Construction of the problem area ontology based on the syntagmatic analysis of external wiki-resources

A.A. Zarubin[1], A.R. Koval[1], V.S. Moshkin[2], A.A. Filippov[2]

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

Abstract

The activities of any large organization requires the work of specialists with a large volume of unstructured information in order to obtain and extract the necessary knowledge to interact with partners, decision-making, etc. An array of unstructured textual information is not adapted to structuring and semantic search. Thus, development of intelligent algorithms and text analysis methods for dynamic generation of the knowledge base contents is needed. Extract of syntagmatic structure of a text and further representation of extracted knowledge in the form of a single unified ontology allows to get access to 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 PrA 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 PrA;
- the need of unified 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], A.S. Kleschev [6], I.P. Norenkov, D.E. Palchunov, S.V. Smirnov [7], D. Bianchini, T.R.Gruber, A.Medche, G. Stumme and others address 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 PrAs 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.

Formally, an 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 ontology contexts, $T = \{T_1, T_2, \quad, T_n\}$ 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. Extracting the core of ontology of the problem area based on the syntagmatic analysis of external wiki-resources

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

The algorithm of extracting the core of the ontology from the external wiki-resources is based on the methods described in [3].

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

The first method of extracting the core of the PrA ontology is based on the Lee algorithm [13]. Concepts are reduced to the initial form (lemmatization). Defining types of relations between concepts is in the process of the 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).

The second method of extracting the core of the domain ontology based on the contents of wiki-resources allows the intelligent system to adapt dynamically to the changes in the domain [14]. Methods of automatic text processing (ATP) in a natural language (NL) can be used in order to extract knowledge from the text of the wiki resource pages.

The ATP process is usually carried out in several steps [15]:

1. Grafematic analysis is the process of initial analysis of the text in a NL. The grafematic analysis presents the input data in a convenient format for further analysis (separation of input text into words, delimiters, etc).
2. Morphological analysis (lemmatization) is a process of transforming the words of the input text to the initial form defining the part of speech, gender, case, etc.
3. Parsing is the process of selecting members of simple sentences and constructing a parse tree.
4. Semantic analysis consists of
   - construction of a semantic tree of sentences,
   - semantic interpretation of words and constructions,
   - definition of semantic relations between elements of the text.

Semantic representation of the text in a NL is the most complete of those that can be achieved only by linguistic methods. The core of the domain ontology can be extended by merging with the semantic tree extracted from wiki-resources. It is necessary to develop a method for translating a parse tree into a semantic tree in order to obtain a semantic tree.

It is necessary to determine the syntactic structure of the sentence for constructing the semantic tree of sentences in a NL. There are several parsing tools of texts in Russian, for example [16, 17, 18]:

- Lingo-Master;
- Treeton;
- Sreda RGTU;
- DictaScope Syntax;
- ETAP-3;
- ABBYY Compreno;
- Tomita-parser;
- AOT etc.

In the context of this work, AOT (tool for constructing a parse tree) was chosen [18]. Let us consider the application of the algorithm for translating a syntactic tree into a semantic tree using the example of a test fragment in Russian:

> *Онтология в информатике — это попытка всеобъемлющей и подробной формализации некоторой области знаний с помощью концептуальной схемы.*

The parse tree for the test fragment is shown in the figure 1.

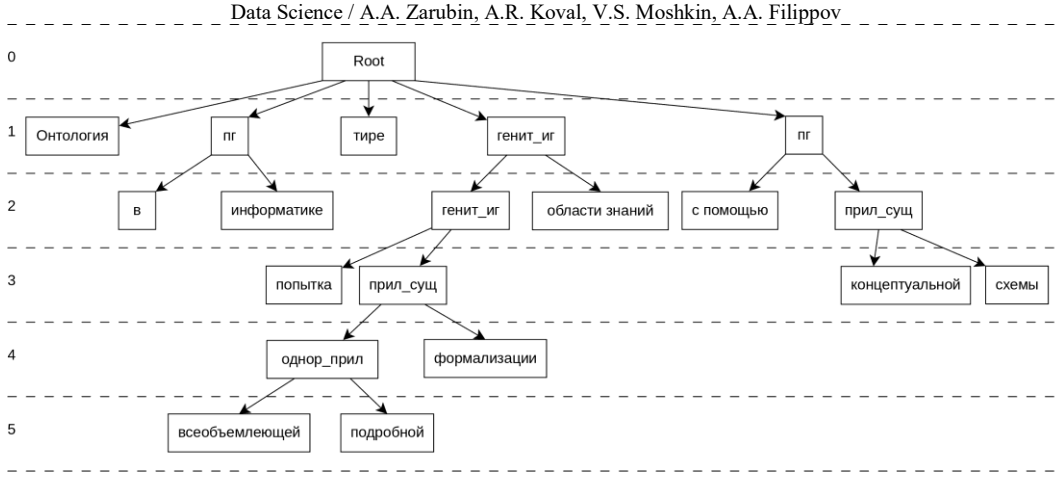

Fig. 1. Example of a parse tree.

Formally, the function of translating a parse tree into a semantic tree can be represented as follows:

$$F^{Sem}:\left\{N_{li}^{Synt},P_j\right\}\rightarrow\left\{N^{Sem},R^{Sem}\right\},$$

where N_{li}^{Synt} is the i-th node of the l - th level of a parse tree. For example, for the parse tree in Figure 1, the first node of the first level is the node “Онтология”, the second one is “пг”, the third one is “тире”, etc. The node of the parse tree can be a member of the sentence, for example, the node “Онтология”. Also, the parse tree node can be a syntactic label that defines the constituent members of the sentence, for example, “пг” (the prepositional group); P_j is the j-th syntagmatic pattern for defining the nodes of the parse tree. The nodes will be translated into nodes and relations of the semantic tree. The syntagmatic pattern is a collection of several words united according to the principle of semantic-grammatical-phonetic compatibility. Formally, syntagmatic pattern can be represented as follows:

$$\left(N_1^{Synt},N_2^{Synt},\quad,N_k^{Synt}\right)\rightarrow\left\{N^{Sem},R^{Sem}\right\},k=\overline{1,K},$$

where N_k^{Synt} is the k-th syntagmatic unit of the pattern corresponding to the node of the parse tree. It is necessary to use all the syntagmatic units included in it in order to use the syntagmatic pattern. Examples of syntagmatic patterns and the results of their use are presented in Table 1;

K – number of syntagmatic units in the pattern;

$\left\{N^{Sem},R^{Sem}\right\}$ are the sets of nodes N^{Sem} and relations R^{Sem} of the semantic tree obtained as a result of translation of the parse tree into a semantic tree. Formally, R^{Sem} can be defined as follows:

$$R^{Sem}=\left\{R_{isA}^{Sem},R_{partOf}^{Sem},R_{associateWith}^{Sem},R_{dependsOn}^{Sem},R_{hasAttribute}^{Sem}\right\},$$

where R_{isA}^{Sem} is a set of transitive relations of hyponymy;

R_{partOf}^{Sem} is a set of transitive relations “part/whole”;

$R_{associateWith}^{Sem}$ is a set of symmetrical relations of association

$R_{dependsOn}^{Sem}$ is a set of asymmetric relations of associative dependence;

$R_{hasAttribute}^{Sem}$ is a set of asymmetric relations describing the attributes of nodes.

Table 1. Examples of syntagmatic patterns and the results of their application.

Initial data	Syntagmatic pattern	Result
попытка-**генит_иг**-*формализации*	{node1}-{**генит_иг**}-{node2} → {node1}-associateWith-{node2}	*попытка*-associateWith-*формализация*
в-**пг**-*информатике*	{node1}-{**пг**}-{node2} → {prevNode}-***getRelation***(node)-{node2}	lastNode-***relation***-*информатика*
тире	{**тире**} → {prevNode}-isA-{nextNode}	lastNode-isA-nextNode
концептуальной-**прил_сущ**-*схемы*	{node1}-{**прил_сущ**}-{node2} → {node2}-hasAttribute-{node1}	*схема*-hasAttribute-*концептуальный*
(всеобъемлющей, подробной) **однор_прил**- *формализации*-	{node1}-{**однор_прил**}-{nodes} → {node1}- hasAttribute-{nodes[1]}, {node1}- hasAttribute-{nodes[2]}, {node1}- hasAttribute-{nodes[…]}, {node1}- hasAttribute-{nodes[count]}	*формализация*-hasAttribute-*всеобъемлющий*, *формализация*-hasAttribute-*подробный*

The algorithm for translating a parse tree into a semantic tree consists of the following steps:

1. Go to the first level of the parse tree.

2. Select the next node of the current tree level.
3. If the node is marked, go to step 2.
4. If the node is not a syntax label, go to step 9.
5. If the node is a syntax label and does not have child elements, go to step 9.
6. If the node is a syntax label and all its child nodes are not syntax labels, go to step 9.
7. If there is a temporary parent node, replace it, otherwise, create a temporary node.
8. If there is a previous node and there is no relation with it, add a temporary relationship with it and go to step 2.
9. Apply the syntagmatic pattern for translation.
10. Mark the processed nodes and go to step 2.
11. Go to the next level of the parse tree and go to step 2.

The resulting semantic tree for the test fragment is shown in Figure 2.

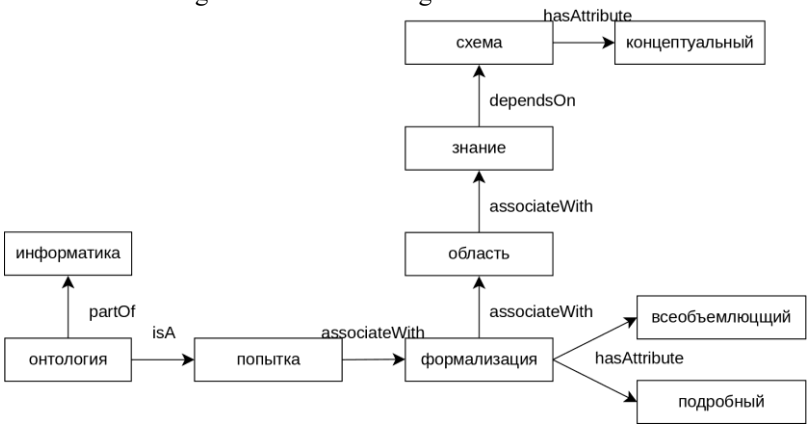

Fig. 2. Example of a semantic tree for a test fragment.

The result semantic tree can be merged with other semantic trees within the text. In addition, the semantic tree can be merged with the domain ontology compiled by an expert. Extending the knowledge base by merging semantic trees retrieved from semi-structured resources allows:

- provide a common terminology space for sharing and understanding by all users;
- determine the exact and consistent meaning of each term.

Ontology is a common terminological basis for complex iterative processes. Figure 3 shows the fragment of the core of the ontology “LAN Administration” extracted from the thematic wiki-resource.

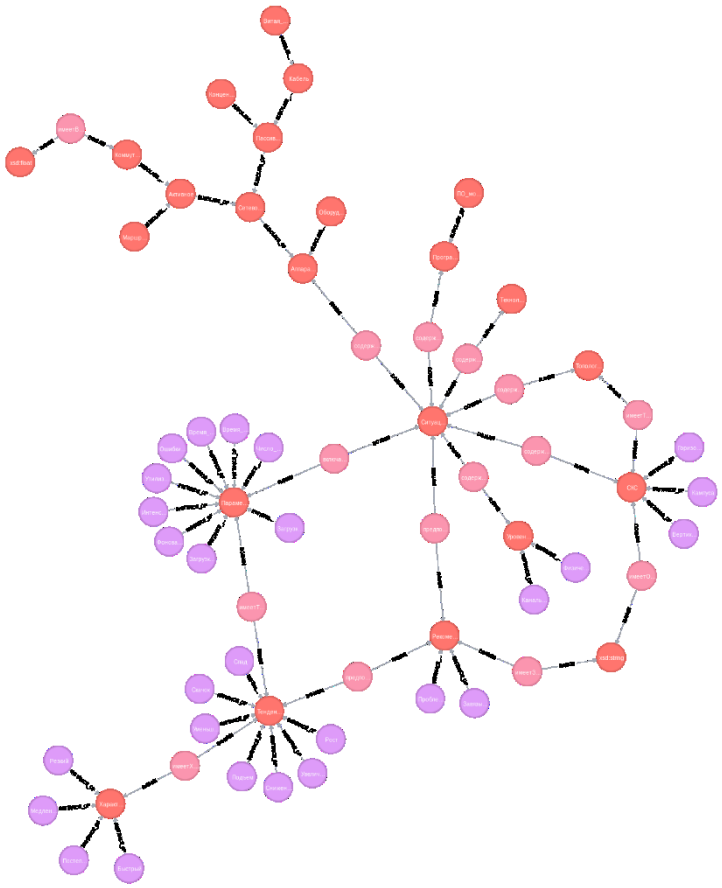

Fig. 3. The fragment of the core of the ontology “LAN Administration”.

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

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

- the thesaurus-based algorithm;
- the internal linkage algorithm [19].

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

The thesaurus-based algorithm. A thesaurus is a reference work that lists words grouped together according to the similarity of meanings (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.

The 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 “HasALemma” property, which has a string value obtained by object name lemmatization.

The supporting ontology object used in the further analysis has the degree of proximity in relation to the input word / word combinations that is calculated by the following equation:

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

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 4.

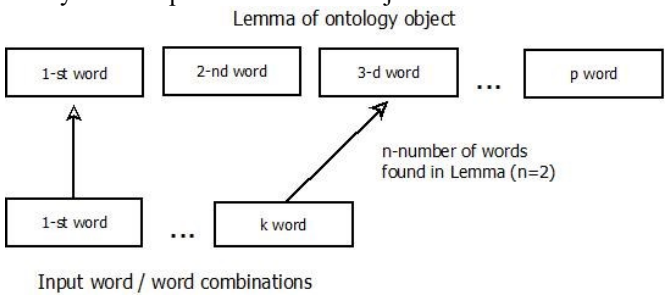

Fig. 4. Finding the supporting ontology object.

Each object in the ontology has an “IsATerm” 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 equation:

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

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 “IsATerm” 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 the following one.

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

where $R_i \in R_{add}$, $t_j \in T$, 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 “IsAPartOf”.

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

5. 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 2.

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

The internal linkage algorithm extracts less wrong terms in 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 PrA features and can extract general scientific terms and terms from other problem areas. Moreover, statistical methods are more focused on the frequency of tokens without reference to the PrA and can extract general scientific terms and terms of other problem areas.

Table 2. 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

6. 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 methods described in the article. The methods were developed by combining linguistic algorithms of terminology extraction from large text corpora in the process of syntagmatic analysis and extracting the core of the ontology from external wiki-resources.

Acknowledgements

This paper has been approved within the framework of the Federal Targeted Programme for Research and Development in Priority Areas of Development of the Russian Scientific and Technological Complex for 2014-2020, Government Contract No. 14.607.21.0164 on the subject “Development of architecture, methods and models in order to build software and hardware complex for semantic analysis of semi-structured information resources on the Russian element base” (Application Code “2016-14-579-0009-0687”).

References

[1] Bova VV, Kureichik VV, Nuzhnov EV. Problems of representation of knowledge in integrated systems of support of administrative decisions. News of SFedU 2010; 108(7): 107–113.
[2] Gavrilova TA. Ontological approach to knowledge management in the development of corporate information systems. Artificial Intelligence News 2003; 2(56): 24–29.
[3] Vagin VN, Mikhailov IS. Development of the method of integration of information systems based on metamodelling and ontology of the subject domain. Software Products And Systems 2008; 1: 22–26.
[4] Gribova VV, Kleschev AS. Managing the design and implementation of the user interface based on the ontology. Management 2006; 2: 58–62.
[5] Zagorulko YuA. Construction of scientific knowledge portals based on ontology. Computational Technologies 2007: 12: 169–177.
[6] Kleshchev AS. The role of ontology in programming. Part 1. Analytics. Information Technologies 2008; 10: 42–46.
[7] Smirnov SV. Ontological modeling in situational management. Ontology of Design 2012; 2(4): 16–24.
[8] Golenkov VV, Guliakina NA. Semantic technology of component design of knowledge-driven systems. Fifth International Scientific and Technical Conference "OSTIS". Minsk, 2015: 57–78.
[9] Namestnikov AM, Filippov AA. Implementation of the clustering system for conceptual indexes of project documents. Automation of management processes 2011; 3(25): 46–50.
[10] Namestnikov AM, Filippov AA, Avvakumova VS. An ontology based model of technical documentation fuzzy structuring. CEUR Workshop Proceedings, SCAKD 2016; 1687: 63–74.
[11] Shestakov VK. Development and maintenance of information systems based on ontology and Wiki-technologies. 13-th all-Russian Scientific Conference “RCDL-2011”. Voronezh, 2011: 299–306.
[12] Hepp M, Bachlechner D, Siorpaes K. Harvesting Wiki Consensus – Using Wikipedia Entries as Ontology Elements. Proceedings of the First Workshop on Semantic Wikis – From Wiki to Semantics. Annual European Semantic Web Conference (ESWC), 2006: 124–138.
[13] Subkhangulov RA. Ontologically-oriented method of searching for project documents, Automation of management processes 2012; 4(30): 83–89.
[14] Konstantinova NS, Mitrofanova OA. Ontologies as a knowledge storage system. Portal "Information and Communication Technologies in Education". URL: http://www.ict.edu.ru/ft/005706/68352e2-st08.pdf (21.03.2017).
[15] Sokirko AV. Semantic dictionaries in automatic processing: issertation for the degree of candidate of technical sciences. State Committee of the Russian Federation for Higher Education Russian State Humanitarian University. Moscow, 2001.

[16] Boyarskiy KK, Kanevskiy YeA. Semantico-syntactic parser Semsin, Scientific and Technical Herald of Information Technologies. Mechanics and Optics 2015; 5: 869–876.
[17] Artemov MA, Vladimirov AN, Seleznev KYe. Survey of natural text analysis systems in Russian. Scientific journal Bulletin of Voronezh State University. URL: http://www.vestnik.vsu.ru/pdf/analiz/2013/02/2013-02-31.pdf (22.02.2017).
[18] Automatic text processing. Automatic word processing. URL: http://aot.ru(22.02.2017).
[19] Yarushkina N, Moshkin V, Klein V, Andreev I, Beksaeva E. Hybridization of Fuzzy Inference and Self-learning Fuzzy Ontology Based Semantic Data Analysis. Proceedings of the First International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’16), 2016: 277–285.