Construction of ontology of problem area based on the syntagmatic analysis of text documents

A.A. Zarubin[a], A.R. Koval[a], V.S. Moshkin[b], A.A. Filippov[b]

[a] *The Bonch-Bruevich Saint - Petersburg State University of Telecommunication, 191186, 61 Moika street, Saint - Petersburg, Russia*
[a] *Ulyanovsk State Technical University, 432027, 32 Severny Venetz street, Ulyanovsk, Russia*

Abstract

The activities of any large organization requires the work of specialists with a large volume of unstructured information to obtain and extract the necessary knowledge to interact with partners, decision-making and so on. An array of unstructured textual information is not adapted to the structuring and semantic search. Thus, development intelligent algorithms and text analysis methods to dynamically generate the contents of a knowledge base is needed. Extract of syntagmatic structure of the text and further representation of extracted knowledge in the form of a single unified ontology allows you to access the knowledge base for solving complex problems.

Keywords: ontology; knowledge base; syntagmatic analysis; text resource

1. Introduction

In the process of any large modern organization activity, it is necessary to make urgent management decisions timely that requires specialists to have deep knowledge of the problem area (PrA). Moreover, they should be able to use different decision support systems and tools for work with knowledge.

The desire to automate and speed-up the process of obtaining necessary knowledge about the problem area drives the need in the unified multipurpose toolkit for knowledge management that does not require a user to have some additional skills in the field of knowledge engineering and ontological analysis.

Thus, one can identify a number of scientific problems besetting modern organizations. In order to be solved, such problems require the systematic approach and include the following ones:

- the need of developing the semantic basis for representation of electronic information storage content;
- the lack of integrative conceptual models using different approaches to the storage of knowledge about the PA;
- the need of unifying the automated processing of the stored knowledge;
- the need of simultaneous use of multi-aspect contexts of the PrA under consideration;
- the need of solving the problem of tracking the clarity of human reasonings.

Thereby, nowadays, the actual problem is providing specialists of a wide range of organizations with a universal tool allowing to address the knowledge management challenges [1]. Furthermore, the tool should not require some extra training of users.

At the moment, the ontological approach is most often used for organization of knowledge bases of expert systems. A lot of Russian and foreign researchers such as T.A. Gavrilova [2], V.N. Vagin [3], V.V. Gribova [4], Yu.A. Zagorulko [5]**Ошибка! Источник ссылки не найден.**, A.S. Kleschev [6], I.P. Norenkov, D.E. Palchunov, S.V. Smirnov [7], D. Bianchini, T.R.Gruber, A.Medche, G. Stumme and othersaddress the problem of integration and search of information in order to provide management decision support on the basis of an ontology.

In a broad sense, ontologies are models representing knowledge within the individual contexts of the PrA in the form of semantic information-logical networks of interrelated objects where the PrA concepts with properties and relations between objects are the main elements.

Ontologies serve as integrators proving the common semantic basis in the processes of decision-making and data mining, and the unified platform for combination of different information systems [8] [9].

2. Formal model of knowledge base

The knowledge base (KB) represents the storage of knowledge of different PAs and contexts in the form of an applied ontology. The PrA ontology context is a specific state of the KB content that can be chosen from a set of the ontology states. The state was obtained as a result of either versioning or constructing the KB content from different points of views [10].

Formally, the ontology can be represented by the following equation:

$$O = \left\langle T, C^{T_i}, I^{T_i}, P^{T_i}, S^{T_i}, F^{T_i}, R^{T_i} \right\rangle, i = \overline{1,t},$$

where t is a number of the ontology contexts, $T = \{T_1, T_2, \dots, T_n\}$ is a set of ontology contexts, C^{T_i} is a set of ontology classes within the i -th context, I^{T_i} is a set of ontology objects within the i -th context, P^{T_i} is a set of ontology classes properties within the i -th context, S^{T_i} is a set of ontology objects states within the i -th context, F^{T_i} is a set of the PrA processes fixed in the ontology within the i -th context, R^{T_i} is a set of ontology relations within the i -th context defined as:

$$R^{T_i} = \left\{ R_C^{T_i}, R_I^{T_i}, R_P^{T_i}, R_S^{T_i}, R_{F_{IN}}^{T_i}, R_{F_{OUT}}^{T_i} \right\},$$

where $R_C^{T_i}$ is a set of relations defining hierarchy of ontology classes within the i -th context, $R_I^{T_i}$ is a set of relations defining the 'class-object' ontology tie within the i -th context, $R_P^{T_i}$ is a set of relations defining the 'class-class property' ontology tie within the i -th context, $R_S^{T_i}$ is a set of relations defining the 'object-object state' ontology tie within the i -th context, $R_{F_{IN}}^{T_i}$ is a set of relations defining the tie between $F_j^{T_i}$ process entry and other instances of the ontology within the i -th context, $R_{F_{OUT}}^{T_i}$ is a set of relations defining the tie between $F_j^{T_i}$ process exit and other instances of the ontology within the i -th context.

3. Extraction the core of ontology of problem area based on the syntagmatic analysis of external wiki-resources

Wiki-resources formed a large number of users. Thus, applying of the automated methods for extraction the core of ontology based on the knowledge contained in the Wikipedia, can reduce the degree of subjectivity and increase the number of experts involved in ontology building process [11] [12].

Algorithm of extraction the core of ontology from the external wiki-resources based on the methods described in [3].

Problem area features in the wiki-resource represented as a hierarchy of associated hyperlinked HTML-pages having a certain semantics. The core of the ontology automatically extracted from external wiki-resources during of data mining. The core of the ontology can be expanded in the process of syntagmatic analysis of set of thematic text documents.

Extraction the core of ontology of problem area based on the syntagmatic analysis of external wiki-resources consists of sequence steps [13]:

1. The expert chooses the page of wiki-resource that describes the root concept of PrA.
2. Set of references to pages that describe other concepts is extracted from the selected page.
3. These pages are analyzed to check for references to the previous page.
4. The concepts of PrA extracted by the analysis of pages of wiki-resource added to the core of ontology.
5. Checking the condition of the existence of the route between the concepts derived in the first stage.
6. If the condition is satisfied, then the algorithm stops, otherwise the steps from the second to the fifth apply to the concepts extracted on the third step.

Concepts are reduced to the initial form (lemmatization). Defining types of relations between concepts is in the process of syntagmatic analysis of terms located on the right and the left of reference defines the concept. The rules for determining the type of relations are presented in the form of syntagmatic patterns (patterns contain a sequence of words).

Figure 1 shows the fragment of the core of ontology «LAN Administration» extracted from the thematic wiki-resource.

4. Construction of ontology of problem area based on the syntagmatic analysis of text documents

In the course of solving the problem of automated ontology expansion we developed two algorithms for terms extraction from domain texts using existing ontology core:

- thesaurus-based algorithm;
- internal linkage algorithm [14].

The main feature of the developed algorithms is term extraction from text documents by matching syntagmatic patterns with the lemmas of the objects from the core of ontology. Syntagmatic patterns are extracted by morphological analysis of text documents.

Thesaurus-based algorithm. A thesaurus is a reference work that lists words grouped together according to the similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words, and generally lists them in alphabetical order. Any ontology is a complicated version of the thesaurus.

Thesaurus approach assumes search of lemmas from the input words and their combinations among the terms defined in the ontology. For this purpose, each ontology class has a "HasLemma" property, which has a string value obtained by object name lemmatization.

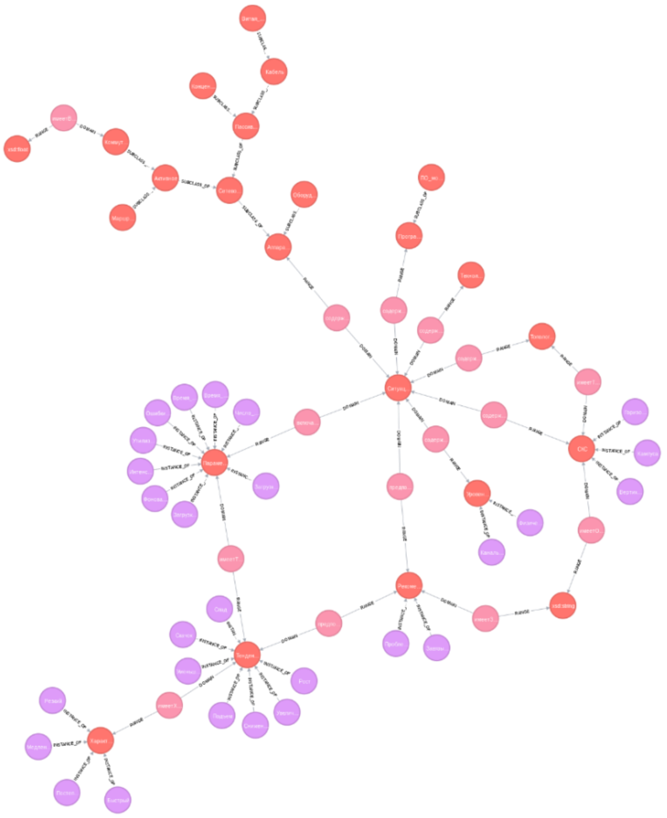

Fig. 1. The fragment of the core of ontology «LAN Administration».

The suppporting ontology object, used in further analysis, has the degree of proximity in relation to the input word / word combinations, is calculated by the following formula:

$$k_t = \max_{i=1}^{m} \frac{n_i}{p_i}, \qquad (1)$$

where m is the number of all ontology objects, n_i is the number of words from the input sequence, contained in the lemma of the current ontology object, p_i is the number of words in the current ontology object.

The process of assessing the proximity of the input words to the subject area terms is shown on Figure 2.

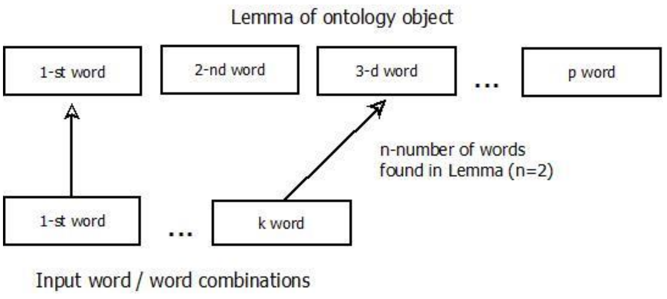

Fig. 2. Finding the supporting ontology object.

Each object in the ontology has an "IsTerm" property of boolean type. The degree of proximity of input words to the terms of domain, according to the Thesaurus algorithm, is calculated by the following formula:

$$k_{Ont} = \frac{k_t}{c+1}, \qquad (2)$$

where k_t is the result of the first step of the analysis, c is the number of relations between the supporting ontology object and the nearest object with the true “IsTerm” value.

Internal linkage algorithm. The developed metrics allows extracting terminology by not only defining the termhood of single words, but also comparing the terms from the text with ontology objects and lemmas combinations of those objects, using Radd relations. The Internal linkage algorithm is the implementation of this.

$$t_1 + R_1 + t_2 + R_2 + \ldots + R_m + t_n, \qquad (3)$$

where $R_i \in R_{add}$, $t_{\mathrm{j}} \in T$, R_{add} is a set of relations that allow expanding the set of objects of the described domain through a combination of related objects lemmas. For example: properties «IsRelatedWith» and «IsPartOf».

Thus, extracted terms that are part of other terms, consisting of more words, are not considered as terms in order to avoid redundancy.

5. The architecture of the knowledge base

The KB consists of some modules, which interact closely among themselves (Fig. 3):

- the module for managing the KB content;
- the module for import/export of the KB content from/to OWL format of the PrA ontology description;
- the module for organizing mechanisms for import content from external wiki-resources;
- the module for organizing the inference according to the KB content.

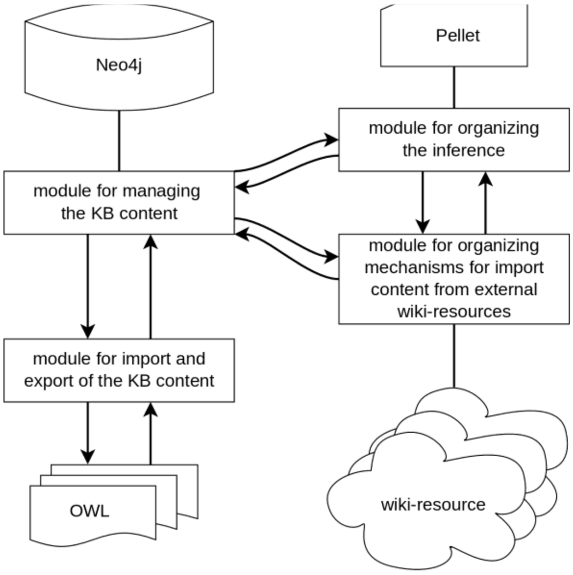
Neo4j
Pellet
module for organizing the inference
module for managing the KB content
module for organizing mechanisms for import content from external wiki-resources
module for import and export of the KB content
OWL
wiki-resource

Fig. 3. The knowledge base architecture.

In order to develop modules of the KB, Java programming language and Spring Boot framework were used [15] . Such development tools have the following advantages:

- high development rate;
- existence of documentation and active community of developers;
- platform independence;
- advanced infrastructure.

Neo4j [16] graph database is used as storage of ontologies for a module for managing the KB content. It has the following advantages:

- native format for graph storages;
- one database instance can serve graphs with billions of nodes and relations;
- it can process graphs that do not have enough space in RAM.

Modules are performed in the Jetty servlets container with the modular architecture that allows to use only needed functions, thereby, it reduces the performance load on the server. Also Jetty is highly scalable for performing a lot of connections with significant downtime between the queries. It also allows to serve a lot of users [17].

In order to develop means for interaction with modules, the REST (Representational State Transfer) [18] mechanism was used [19]. It this case, the remote procedure call represents a simple HTTP request (GET, POST, PUT, etc.), and necessary data are transmitted as parameters of the request. The main benefits of REST are the following ones:

- high performance due to the use of cash;
- scalability;
- integration system transparency;
- simplicity of interfaces;
- portability of components;
- modification simplicity.

For inference the Pellet [20] reasoned was used. It has the following [21]:

- soundness;
- completeness;
- SROIQ(D) expressivity support;
- incremental classification;
- SWRL rules support;
- justifications;
- ABox reasoning.

All the above resources, applications, and technologies are free.

6. Experiments

The text volume of about 62000 words from "LAN Administration" PrA was analyzed to assess the accuracy of the term extraction. OWL-ontology consisted of 261 classes and 46 relations.

Precision (P), Recall (R) and F_1 measures were used to assess the effectiveness of the algorithms for each category of tokens. Experiments on term extraction using the most frequently applied statistical methods: Frequency, TF*IDF, C-Value were also carried out. Results are presented in Table 1.

Table 1. Term extraction using statistical and syntagmatic methods

Amount of words	Terms	Candidates	Right	P	R	F_1
Internal linkage algorithm						
1	294	168	134	0,80	0,46	0,58
2	631	431	372	0,86	0,59	0,70
3	361	370	327	0,88	0,91	0,89
Frequency						
1	294	134	123	0,92	0,42	0,58
2	631	469	347	0,74	0,55	0,63
3	361	334	267	0,80	0,74	0,77
TF*IDF						
1	294	147	138	0,94	0,47	0,63
2	631	456	328	0,72	0,52	0,60
3	361	277	166	0,60	0,46	0,52
C-Value						
1	294	120	112	0,93	0,38	0,54
2	631	789	316	0,40	0,50	0,44
3	361	295	162	0,55	0,45	0,50

Thus, statistical methods showed significantly better results when retrieving one term tokens. Internal linkage algorithm first extracts terms related to existing knowledge base terms.

Internal linkage algorithm extracts less wrong terms In the case of two and three term tokens. Statistical methods are more focused on the frequency of occurrences of phrases, regardless of the reference to the Problem area features and can extract general scientific terms and terms from other problem areas. Statistical methods are more focused on the frequency of tokens without reference to the problem area and can extract general scientific terms and terms of other problem areas.

7. Conclusion

The use of mathematical and statistical approaches to the building of domain ontologies by extracting knowledge from text documents does not take into account morphological, semantic and syntagmatic features used in the text of linguistic forms. The methods of syntagmatic analysis allows:

- to reduce all synonyms for the same concept;
- to include polysemous words for different concepts;
- to use the connections between the concepts and the appropriate terms to generate a new ontology entities.

Thus, the experimental results suggest a high efficiency of the described in the article methods. These methods were developed by combining linguistic algorithms of terminology extraction from large text corpora in the process of syntagmatic analysis and extraction the core of ontology from external wiki-resources.

Acknowledgements

This paper has been approved within the framework of the federal target project “R&D for Priority Areas of the Russian Science-and-Technology Complex Development for 2014-2020”, government contract No 14.607.21.0164 on the subject “The development of architecture, methods and models to build software and hardware complex semantic analysis of semi-structured information resources on the Russian element base” (Application Code « 2016-14-579-0009-0687 »).

References

[1] Bova, V.V. Problemy predstavleniia znanii v integrirovannykh sistemakh podderzhki upravlencheskikh reshenii / V.V. Bova, V.V. Kureichik, E.V. Nuzhnov // Taganrog: Izvestiia SFedU – 2010. – Vol. 108 (7). – P. 107-113.

[2] Gavrilova, T.A. Ontologicheskii podkhod k upravleniiu znaniiami pri razrabotke korporativnykh informatsionnykh sistem / T.A.Gavrilova // Novosti iskusstvennogo intellekta – 2003. – Vol. 2(56). – P. 24-29.

[3] Vagin, V.N. Razrabotka metoda integratsii informatsionnykh sistem na osnove metamodelirovaniia i ontologii predmetnoi oblasti / V.N. Vagin, I.S. Mikhailov // Programmnye produkty I sistemy – 2008. – Vol. 1. – P. 22-26.

[4] Gribova, V.V. Upravlenie proektirovaniem i realizatsiei polzovatelskogo interfeisa na osnove ontologii / V.V. Gribova, A.S. Kleschev // Problemy Upravleniia – 2006. – Vol. 2. – P. 58-62.

[5] Zagorulko, Yu.A. Postroenie portalov nauchnykh znanii na osnove ontologii / Yu.A. Zagorulko // Vychislitelnye tekhnologii – 2007. – Vol. 12. – P. 169-177.

[6] Kleschev, A.S. Rol ontologii v programmirovanii. Chast 1. Analitika / A.S. Kleschev // Informatsionnye tekhnologii – 2008. – Vol. 10. – P. 42-46.

[7] Smirnov, S.V. Ontologicheskoe modelirovanie v situatsionnom upravlenii / S.V. Smirnov // Ontologiia proektirovaniia – 2012. – Vol. 2 (4). – P. 16-24.

[8] Golenkov, V.V. Semanticheskaia tekhnologiia komponentnogo proektirovaniia sistem, upravliaemykh znaniiami/ V.V. Golenkov, N.A. Guliakina // Materialy V mezhdunarodnoi nauchno" tehnicheskoi konferntsii OSTIS" – 2015. – Minsk – P. 57-78.

[9] Namestnikov, A.M. Realizatsiya sistemy klasterizatsii kontseptual'nykh indeksov proyektnykh dokumentov / A.M. Namestnikov, A.A. Filippov // Avtomatizatsiya protsessov upravleniya – 2011. – Vol.3 (25).– P. 46-50.

[10] Namestnikov, A.M An ontology based model of technical documentation fuzzy structuring / A.M Namestnikov, A.A. Filippov, V.S. Avvakumova // CEUR Workshop Proceedings, SCAKD 2016 – Moscow – 2016. – Vol.1687.– P. 63-74.

[11] Shestakov, V.K. Razrabotka i soprovozhdeniye informatsionnykh sistem, baziruyushchikhsya na ontologii i Wiki-tekhnologii / V.K. Shestakov// Tr. 13-y vserossiyskoy nauchn. konf. «RCDL-2011» – Voronezh – 2011. – P. 299-306.

[12] Hepp, M. Harvesting Wiki Consensus – Using Wikipedia Entries as Ontology Elements / Hepp M., Bachlechner D., Siorpaes K. // Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics, Annual European Semantic Web Conference (ESWC 2006) – 2006. – P. 124–138.

[13] Subkhangulov, R.A. Ontologicheski-oriyentirovannyy metod poiska proyektnykh dokumentov / Subkhangulov R.A. // Avtomatizatsiya protsessov upravleniya –2012. – Vol. 4 (30). – P.83 – 89.

[14] Yarushkina, N. Hybridization of Fuzzy Inference and Self-learning Fuzzy OntologyBased Semantic Data Analysis / Yarushkina N., Moshkin V., Klein V., Andreev I., Beksaeva E. // Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16) – 2016. – P. 277–285.

[15] Spring Boot Framework [Electronic resource]. — Access mode: https://projects.spring.io/spring-boot (9.01.2017).

[16] Neo4j [Electronic resource]. — Access mode: https://neo4j.com/product (10.01.2017).

[17] Greg Wilkins Jetty vs Tomcat: A Comparative Analysis [Electronic resource]. — Access mode: http://www.webtide.com/choose/jetty.jsp (10.01.2017).

[18] Representational state transfer [Electronic resource]. — Access mode: https://en.wikipedia.org/ wiki/Representational_state_transfer (9.01.2017).

[19] James Lewis, Martin Fowler Microservicesa definition of this new architectural term [Electronic resource]. — Access mode: http://martinfowler.com/articles/microservices.html (10.01.2017).

[20] Pellet Framework [Electronic resource]. — Access mode: https://github.com/stardog-union/pellet (10.01.2017).

[21] Dentler, K. Comparison of reasoners for large ontologies in the OWL 2 EL profile / Dentler, K., Cornet R., Ten Teije A., De Keizer N. // Semant. Web – 2011. –Vol. 2 – P. 71-87.