Methods and formal models of intelligent analysis of weakly structured data

A V Rabin¹, A A Petrushevskaya¹ and O V Sinitsin²

¹Saint-Petersburg State University of Aerospace Instrumentation (SUAI), ul. Bolshaya Morskaya, 67, lit. A, St. Petersburg, 190000, Russia
²RBS Consulting LLC, Sinopskaya nab., 52, St. Petersburg, 191124, Russia

E-mail: alexey.rabin@guap.ru, aap@guap.ru, seo@grbs.spb.ru

Abstract. The article presents a formal presentation of four ontologies of knowledge bases of the software environment: ontology of a problem area, linguistic ontology, ontology of precedents, ontology of rules for analyzing texts in natural language. When developing the knowledge base of the software environment, it is necessary to formulate requirements for the model of knowledge representation, the ontology. The development of a unified environment for semantic analysis of flows of weakly structured information that implements modern intelligent algorithms for processing text information will greatly facilitate the decision-making process by a specialist in the time constraint mode, due to the possibility of using a single unified bank of expert knowledge in the work of the question-answer system, and will also allow an automated semantic verification of information flows in order to provide information safety organization.

1. Introduction
Currently, the problem of software intellectualization is quite acute. This fact is caused by limitations related to the impossibility of further reducing the semiconductor manufacturing process; the solution in this case can be either the development of a new generation of computing systems based on completely different approaches to production technology or the use of new approaches to software development in order to expand the boundaries of possibilities last one.

The development of a unified environment for semantic analysis of flows of weakly structured information that implements modern intelligent algorithms for processing text information will greatly facilitate the decision-making process by a specialist in the time constraint mode, due to the possibility of using a single unified bank of expert knowledge in the work of the question-answer system, and will also allow an automated semantic verification of information flows in order to provide the organization of information safety.

At present, mankind has a huge amount of information presented in a structured, weakly structured and unstructured form. If information in a structured form can be processed and analyzed using mathematical and statistical methods, then information presented in a weakly structured and unstructured form requires a different approach to processing and analysis. Modern methods of analyzing this kind of information are based on approaches from the following areas of data science [1-3]:

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd
• Data Mining – extracting knowledge and facts from «raw» (unstructured, raw) data.
• Machine learning – methods that allow you to «teach» a computer system to identify explicit and implicit patterns in the «raw» data.
• Big data – analysis of a large amount of data, actually containing various explicit and implicit patterns.
• Semantic analysis – methods that allow a computing system to understand the meaning embedded in the data.
• Analysis of texts in a natural language (Native Language Processing, NLP) – methods that allow a computing system to understand the meaning of text in a natural language.
• Text Mining – extracting knowledge and facts from texts in a natural language, etc.

2. Formal model for representing the knowledge base of the software environment
The quality of the software environment depends on the quality of the knowledge base training. Currently, more than a dozen methods for assessing the quality of ontology are known [5, 6].

Existing methods for assessing the quality of ontologies solve the following problems:

• assessment of the completeness and accuracy of the dictionary of the subject area;
• assessment of the adequacy of the structure in terms of taxonomy, relationships, etc.
• performance evaluation when using ontologies in applications.

The ontology quality assessment process can be applied to various stages of the software environment.

By the degree of automation, all methods for assessing the quality of ontologies can be divided into the following groups:

• automatic;
• semi-automatic;
• manual.

Existing methods for assessing the quality of ontologies are based on the analysis of the following aspects of ontology:

• structure;
• vocabulary;
• the effectiveness of practical use.

Methods for assessing the quality of ontologies can be divided into the following classes:

• based on data (data-driven);
• expert opinions;
• exploration of use cases;
• comparison with the standard;
• studies of the topology of the graph of ontology.

The main methods for assessing the quality of ontologies at the moment are methods based on the analysis of the perception of the structure (topology of the graph) of the ontology by a person.

When developing the knowledge base of the software environment, it is necessary to formulate requirements for the model of knowledge representation, i.e. the ontology.

Formally, the ontology of the knowledge base of the software environment can be represented as the following expression:
3. The ontology of the problem domain

The ontology of the problem domain can be represented by the following expression:

\[ O^{PA} = \{ C^{PA}, V_{al}(C^{PA}), R^{PA} \}, \]  

where
- \( O^{PA} \) is the ontology of the problem area;
- \( O^L \) is the linguistic ontology (linguistic basis);
- \( O^F \) is the ontology of precedents;
- \( O^{NLP} \) is the ontology of rules for the analysis of texts in natural language (NLP ontology);
- \( R^A \) means the unidirectional association relationships between ontology components.

The knowledge base \( O \) consists of the ontologies shown in figure 1.

![Figure 1. Formal representation of the ontology of the knowledge base of the software environment.](image-url)

### 3. The ontology of the problem domain

The ontology of the problem domain can be represented by the following expression:

\[ O^{PA} = \{ C^{PA}, V_{al}(C^{PA}), R^{PA} \}, \]  

where
- \( C^{PA} = \{ C_1^{PA}, C_2^{PA}, K, C_c^{PA} \} \) is the set of concepts of the ontology of the \( O^{PA} \) subject area;
- \( V_{al}(C^{PA}) = \{ V_{al}(C_1^{PA}), V_{al}(C_2^{PA}), K, V_{al}(C_c^{PA}) \} \) is the set of values of the properties of the concepts of the ontology of the domain \( O^{PA} \);
- \( R^{PA} = \{ R^{PA}_S, R^{PA}_P \} \) is the set of relations between the concepts of the ontology of the \( O^{PA} \) subject area, in which:
  - \( R^{PA}_S \) is the relation of the kind «genus-species» («class-subclass») between the concepts of \( C^{PA} \) ontology of the \( O^{PA} \) subject area;
  - \( R^{PA}_P \) is a part-whole relationship between the concepts of the ontology of the \( O^{PA} \) subject area.

The \( O^{PA} \) subject area ontology is formed by an expert in the current subject area. The ontology of the \( O^{PA} \) subject area contains a hierarchy of concepts of the considered problem area, which defines the semantic context for the operation of the software environment. The concepts in the set \( C^{PA} \) of the ontology of the \( O^{PA} \) subject area are interconnected by two types of relations: the relation \( R^{PA}_S \) of the kind «genus-view» («class-subclass») and the relation \( R^{PA}_P \) of the type «part-whole». These types of relationships allow you to describe the laws of existence and the relationship between the entities of the subject area. Thus, the ontology of the \( O^{PA} \) subject area allows you to take into account the features of the problem area, in which the functioning of the software environment occurs, as well as correctly interpret the concepts of the subject area under consideration.

### 4. Linguistic ontology

The linguistic ontology \( O^L \) (linguistic basis) can be represented by the following expression:

\[ O^{PL} = \{ W^L, R^L \}, \]  

where
\[ W^L = \{W^L_1, W^L_2, K, W^L_n\} \] is the set of linguistic units (terms and syntagmatic patterns), cured from various information storages of the enterprise (databases, corporate and external Internet resources, various cases of text documents) in the process of learning the knowledge base;

\[ R^L = \{R^L_1, R^L_2, K, R^L_n\} \] is the set of synonymy relations between linguistic units \( W^L \) of the linguistic ontology \( O^L \).

The linguistic ontology of \( O^L \) is formed in the process of automated learning of the knowledge base by extracting from the information resources of the organization, containing text weakly structured and unstructured data, sets of linguistic units \( W^L \): terms and syntagmatic patterns. The term refers to a word extracted from a text in a natural language, which is a 1-gram or one-word term. A syntagmatic pattern is understood as a combination of several words (\( n \)-grams, where \( n \geq 1 \)), united by the principle of semantic-grammatical-phonetic compatibility. For example, the syntagmatic pattern «building a knowledge base» will allow you to find sentences containing the following \( n \)-grams:

- building a knowledge base;
- building a corporate knowledge base;
- building a fuzzy knowledge base, etc.

Moreover, each \( i \)-th linguistic unit \( W^L_i \) of the linguistic ontology \( O^L \) is related by the ratio of the unidirectional association \( R \) with the \( j \)-th concept \( C^L_j \) of the ontology of the \( O^PA \) subject area.

Thus, in the process of automated learning of the knowledge base, a semantic basis is formed, consisting of the entities of the ontology of the \( O^PA \) subject area and the linguistic ontology of \( O^L \), connected by the relationship of a unidirectional association. Based on the semantic basis obtained in this way, the tasks of semantic analysis of weakly structured and unstructured text resources are solved.

5. Precedent ontology

The ontology of \( O^F \) precedents can be represented by the following expression:

\[ O^F = \{S^F, P^F, E^F, R^F\}, \quad (4) \]

where

- \( S^F = \{S^F_1, S^F_2, K, S^F_n\} \) is the set of logical rules in the format of the Semantic Web Rule Language (SWRL) ontology of use-cases \( O^F \);
- \( P^F = \{P^F_1, P^F_2, K, P^F_n\} \) is the set of precedents (potential solutions) of the ontology of precedents \( O^F \);
- \( E^F = \{E^F_1, E^F_2, K, E^F_n\} \) is the set of estimates of the effectiveness of the adopted solution to the problem of the ontology of use cases \( O^F \);
- \( R^F = \{R^F_1, R^F_2, K, R^F_n\} \) is the set of unidirectional association relations between the \( i \)-th logical rule \( S^F_i \) and the \( j \)-th precedent \( P^F_j \) of the ontology of the use-cases \( O^F \).

The precedent ontology \( O^F \) is generated automatically in the process of interacting with users of the system, each response to a user request from the software environment can be marked as «useful» or «erroneous». In this case, on the basis of linguistic units extracted from the user’s request, the \( i \)-th logical rule \( S^F_i \) of the case precedent ontology \( O^F \), is formed, which allows you to explicitly determine the potential subset of precedents \( \tilde{P}^F \) of precedents \( O^F \), such that \( \tilde{P}^F \) \( R \) \( P^F \).

Thus, the sets of precedents \( P^F \) and logical rules \( S^F \) are formed, which can be considered as answers that satisfy the needs of users of the software environment. There is also the possibility of manually editing the ontology of use cases \( O^F \).

6. Ontology of rules for the analysis of texts in natural language

The ontology of the rules of the analysis of texts in natural language (NLP ontology) \( O^{NLP} \) can be represented by the following expression:
where
\[ N^{NLP} = \{N_1^{NLP}, N_2^{NLP}, \ldots, N_n^{NLP}\} \]

is the set of the rules for the analysis of texts in natural language (NLP rules) of the NLP ontology \( O^{NLP} \);

\[ M^{NLP} = \{M_1^{NLP}, M_2^{NLP}, \ldots, M_{M}^{NLP}\} \]

is the set of the morphological features (genus, species, case, etc.) of the NLP ontology \( O^{NLP} \);

\[ R^{NLP} = \{R_1^{NLP}, R_2^{NLP}, \ldots, R_{R}^{NLP}\} \]

is the set of unidirectional association relations between the \( i \)-th NLP rule \( N_i^{NLP} \) and the \( j \)-th morphological attribute \( M_j^{NLP} \) and the NLP-ontology \( O^{NLP} \).

The NLP ontology \( O^{NLP} \) is formed by an expert linguist during the development and commissioning of a software environment. This ontology allows the analysis of texts in natural language to carry out the function of semantic analysis of weakly structured and unstructured text resources.

\[ O^{NLP} = \{N^{NLP}, M^{NLP}, R^{NLP}\}, \quad (5) \]

**Figure 2.** An illustrative example of an ontology fragment extended by the knowledge base learning subsystem.

Figure 2 shows a fragment of the ontology of the \( O^{FA} \) subject area of the knowledge base, formed manually by an expert based on an analysis of a fragment of the «OSI Network Model». As can be seen from this figure, a fragment of the ontology includes the concept of «OSI Network Model» associated with the «part-whole» relationship with the following concepts: «Physical Layer», «Link Layer», «Network Layer», etc.

The concept of «Physical Layer» is associated with the relationship of synonymy with the concept of «1st layer of the OSI model» and the inclusion relation with the concept of «Physical layer devices».

- «Концентратор» («Hub» on Russian), associated with the relationship of synonymy with the concepts of «Хаб» («Hub» on Russian) and «Hub»;
- «Повторитель» («Repeater» on Russian), associated with the relationship of synonymy with the concepts of «Репитер» («Repeater» on Russian) and «Repeater».

The test question-answer basis formed by the expert contains the question «List the devices of the 1st layer of the OSI model», which has a set of syntagmatic patterns as a standard answer.

**7. Conclusions**

When solving complex problems in the condition of uncertainty, the user of the software environment should receive comprehensive support from the system, having the opportunity to choose a solution not only through the use of expert knowledge stored in the knowledge base, but also taking into
account the experience of solving this problem in the past. Such experience can be formalized in the form of the set of precedents for the ontology of precedents, which should be processed and taken into account as part of the solution of the problem.

The developed algorithms for the structural analysis of the knowledge base ontology, as well as the analysis of the degree of correspondence of the system response to the standard response, allow the expert to automatically evaluate the quality of the knowledge base ontology obtained from the point of view of cognitive ergonomics and, if necessary, make adjustments.

Acknowledgments

The paper was prepared with the financial support of the Ministry of science and higher education of the Russian Federation in the course of the applied research «The comprehensive project to create high-tech production of software tools for automatic analysis of documentation on paper and digital media using semantic-cognitive technologies for cataloging weakly structured information» (unique identifier of the project 02.G25.31.0305, Decree of the Government of the Russian Federation N 218, 09.04.2010).

References

[1] Gao M, Liu C 2005 Extending OWL by fuzzy description logic. (Hong Kong: 17th IEEE international conference on tools with artificial intelligence. IEEE Press) pp 562–7
[2] Falbo R A, Quirino G K, Nardi J C, Barcellos M P, Guizzardi G and Guarino N 2016 An ontology pattern language for service modeling. (Pisa: 31st annual ACM symposium on applied computing) pp 321–6
[3] Bobillo F, Straccia U 2008 Fuzzy DL: an expressive fuzzy description logic reasoner. (Hong Kong: 17th IEEE international conference on fuzzy systems. IEEE Press) pp 923–30
[4] Serrano-Guerrero J, Olivas J A, de la Mata J and Garces P 2005 Physical and Semantic Relations to Build Ontologies for Representing Documents. (Beijing: Eleventh International Fuzzy Systems Association World Congress IFSA) I 503–8
[5] Zarubin A A, Koval A R, Moshkin V S and Filippov A A 2017 Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources Retrieved from: http://ceur-ws.org/Vol-1903/paper26.pdf
[6] Renu R S, Mocko G and Koneru A 2013 Use of big data and knowledge discovery to create data backbones for decision support systems Procedia Computer Science 20 446–53
[7] Ltifi H, Kolski C, Ayed M B and Alimi A M 2013 A human-centred design approach for developing dynamic decision support system based on knowledge discovery in databases Journal of Decision Systems 22 69–96
[8] Rubiolo M, Caliusco M L, Stegmayer G, Coronel M and Fabrizi M G 2012 Knowledge discovery through ontology matching An approach based on an Artificial Neural Network model Information Sciences 194 107–19
[9] Rajpathak D, Chougule R and Bandyopadhyay P 2012 A domain-specific decision support system for knowledge discovery using association and text mining (Hong Kong: Knowledge and Information Systems) 31 405–32