2020/>
- [2] Ranjan S, Sood S, Verma V. Twitter Sentiment Analysis of Real-Time Customer Experience Feedback for Predicting Growth of Indian Telecom Companies. In: Proceedings - 4th International Conference on Computing Sciences, ICCS 2018. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 166–74.
- [3] Mariani M, Di Fatta G, Di Felice M. Understanding Customer Satisfaction with Services by Leveraging Big Data: The Role of Services Attributes and Consumers' Cultural Background. *IEEE Access*. 2019;7:8195–208.
- [4] Ahmouda A, Hochmair HH, Cvetojevic S. Using Twitter to Analyze the Effect of Hurricanes on Human Mobility Patterns. *Urban Sci*. 2019;3(3):87.
- [5] Tenkanen H, Di Minin E, Heikinheimo V, Hausmann A, Herbst M, Kajala L, et al. Instagram, Flickr, or Twitter: Assessing the usability of social media data for visitor monitoring in protected areas. *Sci Rep*. 2017 Dec 14;7(1):17615.
- [6] Arthur R, Boulton CA, Shotton H, Williams HTP. Social sensing of floods in the UK. *PLoS One*. 2018;13(1).
- [7] Ludwig T, Reuter C, Pipek V. Social Haystack: Dynamic Quality Assessment of Citizen-Generated Content during Emergencies. *ACM Trans Comput Interact*. 2015;22.
- [8] Asur S, Huberman BA. Predicting the future with social media. In: Proceedings - 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010. 2010. p. 492–9.
- [9] Haryani CA, Hidayanto AN, Budi NFA, Herkules. Sentiment Analysis of Online Auction Service Quality on Twitter Data: A case of E-Bay. In: 2018 6th International Conference on Cyber and IT Service Management, CITSM 2018. IEEE; 2019. p. 1–5.
- [10] Bykau S, Korn F, Srivastava D, Velegakis Y. Fine-grained controversy detection in Wikipedia. In: Proceedings - International Conference on Data Engineering. 2015. p. 1573–84.
- [11] Ouyang S, Li C, Li X. A peek into the future: Predicting the popularity of online videos. *IEEE Access*. 2016;4:3026–33.
- [12] Mensah S, Hu C, Li X, Liu X, Zhang R. A Probabilistic Model for User Interest Propagation in Recommender Systems. *IEEE Access*. 2020;8:108300–9.
- [13] Whiting A, Williams D. Why people use social media: a uses and gratifications approach. *Qual Mark Res An Int J*. 2013 Aug 30;16(4):362–9.
- [14] Viviani M, Pasi G. Credibility in social media: opinions, news, and health information—a survey. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2017 Sep 1;7(5):e1209.
- [15] David CC, San Pascual RS, Torres ES. Reliance on Facebook for news and its influence on political engagement. *PLoS One*. 2019 Mar 1;14(3).
- [16] Polk T, Johnston MP, Evers S. Wikipedia Use in Research: Perceptions in Secondary Schools. *TechTrends*. 2015 May 1;59(3):92–102.
- [17] Syed-Abdul S, Fernandez-Luque L, Jian WS, Li YC, Crain S, Hsu MH, et al. Misleading health-related information promoted through video-based social media: Anorexia on youtube. *J Med Internet Res*. 2013 Feb 13;15(2):e30.
- [18] Goodman J, Carmichael F. US election 2020: “Rigged” votes, body doubles and other false claims

- [internet]. BBC News. 2020 [cited 2020 Oct 25]. Available from: <https://www.bbc.com/news/54562611>
- [19] Lewandowski E, Specht H. Influence of volunteer and project characteristics on data quality of biological surveys. *Conserv Biol*. 2015;29(3):713–23.
- [20] Redman TC. Data quality for the information age. Artech House; 1996. 303 p.
- [21] Warth J, Kaiser G, Kügler M. The impact of data quality and analytical capabilities on planning performance: insights from the automotive industry. In: *Wirtschaftsinformatik Proceedings*. 2011.
- [22] Laranjeiro N, Soydemir SN, Bernardino J. Testing web applications using poor quality data. In: *Proceedings - 7th Latin-American Symposium on Dependable Computing, LADC 2016*. 2016. p. 139–44.
- [23] Batini C, Scannapieco M. *Data quality : concepts, methodologies and techniques*. Springer; 2006. 262 p.
- [24] ISO. ISO/IEC 25012:2008 Software engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Data quality model [internet]. ISO; 2008 [cited 2019 Jan 9]. Available from: <https://www.iso.org/standard/35736.html>
- [25] Wang RY, Strong DM. Beyond accuracy: What data quality means to data consumers. *J Manag Inf Syst*. 1996;12(4):5–34.
- [26] Arolfo F, Vaisman A. Data Quality in a Big Data Context. In: *ACM SIGMOD Record*. Springer International Publishing; 2018. p. 159–72.
- [27] Cai L, Zhu Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Sci J*. 2015 May 22;14(0):2.
- [28] Batini C, Rula A, Scannapieco M, Viscusi G. From data quality to big data quality. *J Database Manag*. 2015;26(1):60–82.
- [29] DAMA UK. The Six Primary Dimensions for Data Quality Assessment - Defining Data Quality Dimensions [Internet]. 2013 [cited 2019 Jan 9]. Available from: <https://www.dqglobal.com/wp-content/uploads/2013/11/DAMA-UK-DQ-Dimensions-White-Paper-R37-1.pdf>
- [30] Immonen A, Pääkkönen P, Ovaska E. Evaluating the Quality of Social Media Data in Big Data Architecture. *IEEE Access*. 2015;3:2028–43.
- [31] Bayraktarov E, Ehmke G, O'Connor J, Burns EL, Nguyen HA, McRae L, et al. Do big unstructured biodiversity data mean more knowledge? *Front Ecol Evol*. 2019;7(JAN).
- [32] Sadiq S, Indulska M. Open data: Quality over quantity. *Int J Inf Manage*. 2017;37(3):150–4.
- [33] Hall C. Valve fought more than 40 'review bombs' on Steam in 2019 - Polygon [Internet]. Polygon. 2020 [cited 2020 Oct 12]. Available from: <https://www.polygon.com/2020/2/6/21126787/steam-review-bombs-policy-effectiveness-valve>
- [34] Kuchera B. The anatomy of a review bombing campaign - Polygon [Internet]. Polygon. 2017 [cited 2020 Oct 12]. Available from: <https://www.polygon.com/2017/10/4/16418832/pubg-firewatch-steam-review-bomb>
- [35] Hawkins J. Yelp vs Google: How they deal with fake reviews [internet]. 2018 [cited 2020 Oct 12]. Available from: <https://searchengineland.com/yelp-vs-google-how-do-they-deal-with-fake-reviews-307332>
- [36] The Guardian. How TripAdvisor changed travel [internet]. 2018 [cited 2020 Oct 12]. Available from: <https://www.theguardian.com/news/2018/aug/17/how-tripadvisor-changed-travel>
- [37] Batini C, Scannapieco M. *Data and Information Quality: Dimensions, Principles and Techniques*. Cham: Springer International Publishing; 2016. 500 p. (Data-Centric Systems and Applications).
- [38] Vincent N, Johnson I, Sheehan P, Hecht B. Measuring the Importance of User-Generated Content to Search Engines. Vol. 13, *Proceedings of the International AAAI Conference on Web and Social Media*. 2019 Jul.
- [39] Brunt CS, King AS, King JT. The influence of user-generated content on video game demand. *J Cult Econ*. 2020 Mar 1;44(1):35–56.
- [40] Strong DM, Lee YW, Wang RY. Data quality in context. *Commun ACM*. 1997;40(5):103–10.
- [41] Moraga C, Moraga MA, Calero C, Caro A. SQuaRE-aligned data quality model for web portals. In: *Proceedings - International Conference on Quality Software*. 2009. p. 117–22.
- [42] Redman TC, Fox C, Levitin A. Data and data quality. *Understanding Information Retrieval Systems: Management, Types, and Standards*. 2011. 269–284 p.
- [43] Cichy C, Rass S. An overview of data quality frameworks. *IEEE Access*. 2019;7:24634–48.
- [44] Lin S, Gao J, Koronios A, Chanana V. Developing a data quality framework for asset management in engineering organisations. *Int J Inf Qual*. 2007;1(1):100–26.
- [45] Bordogna G, Carrara P, Criscuolo L, Pepe M, Rampini A. On predicting and improving the quality of Volunteer Geographic Information projects. Vol. 9, *International Journal of Digital Earth*. Taylor & Francis; 2016. p. 134–55.
- [46] Alabri A, Hunter J. Enhancing the quality and trust of citizen science data. In: *Proceedings - 2010*

- 6th IEEE International Conference on e-Science, eScience 2010. IEEE; 2010. p. 81–8.
- [47] Lee D. Big Data Quality Assurance Through Data Traceability: A Case Study of the National Standard Reference Data Program of Korea. IEEE Access. 2019;7:36294–9.
- [48] Kaur J, Singh J, Sehra SS, Rai HS. Systematic literature review of data quality within openstreetmap. In: Proceedings - 2017 International Conference on Next Generation Computing and Information Systems, ICNGCIS 2017. 2018. p. 159–63.
- [49] Chin MJ, Babashamsi P, Yusoff NIM. A comparative study of monitoring methods in sustainable pavement management system. In: IOP Conference Series: Materials Science and Engineering. 2019.
- [50] Xiong J, Chen X, Tian Y, Ma R, Chen L, Yao Z. MAIM: A Novel Incentive Mechanism Based on Multi-Attribute User Selection in Mobile Crowdsensing. IEEE Access. 2018;6:65384–96.
- [51] Wei X, Wang Y, Tan J, Gao S. Data Quality Aware Task Allocation with Budget Constraint in Mobile Crowdsensing. IEEE Access. 2018 Aug 30;6:48010–20.
- [52] Pang L, Li G, Yao X, Lai Y. An Incentive Mechanism Based on a Bayesian Game for Spatial Crowdsourcing. IEEE Access. 2019;7:14340–52.
- [53] Pipino LL, Lee YW, Wang RY. Data Quality Assessment. Commun ACM. 2002;45(4):211–8.
- [54] Smith M, Szongott C, Henne B, Von Voigt G. Big data privacy issues in public social media. IEEE Int Conf Digit Ecosyst Technol. 2012;
- [55] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage. 2015 Apr 1;35(2):137–44.
- [56] Chen M, Mao S, Liu Y. Big data: A survey. In: Mobile Networks and Applications. Kluwer Academic Publishers; 2014. p. 171–209.
- [57] Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans Knowl Data Eng. 2007 Jan;19(1):1–16.
- [58] Blake R, Mangiameli P. The effects and interactions of data quality and problem complexity on classification. J Data Inf Qual. 2011;2(2).
- [59] Heinrich B, Klier M, Schiller A, Wagner G. Assessing data quality – A probability-based metric for semantic consistency. Decis Support Syst. 2018;110:95–106.
- [60] Firmani D, Mecella M, Scannapieco M, Batini C. On the Meaningfulness of “Big Data Quality” (Invited Paper). Data Sci Eng. 2016;1(1):6–20.
- [61] Reiss J, Sprenger J. Scientific Objectivity. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. Winter 2011. Metaphysics Research Lab, Stanford University; 2017.
- [62] Levenshtein VI. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soc physics, Dokl. 1965;10:707–10.
- [63] Krejcie R V, Morgan DW. Determining Sample Size for Research Activities. Educ Psychol Meas. 1970;30(3):607–10.
- [64] Hulitt E, Vaughn RB. Information system security compliance to FISMA standard: A quantitative measure. Telecommun Syst. 2010;45(2–3):139–52.
- [65] Senarath A, Grobler M, Arachchilage NAG. A model for system developers to measure the privacy risk of data. In: HICSS. 2019.
- [66] Yang F, Feng J, Fabbriozio G Di. A Data Driven Approach to Relevancy Recognition for Contextual Question Answering [Internet]. 2006 [cited 2020 May 15]. Available from: <http://www.ask.com/>
- [67] Cross I, Joana P. Evaluating the Usability of Aggregated Datasets in the GIS4EU Project [Internet]. 2010 [cited 2020 May 15]. Available from: <https://www.directionsmag.com/article/2130>
- [68] Wrabetz J. Measuring the economic value of data [internet]. Network World. 2017 [cited 2020 May 17]. Available from: <https://www.networkworld.com/article/3221387/measuring-the-economic-value-of-data.html>
- [69] Gallaher SM. An Approach For Measuring The Confidentiality Of Data Assured By The Confidentiality Of Information Security Systems In Healthcare Organizations [Internet]. University of Central Florida; 2012 [cited 2020 May 17]. Available from: <http://library.ucf.edu>
- [70] Yao MX. Granularity measures and complexity measures of partition-based granular structures. Knowledge-Based Syst. 2019 Jan 1;163:885–97.
- [71] Lu X, Horn AL, Su J, Jiang J. A Universal Measure for Network Traceability. Omega (United Kingdom). 2019 Sep 1;87:191–204.
- [72] Cheney J, Chiticariu L, Tan W-C. Provenance in Databases: Why, How, and Where. Found Trends R Databases. 2009;1(4).
- [73] Dexun J, Peijun M, Xiaohong S, Tiantian W. Functional Over-Related Classes Bad Smell Detection and Refactoring Suggestions. Int J Softw Eng Appl. 2014 Mar 31;5(2):29–47.
- [74] Gunning R. The technique of clear writing. Toronto: McGraw-Hill; 1952.
- [75] Musto J, Dahanayake A. Improving Data Quality, Privacy and Provenance in Citizen Science Applications. In: Frontiers in Artificial Intelligence and Applications. IOS Press; 2020. p. 141–60.
- [76] Musto J, Dahanayake A. An Approach to Improve the Quality of User-Generated Content of Citizen

Science Platforms. *ISPRS Int J Geo-Information*. 2021 Jun 25;10(7):434.

- [77] Teorey T, Lightstone S, Nadeau T, Jagadish HV. Business Intelligence. In: *Database Modeling and Design*. Elsevier; 2011. p. 189–231.
- [78] Fox TL, Guynes CS, Prybutok VR, Windsor J. Maintaining Quality in Information Systems. *J Comput Inf Syst*. 1999;40(1):76–80.

Personalized Virtual Campus Journey Adaptation to User Controlled Experience

Bakhtiyor ESANOV^{a,1} and Ajantha DAHANAYAKE^a
^a*Lappeenranta-Lahti University of Technology LUT*

Abstract. Increasingly, educational institutes are migrating into mobile platforms and mobile app technology to communicate, advertise, and disseminate education-related information to their stakeholders using virtual campus journey mobile apps. Campus journey mobile apps generally provide standardized generic customization to their user base, incorporating a list of favorite touchpoints based on the users' frequent behaviors. In the literature, personalization, and customization, definitions are mixed-up and inter-changed with no proper separation of these two concepts. In this research, the personalized virtual journey is examined according to the user's preference. It includes creating and updating the personalized virtual campus journey path as an activity of the user and having it as an integral part of the personalized virtual campus journey application. This research presents the concept structure, design, implementation, and evaluation results of the personalized virtual campus journey mobile app development according to the user's preferences with the user given the ability to control his or her own virtual journey experience.

Keywords. Virtual journey, virtual campus, virtual campus app design, personalization, explicit user-controlled experience.

1. Introduction

A 2020 global survey among e-commerce decision-makers revealed that 78 percent of global e-commerce companies are planning on investing more into the realization of personalization [1]. As of October 2020, nearly 4.66 billion people (59 percent of the global population) are active internet users. Ninety-one percent (91%) of these users use mobile channels to access the internet [2], making mobile devices the most used devices in internet traffic.

The virtual campus journey mobile applications have appeared in the market since 2011 [3]. The virtual campus journey apps run on mobile devices, smartwatches, desktop/laptop computers, and tablet devices. Recent developments in software engineering have made it possible for software engineers to design and build apps that can run on different devices and platforms and share the same logic across the devices and platforms.

The virtual campus journey apps have become an integral part of the campus population. The users and stakeholders of virtual campus journey applications are staff,

¹ Corresponding Author, Bakhtiyor Esanov, School of Engineering Science, Lappeenranta-Lahti University of Technology LUT, Yliopistonkatu 34, 53850, Lappeenranta, Finland; E-mail: bakhtiyor.esanov@student.lut.fi.

students, parents, guests, visitors, and prospective students. Virtual journey mobile app is bringing all campus services into a single app. Getting all services into a single place is a difficult task for software engineering. A campus has many services such as student services, library, booking room, cafes/restaurants, events, and many more. University may have more than one campus, and campus(es) can be located in different cities or countries. Each campus has its services that make the personalization of the virtual campus journey complex. Therefore, few universities have mobile applications [3]. Because planning, designing, and building an excellent virtual campus journey app is challenging. Software engineers need to collaborate with students, staff, university administration, visitors/guests of the university, prospective students, and parents in such a successful implementation path. One of the main challenges is that the university's services are legacy systems. This study closely investigated the personalization of the virtual campus journey mobile apps. For this purpose, testing and comparing mobile applications of Times Higher Education World University Ranking's [4] first top 20 universities are considered. Next, mobile applications for both IOS and Android devices are conducted, including a virtual campus journey personalization approach. This research presents a novel concept of virtual campus journey apps that is personalized according to individual user preferences giving control to users to organize their preferred virtual journey. The implementation is limited to the IOS and Android version of the mobile app.

The following research question articulates the research:

- I. What constitutes a personalized virtual campus journey mobile app that overcomes the problems in designing personalization?

The research methodology followed the Design Science Research (DSR) framework [5]. The paper consists of a review of related works, foundational principles of the conceptualization of services systems engineering and touchpoint design, designing and building a prototype of virtual campus journey mobile app implementing a personalized virtual campus journey, feature evaluation, benchmarking leading to a comparison of mobile educational apps, and conclusions.

2. Related Works

Mobile applications provide additional functionality with each new version release. Displaying the many features of the application at the same time makes it challenging to use it. Every user has their preferences. Some features of the application might not be relevant to the user's interest. Users may not want to see some features in their application interface or screen. There is a need to manage the interface, providing users with easy access to the features they use, making personalization increasingly important [6]. The classical characterization of personalization is presented in table 1.

Table 1. The typical classical characterization of personalization

Characteristics	Action
Customize the appearance[7] [8]	The user chooses a color cover or attaches a sticker on the phone; the change persists in the system [7]. Having a customizable appearance with changeable skins or color schemes would make the app feel more personal[8].

Set preferences [8]	Recommending products based on preferences of similar-minded consumers[7].
Receiving notification on those that are of interest	The user indicates which areas she/he is interested in; the system memorizes this and will notify the user about items that fall within these defined categories [7]
Personalizing ringing tone	The user selects, orders, or composes the tone; the system memorizes the change [7].
Changing font size	The user enlarges the font size or increases the contrast of the front; the system memorizes the change. [7]

Personalization is a process of changing the system's functionality, interface, and information content to increase its user's relevance. Personalization is initiated by the system or user [7]. The system saves the user's preferences and adapts them to the user's needs. Adaption of the system is essential in the personalization process [5].

There are many types of personalization approaches. According to Li et al. [9] the types of personalization approaches are:

- Content-based personalization recommends items similar to those a user liked in the past.
- Collaborative personalization recommends items that other similar people have liked.
- Hybrid personalization is the combination of content-based personalization and collaborative personalization approaches.
- Social network-based personalization collects information from people in the same social network as the focal consumer and then recommends items liked by these people to the consumer [9].

Orji et al. [10] compare two types of personalization approaches in their research. These approaches are system-controlled personalization and user-controlled personalization. The difference between these two approaches depends on who controls the personalization. The adaptive interfaces are system-controlled, whereas adaptable interfaces are user-controlled [5]. The system's adaptability to user preferences with user-controlled interfaces is an essential contributor to virtual campus journey personalization.

3. Concept Structure: Touchpoints and Service Engineering fundamentals

The research [3] introduced the conceptual model of a virtual campus journey. The study [11] presented a survey conducted among the LUT University students and staff to identify the user preferences of a virtual campus journey. A total of 75 responses are collected from this survey involving university staff, students, visitors, parents, and software developers' close collaboration. These research initiatives and use cases contributed to the formulation of touchpoints and service engineering fundamentals.

Universities have many services and more than one campus location offering different services. Examples of university services are student service, student union, library, courses, timetables, Learning Management System (LMS), news services, university mail service, restaurants/cafes/dining places. These services are located on different servers and domains. When designing a personalized virtual journey app, service systems engineering is an important conceptual underpinning [12].

A **virtual campus journey** is the discovery process and orchestration of the essential services offered at an education institute for communication, information retrieval, and dissimulation.

- Virtual campus journey is a mobile app implemented at the root level of the mobile device's interface, just like any other mobile app.

A **touchpoint** is a point of interaction offered to the user at the mobile app's user interface.

- A touchpoint is either an **atomic** or **composite service touchpoint** and is a feature of a mobile app.
- Atomic touchpoints provide access to a virtual atomic service that composes a physical service.
- Atomic service consists of only one single virtual service and a single touchpoint.
- Composite services can consist of more than one virtual service but consist of a single touchpoint, a composite service touchpoint.
- Therefore, an atomic service touchpoint represents a single service, whereas a composite service touchpoint opens up the touchpoints of the services bundled in the composite service.

The **virtual campus journey app** is a **composite service touchpoint** at the root of the mobile device's user interface. It guides the access to physical services that represent the educational services to respective stakeholders.

The **virtual campus journey's virtual space** consists of touchpoints and virtual services defined in the website of an institute and their seamless orchestration to offer physical service available at the institute.

The **physical space** of an educational institute accommodates the actual physical services of an educational institute.

The **virtual space mapping to the physical space** is accommodated by the coordination between the touchpoints and virtual services.

A **physical service** is an offering to a stakeholder (customer, consumer, etc.) that the stakeholder initiates via a physical interface. In so doing begins one or more underlying actions, whose normal and expected result is a delivered response [12]. Every physical service has an associated business process in any organization. A physical service has defined clients or stakeholders. A campus has many services

- The stakeholder can be internal or external to the organization that provides the service.

The **virtual services** corresponding to physical services are made available via a website/webpage of today's organizations via a web interface.

A **virtual educational service** can be of **virtual atomic service** implemented in the virtual space (website) corresponding and interacting with the educational institutes' physical space of service offerings.

The **composite virtual services** are services bundles of aggregated services mapping to a composite virtual touchpoint.

- Composite virtual touchpoints expand into a set of atomic touchpoints during the interaction with a user.
- The **Atomic virtual services** map to atomic touchpoints to reach an independent service via virtual space.

This way of conceptualization allows the advantage of defining virtual services to be acquired in Webservices and APIs.

Figure 1 shows the simple mobile touchpoints and virtual services tree.

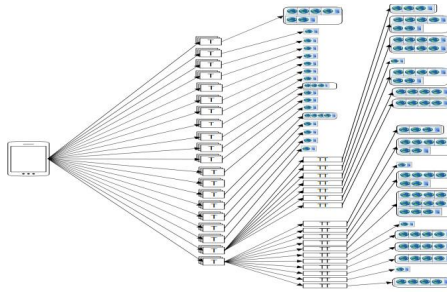


Figure 1. Mobile touchpoints and services tree.

Virtual services are subject to the following characteristics [13]:

- Loosely coupled: no direct dependencies among individual virtual services.
- Abstract: a virtual service hides its logic from the outside world.
- Reusable: support potential reuse.
- Composable: comprise of other virtual services to form a composite virtual service.
- Stateless: not maintains state information specific to an activity.
- Discoverable: let a virtual service consumer discover and understand service descriptions.
- Virtual services flow the data to touchpoints.
- The user uses services by interacting with touchpoints. Therefore, the features are considered touchpoints.

Through service discovery and composition, SOA-based application development identifies three stakeholders [13]: A service provider (or developer) is the party who develops and hosts the service. A service consumer is a person or program that uses a service to build an application. A service broker helps service providers publish and market their services and helps service consumers discover and use the available services. Figure 2 illustrates the summary of the service systems engineering and touchpoint association at the user interface to accommodate the user preference integration foundation to virtual campus journey apps.

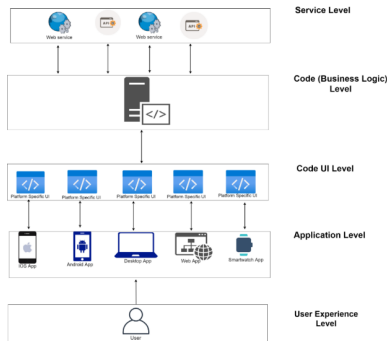


Figure 2. Virtual Journey Software Architecture Levels

The concept of personalization is defined as "a process that changes the functionality, interface, information access, content, or distinctiveness of a system to increase its relevance to an individual or a category of individuals" [18]. The core elements of personalization are:

- "a purpose or goal of personalization."
- "what is personalized." Four aspects of information systems may be personalized: the information (content), the presentation of information (user interface), the delivery method (channel), the action
- "the target of personalization." The target can be a group of individuals or a specific individual.

Personalization enables users to acquire information specific to their "needs, goals, knowledge, interests or other characteristics" [19]. The personalization principle of the virtual campus journey app is based on the following criterion:

- **The goal of personalization:** is to allow the user to select and orchestrate a virtual journey according to the user's preferences of services and a user being able to control his or her actions of controlling the revision, updates, and order of services self.
- **What is personalized:**
 - *The user interface* of the virtual journey app allows the user to select useful and dynamically adaptable touchpoints according to user needs.
 - *Actions:* of drag-drop, create a journey with favorite order of services, update and revisions of favorite services, activate an improved personalized user experience.
 - *Content:* the services that have a use for the user selected by the user
 - *Channel:* the delivery method as touchpoints
 - **The target of personalization:** is to provide a personalized user experience that supports the user preference adaptation adaptable to user-controlled actions in a dynamic user interface by:
 - Settings
 - Layout

- Navigation home screen
- Dynamic, adaptable Touchpoints

Therefore, this approach's novelty can dynamically adapt the user interface at the touchpoint level – which is the user-experience level. Then service system adapts according to the user's preference and provides a seamless and improved user experience.

4. LUT Prototype: The LUT Mobile App

This section describes how the above-described novel personalized service and touchpoint engineering is integrated into the implementation.

4.1 Technical Details

Xamarin.Forms framework is used for designing the virtual journey mobile app. Xamarin.Forms[14] is an open-source mobile UI framework Microsoft for building iOS, Android, & Windows apps with .NET from a single shared codebase using C# programming language. Xamarin.Forms share the code across the platforms. Business Logic, Platform APIs, and User Interface codes are shared across the platforms by Xamarin Forms. The main reason for choosing Xamarin.Forms is designing and building the same mobile app for IOS and Android devices. The design uses the Model-View-ViewModel (MVVM) software architectural pattern [15].

4.2 Platform decisions

The mobile apps for IOS and Android versions are created. The application runs on both platforms. Because IOS has a different user interface than Android, a decision is made to create a cross-platform mobile app to analyze IOS users' preferences at the same time. IOS and Android share most of the mobile app marketplace.

4.3 Platform Specific UI.

Platform-specific UIs have resources such as photos, which have been used for the application. In this phase, touchpoints are created according to the platform. This is a crucial point for developers. A cross-platform solution can be good practice; however, these solutions do not give 100% representative native user interface and native codes for the platform.

USE CASE: LUT University has two campuses in different cities. Each campus has services that are reachable within that campus.

Personalization of mobile app can be in 3 ways in a mobile app:

- Navigation Home Screen (NHS) Personalization (Figure 5)
- Touchpoint personalization which happen in touchpoint, user can select services in this touchpoint; for example, users can select their preferred touchpoints representing news, food, and events. (Figure 6)
- Settings-personalization happens in settings (Figure 3)

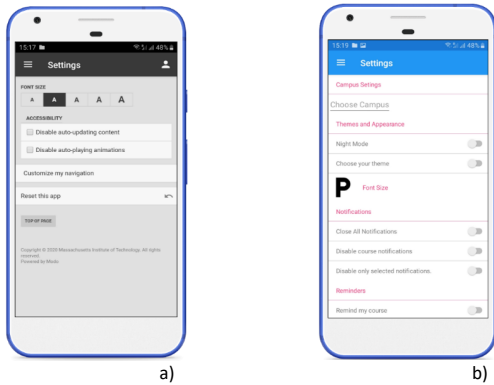


Figure 3. Comparison of ordinary Settings section with LUT mobile a) Ordinary settings in benchmarked mobile apps b) Settings section in virtual tour journey – prototype.

Figure 4 shows the Navigational Home Screen interface.

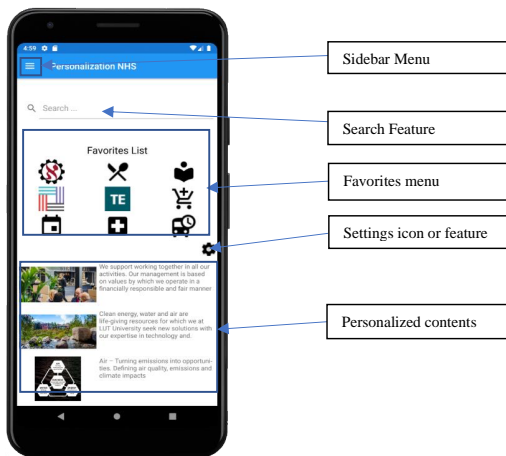


Figure 4. Navigational Home Screen

Touchpoints are considered as features according to the above-formulated foundation.

The navigation home screen is the main interface where users can see important touchpoints.

The settings button is where all the touchpoints are located. The user triggers the setting icon by clicking on it, and a list of touchpoints appears. **Figure 5** shows the home screen touchpoint personalization. In the first step, the user clicks the settings icon, and all touchpoints appear. In the second step, the user can drag and drop touchpoint(s) to the favorites list, remove touchpoint(s), adjust the position of touchpoints in favorites, and remove touchpoints from the favorites list. In the final step, the user clicks the "Done" button to save the changes.

Personalized content is the content that appears according to the user's preferences.

Sidebar Menu is a menu where the settings and other essential touchpoints are located.

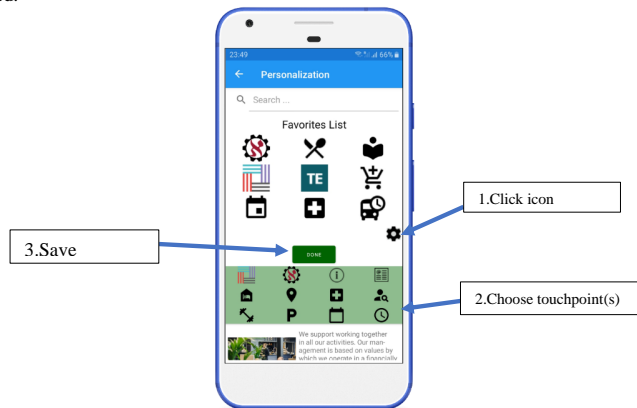


Figure 5. Navigational Home Screen Personalization of LUT Mobile

5. Comparison and evaluation

The virtual journey mobile app prototype is evaluated and compared to other universities' mobile apps.

A close look at mobile apps with touchpoints such as courses, campus maps, directories, timetables, and libraries to be considered for this evaluation. Mobile apps not related to education purposes removed from the list. A decision is made to compare the first 20 universities of Times Higher Education World University Ranking 2021[4] campus journey mobile apps. The universities selected have their mobile apps for the use of their students. In the end, a total of 14 universities' mobile apps are selected. The main reason for selecting top-ranking universities is that most universities do not own their virtual campus journey apps. Few universities have mobile apps in Finland, and the data on personalization is not enough for making the comparison because they have not considered personalization in their present implementations.

The selected universities are compared against the LUT Mobile app prototype. Table 2 shows the comparison criteria, mobile apps, and evaluation against other virtual campus journey apps.

Comparison criterion:

- **Home Screen Menu type:** 8 out of 14 benchmarked universities have Grid layout interfaces. Grid [16, 17] layout organizes rows and columns, which can have proportional or absolute sizes.
- **Settings:** The survey in previous research [11] is conducted among university students and staff and is informed whether a mobile app should accommodate a settings section. Surprisingly 81.34% of respondents agree that every mobile application needs to contain the settings section. Therefore, a settings personalization is included in the LUT Mobile prototype as the criterion for the comparison.
- **Navigational Home Screen Personalization:** Navigational Home Screen Personalization helps to personalize the mobile app's main screen, whether the user can see when he/she opens the mobile app every time.
 - o Some mobile apps (MIT Mobile, Princeton Mobile, Harvard College, and Berkeley) have my favorites option for personalizing Navigational Home Screen.
 - My favorites can provide a list where user can remove, add his/her preferred services and these selected services appear at the top of the Home screen or Sidebar menu.
 - However, the user can not adjust the position.
 - o U of T Mobile has a different design for Home screen personalization. In U of T Mobile, the user can remove, add new to the home screen from a list of available apps with the extra feature of changing the position of the list through the settings. It does not support dynamic adjustment at the user's touchscreen.
 - o Imperial Mobile Navigational Home, Screen level personalization, offered to add, remove, and adjust the list on the same page using extra functionality.
 - o LUT Mobile's Navigational Home Screen is shown in **Figure 4**. LUT Mobile Navigational Home Screen is a touchscreen where the user can drag-drop, add new touchpoints, adjust touchpoints on the same screen to favorite touchpoints, and give the user the control to personalize his/ her most preferred educational services.
- **Touchpoint personalization:** The bookmarking of the touchpoints is not considered personalization. Touchpoint personalization must include dynamic flexibility and adaptability as described in the personalization principles in section 3. Giving them control of personalization to be controlled by the user. The drag-drop functionality of touchpoints that is capable of:
 - changing the order of the favorite touchpoints,
 - add new touchpoints that are available in the Mobile Virtual Journey App
 - remove unnecessary touchpoints and
 - making the favorite touchpoints available each time a user opens the virtual journey app.
 - Further personalization of services with a touchpoint through the touch screen.

Figure 6 shows the news touchpoint of the LUT mobile app. Users can add, remove news categories in news touchpoints, and users' preferred news appear at touchpoints. **Figure 7** shows the steps of touchpoint personalization.

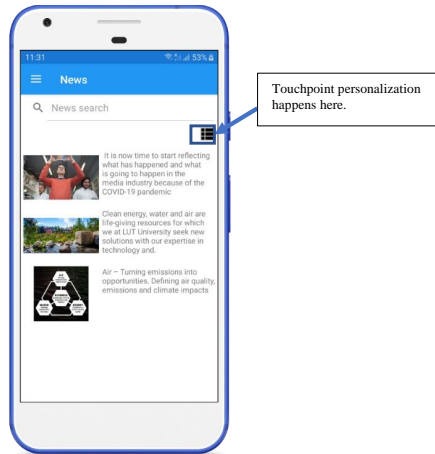


Figure 6. LUT Mobile News Touchpoint

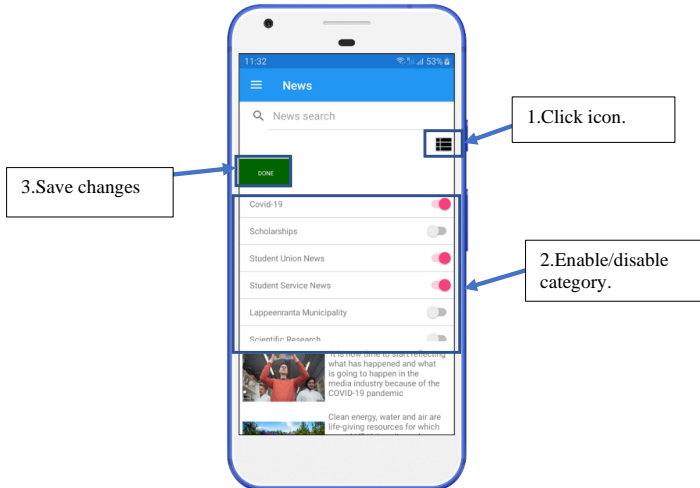


Figure 7. LUT Mobile News Touchpoint Personalization

Table 2. Personalization Comparison

University Name	Mobile App Name	Settings Personalization	Navigational Home Screen Personalization	Dynamic and Adaptable Touchpoint Personalization	Partial Adjustment of Touchpoints	(My)Favorites as Personalization	Home Screen Menu type
Stanford University	Stanford Mobile	No	No	No	No	No	Mix/Static
Harvard University	Harvard College	Yes	No	No	No	Yes	Mix/Static
Massachusetts Institute of Technology	MIT Mobile	Yes	No	No	No	Yes	Grid/Static
University of California, Berkeley	Berkeley	Yes	No	No	No	Yes	Grid/Static
Yale University	Yale	No	No	No	No	No	Mix/Static
Princeton University	Princeton Mobile	Yes	No	No	No	Yes	Grid/Static
The University of Chicago	Uchicago	Yes	No	No	No	No	Grid/Static
Imperial College London	Imperial	Yes	Yes	No	No	No	Grid/Static
Johns Hopkins University	JHUMobile	Yes	No	No	No	No	Grid/Static
University of Pennsylvania	Penn Mobile	Yes	No	No	No	No	Mix/Static
ETH Zurich	ETH Zurich	No	No	No	Yes	No	Mix/Static
UCL	UCL Go!	Yes	Yes	No	No	No	Grid/Static
University of Toronto	U of T Mobile	No	Yes	No	No	Yes	Grid/Static
Duke University	Duke Mobile	No	No	No	No	No	Mix/Static
LUT University (Prototype)	LUT Mobile	Yes	Yes	Yes	Yes	Yes	Grid/Dynamic

6. Conclusion

This research introduces the concept structure, design, implementation, and evaluation results of the personalized virtual campus journey application design and development with dynamic user-controlled adaptation to user's preferences. Within this

research, LUT Mobile virtual campus journey mobile app is designed and developed. An evaluation is conducted by comparing the personalization criteria of the created mobile app with other virtual campus journey mobile apps. It is observed that most of the virtual campus journey mobile apps are not capable of providing a dynamic user-controlled personalization.

This research's follow-up realizes the Personalized Virtual Campus Journey mobile app at LUT University based on the presented study and prototype. The test results of LUT Mobile's usability will be published in an upcoming research article.

References

- [1] Statista, *Global e-commerce companies investing in personalization 2020* | Statista. [Online]. Available: <https://www.statista.com/statistics/1174164/investing-personalization-ecommerce-companies-worldwide/> (accessed: Jan. 14 2021).
- [2] Statista, *Internet users in the world 2020* | Statista. [Online]. Available: <https://www.statista.com/statistics/617136/digital-population-worldwide/> (accessed: Jan. 14 2021).
- [3] B. Esanov and A. Dahanayake, "Visitor Journey Application Development for Omni-Channels," in *Frontiers in Artificial Intelligence and Applications, Information Modelling and Knowledge Bases XXXII*, M. Tropmann-Frick, B. Thalheim, H. Jaakkola, Y. Kiyoki, and N. Yoshida, Eds.: IOS Press, 2020.
- [4] Times Higher Education (THE), *World University Rankings 2021*. [Online]. Available: https://www.timeshighereducation.com/world-university-rankings/2021/world-ranking#!/page/0/length/25/sort_by/rank/sort_order/asc/cols/stats (accessed: Jan. 15 2021).
- [5] A. Hevner and S. Chatterjee, "Design Science Research in Information Systems," in *Integrated Series in Information Systems, Design Research in Information Systems*, A. Hevner and S. Chatterjee, Eds., Boston, MA: Springer US, 2010, pp. 9–22.
- [6] L. Findlater and J. McGrenere, "A Comparison of Static, Adaptive, and Adaptable Menus," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*, vol. 6, no. 1, pp. 89–96, 2004, doi: 10.1145/985692.985704.
- [7] J. Bloom, "Personalization: A Taxonomy," in *CHI '00 Extended Abstracts on Human Factors in Computing Systems (CHI EA '00)*, pp. 313–314, 2000, doi: 10.1145/633292.633483.
- [8] S. Garrido et al., "Young People's Response to Six Smartphone Apps for Anxiety and Depression: Focus Group Study," *JMIR mental health*, vol. 6, no. 10, e14385, 2019, doi: 10.2196/14385.
- [9] S. Li and E. Karahanna, "Peer-Based Recommendations in Online B2C E-Commerce: Comparing Collaborative Personalization and Social Network-Based Personalization," in *2012 45th Hawaii International Conference on System Sciences*, Maui, HI, USA, 012012, pp. 733–742.
- [10] R. Orji, K. Oyibo, and G. F. Tondello, "A Comparison of System-Controlled and User-Controlled Personalization Approaches," in *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, Bratislava Slovakia, 07092017, pp. 413–418.
- [11] B. Esanov and A. Dahanayake, "Virtual Campus Journey: Personalization vs Customization," *Lecture Notes in Networks and Systems (2021, in press)*, 2020.
- [12] Keen Peter G.W. and Sol Henk G., "Decision Enhancement Services: Rehearsing the Future for the Decisions that Matter," pp. 1–184, 2008, doi: 10.3233/978-1-58603-837-3-i.
- [13] S. Yau and H. An, "Software Engineering Meets Services and Cloud Computing," *Computer*, vol. 44, no. 10, pp. 47–53, 2011, doi: 10.1109/MC.2011.267.
- [14] Microsoft, *Cross-platform with Xamarin | .NET*. [Online]. Available: <https://dotnet.microsoft.com/apps/xamarin/cross-platform> (accessed: Jan. 23 2021).
- [15] C. Anderson, "The Model-View-ViewModel (MVVM) Design Pattern," *Pro Business Applications with Silverlight 5*, pp. 461–499, 2012, doi: 10.1007/978-1-4302-3501-9_13.
- [16] *Layouts | Android Developers*. [Online]. Available: <https://developer.android.com/guide/topics/ui/declaring-layout> (accessed: Jan. 14 2021).
- [17] Davidbritch, *Xamarin.Forms Grid - Xamarin*. [Online]. Available: <https://docs.microsoft.com/en-us/xamarin/xamarin-forms/user-interface/layouts/grid> (accessed: Jan. 14 2021).
- [18] H. Fan and M.S. Poole, "What is personalization? Perspectives on the design and implementation of personalization in information systems," *Journal of Organizational Computing and Electronic Commerce*, vol. 16, no. 3-4, pp. 179-202, 2006.

- [19] A. Zimmermann, M. Specht and A. Lorenz, "Personalization and context management," *User modeling and user-adapted interaction*, vol. 15, no. 3-4, pp. 275-302, 2005.

The Geo CPS Platform for Designing Smart Cities

Wanglin YAN^{a,1}, Yasushi KIYOKI^a and Yoshifumi Murakami^b

^a*Graduate School of Media and Governance, Keio University*

^b*Ad-sol Nissin Corporation*

Abstract: The authors have developed the Geo CPS platform, which incorporates the advantages of cyber-physical systems, geographic information systems, and tangible user interfaces, and provides a platform to connect the three components for interactive sensing, processing and actuation in smart city development. It can be also applied for education in environmental and disaster management, as a tool for technical training, or as a testbed for business solutions.

Keywords: IoT, Tangible GIS, smart city, operating structure

1. Introduction

The widespread use of the Internet and mobile devices enables the Internet of Things (IoT) and the Internet of Everything (IoE) to capture the appearance and movement of people and objects anytime and anywhere [1]. From this information, it is possible to analyze movements and patterns of people and objects, and predict what will happen next. This information can then be fed back into the physical space to support human decision-making in real time. For example, drones can automatically take off and land anywhere, anytime, and collect and process images in real time. Self-driving cars are expected to appear on public roads in the near future. These examples foreshadow the arrival of a new type of information system, referred to as a cyber-physical system (CPS), where the cyber and physical spaces work together seamlessly in real time [2, 3, 4].

The CPS concept is being actively discussed in areas as diverse as civil engineering and construction, urban management, and transportation [5]. However, compared to its original use in factories or other closed environments, the use of CPS in open environments tends to be more complex, because most of the work is done in outdoor conditions. For this reason, the development of CPS solutions needs to involve intensive communications from the initial stage, which calls for innovative platforms and user interfaces.

The most common interfaces for geographic information are analog maps and graphic user interfaces (GUI) on electronic monitors. Spherical projection devices such as the digital globe at Japan's National Museum of Emerging Science and Innovation (MIRAIKAN) have also been developed, but they have not gone much beyond the realm of flat electronic display devices. Recently, GIS content is being projected on topographical models, a valuable way to stimulate discussion around a model in a

¹ Corresponding Author, Wanglin Yan, Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa City, Kanagawa 252-0816, Japan; E-mail: yan@sfc.keio.ac.jp.

realistic manner. However, without the ability to respond simultaneously to discussions, that approach is not as interactive and easy to operate as it could be. To address those limitations, tangible user interfaces (touchable user interfaces, or TUIs for short) have been developed that combine 3D models and map projections [6, 7]. TUIs are 2D, 2.5D, or 3D mockups that utilize augmented reality to provide an interactive space where multiple users can discuss their ideas intuitively in an environment that blends the cyber and physical worlds [8,9,10].

While CPS aims to monitor information and control devices in physical space, TUI attempts to promote multiple-user communication with a mock-up of the physical space. Although they work at different scales, CPS and TUI share the same basic idea of connecting the cyber and physical spaces. With this in mind, the authors have developed the Geo CPS platform [11], which incorporates the advantages of CPS, GIS, and extended TUI (XTUI) [12], and integrates data sensing and actuation of CPS in the physical space, processing of GIS data in the cyber space, and communication with XTUI in the social space (a meeting room). Essentially, the Geo CPS platform is an integration of the three spaces to form a systematic environment for participatory smart city system design and communications. It can be applied for education in environmental and disaster management, as a tool for technical training using GIS and CPS, or as a testbed for business solutions.

Based on the concepts proposed in previous publications [11, 12], this article clarifies the architecture of the Geo CPS platform and describes the functions of the operating structure which we refer to as iSPA. The latest experiment in an Urban Living Lab for a smart city project is presented in the last section.

2. Concept of the Geo CPS Platform

In the geographic dimension, information is obtained in physical space by various tools at different heights, from satellite to ground level. The data is managed in cyber space by databases, analyzed by GIS, or modelled by CAD, and then passed on to graphic devices or printers as illustrated in Figure 1. With the advent of IoT devices and ever-increasing speed of computer processing, the cycle of observation, storage, processing, and reproduction is becoming faster and finer. The deep learning and structuring of this accumulated data allow us to acquire more sophisticated knowledge and to respond interactively between the cyber and physical spaces.

The technological elements in the Geo CPS platform are not entirely new. Systems to obtain real-time location information from users in the physical space and overlay it with geographic information in the cyber space have been around for a long time. Car navigation systems are the best example of this. Vehicle location, traffic conditions, and the surrounding environment in the physical space are acquired and monitored online, processed and analyzed in real time in the cyber world, and used to direct the actions of people and moving objects. Within Geo CPS, Yan and Sakairi (2019) made a distinction between pseudo-CPS and true-CPS, with the former imposing layers of geospatial datasets on-screen only, and the latter combining those layers into one digital model [11]. The Geo CPS platform integrates the physical space and cyber space. To do this well, the interface between cyber and physical spaces is critical. To cover this requirement, an eXtended TUI (referred to as XTUI) was introduced [12].

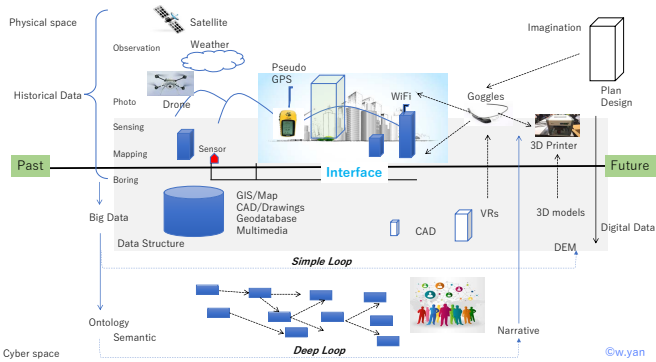


Figure 1. The physical world and cyber world in geographic dimensions (Revised by authors from previous version [11])

3. System Architecture and Functions

3.1. System Architecture

To utilize the respective advantages of the three spaces, the Geo CPS platform integrates the physical space with sensors, the cyber space in geodatabase, and the social space of XTUI for communications in system development. Figure.2 shows the system architecture. By analyzing the system components, we discover that these three spaces can each be conceptualized in four layers: platform, network, applications, and users. The physical space is built upon physical terrain, infrastructure, private and public services, and users. The cyber space consists of computer hardware and software, the Internet, intelligent algorithms, and operators. The XTUI is composed of tangible tables, network connections, intuitive interfaces, and multiple participants.

The uniqueness of this structure is that the three spaces are connected to each other by key layers, namely, the network infrastructure and the Internet as the layer between physical and cyber spaces; topology, between physical and social spaces; and digital content and algorithms, between cyber and social spaces. These key layers express the advantages of the three spaces and their specific roles in the platform. Meanwhile, communications between the spaces are realized by iSPA, an extension of the concept of SPA [13], where, “S” stands for sensing in physical space, “P” for processing in the cyber space, “A” for actuation by cyber space or actions inspired by communications in social space, and “i” for intuitive and interactive communication within the cyber-physical-social system. For example, iSPA gathers multi-dimensional information from the local environment and delivers it to the Internet in real time. The information is organized in a layer-based geo-database for cross-scale and cross dimensional analysis and simulation. The contents and algorithms are presented via the XTUI system.

With the Geo CPS platform, communications in a traditional CPS and sensing network can be mocked up in a tangible landscape and visualized in a 3D model.

Modification of the mockup can be detected automatically with intelligent algorithms in cyber space. In this way, for town planners, system developers and participants of the project, the design of smart city systems could be simulated in advance of installation similar with the physical mock-up of an urban and architectural development project before construction.

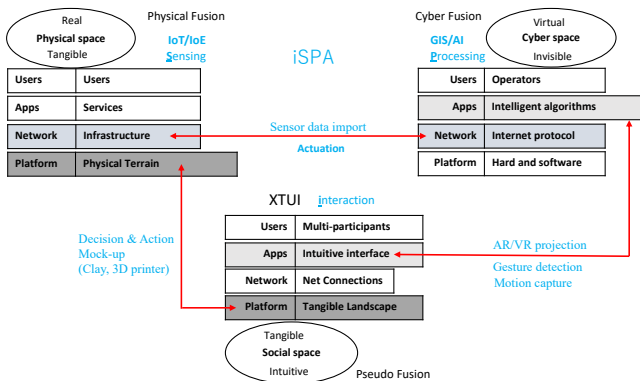


Figure 2. Geo CPS platform's system architecture (Revised by author from previous version [12])

3.2 System Functions

Figure 3 shows the functions of the Geo CPS operating structure. For this depiction of iSPA (left), the sensing and actuation layers are placed at the bottom while processing and interaction are placed at the top. The gray boxes represent classes of functions in which sensors and devices are grouped by network and communication protocols. Functions related to cyber space (e.g., data management, data mining, machine learning) are in close proximity to GIS, map projection and AR/VR (augmented reality, virtual reality) [14, 15]; XTUI brings user services visually to the platform on top of 2.5D and 3D mockups of the physical space. The tangible landscape is enhanced by GIS and AR/VR simulations. Tangible objects could be mockups of buildings and facilities, etc. Multiple participants can touch the 2.5D and 3D objects to explore the depth of the cyber space using interfaces such as REST and SPARQL and customized application programming interfaces (APIs).

User services could be as diverse as transportation planning, energy and environmental management, innovation, social revitalization, and human health and security. Issues in each of these fields can be expressed in terms of applications in the physical space by integrating information from the inside and interaction with the outside. Work scenarios such as water flow, object detection, thermal simulation, and urban access assessment provide some basic tools for solution development. On the other hand, the Geo CPS platform with XTUI can be used to study ideas, feasibility, and applicability.

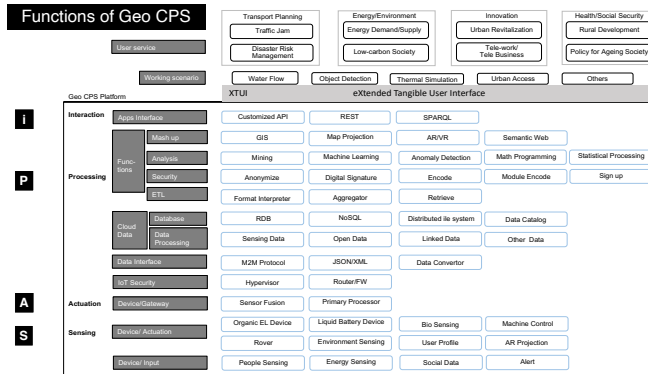


Figure 3. System functions in the Geo CPS platform (Revised by author from previous version [12])

4. Prototype and Social Experiment

In urban planning and design, map projection is often used as a tool to support public participation. In those cases, TUI is used as a tool for visualizing cyber content. For example, Maquil et al. reported on the design and implementation of Geospatial Tangible User Interfaces (GTUIs) to support stakeholder participation in collaborative urban planning [15, 18].

The authors are experimenting with the Geo CPS platform by bringing it into our Urban Living Lab. An Urban Living Lab (ULL) is a geographic or institutional place or approach where researchers, citizens, businesses, and governments come together to collaborate, and it is often set up and used for various types of planning and designing, as well as for testing corporate products and solutions [16, 17]. Recently, the concept of the Urban Living Lab has been proposed for use in participatory system design.

The City of Yokohama has developed as a port city, and its waterfront area has played a major role as a new commercial center in the metropolitan region. In the western part of the city, large residential areas have been built as bedroom communities for Tokyo. Many of these residential areas were planned and constructed during the postwar period of rapid economic growth, but the population is now declining, and infrastructure is aging, so the time has come to upgrade infrastructure. To examine this situation, the City of Yokohama partnered with Tokyu Corporation (a major private railway company) to open the WISE Living Lab (WLL) in 2017 on a property owned by the company near Tama Plaza Station in an urban planning effort for the next-generation suburban town (NST). WLL is a place where residents, government, companies, and universities can communicate about local issues, think together about solutions, and co-create the outcomes. The WISE acronym encapsulates the vision of Well-being, Intelligent, Smart and Ecology. We participated in this project and launched a collaborative initiative with

stakeholders to redesign the living environment in a suburban city. A Geo CPS prototype was set up at WLL to improve users' understanding of the local environment (Figure 4).

The physical space in this case is the outdoor community itself, the cyber space is the cloud GIS server, and the social space is the WLL supported by the XTUI. The XTUI consists of a personal computer, a projector, an infrared camera to detect gestures, and a designed frame. Temperature and air quality sensors are installed in the physical indoor and outdoor areas of the living lab. Three types of sensors are installed: the commercially available Air Egg, one we built based on a Raspberry PI single-board computer, and the coin-sized Leafony developed by the Trillion Node Study Group. The sensing data is sent to cloud data storage through LoRa via an IoT gateway. The data is processed in GIS for mapping and visualization. The processed data will be presented in a cyber space in charts and maps on the project website. Meanwhile, the maps are projected onto a 3D model of the region through the Internet. This allows participants to know the status of the environment on the spot, and the results can be reflected in the XTUI in real time. Participants can manipulate the ground surface made of clay or sand, 3D printed blocks, buildings, and road pavement. The participant's manipulation is captured by an infrared camera and thermal effect of the modification is projected.

In the cyber system, the walkability of the area is calculated and displayed on a monitor. It is then projected onto a 3D model and displayed with red indicating high accessibility along the road network. The user can relocate facilities by moving tangible icons of the facilities on the table. This interaction is done in real time, so the results of changes can be viewed immediately. Overall, this XTUI is beneficial for users to better understand their living environment and built environment and consider ways to improve them.

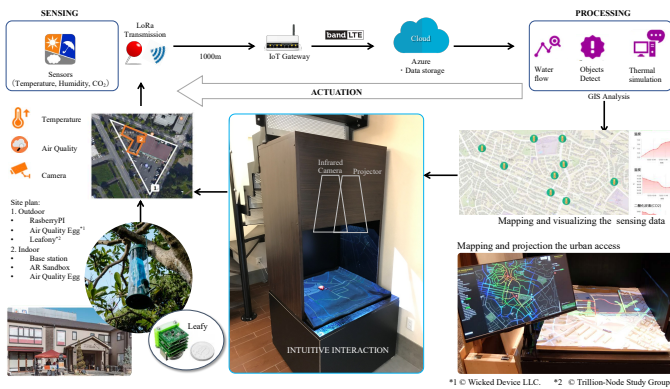


Figure 4. Prototype and experiment
(Revised by author from previous version [12])

5. Conclusions

This article focused on the integration of geo-technology, IoT and TUI, and presents a Geo CPS platform with eXtended TUI (XTUI) that enables intuitive interaction in cyber, physical and social spaces.

Nowadays, elemental technologies such as GIS, AR/VR, and GNSS (global navigation satellite systems) have made significant advances and become mainstream in the information society, and emerging technologies such as IoT and CPS are driving a new wave of industrial innovation in what has been referred to as Society 5.0. The Geo CPS platform aims to bridge gaps between advanced technologies, geospatial industries, and practical social issues. The platform, consisting of CPS, GIS, and XTUI, leverages the advantages of elemental geospatial technologies and provides a new perspective, focusing on interactions in physical, cyber, and social spaces. The application of the Geo CPS platform in the Urban Living Lab exemplifies an implementation model with community involvement that visibly displays the physical environment on the XTUI table and intuitively promotes interactive discussions. Interactions in the Urban Living Lab can create opportunities for system developers and community leaders to jointly identify problems, co-design solutions, and provide benefits to society.

The Geo CPS platform presented here is still at the early stages of development. Not all its functions have yet been demonstrated and the sensor network is still evolving with the latest availability of sensors, but the design concept is innovative and we anticipate that it will be able to address many geospatial challenges in smart cities, including social infrastructure management, urban planning, and urban disaster risk response.

Acknowledgements

This article is the result of research conducted jointly since 2017 by the SFC Institute of Keio University and Ad-sol Nissin Corporation. The prototype has been demonstrated annually at the SFC Open Research Forum. A community trial was conducted by the WISE Living Lab of Tokyu Co. Ltd., with support from the City of Yokohama. The research is also an experiment supported by the grant of Kakenhi titled “Educational practice of ICT-based citizen science for biodiversity conservation through domestic and international collaboration”.

References

1. Bosch. (2018). 7 factors for getting the most value from your Geo IoT project. <https://www.bosch-si.com/geo-iot/geo-iot/homepage-geo-iot.html>. Last accessed 2018/03/03.
2. Hu, L., Xie, N., Kuang, Z., & Zhao, K. (2012). Review of Cyber-Physical System Architecture. 2012 IEEE 15th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops, 25–30. <https://doi.org/10.1109/ISORCW.2012.15>
3. Lee EA, Seshia SA. (2015). Introduction to Embedded Systems, A Cyber-Physical Systems approach Second Edition, E. A. Lee and S. A. Seshia, Berkeley, USA.
4. Liu, Y., Peng, Y., Wang, B., Yao, S., Liu, Z., & Concept, A. (2017). Review on Cyber-physical Systems. IEEE/CAA Journal of Automatica SINICA, 4(1), 27–40.
5. Yang Lu (2017) Yang Lu (2017) Cyber Physical System (CPS)-Based Industry 4.0: A Survey, Journal of Industrial Integration and Management VOL. 02, NO. 03. <https://doi.org/10.1142/S2424862217500142>
6. Hofstra, H., Scholten, H., Zlatanova, S., & Scotta, A. (2008). Multi-user tangible interfaces for effective decision-making in disaster management. In S. Nayak & S. Zlatanova (Eds.), *Remote Sensing and GIS*

Technologies for Monitoring and Prediction of Disasters (pp. 243–266). Berlin Heidelberg: Springer-Verlag.

7. Monostori, L., Kádá, B., Bauernhansl, T., Kondoh, S., Kumara, S., Reinhart, G., Kumara, S. (2016). Cyber-physical systems in manufacturing. *CIRP Annals*, 65(2), 621–641.
8. Ishii, H., & Ullmer, B. (1997). Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proceedings of CHI '97*, March 22-27, 1997, (pp. 234–241). New York, New York, USA.: ACM Press. <https://doi.org/10.1166/jnn.2017.14175>.
9. Ishii, H., Ratti, C., Piper, B., Wang, Y., Biderman, A., & Ben-Joseph, E. (2004). Bringing clay and sand into digital design - continuous tangible user interfaces. *BT Technology Journal*, 22(4), 287–299. <https://doi.org/10.1023/B:BTTJ.0000047607.16164.16>
10. Harmon, Brendan A., Petrasova, A., Petras, V., Mitasova, H., & Meentemeyer, R. (2018). Tangible topographic modeling for landscape architects. *International Journal of Architectural Computing*, 16(1), 4–21. <https://doi.org/10.1177/1478077117749959>
11. Yan, W., & Sakairi, T. (2019). Geo CPS: Spatial challenges and opportunities for CPS in the geographic dimension. *Journal of Urban Management*, 8(3), 331–341. <https://doi.org/10.1016/j.jum.2019.09.005>.
12. Yan W, Y Murakami, A Yasuda, T Makihara, R Fujimoto, and S Nakayama (2021) Developing the eXtended Tangible User Interface as an Experimental Platform for Geo CPS, in M Sakurai and R Shaw Eds. *Emerging Technologies for Disaster Resilience - Practical Cases and Theories*, Springer Nature (in press).
13. Kiyoki Yasushi, Petchporn Chawakitchareon, Somporn Rungsupa, Xing Chen and Kittiya Samlansin (2020) A Global & Environmental Coral Analysis System with SPA-based Semantic Computing for Integrating and Visualizing Ocean-Phenomena with “5-Dimensional World-Map”. In B. Thalheim, M. Tropmann-Frick, H. Jaakkola, Y. Kiyoki (eds). *Proceedings of the 30th International Conference on Information Modelling and Knowledge Bases (EJC 2020)*, June 8-9, 2020, Hamburg, Germany.
14. Maquil, V., De Sousa, L., Leopold, U., & Tobias, E. (2015). A geospatial tangible user interface to support stakeholder participation in urban planning. In *GISTAM 2015 - 1st International Conference on Geographical Information Systems Theory, Applications and Management, Proceedings* (pp. 113–120). SCITEPRESS.
15. Maquil, V., Leopold, U., De Sousa, L. M., Schwartz, L., & Tobias, E. (2018). Towards a framework for geospatial tangible user interfaces in collaborative urban planning. *Journal of Geographical Systems*, 20(2), 185–206. <https://doi.org/10.1007/s10109-018-0265-6>
16. McCormick, K. and Hartmann, C. (2017) The emerging landscape of urban living labs: Characteristics, practices and examples. <https://doi.org/10.1371/journal.pone.0099587>.
17. Thinyane, M., Terzoli, A., Thinyane, H., Hansen, S., & Gumbo, S. (2012). Living Lab Methodology as an Approach to Innovation in ICT4D: The Siyakhula Living Lab Experience. *IST-Africa* 2012, 1–9.

Global Coral Health Levels Analysis Database with Semantic Computing and 5D World Map

Piyaporn NURARAK^{a,1}, Yasushi KIYOKI^a,
Petchporn CHAWAKITCHARON^b and Yasuhiro HAYASHI^c

^a *Graduate School of Media and Governance, Keio University, Japan*

^b *Environmental Engineering Association of Thailand, Thailand*

^c *Faculty of Data Science, Musashino University, Japan*

Abstract. This global warming and climate change affect not only all living things but also affect many non-living things. Furthermore, It caused extreme disasters that become impossible to ignore. Coral bleaching is the one to show ocean warming due to climate change. This paper presents the analysis and visualization of the coral health levels database by using 5D World Map system. Coral health levels are analyzed using a coral-knowledge image that includes coral with a coral health chart. We use image processing and color semantic distance to interpreting coral health levels. We have implemented an actual space integration system to access environmental information resources with coral health levels and image analysis that the results have been shown on the 5D world map. As for the experiment study, coral health levels are located in the ocean close to Thailand's islands as Ko Ha (Five Island), Ko Bon, Ko Hin Ngam, Ko Tarutao, Ko Thalu, and Ko Samaesarn.

Keywords. Coral Health Levels, Semantic computing, Coral health levels database, Coral-knowledge image, Image processing, Global environmental analysis, 5D World Map.

¹ Corresponding Author, Graduate School of Media and Governance, Keio University, Shonan Fujisawa Campus, 5322 Endo, Fujisawa, Kanagawa, 252-0882, Japan; Email: pnurarak@sfc.keio.ac.jp.

1. Introduction

This global warming and climate change affect not only all living things but also affect many non-living things. Furthermore, It caused extreme disasters such as stronger storms, rising seas, and ocean warming become impossible to ignore. Global warming is very close to us because everything in this world is dependent on each other. That is why everyone on this planet should start to be aware and understand the impacts of the future. Coral bleaching is the one to show ocean warming due to climate change. Our motivation is promoting healthy reefs by engaging the global community in monitoring coral health and coral bleaching with 5D World Map System. The 5D World Map System proposed by Y. Kiyoki and S. Sasaki in [1-3] is globally utilized as a Global Environmental Semantic Computing System for disaster, natural phenomena, ocean-water analysis with local and global multimedia data resources.[4-8]

This paper presents the analysis and visualization of the coral health levels database by using 5D World Map system. Coral health levels are analyzed using an image that includes coral with a coral health chart. We use image processing and color semantic distance to interpreting coral health levels. Our environmental-semantic computing system realizes integration and semantic-search among environmental-semantic spaces with coral health levels and image databases. We have implemented an actual space integration system for accessing environmental information resources with coral health levels and image analysis. We clarify the feasibility and effectiveness of our method and system by showing several experimental results for coral health levels analysis databases.

This paper is organized as follows: Section 2 shows the researches related to our study. Section 3 shows the proposed methods of the system. Section 4 shows Results and Discussion. Section 5 presents the conclusion and future work and the acknowledgment in section 6.

2. Related Works

In this section, we introduce the works that related to our study

2.1 5D World Map System

The 5D world map system is globally utilized as a global environmental semantic computing system for disaster, natural phenomena, ocean-water analysis with local and global multimedia data resources. This system is a collaborative knowledge sharing, analyzing, searching, integrating, and visualizing data, such as images, videos, audio, etc., onto 5D time-series which temporal (1 Dimension), spatial (2-4 Dimension), and semantic (5 Dimension)[5]. In addition, this system integrates and visualizes the analyzed results as multi-dimensional axes dynamic historical atlas. The main feature of this system is to dynamically create various context-dependent patterns of environmental according to a user's viewpoints. This system support composition of images as the image retriever and enables to realize that users find out the images that show the particular situation of the environment.

2.2 Coral-Knowledge Image

Coral-knowledge image[15] in Figure 1 consists of two essential elements:1) Coral image in the center and 2) Coral health chart [10]. The coral health chart is based on the actual colors of bleached and healthy corals, that each color square corresponds to a concentration of symbionts contained in the coral tissue is directly linked to the health of the coral. This chart represents the most common colors of corals and helps our eyes to make an accurate match. The color is divided into 4 groups (B, C, D, and E) and classified into 6 levels (1-6) for each side. The highest level (level 6) represents the best healthy coral, and the lowest (level 1) means the worst healthy coral. The health status and mortality percentages from the coral health chart shows in table 1. [16]



Figure 1. Coral-knowledge image

Table 1 The health status and mortality percentages from the coral health chart

Level	State of health	Healthy percentage	Mortality percentage
1	Worst	16.67	83.33
2	Poor	33.33	66.67
3	Declining	50.00	50.00
4	Fair	66.67	33.33
5	Good	83.33	16.67
6	Best	100.00	0.00

3. Proposed Methods

3.1. Implementation of a prototype system

This section explains our system's concept of creating an automatic database in semantic space to analyze, interpret, and visualize coral health levels using the 5D World Map system.

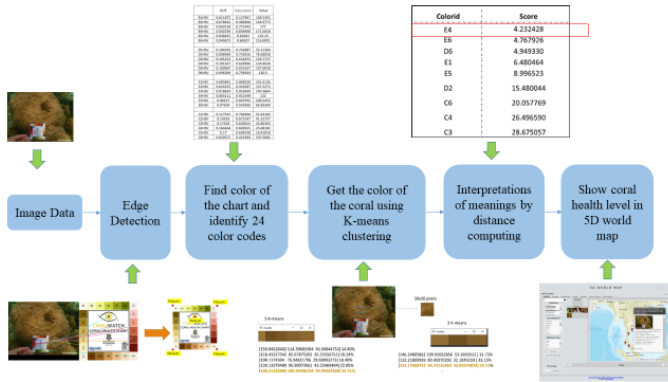


Figure 2. The concept of system

The concept of the system is shown in Figure 2:

Step 1: Image data must be coral-knowledge images [8], including coral images and coral health charts.

Step 2: Edge detection is a process to retrieve and detect the coral health chart by using SIFT algorithm (Scale Invariant Feature Transform) to get 4 corner points and 1 point at the eye of the coral health chart.[15]

Step 3: Find the chart's color and identify 24 color codes as B1-B6, C1-C6, D1-D6, and E1-E6.

Step 4: Getting the color value of coral using k-means clustering to determine the dominant colors in coral images. The highest score has been chosen to represent the coral color.

Step 5. Interpretations of meanings by color distance computing. We are finding color semantic distance in HSV color space; the result gives the meaning of the coral health-level according to the distance ordering of "minimal value" between the color of coral and the closed color code.

Step 6. Show the coral health level in 5D world map.

3.2. Color semantic distance computing.

we use the color semantic distance calculation to show the color distance between 24 color codes in the coral health chart and coral color value. In HSV color space, we use Equation (1) – (4) to find distance 2 pixels $P_0(h_0, s_0, v_0)$ and $P_1(h_1, s_1, v_1)$ as follows:

$$dh = \min(\text{abs}(h_1 - h_0), 360 - \text{abs}(h_1 - h_0)) / 180.0 \quad (2)$$

$$ds = \text{abs}(s_1 - s_0) \quad (3)$$

$$dv = \text{abs}(v_1 - v_0) / 255.0 \quad (4)$$

When dh , ds and dv are the distance between P_0 and P_1 in H (Hue), S (Saturation), and V (value), respectively. Each of these values will be in the range $[0,1]$, we can compute the length of this :

$$\text{distance} = \sqrt{dh^2 + ds^2 + dv^2} \tag{5}$$

In the color distance, the smaller result is the higher similarity.

3.3. K-means Clustering

Given a set of observations (x_1, x_2, \dots, x_n) , where each observation is a d -dimensional real vector, k -means clustering aims to partition the n observations into k ($\leq n$) sets $S = \{S_1, S_2, \dots, S_k\}$ to minimize the within-cluster sum of squares.

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 = \underset{S}{\operatorname{argmin}} \sum_{i=1}^k |S_i| \operatorname{Var} S_i$$

Where μ_i is the mean of points in S_i . This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\underset{S}{\operatorname{argmin}} \sum_{i=1}^k \frac{1}{2|S_i|} \sum_{x,y \in S_i} \|x - y\|^2$$

The equivalence can be deduced from identity

$$\sum_{x \in S_i} \|x - \mu_i\|^2 = \sum_{x \neq y \in S_i} (x - \mu_i)^T (\mu_i - y)$$

In HSV color space, we use k -means clustering to determine the dominant colors in a coral image by using 10×10 pixels in the center of the coral-knowledge image, as shown in Figure 3.

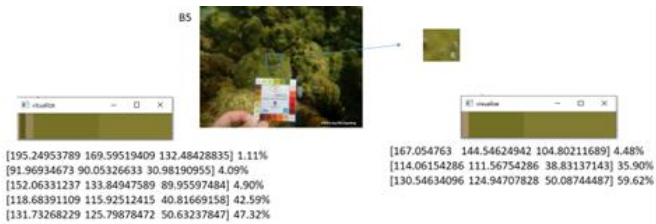










Figure 3. K-means clustering to determine the dominant colors

3.4. Data Sources

The image data that we use in coral health levels analysis are coral-knowledge-images that include coral in the center of the image, and the coral health chart can be up, down, left, or right of the coral. All of the coral-knowledge-images have been taken underwater environment.

Table 2. The description of the coral-knowledge-images

Coral-knowledge-image	Date	Time	Island	Latitude	Longitude
	8/25/2019	14:17	Ko Samaesarn	12.57107812	100.9461504184105
	4/14/2021	8:54	Ko Ha	8.1502651	98.755861
	4/14/2021	12:20	Ko Bon	7.7564336	98.332116
	11/14/2020	15:11	Ko Hin Ngam	6.5149976	99.2624372
	11/13/2020	15:52	Ko Tarutao	6.5913175	99.6564092
	11/13/2020	11:45	Ko Thalu	11.0762275	99.560394
	8/25/2019	14:17	Ko Samaesarn	12.57107812	100.9461504184105
	4/14/2021	12:40	Ko Bon	7.7564336	98.332116

3.4. Description of study areas

The study areas of coral health levels are located in the ocean close to Thailand's islands as Ko Ha (Five Island), Ko Bon, Ko Hin Ngam, Ko Tarutao, Ko Thalu, and Ko Samaesarn. The descriptions of the coral-knowledge-images are shown in Table 2.

3.5. Data Structure

We collected 311 files in CSV form and added semantic and spatiotemporal metadata such as category, location, date, and description for each image data. The data structures are based on 5D world map system, as shown in Figure 4.

ID	File	Kind	Category	Location	Date	Description	User	Actions
77228	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:54:00	Coral Health Level: 4, warning, for heat h, n metadata, CAD4	phurarak	View Edit Download Delete
77227	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:48:00	Coral Health Level: 4, warning, for heat h	phurarak	View Edit Download Delete
77226	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:27:00	Coral Health Level: b leach, bleaching D3, warning, declining he alth	phurarak	View Edit Download Delete
77225	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:28:00	Coral Health Level: 0, good health	phurarak	View Edit Download Delete
77224	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:25:00	Coral Health Level: b leach, bleaching B3, warning, declining he alth	phurarak	View Edit Download Delete
77222	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Ha (Five Island), Kho Thong, Muean Krabi District, Krabi 81000, Thailand	2021-04-14 08:25:00	Coral Health Level: 2, poor health	phurarak	View Edit Download Delete
77221	[Image]	image	{biodiversity los s7 coral reef enviro nment, ocean pollu sion, water pollution n}	Ko Bon, Rawai, Phu Hill, Thailand	2021-04-14 12:20:00	Coral Health Level: 0, good health	phurarak	View Edit Download Delete

Figure 4. The data structures are based on 5D world map system

4. Results and Discussion

To clarify the feasibility and effectiveness of our method and system by (1) showing several experimental results for coral health levels analysis databases and (2) showing several experimental results for coral health levels analysis with a semantic interpretation method and 5D World Map.

4.1. Coral Health Levels Analysis Databases

we have implemented our coral health levels analysis with coral-knowledge-base-image retrieval, using k-means to determine the dominant colors in a coral image and color semantic distance computing system for coral-knowledge image datasets. Some of the results of the experiment are shown in Figure 5. The accuracy of the experiment is 85.0%.

4.2. Coral Health Levels Analysis with Semantic-Interpretation Method and 5D World Map.

We realized the mapping and visualization results of coral health levels analysis in 5D World Map and applied the coral-knowledge image. Our experimental results have shown a semantic associative computing system based on semantic computing in the global environmental analysis shown in Figures 6-9.









coral id	Human eyes	K-means in HSV color space	coral-knowledge image
P8250068-E1	E1	D1	
P8250135-C2	C2	B2	
P8250128-E3	E3	E3	
P8250025-E4	E4	E4	
P8250048-E4	E4	E4	
P8250028-B5	B5	B5	
P8250168-E5	E5	E5	
P8250035-D6	D6	D6	

Figure 5. The result of the coral health levels analysis experiments

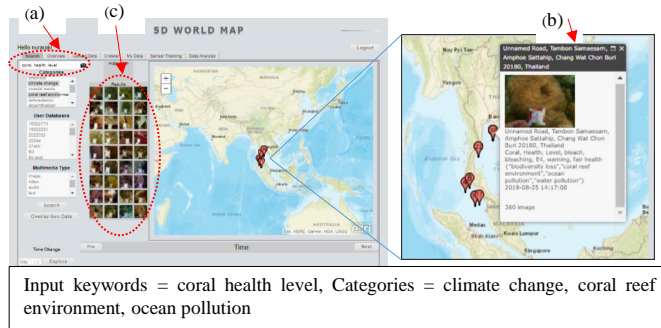


Figure 6. Global coral health levels analysis in 5D world map system: (a) Keywords search by coral health levels, (b) Spatiotemporal analysis (global overview of geographical distribution and the time-series), and (c) Example of coral-knowledge images

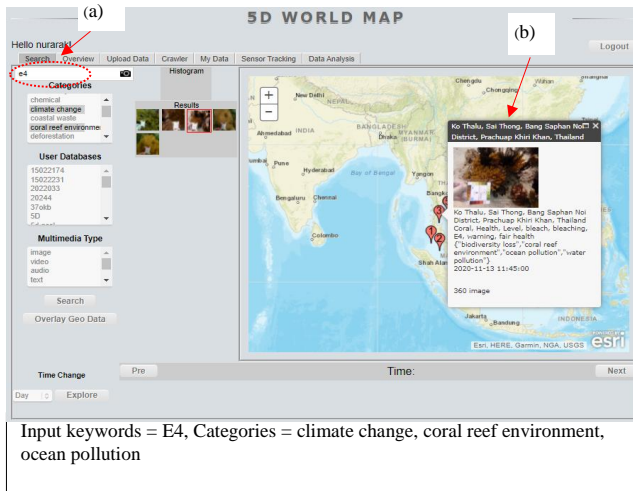


Figure 7. Global coral health levels analysis in 5D world map system: (a) Keywords search by coral health levels, (b) Spatiotemporal mapping in the global overview of geographical distribution



Figure 8. Global coral health levels analysis in 5D world map system: (a) Keywords search by coral health, (b) Spatiotemporal mapping in the global overview of geographical distribution

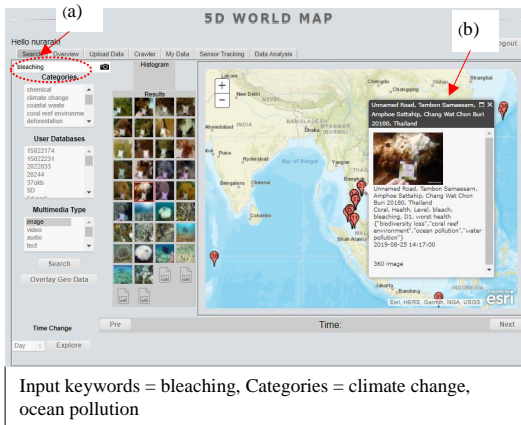


Figure 9. Global coral health levels analysis in 5D world map system: (a) Keywords search by coral health, (b) Spatiotemporal mapping in the global overview of geographical distribution

5. Conclusion

We have presented a new environmental-semantic computing system for coral health levels analysis database and coral-knowledge image spaces with 5D World Map. The main feature of our approach is to realize integration and semantic-search among global environmental-semantic spaces based on the concept of coral health levels analysis space. This system enables creating an automatic coral health level analysis in semantic space for the ocean environment. It makes artificial intelligence vision cognitive functions using image processing with semantic computing instead of the human eye.

As future work, we will extend the coral health level analysis semantic space realized onto the 5D world map system to global knowledge-sharing on the ocean environment worldwide.

6. Acknowledgement

This work is supported by Multimedia Database Laboratory (MDBL), Graduate School of Media and Governance, Keio University. We thank the MDBL members for their valuable comments and suggestions. We also appreciate Ms. Veranuch Chawakitchareon for her activities in oceans to take coral photographs.

References

- [1] Yasushi Kiyoki, Shiori Sasaki, Nhung Nguyen Trang, Nguyen Thi Ngoc Diep, "Cross-cultural Multimedia Computing with Impression-based Semantic Spaces," *Conceptual Modelling and Its Theoretical Foundations, Lecture Notes in Computer Science*, Springer, pp.316-328, March 2012..
- [2] Shiori Sasaki, Yusuke Takahashi, Yasushi Kiyoki: "The 4D World Map System with Semantic and Spatiotemporal Analyzers," *Information Modelling and Knowledge Bases*, Vol. XXI, IOS Press, 18 pages, 2010.
- [3] Totok Suhardijanto, Yasushi Kiyoki, Ali Ridho Barakbah: "A Term-based Cross-Cultural Computing System for Cultural Semantics Analysis with Phonological-Semantic Vector Spaces," *Information Modelling and Knowledge Bases XXIII*, pp.20-38, IOS Press, 2012.
- [4] Yasushi Kiyoki, Xing Chen, Shiori Sasaki and Chawan Koopipat, "Multi-Dimensional Semantic Computing with Spatial-Temporal and Semantic Axes for Multi-spectrum Images in Environment Analysis", to appear in *Information Modelling and Knowledge Bases (IOS Press)*, Vol. XXVI, 20 pages, March 2016.
- [5] Chalisa Veessommai, Yasushi Kiyoki, Shiori Sasaki and Petchporn Chawakitchareon. "Wide-Area River-Water Quality Analysis and Visualization with 5D World Map System", *Information Modelling and Knowledge Bases*, Vol. XXVII, pp.31-41, 2016.
- [6] Jinnika Wijitdechakul, Yasushi Kiyoki, Chawan Koopipat, "An environmental-semantic computing system of multispectral imagery for coral health monitoring and analysis", *Information Modelling and Knowledge Bases XXX*, Vol.312, pp.293 – 311, IOS Press, 2019.
- [7] Yasushi KIYOKI, Xing CHEN, Chalisa VEESOMMAI, Shiori SASAKI, Asako URAKI, Chawan KOOPIPAT, Petchporn CHAWAKITCHAREON and Aran HANSUEBSAI, "An Environmental-Semantic Computing System for Coral-Analysis in Water-Quality and Multi-Spectral Image Spaces with "Multi-Dimensional World Map", *Information Modelling and Knowledge Bases*, 2018.
- [8] Yasushi Kiyoki, Petchporn Chawakitchareon, Sompop Rungsupa, Xing Chen, and Kittiya Samlansin "A Global & Environmental Coral Analysis System with SPA-based Semantic Computing for Integrating and Visualizing Ocean-Phenomena with "5-Dimensional World-Map"", *Information Modelling and Knowledge Bases XXXII*, 2020.
- [9] Mahshid Oladi, Mohammad Reza Shokri and Hassan Rajabi-Maham, "Application of the coral health chart to determine bleaching status of *Acropora* downingi in a subtropical coral reef. *Ocean Science Journal* volume 52, pp. 267–275. 2017.
- [10] <https://coralwatch.org/index.php/product/coral-health-chart/>

- [11] Kiyoki, Y., Kitagawa, T., and Hayama, T., "A metadata system for semantic image search by a mathematical model of meaning," *ACM SIGMOD Record*, vol.23 no.4 pp.34-41, 1994.
- [12] Kiyoki, Y., Chen, X., Sasaki, S., and Koopipat, C., "Multi-Dimensional Semantic Computing with Spatial-Temporal and Semantic Axes for Multi-spectrum Images in Environment Analysis", *Information Modelling and Knowledge Bases*, Vol. XXVII, IOS Press, pp.14-30, 2016.
- [13] Sasaki, S. and Kiyoki, Y., "Analytical Visualization Function of 5D World Map System for Multi-Dimensional Sensing Data", *Information Modelling and Knowledge Bases*, Vol. XXIX, IOS Press, pp. 71-89, 2018.
- [14] Nguyen, D. T. N, Sasaki, S., and Kiyoki, Y., "5D World PicMap: Imagination-based Image Search System With Spatiotemporal Analyzer", *Proceeding of The IASTED e-society 2011 Conference*, Avila, Spain, pp. 271- 278, 2011.
- [15] Piyaporn Nurarak, Yasushi Kiyoki, Petchporn Chawakitchareon and Yasuhiro Hayashi, "A Coral-Color Analysis System for Observing Environmental Situation and Change with K-means Clustering and Semantic Classification", *Thai Environmental Engineering Journal* Vol. 35 No. 1: pp.63-73, 2021.
- [16] Kittiya Samlansin, Petchporn Chawakitchareon and Sompop Rungsupa., "Effects of Salinity and Nitrate on Coral Health Levels of *Acropora* sp.", *Thai Environmental Engineering Journal* Vol. 34 No. 1: pp.19-26, 2020.

Human-Health-Analysis Semantic Computing & 5D World Map System

Yasushi Kiyoki ^{a,1}, Koji Murakami ^{b,c}, Shiori Sasaki ^a, Asako Uraki ^a,
^a*Graduate School of Media and Governance, Keio University, Japan*
^b*SFC Research Institute, Keio University, Japan*
^c*PreventScience Co.,Ltd*

Abstract. Semantic space creation and computing are essentially significant to realize semantic interpretations of situations and symptoms in human-health. We have presented a semantic space creation and computing method for domain-specific research areas. This method realizes the semantic space creation with domain-oriented knowledge and databases. This paper presents a semantic space creation and computing method for “Human-Health Database” with the implementation process for “Human-Health-Analysis Semantic Computing”. This paper also presents a new knowledge base creation method for personal health data for preventive care and potential risk inspection with global and geographical mapping and visualization in 5-Dimensional World Map System. This method focuses on analysis of personal health and potential-risk inspection and provides a set of semantic computing functions for semantic interpretations of situations and symptoms in human-health. This system is applied to “Human-Health-Analysis Semantic Computing” to realize world-wide evaluation for (1) multi-parameterized personal health/bio data, such as various biomarkers, clinical physical parameters, lifestyle parameters, other clinical / physiological or human health factors, etc., for a health monitoring, and (2) time-series multi-parameter health/bio data in the national/regional level for global analysis of potential cause of disease. This Human-Health-Analysis Semantic Computing method realizes a new multidimensional data analysis and knowledge sharing for a global-level health monitoring and disease analysis. The computational results are able to be analysed by the time-series difference of the value of each place, the differences between the values of multiple places in a focused area, and the time-series differences between the values of multiple places to detect and predict a potential-risk of diseases.

Keywords. Semantic Search, Semantic Computing, Medicine, Medical Data, Big Data, Biographical Data, Vital Data, Sensing, AI, Cyber-Physical System, Visualization, Data Mining, Warning, SPA, Sensing, Processing, Actuation, SDGs

1. Introduction

Semantic computing is a promising approach to realize Human-Health-Analysis and Health caring. Various medical analysis methods for medical-domain information resources have been proposed.

We have proposed semantic computing methods with a multi-dimensional semantic space creation architecture representing domain-specific knowledge and databases [1][2][3]. We have designed a domain specific knowledge space creation process for our semantic computing. In this paper, we apply this semantic computing method to Human-Health-Analysis and Health caring, as “Human-Health-Analysis Semantic Computing Method”.

The main objective of this method is to realize a semantic-space creation process and a semantic computing environment for semantic interpretations of situations and symptoms in human-health. This method uses domain-specific knowledge resources in “Human-Health Database”, and creates various semantic spaces, consisting of domain specific knowledge expressions. The variety of semantic spaces is derived from parameters which a Human-Health database designer defines in the semantic-space creation process. Those parameters are expressed in two categories: (1) expertized-level determining the expertized-level parameters of semantic space, (2) sub-domain-level determining sub-domain-specific parameters for defining semantic subspaces consisting of ‘sub-domain’ features. By using these parameters, this method controls the scope of the analysis, according to the objectives of the semantic analysis and interpretations.

In our “Human-Health-Analysis Semantic Computing”, domain-specific knowledge is formalized as a featured-vector matrix in a semantic space, with expertized-level and sub-domain-level parameters. The domain-specific knowledge for Human-Health-Analysis is utilized for domain-specific semantic space creation. The matrix data structure of the domain-specific knowledge is expressed in the Human-Health semantic space, and human-health data are mapped on to this semantic space. The semantic computing is applied to this space and realizes semantic interpretations of situations and symptoms in human-health with the relationships between disease and patients.

We have presented a concept of “Semantic Computing System [1][2][3]” and “5 Dimensional World Map System” for realizing global environmental data-analysis, integration, search and visualization for various issues in the world-wide scope. The main feature of 5D World Map System is to provide a platform of collaborative work for users to perform a global analysis for sensing data in a physical space along with the related multimedia data in a cyber space, on a single view of time-series maps based on the spatiotemporal and semantic correlation calculations. This paper applies 5D-World Map System to human-health analysis and support.

This paper presents a “Human-Health-Analysis Semantic Computing Method” with semantic computing to realize semantic computing for biologic data, such as height, weight, BMI, blood pressure, blood glucose, blood protein, urine metabolites, images, personal genes, mRNA, etc. in the “*multi-dimensional-biologic-parametrized semantic-space*”. The “Human-Health-Analysis Semantic Computing realizes semantic associative computing of semantic equivalence, similarity and difference between

"multi-dimensional-biologic data". We apply this system to human-health analysis and support, as a new platform of human-health care and support.

2. Human-Health-Analysis with Semantic Computing

2.1. Human-Health Semantic Space Creation and Semantic Computing

A semantic computing method, the Mathematical Model of Meaning (MMM) [1][2][3], is applied to human-health databases to realize health-health analysis in the earlier stage in disease, as shown in Figure 1. Although the survival rate of cancer patients is improving due to the emergence of new drugs and the development of new treatments, a worldwide perspective suggests that the number of cancer patients will continue to increase. There is no change in the situation where the foundation of a healthy healthcare measure is required. It has been proven in many cancers that the earlier the stage, the better the survival rate of cancer patients. The fact is that future cancer control is focused on prevention and early detection.

Process-1: Semantic Space Creation of " n -dimensional health-analysis universe"

(Step-1-1): A set of " $m1$ patient-data" is given as n -parameters (human biological data (various biomarkers, clinical physical parameters, lifestyle parameters, other clinical / physiological or human health factors, etc.), and defined as the "patient-data matrix, and each patient-data element is characterized by n features. That is, an m by n matrix $M1$ is defined as "patient-data matrix" in the "orthogonal semantic space S " as " n -dimensional health-analysis universe".

(Step-1-2) A set of " $m2$ non-patient-data" is given as n -parameters (human biological data (various biomarkers, clinical physical parameters, lifestyle parameters, other clinical / physiological or human health factors, etc.), and defined as the "non-patient-data matrix, and each non-patient-data element is characterized by n features. That is, an $m2$ by n matrix $M2$ is defined as "non-patient-data matrix" in the "orthogonal semantic space S " as " n -dimensional health-analysis universe".

Process-2: Context-specific distance computing in the context-dependently-selected subspace from the semantic space S :

(Step-2-1) Specific parameters, corresponding to a specific-context to be analysed, are selected from n -parameters" in S , and the context is characterized to make analysis in a context-dependent way.

(Step-2-2) A "**selected subspace**", from the orthogonal semantic space, is extracted to be analysed according to a context. The **subspace is projected according to the given "context" from the semantic space S** , which are given as "**context**" represented by "**subset of parameters**."

(Step-2-3) The metric for computing the distance between data elements in $m1$ and $m2$ is defined in the semantic space, and the highly-correlated data elements between $m1$ and $m2$, to the given “context” are extracted, as the distance in the selected subspace.

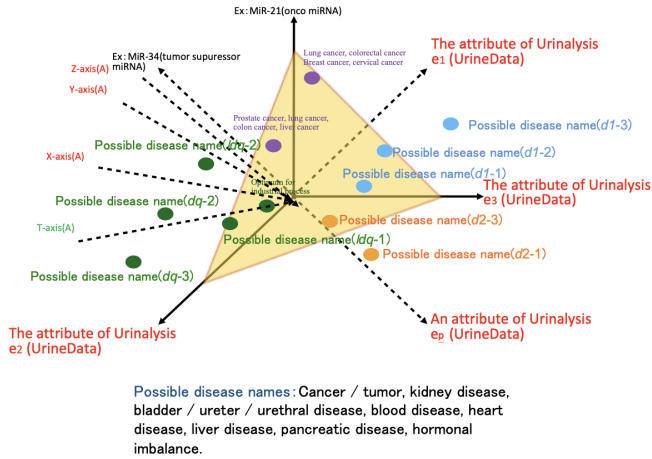


Figure 1. Human-Health Semantic Space Creation for Semantic Computing

3. 5D World Map System for Human-Health-Analysis

One of the important applications of the semantic computing system is “Human-Health Analysis”, which aims to evaluate various situations and symptoms in human-health.

We have proposed and introduced the global knowledge sharing and processing system architecture as “5 Dimensional World Map System [4][6][8],” with a multi-visualized and dynamic knowledge integration and representation functions.” This system has been applied to a lot of knowledge applications, as global environmental analysis and visualization. The basic space of this system consists of a temporal (1st dimension), spatial (2nd, 3rd and 4th dimensions) and semantic dimensions (5th dimension, representing a large-scale and multiple-dimensional semantic space that is based on our semantic associative computing system (MMM [1][2][3][4]). This space memorizes and recalls various multimedia information resources with temporal, spatial and semantic correlation computing functions, and realizes a 5D World Map for dynamically creating temporal-spatial and semantic multiple views applied for various “environmental multimedia information resources.”

As an important application of 5 Dimensional World Map System, we have constructed “5D World Map System for Human-Health-Analysis” for globally sharing and analyzing human health situations with semantic computing functions,

applied to “human health data sharing,” as a new platform of collaborative human-health analysis [4][6]. This platform enables to create a remote, interactive and real-time human health and academic research exchange among different areas.

3.1. 5D World Map System for Integration of Spatiotemporal and Semantic Computing

As an important human health system, we have proposed a multi-dimensional data mining and visualization system. In the design of this global human-health analysis system, we focus on how to search and analyze human health data related to human health situations, according to contexts and analysis-points.

“**5D World Map System for Human-Health-Analysis**” is structured, as shown in **Figure 2**, for sharing and analyzing human health situations and stimulus with semantic functions applied to “human health databases,” as a new platform of collaborative human-health analysis. This platform enables to create a remote, interactive and real-time human-health data-sharing and academic research exchange among different research activities.

We have already had track records in medical applications. We focus on simple urine sampling screening based on the biological characteristics of cancer, and realize new risk presentation methods by analyzing indices related to multiple parameters in a multidimensional manner. Urine specimens do not cause pain during collection and do not carry the physical risk of exposure to radiation for measurement. Based on these measurement results, we introduced the analysis concept of the 5D World Map, which is also used for regional disaster prediction, and succeeded in developing a new risk prediction. We propose a new concept of disease risk prediction by fusion of biology and information technology.

This system is constructed by SPA (Sensing-Processing-Actuation) process concept as shown in **Figure 2**.

The essential features of the system are described as followings:

(1) The system sets human biological data (various biomarkers, clinical physical parameters, lifestyle parameters, other clinical / physiological or human-health factors, etc.) as a target data and represents them as a n -dimensional MATRIX.

(2) The system sets the origin of each parameter as a neutral “semantic origin point” and represents the n types of parameters as an n -dimensional space to measure the similarity of n dimensions comprehensively and dimensionally by aggregating the parameter values at the origin for multiple parameters.

(3) The system calculates the distance in the n -dimensional space among the analysis subjects, the existing diseased patients and the healthy subjects as the disease occurrence risk degree of the subjects.

(4) The system visualizes and represents the risk of disease occurrence of the subjects as a Risk Person Discovery Method device using the graph of each dimension value as a polygon.

While the general existing methods in medical field grasps the tendency of the entire subjects’ data distribution by statistics, this method realizes more detailed and fine-grain prediction and diagnosis because of the features described above. As described in the next section, the system consists of 3-phase checking process: 1st Phase: Finding similar person with principal parameters, 2nd Phase: Rule-based analysis with secondary-essential parameters, and 3rd Phase: Time-series analysis.

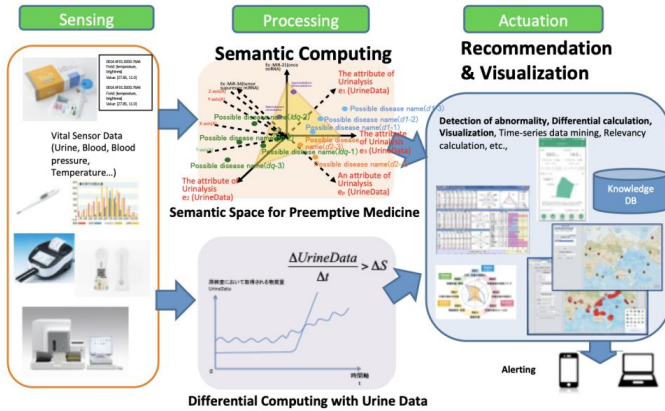


Figure 2. SPA process of 5D World Map System for Preemptive Care with analytical visualization

5D World Map System [4][6][8][13] has been providing various functionalities to share and visualize various types of multimedia data [8][11][13]. A combination of the analysis and visualization functions for multimedia and real-time sensing-data of 5D World Map System has been proposed to make environmental analysis much richer and deeper, which contributes to activities of collaborative environmental knowledge creation [5][6]. Also, a multi-dimensional and multi-layered visualization and Monitoring-Analysis-Warning functions of 5D World Map System for building disaster resilience has been proposed for monitoring Sustainable Development Goals in United Nations ESCAP [8].

The feature of the proposed system is in its dynamics of dimensional selection and visualization.

In the system, every parameter value is expressed in a dimensional way and the correlation calculation is performed in each dimension. The domain of target data (Eg. a set of healthy cases, a set of disease cases, a set of cases in a specific area, a set of cases with the same age, etc.) is also dynamically selected by dimensional selection.

3.2. "5D World Map System for Human-Health-Analysis"

We have presented several application systems with 5D World Map and semantic analytical visualization for environmental monitoring. In paper [5][6][7][8], we presented river-water quality analysis and its visualization on 5D World Map System. As one of the important experiments by applying SPA concept in 5D World Map System, we introduced a seawater quality analysis and its visualization on 5D World Map System in [9][10][11]. The experiment conducts 1) seawater quality sensor data in 12 criteria to find polluted point of area after flash flood as physical data, and 2) text article of warning message from local government as cyber information. The experiment coordinates the physical data and cyber information on 5D World Map. The important meaning of the

experiment is that the combination of SPA in semantic analysis on 5D World Map are effective to make actuation of warnings for users for protecting sea creatures, and it also can be applicable to make actuation of warnings directly for users and indirectly for lifeguards sectors in local government to protect the human in seawater. An example result of the experiment is shown in **Figure 3**.

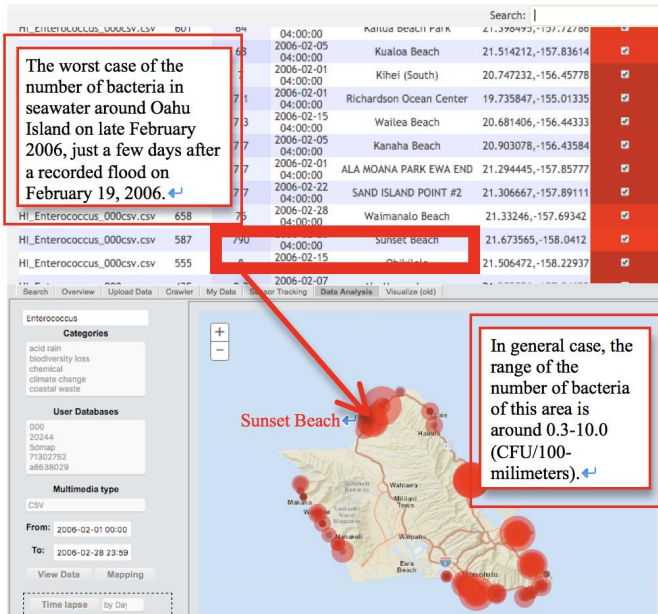


Figure 3. Semantic space for preemptive medicine based on urine data (An example result of the experiment of environmental analysis by applying SPA in 5D World Map System: Water-quality data in Hawaii mapped onto 5D World Map in case of a few days after the heavy flood on February 2006. (This figure shows that an experiment of bacterial data (CFU/100-millimeters) visualization.) [11])

3.3. Experiments of Seawater-quality Analysis with 5D World Map System for Seawater-Quality Data in Hawaii-Islands

We realized the experimental system using the SPA functions in 5D World Map and applied the collected seawater quality data from points Hawaii-Islands. The major issues of seawater quality in Hawaii-Islands are followings, issue-1) Issue of seawater pollution after flash flood for sea creatures, and issue-2) Issue of seawater pollution after flash flood for human.

For the issue-1, we integrate the following data sources, flash flood warning text from Hawaii State, and water quality data period from the time of warning issued to the

time of 1 week after the warning. The warnings are issued from NOAA National Weather Service Weather Forecast Office. We can upload them to 5D as text data with time stamp and location data. In typical case of pollution for sea creatures around coastal area of seawater, to keeping the appropriate pH value is one of the requirements to keep high survival ratio of sea creatures like corals at around coastal area. We realized that to visualize pH values right after the period of flash flood. The snapshot of output of our system is shown in **Figure 4**. By this experiment, we can see the polluted point of area after flash flood on 5D map and it can be applicable to make actuation of warnings for users at institutes of NGO, NPO that are doing to protect the sea creatures.

For the issue-2, to find the period of risk for human in seawater from bacterial pollution, it is mostly impossible by using only by the visual aspect from human (ex. Turbidity). For the issue, we integrate the following data sources, flash flood warning text from Hawaii State, and water quality data period from the time of warning issued to the time of 1 week after the warning. The warnings are issued from NOAA National Weather Service Weather Forecast Office. We can upload them to 5D as text data with time stamp and location data. In typical case of pollution for human in seawater, to keeping low bacterial value is one of the requirements to keep low ratio of occurring disease of human like swimmers at around coastal area. We realized that to visualize Turbidity and Bacterial values right after the period of flash flood. The snapshot of output of our system is shown in **Figure 3** and **Figure 4**. By this experiment, we can show the time difference of the two values, Turbidity value and bacterial value after flash flood on 5D map and it can be applicable to make actuation of warnings directly for users and indirectly for lifeguards sectors in Hawaii State to protect the human in seawater.

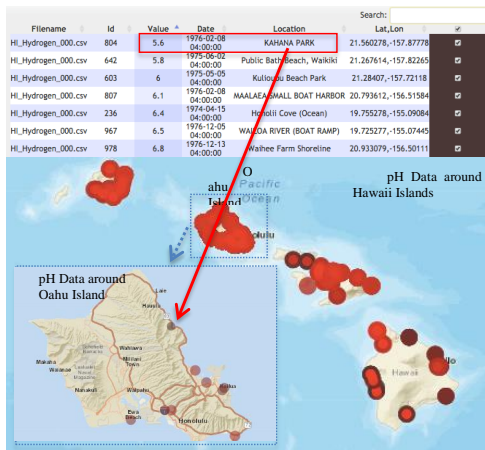


Figure 4. Water-quality data in Hawaii mapped onto 5D World Map. (This figure shows that an experiment of the pH data visualization on the 5D-map by using colors reflected its value. Dark color is expressing lower pH, and it means high hazardness for sea creatures.)

4. Conclusion

We have presented “Human-Health-Analysis Semantic Computing” and “*Human-Health-Analysis Space Creation*” for realizing semantic interpretations of situations and symptoms in human-health. This Human-Health-Analysis Semantic Computing method realizes a new multidimensional data analysis and knowledge sharing for a global-level health monitoring and disease analysis. The computational results are able to be analysed by the time-series difference of the value of each place, the differences between the values of multiple places in a focused area, and the time-series differences between the values of multiple places to detect and predict a potential-risk of diseases.

We have applied this method to biological data resources as a new experiment for biological data. This system enables to create a remote, interactive and real-time global and academic research exchange between remote areas. We have also presented *Human-Health* Computing system and the 5D World Map System, as an international and environmental research platform with Spatiotemporal and semantic analysers.”

As our future work, we will extend our semantic computing system to new international and collaborative research and education for realizing mutual understanding and knowledge sharing on global human-health issues in the world-wide scope.

Acknowledgement

I would like to appreciate Dr. Chalisa Veesommai and Ms. Yoshiko Itabashi for their significant discussions and experimental studies.

REFERENCES

- [1] Yasushi Kiyoki, Takashi Kitagawa, T. and Takanari Hayama: “A metadatabase system for semantic image search by a mathematical model of meaning,” ACM SIGMOD Record, vol. 23, no. 4, pp.34-41, 1994.
- [2] Yasushi Kiyoki, Takashi Kitagawa and Takanari Hayama: “A metadatabase system for semantic image search by a mathematical model of meaning,” *Multimedia Data Management -- using metadata to integrate and apply digital media--*, McGrawHill(book), A. Sheth and W. Klas (editors), Chapter 7, 1998.
- [3] Yasushi Kiyoki and Saeko Ishihara: “A Semantic Search Space Integration Method for Meta-level Knowledge Acquisition from Heterogeneous Databases,” *Information Modeling and Knowledge Bases (IOS Press)*, Vol. 14, pp.86-103, May 2002.
- [4] Yasushi Kiyoki, Shiori Sasaki, Nhung Nguyen Trang, Nguyen Thi Ngoc Diep, “Cross-cultural Multimedia Computing with Impression-based Semantic Spaces,” *Conceptual Modelling and Its Theoretical Foundations, Lecture Notes in Computer Science*, Springer, pp.316-328, March 2012.
- [5] Yasushi Kiyoki: “A “*Kansei*: Multimedia Computing System for Environmental Analysis and Cross-Cultural Communication,” 7th IEEE International Conference on Semantic Computing, keynote speech, Sept. 2013.
- [6] Shiori Sasaki, Yusuke Takahashi, Yasushi Kiyoki: “The 4D World Map System with Semantic and Spatiotemporal Analyzers,” *Information Modelling and Knowledge Bases*, Vol.XXI, IOS Press, 18 pages, 2010.

- [7] Totok Suhardijanto, Yasushi Kiyoki, Ali Ridho Barakbah: "A Term-based Cross-Cultural Computing System for Cultural Semantics Analysis with Phonological-Semantic Vector Spaces," *Information Modelling and Knowledge Bases XXIII*, pp.20-38, IOS Press, 2012.
- [8] Yasushi Kiyoki, Xing Chen, Shiori Sasaki and Chawan Koopipat: "Multi-Dimensional Semantic Computing with Spatial-Temporal and Semantic Axes for Multi-spectrum Images in Environment Analysis", to appear in *Information Modelling and Knowledge Bases (IOS Press)*, Vol. XXVI, 20 pages, March 2016.
- [9] Chalisa Veesommai, Yasushi Kiyoki, Shiori Sasaki and Petchporn Chawakitchareon, "Wide-Area River-Water Quality Analysis and Visualization with 5D World Map System", *Information Modelling and Knowledge Bases*, Vol. XXVII, pp.31-41, 2016.
- [10] Chalisa Veesommai, Yasushi Kiyoki, "Spatial Dynamics of The Global Water Quality Analysis System with Semantic-Ordering Functions", *Information Modelling and Knowledge Bases*, Vol. XXIX, 2018.
- [11] Yasushi Kiyoki, Asako Uraki, Chalisa Veesommai, "A Seawater-Quality Analysis Semantic- Space in Hawaii-Islands with Multi- Dimensional World Map System", 18th International Electronics Symposium (IES2016), Bali, Indonesia, September 29-30, 2016.
- [12] Shiori Sasaki, Koji Murakami, Yasushi Kiyoki, Asako Uraki, "Global & Geographical Mapping and Visualization Method for Personal/Collective Health Data with 5D World Map System", *Proceedings of the 30th International Conference on Information Modelling and Knowledge Bases (19 pages)*, Hamburg, Germany (Online), June, 2020.
- [13] Sasaki, S. and Kiyoki, Y., "Real-time Sensing, Processing and Actuation Functions of 5D World Map System: A Collaborative Knowledge Sharing System for Environmental Analysis", *Information Modelling and Knowledge Bases*, Vol. XXVIII, IOS Press, pp. 220-239, May 2016.

A Conceptual Framework for the Conversion of a Text Document into TIL-Script

Martina Číhalová¹, Marek Menšík²

¹ Palacký University Olomouc, Department of Philosophy, Czech Republic

² VSB-Technical University Ostrava, Department of Computer Science, Czech Republic

martina.cihalova@upol.cz mensikm@gmail.com

Abstract: The paper deals with the introduction of the TILUS tool for the needs of retrieval of appropriate textual information sources and natural language processing. TILUS tool assumed up to the present that all the texts are formalized in TIL-Script, the computational variant of Transparent Intensional Logic (TIL). We outline a general proposal for utilizing Stanford typed dependencies representation for automate conversion of natural language into TIL-Script. In order to be able to correctly solve this problem, we also introduce our general conceptualization which is able to cover the thematic variations of the processed texts.

Keywords: *Information sources, Stanford typed dependencies, Transparent Intensional Logic, TIL-Script, Ontology.*

1 Introduction

In this paper we briefly introduce the current state of development of the TILUS tool for the needs of retrieval of appropriate of textual information sources and natural language processing. The paper is organized as follows. In section 2, there is a brief introduction of TILUS and its functions. TILUS assumes that all the data are previously formalized in TIL-Script, the computational variant of Transparent Intensional Logic (TIL). This automated conversion of natural language texts into TIL-Script was not, however, incorporated within TILUS. The aim of this paper is to outline the basis of a strategy to automate or at least semi-automate the conversion of natural language into TIL-Script. In order to be able to correctly solve this problem, we introduce our general conceptualization which is able to cover the thematical variations of processed texts in section 3. Finally, we outline how to utilize the Stanford dependencies relations for automate conversion of natural language into TIL-Script in section 4. Concluding remarks on further research can be found in Section 5.

2 A brief introduction of TILUS for retrieval of appropriate textual information sources

We are developing the TILUS tool for the needs of retrieval of appropriate textual information sources and natural language processing. This tool has a number of modules that can be divided into two main parts. The first part is the module of logical deduction in order to be able to infer the conclusions of the premises and verify the validity of the arguments. The inference is based on natural deduction. This part was introduced in [1], [2], [3]. The second part concerns machine learning and searching for relevant text sources which was introduced in [4], [5], [6], [7]. Our system currently assumes that all the inputs are transformed into TIL-Script. A module for the automatic transformation of natural language texts into TIL-Script has not, however, been

developed yet.

We are in favour of TIL-Script, the computational variant of Transparent Intensional logic (TIL), because its procedural semantics are close to natural language and have a great expressive power. TIL was developed by Pavel Tichý who introduced its main principles in [8]. The complete work of Tichý can be found in [9]. TIL has further been developed by M. Duží, P. Materna and B. Jespersen, J. Raclavský and a great deal of the current TIL research can be found in [10]. The outcomes of this book are especially relevant to the inter alia conceptual modelling area. TIL-Script as well as TIL is hyperintensional, typed, partial specification language with a ramified hierarchy of types.

The aim of this paper is not to introduce the current system in detail, but to outline a solution for the conversion of text in natural language into TIL-Script. In order to provide however, the reader with an overview of the current state, we first briefly summarize the functionality of the individual modules of the TILUS tool.

The deduction module is a module for the conclusion inference of premises and for verifying the validity of arguments using natural deduction. Due to the fact that TIL-Script is a robust system with great expressive power, our deductive system had to be modified and new deduction rules had to be introduced, such as beta-conversion rule, rules for working with lambda quantifiers, etc.

In the module for ‘machine learning and text source recommendation’, which is crucial for proposals of solutions in this paper, we deal with how to generate an explication in TIL-Script of an atomic concept. This explanation can be understood as a brief description of the sought term in a particular source text. We generate an explication from each source text and present it to the user of our application. The user then selects a particular explication based on his/her preferences and our system then recommends other possible relevant documents. The theoretical framework for the recommendation module design is the method of association rules and FCA theory, where the relevant documents are arranged according to their significance to the user, for more see [11]. First, we need to analyse textual resources to obtain the basis for the formalised TIL-Script constructions. The conversion from natural language to the TIL-Script is a quite demanding process and requires logical-linguistic analysis. All our contributions published up to now therefore presumed cooperation with the Department of Computational Linguistics to obtain the conversion of text into TIL-Script constructions. A set of relevant propositional constructions is then selected from the formalized set; namely those where the concept to be explicated occurs. The scheme for the preparation of the text source is depicted in Fig. 1.

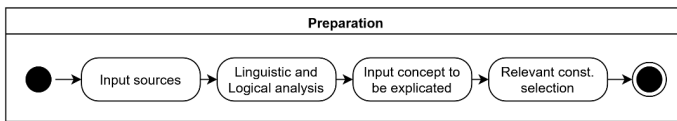


Fig. 1. Pre-processing and formalisation of textual resources

Next, the set of selected constructions serves as input for machine learning techniques, in particular, the *Inductive heuristics* module, for more see [4]. This module produces hypothetical molecular concepts that should explicate the simple concept to be learned. Since there are several resource text documents from which the hypothetical concepts are extracted, we obtain several explications of the learned simple concept. The inductive heuristic modules are schematically specified by the following Fig. 2.

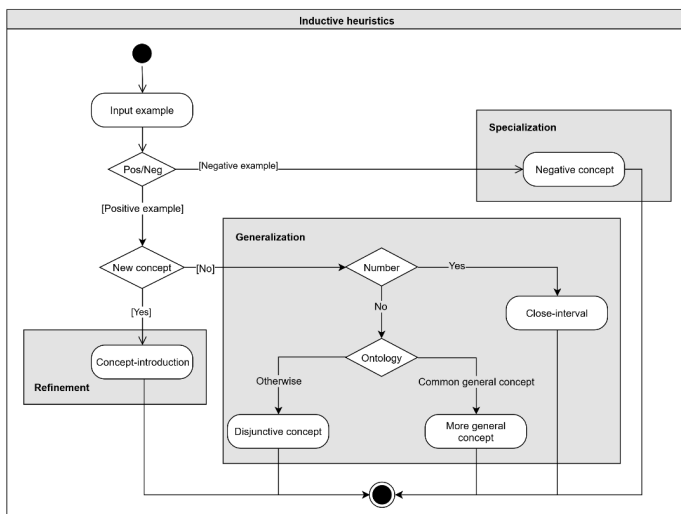


Fig. 2. Inductive heuristics

The last module of *Relevant Source Selection* is still a work in progress. It is the module that deals with hypotheses processing and their evaluation. There are several functionalities that might be realised here. They include, inter alia, filtering out irrelevant sources according to additional user-defined linguistic and logical criteria, searching for inconsistencies among the hypotheses such as *contrarities* and *contradictions*.

As already mentioned, the TILUS tool assumes that all the data are formalized in TIL-Script, the computational variant of TIL. The next step is to automate or at least semi-automate the conversion of natural language into TIL-Script as a crucial part of TILUS tool development. In order to be able to correctly solve this problem, we also introduce our general conceptualization which is able to cover the thematic variations of the processed texts in the following section.

3 Basic ontological types and their logical types

In this section, a general conceptual framework for our approach is proposed.¹ This conceptual framework was designed as a general and basically universal ontology close to natural language. Hence, it can also be successfully utilized for appropriate information retrieval from textual sources.

The starting point for building a general conceptual framework is to distinguish between static objects (static entities), such as particular individuals and necessary relations between their properties, and dynamic entities such as *activities* which are detected by some special type of verbs. The proposed analysis makes use of Tichý's formulation where such verbs are called *episodic verbs*. Tichý [14, pp. 263–296] draws a distinction between *episodic* and *attributive* verbs. Episodic verbs (e.g. *drive*, *tell*, etc.) express the actions of objects or people as opposed to attributive verbs (e.g. *is heavy*, *looks speedy*) that ascribe certain empirical properties. Hence, *activities* are detected just by episodic verbs. Both static and dynamic entities are characterized

¹ This conceptual framework is also introduced in [12] as a general ontology for processes and in [13] as a conceptual framework for logical classification of Wh-questions and possible answers to such questions in a multi-agent system.

by their further specification. We will now, however, focus on analysis of activities and their characteristics in more detail.

Our analysis of activities proceeds from John Sowa's linguistically oriented approach to logical specification of natural language sentences. We focus on his analysis of sentences containing verbs, specifically *episodic verbs* according to Tichý's formulation. This analysis is based on the linguistic theory of verb valency frames.

The semantics of the respective verb is provided via its valency frame. In general, valency is the ability of a verb (or another word class) to bind other formal units, i.e. words, which cooperate to provide its meaning completely. These units are so-called *functors* or *participants* or *case roles*. Thus the valency of a verb determines the number of arguments (*participants*) controlled by a verbal predicate. This ability of verbs (or lexemes in general) results from their meaning rather than from formal aspects. Despite this fact, lexical and grammatical valency can be distinguished. *Grammatical valency* contains information about the formal aspects of a verb, such as the grammatical case, while *lexical valency* contains information about the semantic character of particular participants. Grammatical valency depends on the language that is being analysed. Lexical valency, in contrast, is language independent, since it is established semantically. More details on the theory of valency frames can be found, for instance, in [15]. Attention to lexical valency has to therefore be paid in order to build up ontology independent from the used language. There are several types of lexical verb valency. An impersonal (avalent) verb has no subject or a dummy subject. "It rains" is a typical example. Here the grammatical subject 'it' is only a dummy subject because it does not refer to any concrete object. An intransitive (monovalent) verb has only one argument, the subject S; "John (S) is singing." A transitive (divalent) verb has two arguments, an agent (A) and a patient (P), as in "John (A) kicked the ball (P)." A ditransitive verb has three arguments, etc.² Consider the example of the sentence *John is going to Ostrava by train*. The activity *is going* has the following participants: the actor (*John*), the instrument of transport (*train*) and the final destination (*Ostrava*).

Verb valency frames also determine the obligatory and facultative arguments of a given verb, together with their types. Facultative arguments can be missing, of course. One might consider, for example, the verb *chastise*. This verb has two obligatory participants *who* (agent) and *whom* (patient). In addition, this verb can be connected with other facultative participants which express inter alia locality and time such as in the following sentence: "A teacher chastises a student in the school early in the morning." It would be useful to classify verb participants into types according to their semantics. There are many classifications, however, of the participant types described in the literature. Three approaches to classification, according to the two valency dictionaries for the Czech language VALLEX and VerbaLex³ and John Sowa's approach, are briefly compared in [19].⁴

Sowa uses the term *thematic roles* for the verb valency participants. His summary of all the types of thematic roles can be found in [21, pp. 506-510] or in his web source *Thematic Roles* [22]. Sowa developed the system of conceptual graphs which are specified in [23, p. 187] as the system of logic for representing natural language semantics. Unlike predicate calculus, which was designed for studies in the foundations of mathematics, conceptual graphs were designed to simplify the mapping to and from natural language. They are based on a graph notation for logic first developed by the philosopher and logician C. S. Peirce. A conceptual graph is represented as a labelled bipartite graph. Nodes in set *C* are called concepts and nodes in set *R* are called conceptual relations. Thematic roles are represented by conceptual relations that link the concept of a verb to the concepts of the participants in the *occurrent* expressed by

² For details, see [16].

³ See, for instance [17] and [18].

⁴ A very detailed comparison of these three classifications was provided in [20].

the verb. Apart from the graph notation, there is an equivalent linear notation for conceptual graphs where the boxes are represented by square brackets, and the circles are represented by parentheses. Sowa distinguishes between several types of thematic roles, for instance:

Agent as an active animate entity that voluntarily initiates an action, example: *Eve bit an apple*: [Person: Eve] ← (Agnt) ← [Bite] → (Ptnt) → [Apple],

Theme as an essential participant that may be moved, said or experienced, but is not structurally changed, example: *Billy likes the Beer*: [Person: Billy] ← (Expr) ← [Like] → (Thme) → [Beer: #],

Beneficiary as a recipient that derives a benefit from the successful completion of an event, example: *Diamonds were given to Ruby*: [Diamond: { * }] ← (Thme) ← [Give] → (Benf) → [Person: Ruby],

Destination as the goal of a spatial process, example: *Bob went to Danbury*: [Person: Bob] ← (Agnt) ← [Go] → (Dest) → [City: Danbury],

Instrument as a resource that is not changed by an event, example: *The key opened the door*: [Key: #] ← (Inst) ← [Open] → (Thme) → [Door: #],

Location as an essential participant of a spatial nexus, example *Vehicles arrive at a station*: [Vehicle: { * }] ← (Thme) ← [Arrive] → (Loc) → [Station],

Patient as an essential participant that undergoes some structural change as a result of an event, example: *The cat swallowed a canary*: [Cat: #] ← (Agnt) ← [Swallow] → (Ptnt) → [Canary: #],

etc. For details, see [21, pp. 508-510].

From the logical point of view, we deal with the verb phrases as denoting a function that is applied to its arguments. The number of arguments is controlled by the content verb valency. Dynamic entities relating to *activities* can be characterised by the special relationships in intensions between activities and their participants modelled as functions of TIL-Script. In our background theory (TIL), we view α -intensions as functions mapping possible worlds (of type ω) to type β . Type β is frequently the type of chronology of the elements of type α . These α -chronologies are, in turn, functions mapping time (of type τ) to type α . Thus, α -intensions are usually mappings of type $(\omega \rightarrow (\tau \rightarrow \alpha))$, or in TIL notation $((\alpha\tau)\omega)$, $\alpha_{\tau\omega}$ for short and in TIL-Script notation $((\alpha \text{ Time})\text{World})$. Another frequent type of intension is the *property of individuals*, an object of type $(\text{ot})_{\tau\omega}$, in TIL-Script notation $((\text{Bool Indiv})\text{Time})\text{World}$. Consider, for example, the above mentioned sentence *John is going to Ostrava by train*. The specification of this sentence in TIL-Script is the following one:⁵

```
\wt ['And [['Agent@wt 'John 'Is_going] [['And ['Destination@wt 'Ostrava 'Is_going]
['instrument@wt 'train 'is_going]]]
```

The types are the following:

Agent, Destination / (((Bool Indiv (((Bool Indiv)Time)World))Time)World)

Instrument / (Bool (((((Bool Indiv)Time)World) (((Bool Indiv)Time)World))Time)World)

Ostrava, John / Indiv

Is_going, train / (((Bool Indiv)Time)World)

Now let us return to the conceptual characteristics of *static entities*. Static entities can be characterised by their properties and attributes. From the linguistic point of view, the properties assigned to them are usually denoted by a copular verb + adjective. Hence, the typical form of a sentence, characterizing a static object, is “*S Cv Adj*”, where *S* is the subject, *Cv* a copular

⁵Note the specific notation character @wt. This character expresses an intensional descent to the respective possible world and time to obtain the final value of this intension in this possible world and time.

verb and *Adj* an adjective. Typical copular verbs are *is, am, are, ... , appear, seem, look, sound, smell, taste, feel, become* and *get*. In the conceptual analysis of a given domain, it is useful to distinguish between two essential classes of characteristics of static objects. They are the relatively stable properties of objects and dynamic empirical facts about these objects. The former can be called ‘substantive’ properties and the latter ‘accidental’ properties. One might consider traffic domain, for example, as the domain of interest. The substantial property is then that of being a car with its subsumed properties such as the type of car, namely a *Sedan, Minivan, Hatchback*, etc. According to the ISA hierarchies, of course, each car is a vehicle. The accidental properties of an individual car include, inter alia, the *speed limit, weight, colour, date of manufacture*, etc. In addition, each substantive property is mostly associated with certain accidental attributes of the individuals representing the given substantive property. The substantive property of a car is associated, for example, with the above accidental properties, and also its *owner*, etc. Substantive properties are from the logical point of view the functions whose input is the respective possible world and time and individual while the output is the truth value. The logical type of many accidental properties is the same. However, some accidental properties such as the *weight, speed limit*, etc. are, for example, the functions with a possible world, time and individual as an input and the output is the number.

This conceptual framework is within its universality able to cover the thematical variations of the processed texts. How to achieve, however, the automated or at least semi-automated conversion of the natural language text into TIL-Script? In the following section, we examine the possibility of utilizing Stanford typed dependencies representation for this purpose.

4 The utilization of Stanford typed dependencies representation for the conversion into TIL-Script

Universal Dependencies (UD) is a framework for consistent annotation of parts of natural language texts across different human languages. It was designed to provide a straightforward description of grammatical relations for any user who could benefit from automatic text understanding. UD is an open community effort with over 300 contributors producing nearly 200 treebanks in over 100 languages. It is a project which is developing a cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective.⁶ The annotation scheme is based on the evolution of (universal) Stanford dependencies see [24], Google universal part-of-speech tags, see [25], and the Intersect interlingua for morphosyntactic tagsets, see [26]. The theoretical framework for the types of syntactical relations are Stanford dependencies which map straightforwardly onto a directed graph representation, in which the words in the sentence are nodes in the graph and the grammatical relations are edge labels. This representation contains approximately 50 grammatical relations and in [27], you can find the list of all the types with their definitions and natural language examples. There are some types of grammatical relations with examples below:

acomp: adjectival complement. An adjectival complement of a verb is an adjectival phrase which functions as the complement (like an object of the verb). Example: *She looks very beautiful*, where the term *beautiful* is related to the term *looks*. The type of relation is *acomp*.

advmod: adverb modifier. An adverb modifier of a word is a (non-clausal) adverb or adverb-headed phrase that serves to modify the meaning of the word. Examples: *Genetically modified food*: *advmod(modified, genetically)*, *less often*: *advmod(often, less)*.

attr is a relation intended for the complement of a copular verb such as “to be”, “to seem”, “to appear”.

nsubj: nominal subject. A nominal subject is a noun phrase which is the syntactic subject of a clause. The governor of this relation might not always be a verb: when the verb is a copular verb, the root of the clause is the complement of the copular verb, which can be an adjective or noun. Example: *Clinton defeated Dole*: *nsubj*(defeated, Clinton), *The baby is cute*: *nsubj*(cute, baby).

num: numeric modifier. A numeric modifier of a noun is any number phrase that serves to modify the meaning of the noun with a quantity. Examples: *Sam ate 3 sheep*: *num*(sheep, 3), *Sam spent forty dollars*: *num*(dollars, 40).

agent: agent. An agent is the complement of a passive verb which is introduced by the preposition “by” and does the action. This relation only appears in the collapsed dependencies, where it can replace *prep by*, where appropriate. It does not appear in basic dependencies output. Examples: *The man has been killed by the police* *agent*(killed, police), *Effects caused by the protein are an important* *agent*(caused, protein).

These relations provide a straightforward description of grammatical relations in the analysed text. We will demonstrate that we can also extract some important semantic information from this Stanford specification. We decided to use the tool *Explosion* for the text annotation which is based directly on Stanford dependencies. This is a software company specializing in developer tools for Artificial Intelligence and Natural Language Processing. They are the makers of spaCy, the leading open-source library for advanced NLP and Prodigy, an annotation tool for radically efficient machine teaching. The next step will be the precise specification of the rules for the exact syntactical relations conversion into logical types and TIL-Script. There will be a rough outline of how we are going to proceed in the next paragraph.

We are able to syntactically recognize the substantive properties and accidental properties. For comparison, let’s take the sentence *The cat is a fast mammal*. The relation *advmod* (adverb modifier) determines the accidental property and the *attr* relation determines the essential property.

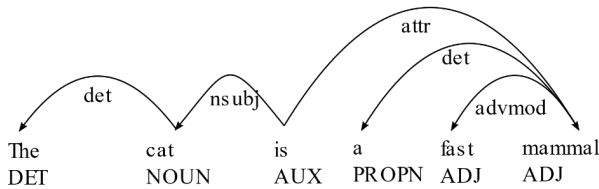


Fig. 3. The output of *Explosion* for the sentence *The cat is a fast mammal*

We are also able to detect the activity and its participants in general on the basis of Stanford typed dependencies relations. Let us take the above mentioned example of the sentence *John goes to Ostrava by train*. The verb valency participants of the activity *going* are the following: *John* is the agent of the activity, *Ostrava* is the final destination and *train* is the instrument of transport. For comparison, the respective output of this sentence from *Explosion* is in Fig. 4.

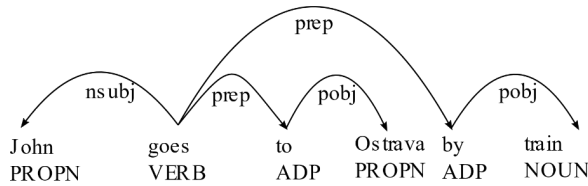


Fig. 4. The output of *Explosion* for the sentence *John goes to Ostrava by train*

We are able to recognize the activity determined by the verb and the agent of this activity which is determined by the proper name *John* and is in the relation *nsubj* with the verb. The terms *Ostrava* and *train* are detected as consecutive relations *prep* (preposition) and *pobj* (object of preposition). We can therefore recognize in an automated way via analysis of dependencies that these terms are the verb valency participants in general, but not the respective categories of these participants. For this purpose, we are considering the possibilities of utilization of the verb valency dictionaries. Another possibility could be querying the user to fill in the particular type of participant manually.

5 Conclusion and future work

The aim of this paper was to outline the basis of the strategy to automate or at least semi-automate the conversion of natural language into TIL-Script, the computational variant of TIL. This strategy was outlined for the currently developed TILUS tool for the needs of retrieval of appropriate textual information sources and natural language processing. To reach this goal, we need to first map the concepts in natural language into the respective logical types. We introduce for these purposes the general conceptualization which is able to cover the thematical variations of the processed texts. Each concept corresponds to respective logical type. We also outline how the Stanford typed dependencies representation can be utilized for automated conversion of natural language into basic ontological types. We are currently working on the precise specification of the rules for this automated conversion of outputs from *Explosion* to TIL-Script. Future research should involve adding the extension based on natural deduction to the explication module.

Acknowledgements

The work on this paper was supported by the project ‘JG_2020_005 Times, Events, and Logical Specification’ of Palacký University and is partially supported by Grant of SGS No. SP2021/87, VSB - Technical University of Ostrava, Czech Republic.

References

1. Duží, M., Menšík, M. (2020): Inferring knowledge from textual data by natural deduction. *Computación y Sistemas*, Vol. 24, No. 1, 2019, pp. 29-48, ISSN: 14055546, doi: 10.13053/CyS-24-1-3345
2. Duží, M., Menšík, M., Pajr, M., Patschka, V. (2019): Natural deduction system in the TIL-script language. In *Frontiers in Artificial Intelligence and Applications*, vol. 312, Endrjúkaite T., Jaakkola H., Dudko A., Kiyoki Y., Thalheim B., Yoshida N. (eds.), pp. 237-255, Amsterdam: IOS Press, doi: 10.3233/978-1-61499-933-1-237
3. Duží, M., Menšík, M. Vich, L. (2012), Deduction system for TIL-2010. *RASLAN*

2012. Brno: Masaryk University, 2012. p. 49-53 ISBN: 978-802630313-8.
4. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Refining concepts by machine learning. *Computación y Sistemas*, Vol. 23, No. 3, 2019, pp. 943–958, doi: 10.13053/CyS-23-3-3242
 5. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Machine Learning Using TIL. In *Frontiers in Artificial Intelligence and Applications*, vol. 321: Information Modelling and Knowledge Bases XXIX, Dahanayake A., Huiskonen J., Kiyoki Y., Thalheim B., Jaakkola H., Yoshida N. (eds.), pp. 344-362, Amsterdam: IOS Press, doi: 10.3233/FAIA200024
 6. Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Seeking relevant information sources. In *Informatics '2019*, IEEE 15th International Scientific Conference on In-formatics, Poprad, Slovakia, pp. 271-276.
 7. Albert, A., Duží, M., Menšík, M., Pajr, M., Patschka, V. (2021): Search for Appropriate Textual Information Sources. In *Frontiers in Artificial Intelligence and Applications*, vol. 333: Information Modelling and Knowledge Bases XXXII, B. Thalheim, M. Tropmann-Frick, H. Jaakkola, N. Yoshida, Y. Kiyoki (eds.), pp. 227-246, Amsterdam: IOS Press, doi: 10.3233/FAIA200832
 8. Tichý, P. (1988): *The Foundations of Frege's Logic*, Berlin, New York, De Gruyter.
 9. Tichý, P. (2004): *Collected Papers in Logic and Philosophy*, V. Svoboda, B. Jespersen (eds.).
 10. Duží, M., Jespersen, B. and Materna, P. (2010): *Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic*. Berlin, Springer, series Logic, Epistemology, and the Unity of Science, vol. 17.
 11. Menšík, M., Albert, A., Patschka, V. (2020): Using FCA for Seeking Relevant Information Source. In *RASLAN 2020*. Brno: Tribun EU, 2020. 144 p. ISBN 978-80-263-1600-8, ISSN 2336-4289.
 12. Číhalová, M. Conceptual Framework for Process Ontology. Submitted to the *Synthese* (An International Journal for Epistemology, Methodology and Philosophy of Science).
 13. Číhalová, M., Duží, M. Modelling dynamic behaviour of agents in a multi-agent system; Wh-questions and answers. Submitted to the *Logic Journal of the IGPL*.
 14. Tichý, P. (1980). The semantics of episodic verbs. *Theoretical Linguistics*, vol. 7, 263-296. Reprinted in (Tichý 2004: 411–446).
 15. Fischer, K., Ágel, V. (2010). Dependency grammar and valency theory. *The Oxford handbook of linguistic analysis*, 223-255.
 16. Dixon, R. M. W. (2000). A Typology of Causatives: Form, Syntax, and Meaning. In R. M. W. Dixon & A. Y. Aikhenvald (eds.), *Changing Valency: Case Studies in Transitivity* (pp. 30-41). New York, NY: Cambridge University Press.
 17. Lopatková, M., Žabokrtský, Z., Kettnerová, V., *Valenční slovník českých sloves*. Praha, Karolinum, 2008. ISBN 978-80-246-1467-0.
 18. Hlaváčková, D., Horák, A. (2006). VerbaLex - New Comprehensive Lexicon of Verb Valencies for Czech. In *Computer Treatment of Slavic and East European Languages* (pp. 107-115). Bratislava, Slovakia: Slovenský národný korpus.
 19. Číhalová, M. (2016): Event ontology specification based on the theory of valency frames. In T. Welzer, H. Jaakkola, B. Thalheim, Y. Kiyoki and N. Yoshida (eds.), *Frontiers in Artificial Intelligence and Applications, Information Modelling and Knowledge Bases XXVII* (pp. 299-313). Amsterdam, Berlin, Tokyo, Washington DC:

IOS Press.

20. Číhalová, M. (2011): *Jazyky pro tvorbu ontologií (Languages for ontology building)*. Ph.D. Thesis, VŠB-Technical university of Ostrava, Ostrava, The Czech Republic.
21. Sowa, J. F. (2000). *Knowledge representation (logical, philosophical, and computational foundations)*. Pacific Grove, CA: Brooks Cole Publishing Co.
22. Sowa, J. F.: Thematic roles. <http://www.jfsowa.com/ontology/roles.htm>. Accessed February 8, 2021.
23. Sowa, J. F. (1991): Towards the expressive power of natural language. In J. F. Sowa (ed.), *Principles of semantic networks. Explorations in the representation of knowledge* (pp. 157-189). San Mateo, CA: Morgan Kaufmann. <https://doi.org/10.1016/C2013-0-08297-7>
24. de Marneffe M., C., Manning, Ch. D. *Stanford typed dependencies manual* https://nlp.stanford.edu/software/dependencies_manual.pdf. Accessed February 8, 2021.
25. Petrov, S., Das, D., McDonald, R. (2012): A universal part-of-speech tagset. In *Proceedings of LREC*.
26. Zeman, D. (2008): Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of LREC*.
27. de Marneffe, M. C., Dozat, T., Silveira, N., Haverinen K. , Ginter, F. , Nivre, J., Manning, Ch. D. Universal Stanford Dependencies: A cross-linguistic typology. https://nlp.stanford.edu/pubs/USD_LREC14_paper_camera_ready.pdf. Accessed February 8, 2021.

Collocation Content Analysis of Privacy Policies of Medical Apps

Boštjan BRUMEN

University of Maribor (www.um.si), Faculty of Electrical Engineering and Computer science, Smetanova 17, Si-2000 Maribor, Slovenia
bostjan.brumen@uni-mb.si

Abstract. Privacy is a fundamental human right and is widely end extensively protected in the western industrialized world. The recent advances in technologies, especially in the use of applications developed and designed for mobile devices, have led to the rise of its abuse on one hand and a higher awareness of the importance of privacy on the other side. Legal texts protecting privacy have attempted to rectify some of the problems, but the ecosystem giants, together with mobile apps developers, adapted. In this paper, we analyze which data mobile apps developers are collecting. We have focused on a sample of apps in the medical field. The research was done using collocations analysis. A relationship between a base word and its collocative partners was sought. The initial visual results have led us to more detailed studies that unveiled some worrying patterns. Namely, applications are collecting data not only about the users but also about their family members, medical diagnoses, treatments, and alike, going well beyond the “need to function” / functionality threshold.

Keywords. privacy, GDPR, collocation, apps, similarity, medical

1. Introduction

Mobile phones have become a part of our everyday life and a potent multifunctional tool [1]. Mobile technologies have changed our habits and behaviors drastically. By the end of 2020, mobile internet traffic was around 50 % of total web traffic, with Africa and Asia reaching about 60 % [2]. Daily routine previously conducted in an offline world has shifted online and increasingly to mobile devices. Reading mail turned to read email, browsing newspapers moved to check the news and journals on the phone, going shopping by car has moved to sites like Amazon, eBay, Rakuten, Apple, Aliexpress, and others. Visiting friends turned into retweeting their tweets and liking their posts. A smartphone has taken over many users’ central functions without a need for profound technological and science skills.

Because of the ease of use and wide availability of services and apps, users increasingly use their smartphones and apps. Developers are developing them to fulfill the users’ needs. However, there are costs associated with developing applications. There are several business models apps developers can use, such as “Free,” “Freemium,” “Subscription,” “Paid,” and “Paymium” [3]. The most popular business model to monetize an app is Free / Freemium (i.e., advertising) because of its easy implementation and wide acceptance by mobile application users [4]. Over half of apps contain advertising [5].

For ads to be effective, they need to “fit” or “match” the target audience (person) as closely as possible. The matching of users’ preferences and ads is done via real-time bidding (RTB) coupled with the app’s data. In the bidding process, the ad technology companies auction off ad space in the apps made available by developers by sharing sensitive users’ data collected by the app, such as location, device ID, cookies, browsing history, and any other data being collected. The sharing is done with many different companies involved in a very complicated process [6].

Sharing of vast amounts of personal data opens many privacy-related questions. Namely, privacy is a human right and is protected in the western world. However, microtargeting the individuals with ads (and other activities) based on their personal data is doable legally, without disclosure and (proper!) informed consent, completely bypassing laws and regulations [7], or at least their intentions to protect the privacy.

Privacy protection is increasingly important in the medical domain [8] as medical data are extremely sensitive and need special protection. However, users using medical apps agree to the terms and conditions and associated privacy policies set by application developers. Users typically do not read privacy policies and hence do not know what they are sharing not only with the application developers but due to the prevailing “free” business model also with (too) many other companies.

In this paper, the research question is which types of personal data medical apps are collecting. We will answer this question using linguistic analysis of privacy policies and visualization techniques to present the findings.

The rest of the paper is organized as follows. Section 2 presents the literature review dealing with medical apps, data collection, and privacy. In Section 3, we describe our research method and deliver the results. In Section 4, we conclude the paper with final remarks.

2. Literature review

First, we give a brief definition and a description of privacy, followed by a description of the advertisement ecosystem in the mobile apps world.

Privacy has many aspects [9, 10], including :

- informational privacy (e.g., confidentiality, anonymity, secrecy, and data security);
- physical privacy (e.g., modesty and bodily integrity);
- associational privacy (e.g., intimate sharing of personal events);
- proprietary privacy (e.g., self-ownership and control over intangibles such as personal identifiers and genetic data, and tangibles, e.g., ownership of objects); and
- decisional privacy (e.g., autonomy and choice in decision-making).

In this paper, we primarily deal with informational and intangibles proprietary privacy. Privacy belongs to fundamental human rights and has a special place in legal texts. It is explicitly stated under Article 12 of the 1948 Universal Declaration of Human Rights and protected by the 1st, 3rd, 4th, and 5th Amendments of the U.S. Constitution [11]. In European Union, it is protected by Article 8(1) of the Charter of Fundamental Rights of the European Union and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU), and several national constitutions [11].

The implementational part of privacy protection is done on the lower level of the hierarchy of laws.

In the European Union, it is protected by General Data Protection Regulation (GDPR) directive [12] – the Regulation (E.U.) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. In the United States, there is no single (federal level) data or privacy protection law [13].

The latest and strictest state law is California’s California Consumer Privacy Act (CCPA), becoming effective on January 1st, 2020. It was amended and extended by the California Privacy Rights Act and will take effect on January 1st, 2023. The law applies to companies that do business in California. CCPA gives users the right to opt-out of the sale of their personal information. Interestingly, there are exceptions to the definition of “sale,” such that not all personal information transfers are sales, e.g., transferring personal information to a service provider is not a “sale” [14].

The above laws are exemplary because they limit or prohibit the unauthorized collection and use (selling) of personal data. They also require the data controllers (collectors) to indicate what type of personal information is being collected and why with special rights vested upon users. Typically, companies disclose their privacy practices in “Privacy Policy.”

However, privacy policies are hard to read lengthy legal texts in practice, and users do not read them. For example, the Norwegian Consumer Council has conducted an “AppFail” campaign. They have downloaded the terms of service and privacy policies for a set of typical apps. Together they exceed the New Testament in length – and would take more than 24 hours to read out loud [15]. The campaign highlighted the absurd length of these agreements. To use apps, the users often need to waive fundamental privacy rights and agree that apps track them, and personally identifiable data can be resold [16].

By using the apps, users consent to data collection and other practices as described in the privacy policies, not actually knowing what and how is being collected and processed.

On the other hand, tech firms need personal data residing in apps to serve users with targeted ads. Marketing campaigns depend on the successfully placed ads [6]. Data need to be shared (in a legal way) with advertisers, and they are asked to bid on an individual ad space within an application.

The bidding process is a multi-level, complex, automated placement of third-party ads known as real-time bidding (RTB) [17]. The partial data flow is depicted in Figure 1 (source: [18]).

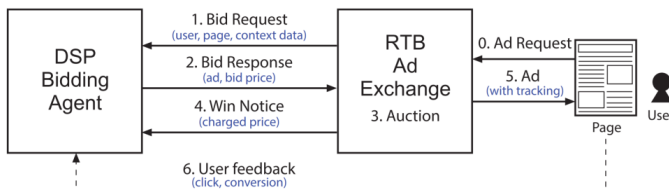


Figure 1: Flow of data and money in an RTB Ad Exchange (Source/Figure from: [18])

In the RTB process, there are several layers of companies involved. It is facilitated by mobile supply-side platforms (SSPs), providing the developers the necessary tools (Software Development Kits, SDKs) that developers build into their apps; the SDKs connect apps to the exchanges. Inside an app, the SSP collects information and enriches it with its own data about the user. The enriched data is sent to ad exchanges that contact demand-side platforms and DSPs to bid for the ad space. DSPs do data matching and possibly bid for the ad space. If the bid is won, a DSP serves a user with an advertisement. The advertiser pays to the DSPs, they pay to the RTB bid exchange (which collaborates closely with SSPs), and the developer gets paid.

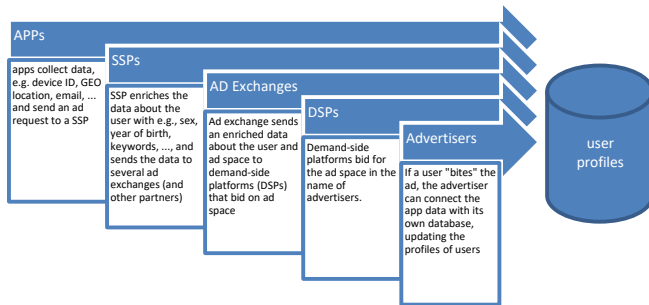


Figure 2: Process flow in a typical ad exchange system

The process of a typical ad exchange system is presented in Figure 2. There are several partners (ad exchanges) competing for advertisement space. Table 1 lists an excerpt of exchanges participating with AdMob, a Google-owned company [19].

Table 1: A partial list of partner RTBs

Open Bidding Partner*	Desktop Web	Mobile Web	Mobile App iOS/Android/Interstitial
Ad Generation	✓	✓	✓
AerServ	✓	✓	✓
Fluct	✓	✓	✓
Improve Digital	✓	✓	✓
Index Exchange	✓	✓	✓
Magnite	✓	✓	✓
Media.net	✓	✓	✓
MobFox	✗	✓	✓
OpenX	✓	✓	✓
PubMatic	✓	✓	✓
ShareThrough	✓	✓	✗
Smaato	✓	✓	✓

Open Bidding Partner*	Desktop Web	Mobile Web	Mobile App iOS/Android/Interstitial
Smart Adserver	✓	✓	✓
Sonobi	✓	✓	✓
Sovrn	✓	✓	✓
SpotX	✓	✓	✓
TripleLift	✓	✓	✓
UnrulyX	✓	✓	✓
Verizon Media	✓	✓	✓
Yieldmo	✓	✓	✓
YieldOne	✓	✓	✓

* Some exchanges participate in Open Bidding but have requested not to be listed in this table.

An RTB system will invade privacy in two ways. Firstly, before the RTB process begins, a myriad of companies have already tracked users, collected their personal information online and offline, and combined them into lengthy user profiles. Again, during the RTB process, a set of companies use these previously acquired profiles to decide how much to pay for the ad space. Secondly, due to the advertisement being displayed, the user leaves some traces behind (e.g., data about click, how long an ad was shown, how the users tracked the ad using eye-tracking algorithms, conversion of clicks to purchases, and many others). The companies involved in the RTB process collect these data and update their profiles, knowing even more about users. The newly updated profiles are feeding the future RTBs. RTB is both a cause of tracking and a means of tracking of personal information [20].

Google controls massive portions of nearly every level of the real-time bidding ecosystem. It is the owner of DoubleClick, an ad network, and AdMob, the largest ad server for the app market [17]. Tech giants send data to advertisers, and advertisers pay to companies. Nevertheless, Google claims it is not selling personal information, as depicted in Figure 3.

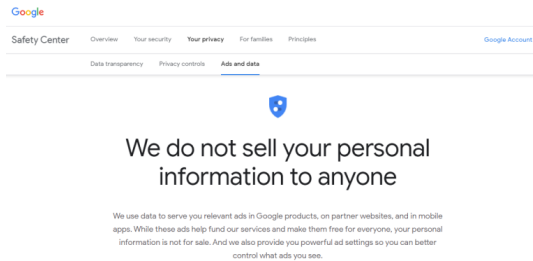


Figure 3: Google's claim about selling personal information (<https://safety.google/privacy/ads-and-data/>, the screenshot was taken 2021-02-05)

It acknowledges that a “sale” is occurring somewhere in this process. Google just insists that they are not selling data, thus breaching the laws [17]. Although facilitating

the RTB process, Google places the responsibility for compliance with the rules upon apps' publishers [21], see Figure 4.

Restrict data processing

When a publisher enables restricted data processing, on the publisher's instruction Google will further limit how it uses data and begin serving non-personalized ads only. Non-personalized ads are not based on a user's past behavior. They are targeted using contextual information, including coarse (such as city-level, but not ZIP/postal code) geo-targeting based on current location, and content on the current site or app or current query terms. Google disallows all [interest-based audience targeting](#) ¹², including demographic targeting and user list targeting when in restricted data processing mode.

Restricted data processing options:

Publishers must decide for themselves when and how to enable restricted data processing mode, based on their own compliance obligations and legal analysis. Two common scenarios are below.

Figure 4: Google's policy on compliance obligations of apps publishers involved in the advertisement business

As described above, a successful advertisement campaign depends on "good" data being collected at the apps' side. A vicious cycle starts with the good quality and quantity of data an app developer makes available in the RTB process. Based on "good" data, "good" ads will be displayed to get the user's attention finally converging to a purchase.

For publishers to get lots of personal data, they need to get users' consent to be rules compliant. The consent is typically hidden in privacy policies. The whole advertisement ecosystem relies on the fact that users neither read privacy policies nor understand or react adequately.

The privacy policies are written in a natural language. A collocation analysis of privacy policies, specifically their parts on personal data collection, can be conducted to extract interesting patterns. A collocation is a set of terms that co-occur more often than would be expected by chance. Collocation can be looked at from three perspectives [22]: a statistical view (co-occurrence), a construction view (a correlation between a lexeme and a lexical-grammatical pattern, or as a relation between a base and its collocative partners), and an expression view (collocations as units of expression). In the present research, we use the statistical and the construction view to extract interesting patterns in the form of keywords [23, 24].

In the following sections, we present a method of analyzing the privacy policies with respect to what medical apps are collecting personal data and the analysis results.

3. Co-occurrence analysis of privacy policies

This section presents the results of the co-occurrence analysis of a sample of free medical apps' privacy policies.

3.1. Data collection and preparation

We browsed the Google Play store in the category “medical” and selected the 50 most popular and free ones (popularity ranked with Google’s algorithms). The exclusion criteria were that the app should have a privacy policy in the first place and be written in English. Out of 50, we ended up with a sample of 32 privacy policies.

We trimmed the irrelevant text of the policies and kept only the parts describing which private data an app is collecting. Next, the documents were tokenized, and part-of-speech details were added (e.g., the word “you’re” was tokenized into “you” and “are”). Punctuations, stopwords, and short words were removed.

3.2. Data processing

We processed and analyzed the data using MathLab® R2020b, using co-occurrence analysis with Text Analysis Toolbox [25].

3.3. Results

Firstly, we calculated single words’ frequencies and visualized the results using the word cloud (Figure 5).

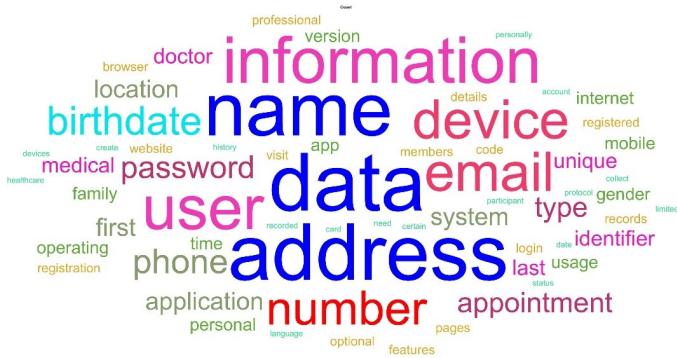


Figure 5: Wordcloud of data collection parts of privacy policies

The visualization of individual words’ frequencies shows that mostly sought-after pieces of information are the user’s email and physical address, phone number, birth date, and (geo)location.

Single-word analysis gives only a glimpse into what words are most frequently used in privacy policies.

Secondly, to put words in context, the co-occurrence analysis was conducted. The results were visualized and manually inspected. The visualization result is presented in Figure 6.

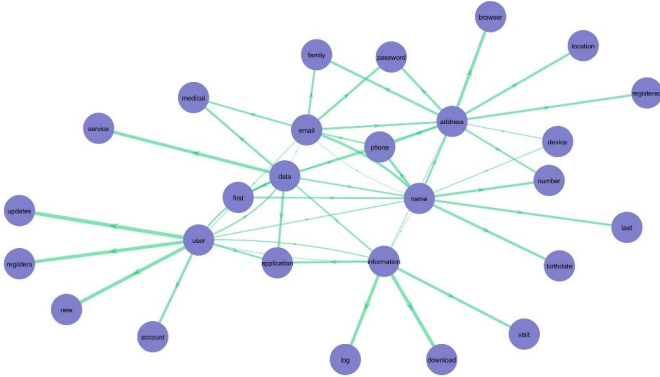


Figure 6: Visualization of co-occurrence analysis of parts of privacy policies

The visualization above shows that apps, not surprisingly, collect information about users, such as names, addresses, email, and alike. However, a detailed inspection unveiled some interesting words worth further investigation. These were “location,” “service,” “device,” “browser,” “password,” “medical,” “family,” and “doctor.”

We separately visualized co-occurrences of the above words in two groups, the first four and the last four words together, see Figure 7 and Figure 8 and, respectively.

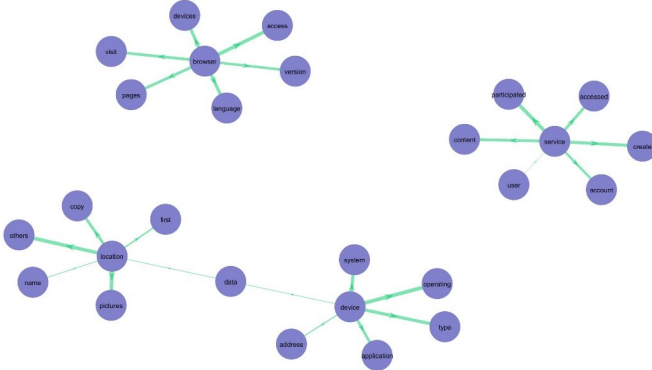


Figure 7: Visualization of co-occurrence analysis, words “location,” “service,” “device,” and “browser.”

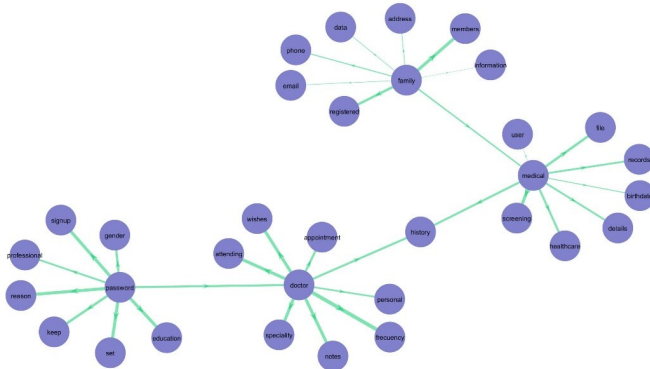


Figure 8: Visualization of co-occurrence analysis, words “password,” “medical,” “family,” and “doctor.”

Here, worrying patterns have shown up. Firstly, it is evident that apps store browser-related data, such as page visits, versions, language used, and devices being used. About devices, they collect their make and type, operating system used, and address (geolocation).

Secondly, apps track the services being used and their content.

Thirdly, passwords are being stored, which poses a considerable security risk [26] and is directly against good password management policies [27]. It is a known fact that users reuse their passwords across several platforms and services. In the RTB ecosystem, where companies share sensitive data, including users’ passwords, such a practice opens a whole new perspective on how securely data are stored and how easy it actually is to gain unauthorized access to users’ data on other platforms.

Fourth, apps collect data about the users and their family members, such as their emails and phone numbers, further enabling the linking of data from various sources.

Finally, medical data are being collected, such as medical files, records, details, and screenings. Closely related are the data about doctor visits and appointments, together with visit frequencies and doctor’s notes taken.

4. Discussion and conclusion

In this paper, we analyzed the texts of a sample of 32 privacy policies of the most frequently used free medical apps in the Google Play store, focusing on privacy policies describing which personal data are being collected and processed by the apps.

We chose free apps due to the prevailing business model in apps markets where developers make their apps available for free and get paid by advertisements being shown inside the apps. A whole ecosystem was built around the advertisements on the web and inside apps, making it a multi-billion-dollar business, predominantly controlled by a single company.

The advertisements are served through a real-time bidding (RTB) process where a multitude of companies are collecting, storing, sharing, and doing business with private

personal data. Due to regulations and restrictions in place, the burden of rules compliance is put on the apps' developers, forcing the users to accept the privacy policies, knowing that users neither read nor understand them. The users actually do not know what type of personal data the applications are collecting and further (for the purpose of advertising) sharing with a myriad of companies.

The paper's novelty is in applying an automated linguistic analysis on privacy policies coupled with visualization techniques to unearth (deliberately!?) hidden information on which types of private data apps are collecting. Best to our knowledge, no previous research on privacy policies was done using the specific linguistic technique combined with visualization.

Based on the co-occurrence analysis of privacy policies, a form of text mining, we have first visualized the results, followed by a manual inspection of simple graphs. Generally, medical apps collect all the "expected" private data, such as user / personal names, gender, email, and physical addresses.

However, a detailed analysis has revealed that apps also collect and store passwords (which is a hazardous practice since users reuse passwords across many websites, services, and apps).

Additionally, medical apps collect and process sensitive medical details, such as medical records, files, notes, doctor's appointments with their frequencies, and alike. What is incredibly worrying is that apps collect data about the app users' family members and their medical data.

Sadly, lenient regulations are not protecting users in the app markets. A simple remedy would require putting personal data on a forbidden list, making trafficking with personal data a criminal offense. It cannot be expected that users opt-out of advertising in every single app they are using. Big players in the market have made the process of protecting personal data deliberately tricky, if not impossible.

Acknowledgment

The author acknowledges the financial support from the Slovenian Research Agency (research core funding No. P2-0057) and the University of Maribor (www.um.si, core funding).

References

1. Brumen, B., et al., *Use of Mobile Technologies in Tourism: Natural Health Resorts Study*. Mediterranean Journal of Social Sciences, 2020. **11**(4): p. 1.
2. Clement, J. *Mobile internet traffic as percentage of total web traffic in October 2020, by region*. 2020 2021-02-01]; Available from: <https://www.statista.com/statistics/306528/share-of-mobile-internet-traffic-in-global-regions/>.
3. Apple. *Choosing a Business Model*. 2021; Available from: <https://developer.apple.com/app-store/business-models/>.
4. Gui, J., et al. *Truth in Advertising: The Hidden Cost of Mobile Ads for Software Developers*. in *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*. 2015.
5. Ruiz, I.J.M., et al., *Impact of Ad Libraries on Ratings of Android Mobile Apps*. IEEE Software, 2014. **31**(6): p. 86-92.
6. Yuan, S., J. Wang, and X. Zhao, *Real-time bidding for online advertising: measurement and analysis*, in *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*. 2013, Association for Computing Machinery: Chicago, Illinois. p. Article 3.

7. Brumen, B., *Automated Text Similarities Approach: GDPR and Privacy by Design Principles*, in *Information Modelling and Knowledge Bases XXXII*, M. Tropmann-Frick, et al., Editors. 2021, IOS Press. p. 213-226.
8. Brumen, B., et al., *Outsourcing medical data analyses: can technology overcome legal, privacy, and confidentiality issues?* *J Med Internet Res*, 2013. **15**(12): p. e283.
9. Allen, A.L., *Privacy Law and Society*. 1st ed. American Casebook Series. 2007: Thomson West.
10. Allen, A.L., *Privacy and Medicine*, in *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, E.N. Zalta, Editor. 2011, Stanford University: Stanford, CA, USA.
11. Solove, D.J., *A Taxonomy of Privacy*. *U Pa L Rev*, 2006. **154**(3): p. 477-564.
12. EU, *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data*. Official Journal of the European Union, 2016. **L:2016:119**.
13. Chabinsky, S. and P.F. Pittman, USA, in *The International Comparative Legal Guide to: Data Protection 2019, 6th Edition*, N. Catlin, T. Hickman, and D. Gabel, Editors. 2019, Global Legal Group: London, UK.
14. Google, *Helping publishers comply with the California Consumer Privacy Act (CCPA)*. 2021 2021-02-03]; Available from: <https://support.google.com/adsense/answer/9560818?hl=en>.
15. Kaldestad, Ø.H. *250,000 words of app terms and conditions*. 2016; Available from: <https://www.forbrukerradet.no/side/250000-words-of-app-terms-and-conditions/>.
16. Phillips, A.M., *Research Handbook on Law and Courts, in All your data will be held against you: secondary use of data from personal genomics and wearable tech*. 2019, Edward Elgar Publishing.
17. Cyphers, B. *Google Says It Doesn't 'Sell' Your Data. Here's How the Company Shares, Monetizes, and Exploits It*. 2020 Electronic Frontier Foundation, March 19, 2020]; Available from: <https://www.eff.org/deeplinks/2020/03/google-says-it-doesnt-sell-your-data-heres-how-company-shares-monetizes-and>.
18. Zhang, W., et al., *Real-time bidding benchmarking with ipinyou dataset*. arXiv preprint arXiv:1407.7073, 2014.
19. Google, *Introduction to Open Bidding*. 2021 2021-02-05]; Available from: <https://support.google.com/admanager/answer/7128453?hl=en>.
20. Cyphers, B. *Behind the One-Way Mirror: A Deep Dive Into the Technology of Corporate Surveillance*. 2019 Electronic Frontier Foundation, December 2, 2019]; Available from: <https://www.eff.org/wp/behind-the-one-way-mirror#Real-time-bidding>.
21. Google, *Helping publishers comply with the California Consumer Privacy Act (CCPA)*. 2021 2021-02-05]; Available from: <https://support.google.com/adsense/answer/9560818?hl=en>.
22. Gledhill, C.J., *Collocations in science writing*. Vol. 22. 2000: Gunter Narr Verlag.
23. Wartena, C., R. Brussee, and W. Slakhorst. *Keyword extraction using word co-occurrence*. in *2010 Workshops on Database and Expert Systems Applications*. 2010. IEEE.
24. Matsuo, Y. and M. Ishizuka, *Keyword extraction from a single document using word co-occurrence statistical information*. *International Journal on Artificial Intelligence Tools*, 2004. **13**(01): p. 157-169.
25. MathWorks. *Co-occurrence analysis with Text Analysis Toolbox*. 2020 [cited 2020-02-05; Available from: <https://github.com/mathworks/Co-occurrenceAnalysis-and-visualization>.
26. Brumen, B., *System-Assigned Passwords: The Disadvantages of the Strict Password Management Policies*. *Informatica*, 2020. **31**(3): p. 459-479.
27. Grassi, P.A., et al., *NIST Special Publication 800-63B. Digital Identity Guidelines. Authentication and Lifecycle Management*. 2017, National Institute of Standards and Technology: Gaithersburg, MD, USA.

Improvement of Searching for Appropriate Textual Information Sources Using Association Rules and FCA

Marek MENŠÍK^a, Adam ALBERT^a, Vojtěch PATSCHKA^a Miroslav PAJR^b

^a*Department of Computer Science, FEI,
VŠB - Technical University of Ostrava, 17. listopadu 15, 708 00 Ostrava,
Czech Republic*

^b*Institute of Computer Science,
Silesian University in Opava, Berzučovo nám. 13, 746 01 Opava,
Czech Republic*

Abstract. This paper deals with an optimization of methods for recommending relevant text sources. We summarize methods that are based on a theory of Association Rules and Formal Conceptual Analysis which are computationally demanding. Therefore we are applying the 'Iceberg Concepts', which significantly prune output data space and thus accelerate the whole process of the calculation. Association Rules and the Relevant Ordering, which is an FCA-based method, are applied on data obtained from explications of an atomic concept. Explications are procured from natural language sentences formalized into TIL constructions and processed by a machine learning algorithm. TIL constructions are utilized only as a specification language and they are described in numerous publications, so we do not deal with TIL in this paper.

Keywords. Association Rules, FCA, Iceberg Lattices, Relevant Ordering

1. Introduction

In case of studying certain problematic area, we need to acquire a list of appropriate papers we want to study to have the whole picture of a particular problem. Therefore in [1], [2] and [3], we introduced methods, where we utilize the methods of Association Rules and the Relevant Ordering based on the Formal Concept Analysis as a theoretical background for selecting the most relevant text-sources. Those methods are based on applying the theory of machine learning and concept explications (more in [4]). Because sentences in the natural language must be formalised into a formal language, we chose to avail of the strong system of Transparent Intensional Logic [5].

The main issue we need to deal with is the time complexity. Making the entire Concept Lattice is immensely time consuming so we were seeking for some time-optimization. Numerous approaches exist to the problem, so we chose to utilize *Iceberg Concepts*. The entire process is based on a horizontal space reduction where we cut a significant part of the concept lattice.

The paper is structured as follows. In chapter 2 we briefly introduce the problem of concept explication which is crucial for the next data processing. In chapter 3 we outline the theories applied in Association Rules, Formal Concept Analysis and Iceberg Lattices. The complete process of finding the set of recommended text sources is demonstrated by an example. For a clear comparison, we used the same example as in [2] and [3]. We point at some problems which might occur using our methods in combination with Iceberg Concepts. Chapter 5 concludes our paper.

2. Explication of an atomic concept

Since we deal with the natural-language processing, we use TIL as our background theory. TIL allows us to formalize salient semantic features of the natural language in a fine-grained way. For more details, see [5].

Atomic concepts are explicated by combination of TIL and machine learning. Explications provide understanding and additional useful information about atomic concepts. *Carnapian explication*¹ is the process of refinement of inaccurate or vague expression. The expression, to be refined, is called an *explicandum*; its refinement, obtained by the explication, is called an *explicatum*. For example, a simple expression such as a dog (explicandum) can be refined as “*Dog is a domesticated carnivore*” (explicatum). In terms of TIL, the explicandum is an atomic concept, i.e. an atomic closed construction. The explicatum is a molecular construction describing the explicandum. We also say that the molecular concept is an ontological definition of the object falling under the atomic concept.

For example:

$${}^{\prime}Dog \approx_{exp} \lambda w \lambda t \lambda x [[{}^{\prime}Domesticated \ {}^{\prime}Carnivore]_{wt} x]$$

$$\text{Types: } Domesticated / ((\text{ot})_{wt} (\text{ot})_{wt}); Dog, Carnivore / (\text{ot})_{wt}; x \rightarrow t$$

The algorithm of obtaining explications has been introduced in [4]. It exploits a symbolic method of supervised machine learning adjusted to the natural language processing. The input of the algorithm are sentences in natural language mentioning the expression to be explicated formalised as TIL constructions.

The algorithm, based on Patrick Winston’s work [7], iteratively builds the explicatum using the constructions marked as positive or negative examples. With positive examples, we refine the explicatum by inserting new constituents into the molecular construction or we generalize the explicatum so that can adequately define the explicandum. With negative examples, we specialize the explicatum by inserting new constituents in the negated way. By those constituents we differentiate the explicatum of our expression from similar expression’s explicata.

¹See [6].

3. Theoretical background

3.1. Association Rules

The method of Association Rules extraction has been introduced in [8]. Yet ten years earlier a similar method has been described in [9]. Basically, it is the process of looking for interesting relations among the large amount of data items. The method can be applied in various areas such as market survey or risk management, and a typical application is a market basket analysis. The goal is to discover associations between data items occurring in a dataset that satisfy a predefined minimum support and confidence. The algorithm first extracts *k-frequent* item-sets, i.e. those item-sets whose occurrences exceed a predefined threshold *k* (minimal support). Then a *confidence* of associations among these frequent item sets is computed and compared with predefined *minimal confidence*. Only those associations that exceed the predefined minimal support and confidence are then considered to be interesting results of the data mining method.

To put these ideas on a more solid ground, here are the definitions. First, we need to define The *support* of a given set $\{i_1, \dots, i_n\}$ of data items. It is the probability of an occurrence of the record with all these items in the dataset.

Definition 1 (support, *k-frequent item-set*). Let $I = \{i_1, \dots, i_n\}$ be a set of data items and $D = \{T_1, \dots, T_m\}$ a dataset of records such that each $T_i \subseteq I$. Then *support* of a set of items (item-set) $A \subseteq I$ in D is

$$supp(A) = \frac{| \{t \in D : A \subseteq t\} |}{|D|}$$

The set A is *k-frequent item-set* iff $supp(A) \geq k$.

Remark. By $|S|$ we denote cardinality of set S . Since $|D| = m$, the support of a set A is the ratio that compares the number of records containing all data items from A to the total number m of records in the dataset.

Definition 2 (confidence of an association rule). Let $I = \{i_1, \dots, i_n\}$ be a set of data items and $D = T_1, \dots, T_m$ a dataset of records such that each $T_i \subseteq I$. Then *association rule* is of the form $A \Rightarrow B$, where $A, B \subseteq I$ and $A \cap B = \emptyset$ and $A, B \neq \emptyset$. *Confidence* of the rule $A \Rightarrow B$ is

$$conf(A \Rightarrow B) = \frac{supp(A \cup B)}{supp(A)}$$

Definition 3 (recommendation rule). Let $A \Rightarrow B$ be an association rule, $E = \{e_1, \dots, e_n\}$ the set of all explications, $e \in E$ the user-selected explication, and let $Prop(e_i)$ be the set of all constituents occurring in an explication $e_i \in E$. Then the rule $A \Rightarrow_e B$ is a *recommendation rule* for a given explication e iff:

$$\begin{aligned} & A \subseteq Prop(e) \\ & B \subseteq \left(\bigcup_{i=1}^n Prop(e_i) \right) \setminus Prop(e) \\ & supp(A \cup B) \geq min-supp \\ & conf(A \Rightarrow B) \geq min-conf \end{aligned}$$

Remark. Obviously, to each explication e there can be more than one recommendation rule for the given explication e .

Definition 4 (recommended sources). Let $A \Rightarrow_e B$ be a recommendation rule for an explication e . Let $exp(d,c)$ be an explication of an input atomic concept c extracted from a textual document d . Then the recommended sources dealing with the concept c according to the rule $A \Rightarrow_e B$ is a set of text-sources RS such that

$$RS = \{d : (A \cup B) \subseteq Prop(exp(d,c))\}$$

This method can be applied for instance in e-shops to recommend other products to be bought once a customer inserts into the shopping basket a given set of products. This feature inspired us to apply the method in our system in order to recommend other possible interesting explications of a given concept once a user votes for one of the obtained explications.

3.2. Formal Concept Analysis

Formal Conceptual Analysis² (FCA) was introduced in 1980s by the group of researchers lead by Rudolf Wille and became a popular technique within the information retrieval field. FCA has been applied in many disciplines such as software engineering, machine learning, knowledge discovery and ontology construction. Informally, FCA studies how objects can be hierarchically grouped together with their mutual common attributes. The following definitions of *significant objects* and *relevant ordering* are originally presented in [3].

Definition 5 (formal context). Let B be a non empty finite set of objects, let M be non empty finite set of attributes and let I be a binary relation $I \subseteq G \times M$ called *incidence* that expresses which objects have which attributes. Then (G, M, I) is called *formal context*.

Definition 6 (formal concept, extent, intent). Let (G, M, I) be a formal context, then $\beta(G, M, I) = \{(O, A) | O \subseteq G, A \subseteq M, A^\downarrow = O, O^\uparrow = A\}$ is a set of all formal concepts of the context (G, M, I) where, $O^\uparrow = \{a | \forall o \in O, (o, a) \in I\}$, $A^\downarrow = \{o | \forall a \in A, (o, a) \in I\}$. A^\downarrow is called *extent* of a formal concept (O, A) and O^\uparrow is called *intent* of a formal concept (O, A) .

Definition 7 (significant objects). The set of *Significant objects* of an object e in $\beta(G, M, I)$ is a set $SO(e) = \bigcup_{i=1}^n O_i^e$, where O^e is an extent of a concept $(O, A) \in \beta(G, M, I)$, $e \in O$ and $B \subseteq M$. Hence, the set of significant objects of an object e is the union of all the extents which the object e is an element of.

Definition 8 (relevant ordering). Let $SO(e)$ be a set of all significant objects of an object e , let $\gamma(e)$ be a set of all concepts (O, A) where $e \in O$, i.e.: $\gamma(e) = \{(O^e, (O^e)^\uparrow) | (O^e, (O^e)^\uparrow) \neq (G, B), B \subseteq M, (O^e, (O^e)^\uparrow) \in \beta(G, M, I)\}$, then $\mathbf{a} \sqsubseteq \mathbf{b}$ is in a *relevant ordering*³ iff

$$\max(|(O^a)^\uparrow|) \leq \max(|(O^b)^\uparrow|), a, b \in SO(e), (O^a, (O^a)^\uparrow), (O^b, (O^b)^\uparrow) \in \gamma(e).$$

²More in [10].

³Classical concept ordering is defined as: $(O, A) \sqsubseteq (O_1, A_1)$ iff $A \subseteq A_1$

3.3. Iceberg Concept Lattices

Iceberg Concept Lattices [11] consist only of the top-most concepts of the concept lattice. Iceberg Concept Lattice is defined as follows:

Definition 9 (iceberg concept lattice). Let $(A, B) \in \beta(G, M, I)$ and let $supp(B) \geq min-supp$, then (A, B) is called *frequent concept*. The set of all frequent concepts of the context (G, M, I) is called the *Iceberg Concept Lattice* of the context (G, M, I)

According to the definition, the ICL represents the top part of the lattice as it is shown in Fig. 1.

4. Demonstration by an Example

As an example of recommending relevant information sources based on FCA, we use the same dataset we used in [2]. In our example, we used text sources dealing with the concept of *wild cat*. We obtained 8 explications of the concept from different textual sources (s_1, \dots, s_8) . Therefore each explication describes the concept of *being a wild cat* from the different point of view. To illustrate the basic idea without troubling the reader with too many technicalities, we present just one of those eight explications:

$$e_1 = [Typ-p \ \lambda w \lambda t \ \lambda x [[\leq [Weight_{wt} \ x] \ '11] \ \wedge \ [\geq [Weight_{wt} \ x] \ '1.2]] [Wild \ 'Cat]] \ \wedge \\ [Req \ 'Mammal \ [Wild \ 'Cat]] \ \wedge \ [Req \ 'Has-fur \ [Wild \ 'Cat]] \ \wedge \ [Typ-p \ \lambda w \lambda t \ \lambda x [[\leq \\ [[Average \ 'Body-Length] \ x] \ '80] \ \wedge \ [\geq [[Average \ 'Body-Length] \ x] \ '47]] [Wild \ 'Cat]] \ \wedge \\ [Typ-p \ \lambda w \lambda t \ \lambda x [[= [[Average \ 'Skull-Size] \ x] \ '41.25]] [Wild \ 'Cat]] \ \wedge \ [Typ-p \ \lambda w \lambda t \ \lambda x [[= \\ [[Average \ 'Height] \ x] \ '37.6]] [Wild \ 'Cat]]$$

Explication mentioned above was obtained from text source describing the wild cat from the biological point of view. It contained information such as classification of this specimen (being a mammal), body proportions and appearance of the wild cat.

After obtaining all explications, the user selects one of them which is the most relevant from his point of view. Let e_1 be the case. The entire process of recommendation starts after the explication selection.

From the explications mentioned above, we generate an incidence matrix written in Table 1.

Each row of the table represents one explication and each column represents particular property/attribute. Value 1 in a cell means that the respective explication contains the respective property, 0 otherwise.

The e_1, \dots, e_8 are identifiers of explications.

The columns' numbers in Table 1 represent the following attributes:

- | | |
|---|---|
| 1. 'Mammal | $[[Average \ 'Body-Length] \ x] \ '80]$ |
| 2. 'Has - fur | 7. $\lambda w \lambda t \lambda x [=$ |
| 3. $\lambda w \lambda t \lambda x [\leq [Weight_{wt} \ x] \ '11]$ | $[[Average \ 'Skull-Size] \ x] \ '41.25]$ |
| 4. $\lambda w \lambda t \lambda x [\geq [Weight_{wt} \ x] \ '1.2]$ | 8. $\lambda w \lambda t \lambda x [=$ |
| 5. $\lambda w \lambda t \lambda x [\geq$ | $[[Average \ 'Skull-Height] \ x] \ '37.6]$ |
| $[[Average \ 'Body-Length] \ x] \ '47]$ | 9. $\lambda w \lambda t \lambda x$ |
| 6. $\lambda w \lambda t \lambda x [\leq$ | $Live-in_{wt} \ [\lambda w \lambda t \lambda y [[Mixed \ 'Forrest_{wt} \ y]$ |

- 10. $\lambda w \lambda t \lambda x \lceil \geq \lceil \text{Territory-Size}_{wt} x \rceil \wedge 50 \rceil$
- 11. $\lambda w \lambda t \lambda x \lceil \lceil \text{Ter-Marking}_{wt} x \wedge \text{Clawing} \rceil \vee \lceil \text{Ter-Marking}_{wt} x \wedge \text{Urinating} \rceil \vee \lceil \text{Ter-Marking}_{wt} x \wedge \text{Leaves-Droppings} \rceil \rceil$
- 12. $\lambda w \lambda t \lambda x \lceil \leq \lceil \text{In-Heat-Period}_{wt} x \rceil \wedge 8 \rceil$
- 13. $\lambda w \lambda t \lambda x \lceil \geq \lceil \text{In-Heat-Period}_{wt} x \rceil \wedge 2 \rceil$
- 14. $\lambda w \lambda t \lambda x \lceil \text{Seek}_{wt} x \wedge \text{Mate} \lceil \text{Loud} \wedge \text{Meow} \rceil \rceil$
- 15. $\lambda w \lambda t \lambda x \lceil \text{Pregnancy-Period}_{wt} x \rceil \wedge 65 \rceil$
- 16. $\lambda w \lambda t \lambda x \lceil \leq \lceil \text{Litter-Size}_{wt} x \rceil \wedge 4 \rceil$
- 17. $\lambda w \lambda t \lambda x \lceil \geq \lceil \text{Litter-Size}_{wt} x \rceil \wedge 3 \rceil$

O/A	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
e ₁	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
e ₂	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
e ₃	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
e ₄	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
e ₅	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	1	0
e ₆	0	0	0	0	1	0	0	0	0	0	1	0	0	1	0	1	0
e ₇	1	1	1	0	0	0	0	0	1	0	1	0	0	1	1	0	0
e ₈	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0

Table 1. Incident matrix

The incident matrix and the selected explication e₁ are the common inputs for both the methods of recommendation, namely the one based on Association Rules and the one based on the Relevant Ordering

4.1. Recommendations based on Association Rules

Min-supp = 0.25 Min-conf = 0.66

Assuming the user has chosen the first explication as the basic one, concepts corresponding to the columns 1-8 can occur only in the antecedents of the recommendation rules. The remaining concepts occur only in succedents of the rules..

Item-sets meeting the min-sup condition, i.e. 0.25-frequent item-sets, are the following ones:

- 1. {1}
- 2. {1, 2}
- 3. {1, 2, 3}
- 4. {1, 2, 7}
- 5. {1, 2, 11}
- 6. {1, 2, 11, 15}
- 7. {1, 2, 15}
- 8. {1, 3}
- 9. {1, 7}
- 10. {1, 11}
- 11. {1, 11, 15}
- 12. {1, 15}
- 13. {2}
- 14. {2, 3}
- 15. {2, 7}
- 16. {2, 11}
- 17. {2, 11, 15}
- 18. {2, 15}
- 19. {3}
- 20. {5}
- 21. {5, 11}
- 22. {5, 11, 16}
- 23. {5, 16}
- 24. {7}
- 25. {9}
- 26. {9, 11}
- 27. {10}
- 28. {11}
- 29. {11, 14}
- 30. {11, 15}
- 31. {11, 15, 16}
- 32. {11, 16}
- 33. {14}
- 34. {14, 15}
- 35. {14, 16}
- 36. {15}
- 37. {15, 16}
- 38. {16}

Frequent item-sets which can be transformed into the rules where the antecedent contains only columns 1-8 and succedent 9-17 are these:

{1, 11}, {1, 11, 15}, {1, 15}, {1, 2, 11}, {1, 2, 11, 15}, {1, 2, 15}, {2, 11}, {2, 11, 15}, {2, 15}, {5, 16}

Final recommendation rules found according to our data are presented in table 2:

Rule	RS	Rule	RS
$\{1\} \Rightarrow_{e_1} \{11\}$	{s4}	$\{1,2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}
$\{1\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}	$\{2\} \Rightarrow_{e_1} \{11\}$	{s4,s7}
$\{1\} \Rightarrow_{e_1} \{15\}$	{s4,s7}	$\{2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}
$\{1,2\} \Rightarrow_{e_1} \{11\}$	{a4,s7}	$\{2\} \Rightarrow_{e_1} \{15\}$	{s4,s7}
$\{1,2\} \Rightarrow_{e_1} \{11,15\}$	{s4,s7}	$\{5\} \Rightarrow_{e_1} \{16\}$	{s5,s6, s8}

Table 2. Recommendation rules: min-supp = 0.25, min-conf = 0.6

Based on the first explication, the algorithm proposes other expliciations and thus also textual sources as relevant for the concept of wild cat. According to the rules, the algorithm recommends sources No. 4 and 7 because these documents contain information about territory marking and pregnancy period. The last rule is a recommendation of documents No. 5, 6 and 8; these sources contain information about litter size.

At this point, we can raise the min-supp up to 0.3 and compare the results. We can see that there are approximately 1/3 of all frequent item-sets⁴ compare to the level set up to 0.25.

$$Min-supp = 0.3 \quad Min-conf = 0.66$$

- | | | |
|-----------|-------------|--------------|
| 1. {1} | 5. {5, 16} | 9. {14} |
| 2. {1, 2} | 6. {11} | 10. {15} |
| 3. {2} | 7. {11, 15} | 11. {15, 16} |
| 4. {5} | 8. {11, 16} | 12. {16} |

The frequent item-set which can be transformed into the rule (antecedent contains only columns 1-8 and succedent 9-17) is the only one {5, 16}.

With increased min-supp value, we acquired only one association rule. Therefore the rule recommends documents No. 5, 6 and 8.

Rule	RS
$\{5\} \Rightarrow_{e_1} \{16\}$	{s5,s6, s8}

Table 3. Recommendation rule min-supp = 0.3, min-conf = 0.6

In case of using Association Rules, min-supp adjustment is not always ideal optimization of the computation process. The best optimization for this method would be an optimization of generating of the frequent item-sets.

4.2. Relevant ordering based on FCA

From Table 1, by using FCA, we obtained the following concepts:

⁴It is clear that raising the min-supp number means that the final amount of frequent item-sets can not be higher.

- | | |
|--|--|
| 0. $(\{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8\}, \emptyset)$ | 16. $(\{e_3\}, \{12, 13, 14, 15, 16, 17\})$ |
| 1. $(\{e_1, e_4, e_7\}, \{1, 2\})$ | 17. $(\{e_4, e_5, e_6\}, \{11, 16\})$ |
| 2. $(\{e_1, e_4\}, \{1, 2, 7\})$ | 18. $(\{e_4, e_5, e_7\}, \{11, 15\})$ |
| 3. $(\{e_1, e_5, e_6, e_8\}, \{5\})$ | 19. $(\{e_4, e_5\}, \{11, 15, 16\})$ |
| 4. $(\{e_1, e_7\}, \{1, 2, 3\})$ | 20. $(\{e_4, e_7\}, \{1, 2, 11, 15\})$ |
| 5. $(\{e_1\}, \{1, 2, 3, 4, 5, 6, 7, 8\})$ | 21. $(\{e_4\}, \{1, 2, 7, 11, 15, 16\})$ |
| 6. $(\{e_2, e_4, e_5, e_6, e_7\}, \{11\})$ | 22. $(\{e_5, e_6, e_8\}, \{5, 16\})$ |
| 7. $(\{e_2, e_7\}, \{9, 11\})$ | 23. $(\{e_5, e_6\}, \{5, 11, 16\})$ |
| 8. $(\{e_2, e_8\}, \{10\})$ | 24. $(\{e_5\}, \{5, 11, 15, 16\})$ |
| 9. $(\{e_2\}, \{9, 10, 11\})$ | 25. $(\{e_6, e_7\}, \{11, 14\})$ |
| 10. $(\{e_3, e_4, e_5, e_6, e_8\}, \{16\})$ | 26. $(\{e_6\}, \{5, 11, 14, 16\})$ |
| 11. $(\{e_3, e_4, e_5, e_7\}, \{15\})$ | 27. $(\{e_7\}, \{1, 2, 3, 9, 11, 14, 15\})$ |
| 12. $(\{e_3, e_4, e_5\}, \{15, 16\})$ | 28. $(\{e_8\}, \{5, 10, 16\})$ |
| 13. $(\{e_3, e_6, e_7\}, \{14\})$ | 29. $(\emptyset, \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17\})$ |
| 14. $(\{e_3, e_6\}, \{14, 16\})$ | |
| 15. $(\{e_3, e_7\}, \{14, 15\})$ | |

Conceptual lattice of these formal concepts is visualised in Fig. 1. Concepts marked in SOC area contain only *significant objects*. The nodes with bright numbers represent particular explications. At this point, the reader does not have to care about the vertex colours.

Significant objects of the object (explication) e_1 is the following set:
 $SO(e_1) = \{e_1, e_4, e_5, e_6, e_7, e_8\}$.

The set of all concepts containing explication e_1 as a common object is the following set:

$$\gamma(e_1) = \{(\{e_1\}, \{1, 2, 3, 4, 5, 6, 7, 8\}), (\{e_1, e_4\}, \{1, 2, 7\}), (\{e_1, e_7\}, \{1, 2, 3\}), (\{e_1, e_5, e_6, e_8\}, \{5\}), (\{e_1, e_4, e_7\}, \{1, 2\}), \}$$

Formal concepts mentioned in the set $\gamma(e_1)$ are represented by numbers 1,2,3,4,5 in Fig. 1.

The *relevant ordering*⁵ (defined in chapter 3.2) is represented by the following sequence:

$$e_8(s_8) \sqsubseteq e_6(s_6) \sqsubseteq e_5(s_5) \sqsubseteq e_7(s_7) \sqsubseteq e_4(s_4) \sqsubseteq e_1(s_1)$$

4.3. Iceberg optimization

In this chapter, we deal with optimization of the above described methods. In general, there are thousands of vertexes or Association Rules in our graphs. It is not necessary to make the entire computation, because the majority of the computed data is irrelevant to user. Hence, it is plausible to reduce the space to some reasonable degree. To this end,

⁵More details can be found in [3]

we decided to apply Iceberg Concept Lattices. The whole process of finding the Iceberg Concept Lattice consists of two parts. The first one is finding *k-frequent item sets*. These are the sets of attributes which have the minimal support greater or equal to *k*. Those *k-frequent item-sets* are then used to find concepts and *recommended sources*.

As (Fig. 1) illustrates, there are numerous vertexes that are useless with respect to the selected explication e_1 . Key part of the lattice is highlighted by SOC set. Concept No. 5 represents our explication and concepts No. 1,2,3,4 contain in their *extents* explanations of the above mentioned recommended sources (e_4, e_7, e_5, e_6, e_8).

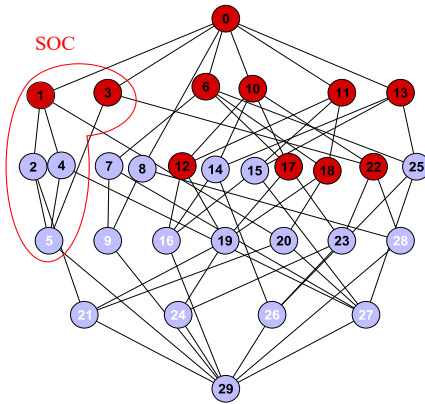


Figure 1. Iceberg lattice of formal concepts - dark vertexes

The entire optimization is based on rising of the *min-support* level. If *min-support* = 0 then *Iceberg lattice* is the same as the standard formal concept lattice. Raising to 0.25, there will be only 38 *frequent item-sets* and 21 concepts. But if we raise the *min-support* to 0.3, then the amount of final *frequent item-sets* will be significantly reduced. In our case, there would be just 12 frequent item-sets and 11 concepts left (highlighted by dark vertexes in Fig. 1).

At this point, with the *Relevant ordering* algorithm, the *Significant objects* of the object (explication) e_1 is the following set: $SO(e_1) = \{e_1, e_4, e_5, e_6, e_7, e_8\}$. The set of all concepts containing our explication e_1 as a common object is the following set:

$$\gamma(e_1) = \{(\{e_1, e_4, e_7\}, \{1, 2\}), (\{e_1, e_5, e_6, e_8\}, \{5\})\}$$

Formal concepts, mentioned in the set $\gamma(e_1)$, are represented by numbers 1, 3 in Fig. 1. As one can realize, the concept No. 5 ($\{e_1\}, \{1, 2, 3, 4, 5, 6, 7, 8\}$) is not in $\gamma(e_1)$. It is not

that important, because the particular explication e_1 was selected by user. Thus the user is aware of existing e_1 and the relevance would be the highest one. At this point we can show the final *relevant ordering*:

Exp.	Intent	DF	RT
e_1	{1,2,3,4,5,6,7,8}	{}	{ s_1 }
e_4	{1,2}	{3,4,5,6,7,8}	{ s_4 }
e_7	{1,2}	{3,4,5,6,7,8}	{ s_7 }
e_5	{5}	{1,2,3,4,6,7,8}	{ s_5 }
e_6	{5}	{1,2,3,4,6,7,8}	{ s_6 }
e_8	{5}	{1,2,3,4,6,7,8}	{ s_8 }

Table 4. Final text sources' ordering. Min-supp = 0.3

As we can see above, the result will be same as the result using the entire conceptual lattice:

$$e_8(s_8) \sqsubseteq e_6(s_6) \sqsubseteq e_5(s_5) \sqsubseteq e_7(s_7) \sqsubseteq e_4(s_4) \sqsubseteq e_1(s_1)$$

As stated earlier, this method of Iceberg Lattices is not generally applicable as the optimization method on Association Rules, because raising the min-support can lead to the loss of an important information. But if we use FCA, the method significantly reduces the data space that is necessary to deal with.

5. Conclusion

In this paper we have summarized two previously proposed methods for text source recommendation. As mentioned in [3], [2], we needed to focus on making recommendation more reliable and computationally more effective. Therefore in this paper, we applied the method exploiting Iceberg Lattices. Recommending method exploiting the Association Rules [2], seeks on the basis of the specified min-supp and min-conf. Using Iceberg Lattices is not suitable as an improvement of this method. Raising min-supp leads to a loss of a large amount of relevant data and information. The best optimization for this method would be an optimization of the frequent item-sets generation.

Recommending method based on FCA [3] utilizes the *Relevant Ordering* of documents. Using Iceberg Lattices as an improvement of this method has proved to be effective, as there is only vertical reduction of data space, thus no loss of information occurred. The loss of information would occur only with a substantial increase in min-supp, thus causing no results to be obtained at all. By an example, we demonstrated how our methods work, and both the methods were implemented in our SW application.

The only obstacle, shared by both the methods, is that they require *explications* of individual atomic concepts from given text sources. It is challenging to automate this process. We are working on the improvement of the transfer of sentences in the natural language into the language of the TIL constructions.

Acknowledgements

This research has been supported by Grant of SGS No. SP2021/87, VSB - Technical University of Ostrava, Czech Republic, "Application of Formal Methods in Knowledge Modelling and Software Engineering IV" and also this work was supported by ESF project 'Zvýšení kvality vzdělávání na Slezské univerzitě v Opavě ve vazbě na potřeby Moravskoslezského kraje' CZ.02.2.69/0.0/0.0/18_05 8/0010238.

References

- [1] Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Seeking relevant information sources. In Informatics'2019, *IEEE 15th International Scientific Conference on In-formatics*, Poprad, Slovakia, pp. 271-276.
- [2] Albert, A., Duží, M., Menšík, M., Pajr, M., Patschka, V. (2021): Search for Appropriate Textual Information Sources. In *Frontiers in Artificial Intelligence and Applications*, vol. 333: Information Modelling and Knowledge Bases XXXII, B. Thalheim, M. Tropmann-Frick, H. Jaakkola, N. Yoshida, Y. Kiyoki (eds.), pp. 227-246, Amsterdam: IOS Press, doi: 10.3233/FAIA200832
- [3] Menšík, M., Albert, A., Patschka, V. (2020): Using FCA for Seeking Relevant Information Source. In *RASLAN 2020*, Brno: Tribun EU, 2020, 144 p. ISBN 978-80-263-1600-8, ISSN 2336-4289.
- [4] Menšík, M., Duží, M., Albert, A., Patschka, V., Pajr, M. (2019): Refining concepts by machine learning. *Computación y Sistemas*, Vol. 23, No. 3, 2019, pp. 943-958, doi: 10.13053/Cys-23-3-3242
- [5] Duží, M., Jespersen, B., Matera, P. (2010): Procedural Semantics for Hyperintensional Logic. *Foundations and Applications of Transparent Intensional Logic*. Berlin: Springer.
- [6] Carnap, Rudolf (1964): *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: University of Chicago Press.
- [7] Winston P. H. (1992): *Artificial intelligence*. 3rd ed., Mass.: Addison-Wesley Pub. Co., 1992.
- [8] Agrawal, R., Imielinski, T., and Swami, A. N. (1993): Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207-216.
- [9] Hájek P., Havránek T., Chytil M. K. (1983): *Metoda GUHA - automatická tvorba hypotéz*. (In Czech. GUHA method; automatic creation of hypotheses). Academia Praha.
- [10] Ganter, B., Wille, R. (1999): *Formal Concept Analysis: Mathematical Foundations*. 1st ed., Berlin: Springer. ISBN 978-3-540-62771-5.
- [11] Taouil, R., Bastide, Y., Lakhal, L. (2001): *Conceptual Clustering with Iceberg Concept Lattices*.

Artizon Cloud: A Multidatabase System Architecture for an Art Museum

Naoki ISHIBASHI ^{a,1}

^a Faculty of Data Science, Musashino University

Abstract.

For dynamically integrating professional knowledge of curators, a multidatabase system architecture for an art museum, *Artizon Cloud* is proposed. A location based data provision is defined in the architecture for visitors. A system and applications are implemented and provided in an actual museum, and heterogeneous archives that were independently implemented as databases with Web UIs are dynamically extracted, integrated, and staged in visitors' devices.

Keywords. multidatabase systems, museum systems, multimedia databases

1. Introduction

By combining art works and advanced knowledge of curators, art museums provide intellectual and emotional experiences to all visitors. Many kinds of devices have been tested to provide the knowledge to users, but a framework to dynamically stage intellectual properties of curators to visitors is not proposed.

In the past years, museums started to develop many kinds of digital archives for photographs, evidencial documents, or exhibitions. Therefore, intellectual properties of artists and curators are now becoming digital archives especially photographs of art works[1,2,3]. Moreover, researches beyond archiving photographs have been proposed to store and share knowledge such as knowledge co-creation on social networking paradigm to maintain a cultural portal[4], archive for museum layouts with visitor behavior analysis[5] and archives for traditional culture and directing new experiences using digital technologies[6].

A multidatabase system[7,8,9] that dynamically integrates the digital archives built by curators to provide services is desired especially for museums which do not employ engineers. By dynamically integrating the archives that represent knowledge of curators and/or art work themselves, the knowledge could automatically provide newer intellectual or entertaining services. For integrating legacy databases, managing heterogeneity is crucial to establish multidatabase systems[10], and a metadatabase approach[11,12,13,14] is observed to fit with museums since following reasons: 1) meta-data domain to share among archives is definable, 2) semantic computing is required for many kinds of art data.

¹Corresponding Author: Naoki Ishibashi, Faculty of Data Science, Musashino University, 3-3-3, Ariake, Kotoku, Tokyo, Japan; E-mail:n-ishi@musashino-u.ac.jp.

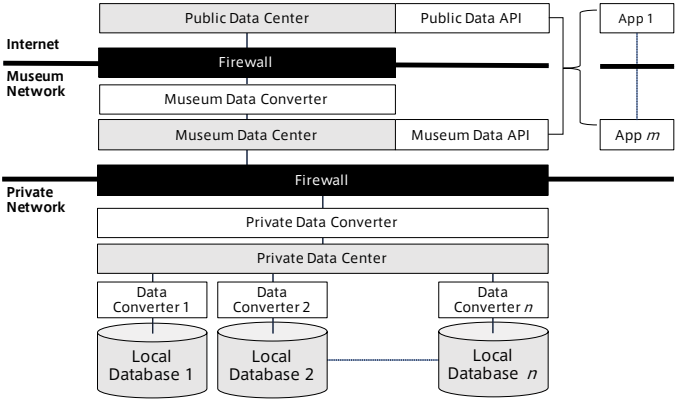


Figure 1. A System Architecture of Artizon Cloud

In this paper, a multidatabase system architecture, *Artizon Cloud*, for a museum is proposed. For an art collection, Curators edit data in archives that are managed in a local network, *Artizon Cloud* dynamically integrates the archives and provide appropriate data to either museum visitors or WWW users.

2. A Multidatabase System Architecture

By defining three distinct areas to provide information, *Artizon Cloud* is designed as Fig. 1 such as Private Network, Museum Network, and Internet.

Curators register and manage art collection data in web-based data archives in a local network with private IP addresses. This Private Network is accessible only by staff members, and the archives are privately maintained. Each archive uses distinct DBMS, and schemas are independently designed and implemented. Data in these archives are converted to Private Data Center, and Private Data Center integrates various data types.

Private Data Center sends a data set that includes information for museum visitors and WWW users through Private Data Converter to Museum Data Center as Fig. 2. Therefore museum visitors can access a lot of art information more than WWW users, such as high-resolution graphic, restricted audio or video data, as well as all information available on WWW. It means a museum with *Artizon Cloud* can provide massive amount of knowledge automatically converted from curators' daily work to the museum visitors.

Museum Data Center sends a set of data for WWW users through Museum Data Converter to Public Data Center, and Public Data Center is a back end to implement WWW applications as well as smartphone applications. Once more, the daily works of the curators are converted to Public Data Center, so *Artizon Cloud* reduces tasks to publish various data to WWW and it also provides infrastructure to develop new museum applications.

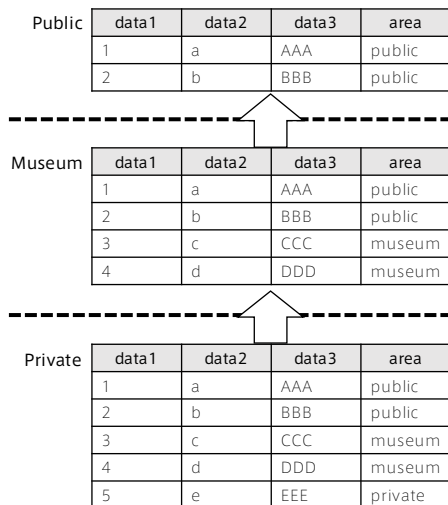


Figure 2. A Data Flow in Artizon Cloud

Many kinds of applications are assumed of Artizon Cloud as shown in Fig. 3. Museum Data Center can provide applications on devices of a group of visitors and applications on museum devices. Furthermore, this architecture could provide capability to quickly develop museum application in an actual museum, so the art museum could become an experimental space for open data innovation[15,16].

3. A Staging Environment

As a reconstruction of Bridgestone Museum of Art that was established in 1952, *Artizon Museum*[17] started in January 2020 in Kyobashi, Tokyo. The scope of the museum extends from antique art works, modern Japanese Western-style paintings, the Impressionists, 20th century art, to contemporary art, and approximately 3,000 art works are managed in a collection. Three floors, approximately 2,100 m^2 in total, are available as galleries, and Wi-Fi is provided to all floors for visitors in a local network with private IP addresses.

4. An Implementation of Artizon Cloud

Artizon Cloud was actually implemented for Artizon Museum, and started to provide services since January 2020. At this point, five archives are connected to Artizon Cloud

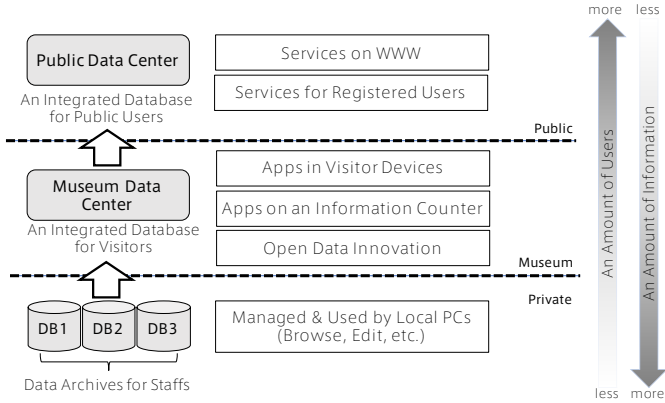


Figure 3. Service Examples of Artizon Cloud

as shown in Fig. 4, and these archives are independently implemented by different organizations. However, only exception is that metadata extracted from Artize are shared with other systems.

4.1. Archives

Following archives are independently implemented, and all these archives are WebDB systems.

4.1.1. Artize

Artize is a collection management system, and a commercial software developed by Nissha Co., Ltd. customized for Artizon Museum. Curators of the museum register data of art works in the collection, and manage a set of relative data, such as names of artists, names of art works, histories of art works, repair records, exhibition logs, photographs, references, etc. Data in Artize include most of detailed information for all art works of the collection, and the data are daily maintained by the curators. The most recent data could be discovered from Artize, and a set of metadata is automatically extracted from Artize as follows: data of 1) art works, 2) artists, and 3) exhibitions. Therefore, Artize is the most significant archive for Artizon Cloud, because it provides these 3 metadata to Private Data Center as well as to other archives.

4.1.2. iDash

iDash is an evidential document archive developed by Governance Design Laboratory, Inc., and it manages evidential documents of art works or artists, repair histories of art works, and publications of the museum.

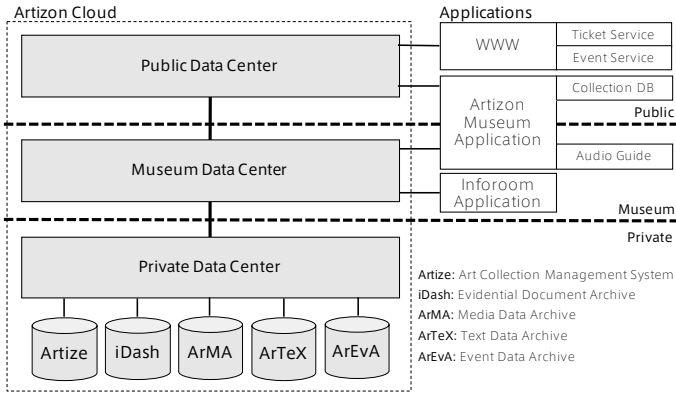


Figure 4. An Implementation of Artizon Cloud

4.1.3. ArMA

ArMA is a media data archive of art works, artists, and exhibitions to manage graphic, audio, or video data, and ArMA was implemented by aery Co. ArMA was initially planned to manage high-resolution photographs as evidences to repair art works, and expanded its role to manage media data such as movies of each exhibition, and audio files for an audio guide in the museum, etc.

4.1.4. ArTeX

ArTeX was implemented by Newphoria Corporation to manage texts written by the curators. The curators of Artizon Museum used to write various text documents using MS Word, and stored data in each PC. To realize unified storage of texts, ArTeX receives MS Word files and extracts texts into DBMS. Any language could be stored using UTF-8, but English, Japanese, Chinese and Korean languages are recently stored.

4.1.5. ArEvA

ArEvA was implemented by Dai Nippon Printing Co., Ltd. to store event data such as exhibitions, lectures, or symposiums. It was designed to store a history of the museum including posters of each exhibition.

4.2. Private Data Center

The sets of data from these archives are automatically extracted and converted to Private Data Center, that we call *MetaDB*. *MetaDB* currently has functions as follows:

- A function to extract, convert, and store metadata for the art works, the artists, and the exhibitions in an ORDB and a file system from Artize

August 2021

- A function to provide the metadata to the archives via Web API
- Functions to edit the metadata such as to set an appropriate zone, private, museum, or public, to each datum via Web UI
- Functions to store the data of each archive in the ORDB and the file system
- Functions to edit the data of the archives such as private, museum, or public via Web UI
- Functions to edit copyrights of the graphics such as the art works, or the posters via Web UI
- A function to convert and transfer the metadata to Museum Data Center
- Functions to convert and transfer the data of the archives to Museum Data Center

4.3. Museum Data Center

The set of data from MetaDB is automatically received and converted to Museum Data Center, that we call *DMZDB*. *DMZDB* currently has functions as follows:

- A function to store metadata for the art works, the artists, and the exhibitions in an ORDB and a file system
- Functions to store the data of the archives in the ORDB and the file system
- Functions to provide a data set of the art works via Web API in the museum
- Functions to provide a data set of the artists via Web API in the museum
- Functions to provide a data set of the exhibitions via Web API in the museum
- Functions to provide a data set of the publications via Web API in the museum
- A function to convert and transfer a data set of the art works to Public Data Center
- A function to convert and transfer a data set of the artists to Public Data Center
- A function to convert and transfer a data set of the exhibitions to Public Data Center

4.4. Public Data Center

Public Data Center is implemented in EC2 and S3 of Amazon Web Services, AWS, and it currently has functions as follows:

- Functions to provide a data set of the artists via Web API
- Functions to provide a data set of the exhibitions via Web API
- Functions to provide a data set of the publications via Web API

5. An Application of *Artizon Cloud*

As a first application of *Artizon Cloud*, *Artizon Museum App*. was released in late 2019[18], and sample screens of the application are presented in Fig. 5.

When a visitor arrives at the museum, their devices are recommended to connect to the Wi-Fi in the museum. Then, these devices are capable to access all information provided from *DMZDB* such as audio guide data of an actor(Fig. 6). In galleries, beacon devices are embedded to specific art works that have audio files to guide, and these devices detect a location of an user to switch a screen of the application.

Since the grand opening of *Artizon Museum* in January 2020, 108,770 people have arrived at *Artizon Museum*, 9,311 downloads were observed for the iOS application,

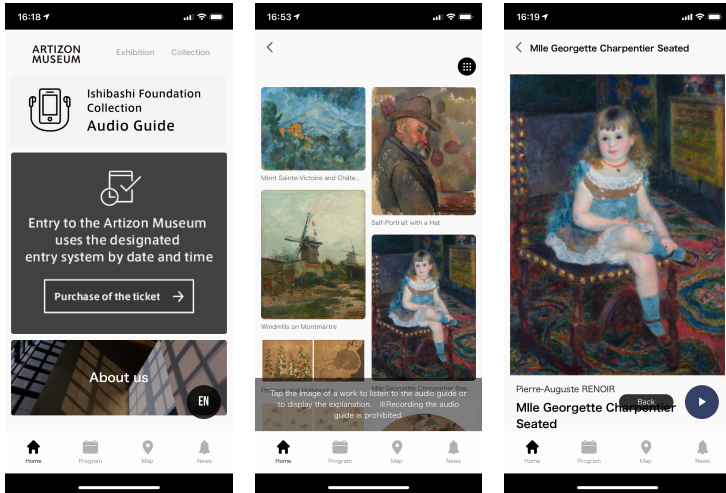


Figure 5. Sample Screens of Artizon Museum App.

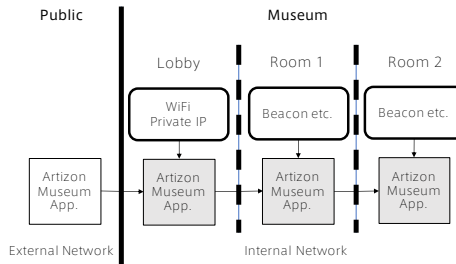


Figure 6. Zone Transitions of Artizon Museum App.

and 4,983 downloads were observed for the Android Application until end of 2020. The numbers of visitors using the application are still small, though many and very positive feedback were observed on various SNSs.

Since Artizon Cloud and the application, Artizon Museum stopped lending devices to visitors, and is encouraging people to bring their own devices and earphones to the museum. Artizon Museum has a strong will to produce intellectual and emotional experiences by providing massive amount of data.

6. Conclusion

A multibase system architecture for an art museum is proposed. Artizon Cloud dynamically integrates independent archives of the art collection, and provide data according to user's location such as in an office, an museum, and elsewhere. The system is implemented and the services are provided to the visitors of the actual museum, Artizon Museum, and iOS and Android applications are currently available.

By applying the implemented system, advanced knowledge of the curators as well as art works themselves are dynamically extracted, integrated and staged to the actual smartphone applications of visitors, the visitors can access audio guides, videos and collection data by plugging in their earphones.

As future works, we are discussing to design methodology of dynamic curation in a cyberspace, an entertainment application for children, and we are also planning to encourage other art museums to join our challenge.

Acknowledgement

Without the vision for the new art museum of Hiroshi Ishibashi, Chairman of Ishibashi Foundation, this project would not even exist. I would like to express my utmost gratitude to him for this vision and his thirst for the future of art museums.

I would like to express my sincere gratitude to Taiji Nishijima, Michiko Kasahara, Tadashi Urami, and Kazunori Yamauchi of Ishibashi Foundation for their great cooperation and support throughout the project period.

Shunsuke Naotsuka, Shoji Kometani, and Tomohiro Kawasaki of Ishibashi Foundation were always my colleagues in the project for years. Without their help, there is no doubt that this project would not have reached its goal.

The members of the curatorial department and the public relations department of Ishibashi Foundation have also shared many meetings and discussions with us. Without the opinions and cooperation of the people on the ground, we would not be able to create a core system for the organization. I would like to express my gratitude to them for their continuous cooperation.

I would also like to express my gratitude to all the organizations that have contributed to the overall system. I would like to express my sincere gratitude to them for their dedication and continuous efforts.

Finally, I would like to express my highest appreciation to Professor Yasushi Kiyoki of Keio University. I am sure that our experience together at Keio University was the initial inspiration for this research. I would like to take this opportunity to thank him.

References

- [1] The Metropolitan Museum of Art: *The Met Collection*, available via WWW, <https://www.metmuseum.org/art/collection> (2021)
- [2] Musée du Louvre: *Atlas database of exhibits*, available via WWW, <http://cartelen.louvre.fr/> (2021).
- [3] Paris Musées: *Les collections en ligne des musées de la Ville de Paris*, available via WWW, <https://www.parismuseescollections.paris.fr/> (2021).

- [4] Sornlerlamvanich, V. and Charoenporn, T.: "Cultural Knowledge Co-Creation on Social Networking Paradigm," *Proceedings of The Second International Conference on Culture and Computing*, Kyoto University, Kyoto, Japan 2011).
- [5] Kovavisaruch, L., Sanpechuda, T., Chinda, K., Kamolvej, P. and Sornlerlamvanich, V.: "Museum Layout Evaluation based on Visitor Statistical History," *Asian Journal of Applied Sciences*, Vol.5, No.3, pp.615-622 (2017).
- [6] Digitized Thailand: *White Paper*, NECTEC, available via WWW, <https://www.facebook.com/digitizedthailand> (2009).
- [7] Batini, C., Lenzerini, M. and Navathe, S.B.: "A comparative analysis of methodologies for database schema integration", *ACM Computing Surveys*, Vol.18, No.4, pp.324-364 (1986).
- [8] Litwin, W., Mark, L. and Roussopoulos, N.: "Interoperability of Multiple Autonomous Databases", *ACM Comp. Surveys*, Vol.22, No.3, pp.267-293 (1990).
- [9] Sheth, A.P. and Larson, J.A.: "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys*, Vol.22, No.3, Special issue on heterogeneous databases, pp.183-236 (1990).
- [10] Zhang, J.: "Classifying approaches to semantic heterogeneity in multidatabase systems," *Proceedings of the 1992 conference of the Centre for Advanced Studies on Collaborative research - Volume 2*, pp.153-173 (1992).
- [11] Kitagawa, T. and Kiyoki, Y.: "The mathematical model of meaning and its application to multidatabase systems," *Proc. 3rd IEEE Int. Workshop on Research Issues on Data Engineering: Interoperability in Multidatabase Systems*, p.130-135 (1993).
- [12] Kiyoki, Y. and Kitagawa, T.: "A metadatabase system supporting interoperability in multidatabases", *Information Modeling and Knowledge Bases*, Vol.5, pp.287-298 (1993).
- [13] Kiyoki, Y., Kitagawa, T. and Hitomi, Y.: "A fundamental framework for realizing semantic interoperability in a multidatabase environment", *Journal of Integrated Computer-Aided Engineering*, Vol.2, No.1, pp.3-20 (1995).
- [14] Kiyoki, Y., Hosokawa, Y. and Ishibashi, N.: "A Metadatabase System Architecture for Integrating Heterogeneous Databases with Temporal and Spatial Operations," *Advanced Database Research and Development Series Vol. 10, Advances in Multimedia and Databases for the New Century, A Swiss/Japanese Perspective*, pp.158-165, World Scientific Publishing (1999).
- [15] Janssen, M., Charalabidis, Y. and Zuiderwijk, A.: "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Information Systems Management*, pp.258-268. (2012)
- [16] Zuiderwijk, A., Janssen, M., and Davis, C.: "Innovation with Open Data: Essential Elements of Open Data Ecosystems", *Information Polity*, pp. 17-33. (2014)
- [17] Artizon Museum: *Artizon Museum*, available via WWW, <https://www.artizon.museum/en/> (2021).
- [18] Artizon Museum: *Artizon Museum Official App*, available via WWW, <https://www.artizon.museum/en/user-guide/application/> (2021).

A Proposal for a Method of Determining Contextual Semantic Frames by Understanding the Mutual Objectives and Situations between Speech Recognition and Interlocutors

Ryosuke KONISHI^a Fumito NAKAMURA^a Yasushi KIYOKI^b

^a*Generic Solution Corporation, Nampoedai-chou, Shibuya-ku, Tokyo, Japan*

^b*Faculty of Environmental Information, KEIO University, Fujisawa, Kanagawa 252, Japan*

Abstract. In this study, we will examine how the speech sounds generated by one-to-one or one-to-n human communication are not only simple exchanges of intentions and opinions, but are also clearly divided into linguistic expression and linguistic understanding. In this course, we will discuss the problems of interaction between interlocutors and the complex interplay of phenomena that occur in individual interlocutors. We propose a method to determine the semantic frame of a dialogue contextually by integrating and calculating the features of speech and the features of the meaning of words.

Keywords. Mathematical Model of Meaning, Integration of heterogeneous information, Natural Language Process, Speech Recognition, Decision Inference,

1. Introduction

The acoustic model in speech recognition technology expresses the relationship between a series of acoustic features, which is a continuous quantity for each generation time, and a series of discrete symbols (words, phonemes, etc.).

The dramatic improvements achieved by deep learning in large-vocabulary continuous speech recognition tasks have attracted much attention. Almost all of the research that had been done on GMM, HMMs in the past has been revised at a rapid pace, and deep learning is now commonly used as the standard for acoustic modeling. Nowadays, deep learning is often used as a standard for acoustic modeling [1,2,3,4]. In the framework of deep learning, the relationship between discrete symbols is modeled on a continuous-valued vector space, and the relationship between variable-length input and output sequences is modeled directly. First, discrete symbols such as words and letters are treated as fixed-length continuous-valued vectors called Distributed Representation or Embedding. Instead of the conventional modeling based on sparse information (e.g., frequency), it is possible to model discrete symbols on a continuous-valued vector space.

It is now possible to capture the relationship between discrete symbols in a vector space of continuous values, rather than modeling based on sparse information (such as frequency).

Second, in addition to the extension techniques, it is now possible to flexibly handle variable-length. In addition to the extension techniques, it is now possible to flexibly handle discrete symbolic sequences of variable length. In addition, it is now possible to flexibly handle discrete symbolic sequences of variable length, from fixed-length inputs (such as Bag-of-Words, which does not consider sequences, or local fixed-length context information) to fixed-length outputs. context information) as well as fixed-length output. It is now possible to treat a discrete symbolic series of variable length as a continuous vector of fixed length [5,6,7,8,9,10].

On the other hand, language models are also being extended in various ways with the advancement of techniques based on deep learning. Methods using deep learning have been reported to be overwhelmingly accurate in various tasks such as language model building [11], proper noun extraction [12], meaning construction based on constructive semantics [13], and reputation classification [14,15], and various other tasks, deep learning methods have been reported to have overwhelming accuracy [16,17].

Traditional word modeling involves word segmentation, parsing, and probabilistic language models. In the field adaptation, it is important to prepare texts in the adaptation field and to perform field adaptation of word segmentation and reading estimation. One of the applications of this method is the "Shabette Concierge" and language processing developed by NTT Docomo, but it is limited to task determination and keyword extraction. However, it is limited to task determination and keyword extraction, and semantic interpretation in simple limited situations [18].

Similar work has been done on slot filling and speech intention understanding. Typical methods include Support Vector Machine modeling using sparse features such as n-grams. Machine modeling using sparse features such as n-grams [19,20].

However, although modeling on discrete space and modeling on continuous space have their own challenges regarding their merits, there is some research on modeling that considers more complex structures as contexts, such as documents and discourse. There are few studies that interpret and determine semantic frames for acoustic and language models [21,22,23].

We propose a method for determining the semantic frame for the process as a result of the speaker's semantic understanding and dialogue in a complex specific situation.

Structure of this paper is organized as follows. First, we outline problem settings in this paper. Next, we briefly describe overall picture of our proposed method. Third, we present the basic theories to embody the components of the method. Fourth, we explain the details of the proposed method, and finally we conclude this paper.

2. Problem Settings

Here, we briefly specify the problem. This paper considers involving communication between consumers and a clerk, such as during a business negotiation or sales promotion. In such cases, the outcome of the communication, such as whether the negotiation was successful, unsuccessful, or put on hold, is considered to depend on the tension and activity of the spoken words and voice. This paper considers the problem of inferring the speaker's semantic understanding using linguistic and acoustic information.

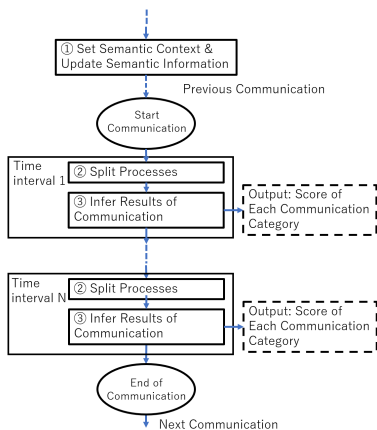


Figure 1. Semantic frame of the proposed method

3. Overview of our method

This section introduces a whole architecture, which is referred to as semantic frame, to solve the problem in section 2. Figure 1 shows the frame. First, our system sets assumed semantic contexts that represent the situation of the communication and updates semantic information that projects linguistic and acoustic information on a semantic space to evaluate the speaker’s semantic understanding. Note that the frequency of the calculations, referred to as the period, is days, weeks, etc., because the update requires a little bit of computational cost for the matrix product and the eigenvalues computation, as you can be seen later.

For each communication, our system splits the entire communication into processes by streaming, where the unit of the segmentation is called the time interval.

For each time interval, the system estimates the result of the communication using linguistic and acoustic information, and it outputs the quantized score for each successful, unsuccessful, or pending communication category. We achieve the following functions by applying the next section’s methodologies:

1. Set the Semantic Contexts and Update Semantic Information
2. Split Conversations
3. Infer the Results of Communication

4. Basic Theories

Here, we explain the proposed method. In this paper, we use the mathematical model of meaning (MMM) to extract semantic information. Moreover, we explain how to process

linguistic and acoustic information because both kinds have to be quantized to calculate the score.

4.1. Mathematical Model of Meaning

The MMM was first proposed to extract semantic information behind data deterministically [24]. Let $X \in \mathbb{R}^{N \times M}$ be a data matrix, where N is the number of data, and M is the number of features. We use 2-norm normalization in each column of the matrix, and denote the resulting normalized matrix by \tilde{X} , i.e., the (i, j) th element of the matrix is

$$\tilde{X}_{i,j} = \frac{X_{i,j}}{\sqrt{\sum_{k=1}^N X_{k,j}^2}}. \quad (1)$$

This is referred to as the fundamental data matrix. The product of the fundamental data matrix $\tilde{X}^T \tilde{X}$ represents a similarity matrix among features, and a subspace spanned using a combination of the eigenvectors extracted from the product represents a semantic space. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M \geq 0$ be the eigenvalues and v_1, v_2, \dots, v_M be the corresponding eigenvectors. Nevertheless, because the similarity matrix is a symmetric matrix, the eigenvalues are non-negative and real, and the number of the combination is 2^M . A context matrix and a threshold are used to determine the subspace from the combination. Let $Y = [Y_1^T, Y_2^T, \dots, Y_t^T] \in \mathbb{R}^{t \times M}$ be the context matrix, and $Q = [v_1, v_2, \dots, v_M] \in \mathbb{R}^{M \times M}$. Then, for the threshold ϵ_{mmm} , MMM determines the index set of the chosen eigenvectors $\Lambda_{\epsilon_{mmm}}$ using

$$\Lambda_{\epsilon_{mmm}} = \{j \in \{1, 2, \dots, M\} | 0 < \epsilon_{mmm} < 1, \frac{(QY)_j}{\|QY\|_{\infty}} > \epsilon_{mmm}\}. \quad (2)$$

Here, $QY = \sum_{i=1}^t Y_i Q$, $(QY)_j$ is the j -th element of QY , and $\|QY\|_{\infty} = \max_{1 \leq j \leq M} |(QY)_j|$. We obtain the semantic projection by arranging the chosen eigenvectors

$$P(Y) = [v_j^T]_{j \in \Lambda_{\epsilon_{mmm}}} \in \mathbb{R}^{|\Lambda_{\epsilon_{mmm}}| \times M}, \quad (3)$$

where $|\Lambda_{\epsilon_{mmm}}|$ represents the cardinality of $\Lambda_{\epsilon_{mmm}}$. When we apply the projection to each datum $g \in \mathbb{R}^M$, the datum is regarded as an element of a semantic space, denoted by $P(Y)g$. In this semantic space, we can measure the distance between the datum and the semantic centroid \bar{D} :

$$\bar{D} = \frac{1}{\|P_Y\|_{\infty}} P_Y, \quad (4)$$

where $P_Y = \sum_{i=1}^t P(Y)Y_i$. Let \bar{D}_j and $(P(Y)g)_j$ be the j -th elements of the vectors; the distance is calculated as the weighted Euclidean distance $dist(\bar{D}, g)$,

$$dist(\bar{D}, g) = \sqrt{\sum_{j=1}^{|\Lambda_{\epsilon_{mmm}}|} c_j (\bar{D}_j - (P(Y)g)_j)^2}, \quad (5)$$

where $c_j = \frac{(P_Y)_j}{\|P_Y\|_{\infty}}$.

4.2. Configuring Linguistic Information

Assuming that we only utilize speech information, we need a system that generates text from speech to apply linguistic techniques. Such a speech recognition technique can be used in the case of both a cloud[25,26] and a premise[27,28]. For the generated text, next we need a linguistic technique to deal with the text quantitatively. In this paper, we utilize the following well-known features as the linguistic features.

- Bag-of-Words(BoW)
- Word embeddings

Let W be a set of words and H the dimension of the embeddings; then, we create a linguistic feature having the dimension of $|W| + H$.

4.3. Configuring Acoustic Information

Famous acoustic features obtained from a speech signal are the following ones [29,30].

- power of speech signal
- fundamental frequency
- Mel-Frequency Cepstrum Coefficient

Note that these heterogeneous features need to be aggregated because the features are calculated differently. For example, the power is obtained for each digitalized timing, while the fundamental frequency is obtained for each window size. In this paper, we summarize each feature by the maximum, minimum, mean, and standard deviation with each common period.

5. Proposed Method

In this section, we explain how to achieve the functions we explained in section 3 one by one.

5.1. Set the Semantic Contexts and Update Semantic Information

In this process, our system sets semantic contexts and updates semantic linguistic and acoustic projections. Algorithm 1 shows an overview of the process.

Let S be a set of a semantic context such as in general cases like a business negotiation and a sales promotion and in specific cases such as a business negotiation in which a customer decides to purchase something, and let C be a set of communication categories such as successes, failures, or reservations. Then, our system first establishes the set S of the semantic context for each period. Next, our system obtains the past linguistic and acoustic fundamental data matrices X_s^l and X_s^a for each element $s \in S$ with the time interval in the row direction and the features in the column direction, where the element of each feature is the one described in section 4.3 and 4.3. Then, our system calculates the eigenvectors $(v_{1s}^l, \dots, v_{M_s}^l)$ and $(v_{1s}^a, \dots, v_{M_s}^a)$ of the product $\tilde{X}_s^{lT} \tilde{X}_s^l$ and $\tilde{X}_s^{aT} \tilde{X}_s^a$ of the normalized matrices X_s^l and X_s^a . Using the combination of the eigenvectors, our system prepares the semantic projections for each communication category $c \in C$. The linguistic

Algorithm 1 Algorithm of Setting the Semantic Contexts and Updating the Semantic Information

Require: ϵ_{mmm} : Threshold to determine the semantic projection, C : Set of communication category

- 1: **procedure** SET THE SEMANTIC CONTEXT AND UPDATE THE SEMANTIC INFORMATION
 - 2: Set the semantic context S
 - 3: **for each** $s \in S$ **do**
 - 4: Construct linguistic and acoustic fundamental data matrices X_s^l and X_s^a in s , respectively.
 - 5: Calculate \tilde{X}_s^l and \tilde{X}_s^a using (1).
 - 6: Calculate eigenvector $(v_{1s}^l, \dots, v_{Ms}^l)$ and $(v_{1s}^a, \dots, v_{Ms}^a)$ of $\tilde{X}_s^l \tilde{X}_s^l$ and $\tilde{X}_s^a \tilde{X}_s^a$, respectively.
 - 7: **for each** $c \in C$ **do**
 - 8: Update linguistic and acoustic context matrices $Y_{s,c}^l$ and $Y_{s,c}^a$ of c in s , respectively.
 - 9: Determine semantic projections $P(Y_{s,c}^l)$ and $P(Y_{s,c}^a)$ by (3).
 - 10: **end for**
 - 11: **end for**
 - 12: **end procedure**
-

and acoustic context matrices $Y_{s,c}^l$ and $Y_{s,c}^a$ of a semantic context s and a communication category c are updated when those features utilized for the previous period are labeled by c . Our system updates the semantic projections $P(Y_s^l)$ and $P(Y_s^a)$ using the context matrices, the eigenvectors, and the threshold ϵ_{mmm} .

Bear in mind that the computational complexity of this process is $O(M^3 + N^2)$, where M is the number of features, and N is the number of data of the fundamental data matrix. Therefore, we assume the period to be days, weeks, and so on.

5.2. Split Processes

In this process, our system determines whether or not the conversation is broken based on a streamed speech signal $S(t)$. Algorithm 2 shows an overview of the process.

Algorithm 2 Algorithm of Split Processes

Require: $S(t)$: Signal of streamed acoustic information, ϵ_a : Threshold of silence length, ϵ_l : Threshold of text distance

- 1: **procedure** SPLIT PROCESSES
 - 2: Aggregate silence interval in $S(t)$.
 - 3: Evaluate silence interval using ϵ_a .
 - 4: Enable speech recognition and generate text $text(t)$ from $S(t)$.
 - 5: Vectorize $text(t)$, denoted by $V(t)$.
 - 6: Evaluate distance between $V(t)$ and previous vector $V(t-1)$ by ϵ_l .
 - 7: Integrate two evaluated results and determine whether or not the conversation is broken.
 - 8: **end procedure**
-

Both linguistic and acoustic determinations can be made on whether or not the conversation is broken. In the acoustic component, our system evaluates the length of the silence interval. Therefore, our system aggregates the length from $S(t)$ and evaluates that the interval exceeds a threshold ϵ_s . In the linguistic component, our system first generates a text $text(t)$ from $S(t)$ using speech recognition. Then, our system transforms $text(t)$ into a vector $V(t)$ using word embeddings. Finally, our system evaluates the distance between $V(t)$ and previous vector $V(t-1)$ and determines whether or not the distance exceeds a threshold ϵ_l . By combining the results, our system determines whether or not the conversation is broken.

5.3. Infer the Results of Communication

In this process, our system estimates the results of the communication in a split process. Algorithm 3 shows an overview of the process.

Algorithm 3 Algorithm of Inferring the Results of Communication

Require: $U(t)$: Signal of acoustic information at time interval t , s : Semantic context, α : Mixing ratio of distance

- 1: **procedure** INFER RESULTS OF COMMUNICATION
 - 2: Convert $U(t)$ into linguistic feature $X^l(t)$ and acoustic feature $X^a(t)$.
 - 3: **for each** $c \in C$ **do**
 - 4: Project the linguistic feature into semantic space in c , denoted by $P(Y_{s,c}^l)X^l(t)$.
 - 5: Calculate the distance between $P(Y_{s,c}^l)X^l(t)$ and the semantic centroid.
 - 6: Project the acoustic feature into the semantic space in c , denoted by $P(Y_{s,c}^a)X^a(t)$.
 - 7: Calculate the distance between $P(Y_{s,c}^a)X^a(t)$ and the semantic centroid.
 - 8: Integrate the two distance by mixing the ratio α .
 - 9: **end for**
 - 10: **end procedure**
-

To estimate the results of communication at time interval t , our system transforms the speech signal into linguistic and acoustic features $X^l(t)$ and $X^a(t)$ using methodologies in 4.2 and 4.3. Moreover, in a semantic context s and communication category c , each type of feature is projected using semantic projection $P(Y_{s,c}^l)$ and $P(Y_{s,c}^a)$. Our system calculates the distance between the projected feature and semantic centroid $dist(D_{s,c}^l, X^l(t))$ and $dist(D_{s,c}^a, X^a(t))$ in (5) for each s and c . Our system integrates the distance using

$$d_{s,c}(X^l(t), X^a(t)) = \alpha \times dist(D_{s,c}^l, X^l(t)) + (1 - \alpha) \times dist(D_{s,c}^a, X^a(t)), \quad (6)$$

where $0 \leq \alpha \leq 1$. Finally, our system estimates the results of communication based on (6).

6. Conclusion

We proposed a method for determining the semantic frame for the process as a result of the speaker's semantic understanding and dialogue in a complex specific situation.

Our future work involves discussing improvements such as feature selection for both linguistic and acoustic features and alternative ways to split processes. In addition, we will conduct numerical experiments and compare the performance between candidates to validate the method.

References

- [1] N. Kanda, "Deep learning based acoustic modeling for speech recognition," *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 73, pp. 31–38, 2017.
- [2] D. Yu, L. Deng, and G. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.
- [3] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," *Twelfth annual conference of the international speech communication association*, 2011.
- [4] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, pp. 30–42, 2011.
- [5] R. Masumura, "Language modeling and spoken language understanding based on deep learning," *THE JOURNAL OF THE ACOUSTICAL SOCIETY OF JAPAN*, vol. 73, pp. 39–46, 2017.
- [6] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE*, vol. 88, pp. 1270–1278, 2000.
- [7] R. Masumura, T. Asami, T. Oba, H. Masataki, S. Sakauchi, and A. Ito, "Combinations of various language model technologies including data expansion and adaptation in spontaneous speech recognition," *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [8] P. Haffner, G. Tur, and J. H. Wright, "Optimizing SVMs for complex call classification," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*, vol. 1, pp. 632–635, 2003.
- [9] S. Yaman, L. Deng, D. Yu, Y. Y. Wang, and A. Acero, "An integrative and discriminative technique for spoken utterance classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 1207–1214, 2008.
- [10] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 234–239, 2012.
- [11] C. D. Manning and H. Schutze, *Foundations of statistical natural language processing*. 1999.
- [12] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *The journal of machine learning research*, pp. 1137–1155, 2003.
- [13] M. Wang and C. D. Manning, "Effect of non-linear deep architecture in sequence labeling," *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 1285–1291, 2013.
- [14] R. Socher, E. H. Huang, J. Pennington, A. Y. Ng, and C. D. Manning, "Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection," *NIPS*, vol. 24, pp. 801–809, 2011.
- [15] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing Natural Scenes and Natural Language with Recursive Neural Networks," *ICML*, 2011.
- [16] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, "Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions," *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 151–161, 2011.
- [17] R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pp. 1201–1211, 2012.
- [18] T. Yoshimura, "Shabette-Concier Service realized by Natural Language Processing," *SLP*, pp. 1–6, 2012.
- [19] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-thought vectors," *arXiv preprint arXiv:1506.06726*, 2015.
- [20] J. Li, M. T. Luong, and D. Jurafsky, "A hierarchical neural autoencoder for paragraphs and documents," *arXiv preprint arXiv:1506.01057*, 2015.
- [21] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," *arXiv preprint arXiv:1511.01432*, 2015.
- [22] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," *arXiv preprint arXiv:1507.04808*, vol. 7, pp. 434–441, 2015.

- [23] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.
- [24] T. Kitagawa and Y. Kiyoki, "A mathematical model of meaning and its application to multidatabase systems," in *Proceedings RIDE-IMS93: Third International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems*, pp. 130–135, 1993.
- [25] <https://cloud.google.com/speech-to-text>, "Google Cloud Speech."
- [26] <https://www.ibm.com/jp-ja/cloud/watson-speech-to-text>, "Watson Speech to Text."
- [27] A. Lee, T. Kawahara, and K. Shikano, "Julius—an open source real-time large vocabulary recognition engine," *EUROSPEECH*, pp. 1691–1694, 2001.
- [28] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," *APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*, pp. 131–137, 2009.
- [29] A. Cheveigné and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," *Seventh European Conference on Speech Communication and Technology*, 2001.
- [30] F. Eyben, K. R. Scherer, B. W. Schüller, J. Sundberg, E. Andr{\'e}, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and T. Khiet, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, pp. 190–202, 2015.