Afruz M. Gurbanova
Institute of Information Technology of ANAS, Baku, Azerbaijan
afruz1961@gmail.com
DOI: 10.25045/jpit.v09.i1.08

AUTOMATED CONSTRUCTION OF THE SEMANTIC NETWORK OF SUBJECT FIELD TERMS TECHNOLOGIES

The article analyzes automated construction technologies of the semantic network of subject field terms. The advantages of using semantic networks in automatic text processing problems are presented. It is proposed to use the terminological vocabulary of the subject domain as the initial information for building a semantic network. The method of calculating the strength of semantic links between terms of the domain is analyzed and software is developed for the application of the method. It is shown that the method of automated construction of a semantic network of terms can be used in designing ontologies when creating corporate knowledge bases.

Keywords: *terminological system, terminological dictionary, semantic relationships, semantic network, ontology.*

Introduction

Automated extraction of knowledge from scientific texts implies not only disclosure of terms, but also the knowledge about the terms. In this regard, it is important to recognize semantic relationships among the terms in the text, as they describe the semantic structure of terminology [1].

The semantic representation of the text makes it easy to make conclusions and decisions, and raises the productivity of the questionnaire, information retrieval, machine translation, and other natural language processes [2]. Semantic relationships are the association among words (semantic words at word level), meanings of expressions, or meanings of sentences (semantic relationships at expression or sentence levels). There are some types of semantic relations at word level such as synonyms, antonyms, homonyms, polysemy (polygamy), metonymy, and etc.

The relationship among the terms of the given subject area during the construction of the terminology system should be taken into account. A simple list of terms is different from the system of terms. The term system is a description of the field of concepts of a specific field of research. It reflects not only the notions, but also the relationships of the field. These relations enter the system as the inter-term relationships. The term system has a semantic structure that transfers the knowledge, while the list of words does include such structure [3].

Thus, the question of establishing a system of inter-term relationships or a system of terminological relationships is directly related to the studies in each area.

The system of terms is a structured set of terms or a complex system of semantic relations among the terms. The description of the terms system is closely related to the classification of the terms and the classification of the relationships among them.

The classification of the Austrian scientist E.Wooster is taken as a basis for linguistics. According to E. Wooster's classification, two groups of inter-terms relations are distinguished:

- ontological relationships;
- logical relationships.

In each category, a personal hierarchy is built. Logical relationships are defined as the proximity and similarity relationships, while ontological relationships are defined as the direct relationships in terms of time and space.

One of the first classifications of the relations among the concepts of the subject field is offered by the Soviet linguist and terminologist T.L.Kondelaki. Several compound classifications of terms and relations are available [4].

More detailed and accurate classifications are focused on the information search engine, which have been emerged due to the combination of logical and philosophical research on knowledge.

There is a sharp inconsistency among the terminology systems of different subject areas,

which are currently being taught in high education institutions. One of the ways to eliminate this inconsistency is the technological approach to the harmonization and implementation of the teaching process. The implementation of the technological approach requires the models and methods for the development of the tools for the automated processing of scientific and educational textual information.

Study of Methods

The textbooks used in the educational process of higher education institutions is analyzed in [5]. The use of ontological modeling for the tools supporting the teaching process is grounded, and possible relationships among the notions of the subject area are considered.

The use of the ontological approach involves analyzing the subject matter, selecting its concepts, determining the relationships among these concepts, and a mathematical description of the rules of logical conclusion taking into account those relationships [6, 7].

Studies show that these procedures are put forth to tackle an issue stemming from the waste of time in intellectual resources. Therefore, the automation of the procedures is very relevant aiming at the problem solution.

The outcomes of some studies were provided in subsequent sections.

Based on the analysis of terminological dictionaries, lexicon is automatically extracted and included in the information-retrieval thesaurus in [8]. The author develops a technology for the automatic processing and editing of the Azerbaijani texts at various language levels (morphological, syntactic, and semantic), logical extracting methods and synthesis of expert systems.

In the study, the terminological data bank architecture is developed related to the terminological data bank of the Azerbaijani language. In addition, it offers the models and methods for compiling, analyzing and correcting the specialized explanatory, terminological and automatic dictionaries, information-retrieval thesaurus for users and information systems [9].

The use of semantic networks in the automatic processing of texts is given in [10].

The author here substantiates the advantages of the representation of the lexicon network as follows:

- Representing the semantic relations among the words;
- Defining the numerical parameters characterizing the semantic structure of the lexicon and the semantic relationship system;
- Identifying the semantic relationships between any two units of the lexical and semantic system;
- Determining the strength of the semantic relationships among the units of the lexical and semantic system.

It should be noted that in [10], the manual construction of the semantic relationships based on the results of the analysis of explanatory dictionary texts is examined. However, this is not a simple matter, and its solution is related to the detection of inter-terms semantic relationships and to the fact that it is quite challenging to keep the track of the link chain.

The approach to the domains of inter-term semantic relationships and automatic determination of their content is based on the use of the syntax pattern and the results of the lexical unit analysis in the text [11].

Note that, in order to support the modeling process of the subject area, the software is developed based on the model of automated development of a large network of semantically related terms of the natural language. SemNet - with 37 million relationships describing the lexical templates, direct relationships analysis, and the rate of semantic relationships and approximately 2.7 million single-word and multi-word terms with the use of several algorithms [12].

A method of automatic construction of the semantic network of terms is developed to construct the ontology of the text using the results of the text source analysis of the information containing the content of the educated discipline [13].

The goal of this study is to develop software for the application of this method and to verify its accuracy.

The method presented in this study aims at the determination of the strength of the semantic relationships among the terms of the subject area. Here, the strength of the relationships characterizes the range of terms in several hypothetical and semantic domains, which is the semantic domain of the educated discipline. Unlike in the concepts of mathematics and physics, semantics in linguistics is conceived as a set of language units combined by some common semantic attributes.

A formal description of the semantic domain of the subject field is denoted by $U = \{x_k\}$. $U = \{x_k\}$ is a set of word forms when describing a part of the content of the educated discipline, where $k = \overline{1, K}$, and K is a cardinal number of the set.

Thus, the semantic domain is $P = (T, S, R)$.

Here, $T = \{t_j\}$ is a set of terms of the subject field $j = \overline{1, J}$; $S = \{s_i\}$ - a set of meaningful content of the terms, $i = \overline{1, I}$; J and I are the cardinal numbers of T and S sets, where $(s_i = \{x_{i_k}\};\ x_{i_k} \in U)$; $R = \{r_{j_1, j_2}\}$ - a set of semantic relationships among the elements of the set $T\{j_1, j_2 \subset J; j_1 \neq j_2\}$, which implies the strength of the relationships among the terms $t_{j_1}\{x\}$ and $t_{j_2}\{x\}$. The elements $t_j\{x\}$ and $s_i\{x\}$ indicate a chain of word forms, and are formed by the elements of the set $U = \{x_k\}$.

It should be noted that, ideally, one term must correspond to one content. In a real situation, a term may have several meanings. In the pedagogical practice, authors often try to give one definition to the term when describing the content of the educated subject. In this case, binary relationships among the elements of the sets T and S are established, resulting in a set of definitions $D = \{d_f\}$.

Binary relationships can be presented as follows:

$$d_f = t_j + s_i$$

The words in the alphabet of a certain language are given. A set $U = \{x_k\}$ here is a chain of alphabet, while the words define a chain of word forms formed from this alphabet.

It should be noted that many linguists refer to the issues related to the methods of detecting the meaning of words and the semantic relationships among them. Unlike the syntax relationships among words, the semantic relationships refer to unobserved objects. It is therefore necessary to use informal or formal methods to detect the semantic relationships. This is based on the correlation dependence between several observed signs of language detection and semantic relationships.

Non-formal methods are intuitive and use the "sense of language" of a native speaker. The use of non-formal methods formulate multiple explanatory dictionaries. These dictionaries sufficiently describe the language system and can be used as a tool for linguistic research in the formal processing of natural language.

There is a hypothesis on the existence of the correlation dependence between the semantic relationships of words and several observed indirect attributes of these relationships based on formal methods.

Nowadays, formal methods for identifying semantic relationships are widely used [11]. According to the authors, the semantic relationships among word forms are identified by the similarity of their lexical meaning.

Based on this, various groups consisting of the words characterized by a certain degree of similarity, correspondence and degree of proximity are formed. For example, synonymic words are characterized by the degree of proximity, these words are considered to be closest to

themselves and equivalent. However, in natural language, such identification does not exist, since a word does not have a precise boundary, which requires the use of the concept of proximity. Semantic relationships among words include the function of contact among the surrounding objects, which, in turn, implies the use of formal methods.

The system of semantic relationships of the lexicon, which imitates the system of relations of the subject field, is not completely defined by the lexical system. The system of semantic relationships in the lexicon is not isomorphic with the system of relationships between the objects and events in the subject area.

The analysis of the methods for solving semantic relationships among the word forms shows that the combination of formal and non-formal methods perform more precise results [10].

These approaches demonstrate that the lexical meaning of terms is defined by the definitions containing their components x_i. Obviously, different relationships among terms are available, such as common, specific, and attached. However, such relationships are not clearly defined in the dictionaries. It is necessary to define the sets of semantically related clusters of the set $T = \{t_j\}$ for a clear description of the relation $R = \{r_{j_1,j_2}\}$ of the semantic domain of the educated discipline. The terms t_{j_1} and t_{j_2}, which have a common semantic component as the elements S_{j_1} and S_{j_2}, are semantically correlated. They also provide the content of the mentioned terms. The more semantic components in the elements S_{j_1} and S_{j_2} correspond, the stronger the semantic relations between the terms t_{j_1} and t_{j_2} is.

To calculate the strength of the semantic relationships among the terms, i.e., the number of elements of the set $R = \{r_{j_1,j_2}\}$, a peer comparison of the components determining each term should be performed and the number of corresponding components should be determined.

The value of the relation strength among the terms can be calculated by the following formula [13]:

$$r_{j_1,j_2} = \frac{card(S_{j_1} \cap S_{j_2})}{card(S_{j_1} \cup S_{j_2})} \quad (1)$$

Here, r_{j_1,j_2} denotes the value of the power of communication between the terms t_{j_1} and t_{j_2}. $card(S_{j_1} \cap S_{j_2})$ is the number of the corresponding components, which define the meaning of the t_{j_1} and t_{j_2}. $card(S_{j_1} \cup S_{j_2})$ is the total number of the components, which define the meaning of the terms t_{j_1} and t_{j_2}.

Application of the method for the automatic construction of semantic domain

It should be noted that the Institute of Information Technologies of the Azerbaijan National Academy of Sciences developed the concept and web-portal of the National Terminological Information System (NTIS) [14]. The web portal provides e-terminology services to citizens, and aims to involve a broad public in the process terminology building. It includes terminological dictionaries covering various fields of science and technology.

The accuracy of the method can be illustrated by applying this method to a fragment of the terminology dictionary (figure 1). The terms and their explanations presented in Figure 1 are taken from the dictionary "Basic terms used in scientific activity" included to NTIS [15].

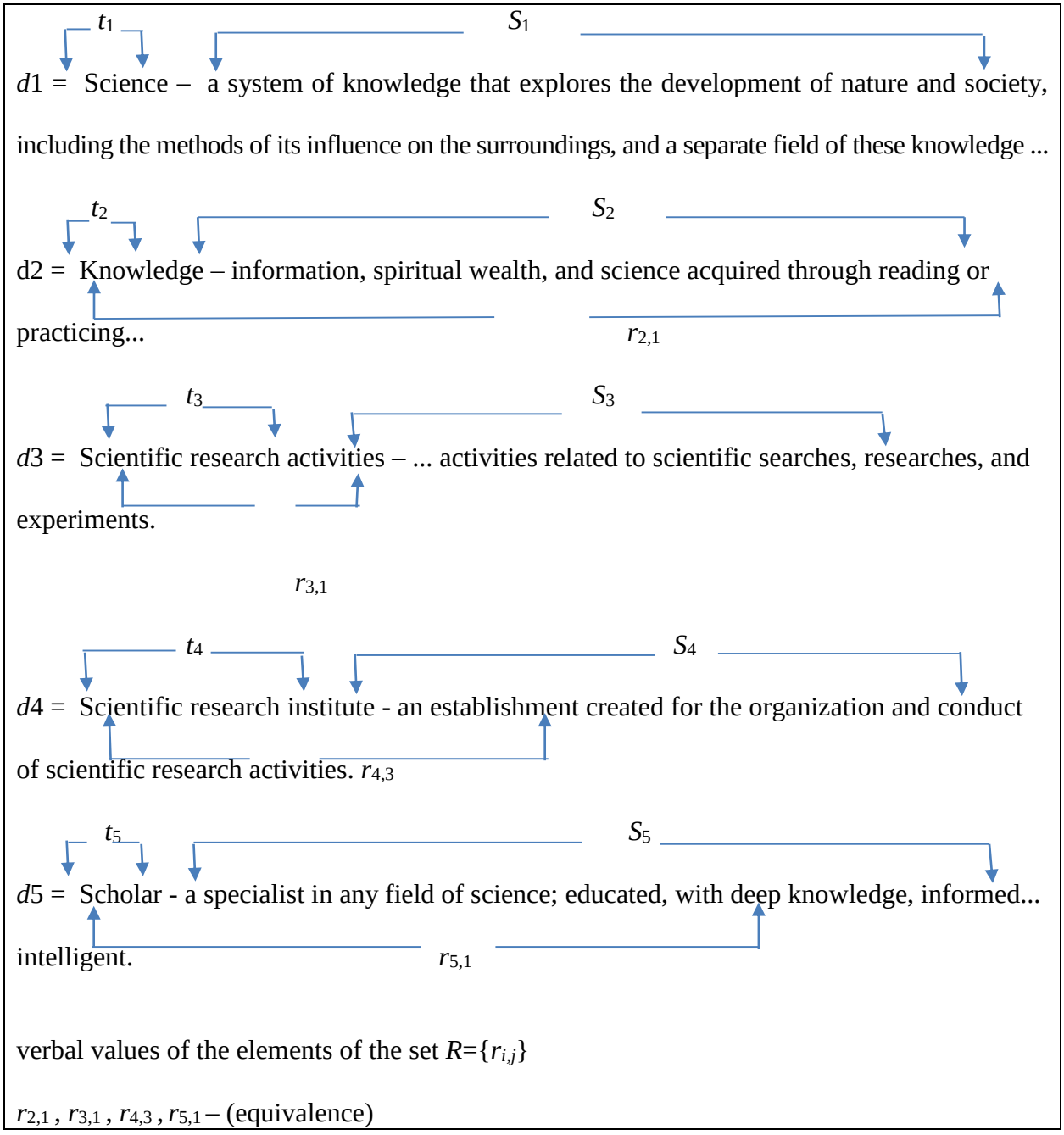

Figure 1. Verbal values of elements of the semantic domain fragment of the subject field

A semantic network is presented in a fragment of the terminology dictionary (Figure 2). The texts of the definitions should be processed prior to the practical use of the formula (1). It should first be filtered and then normalized. At the filtering stage, the word forms that do not affect the meaning of the term are removed. These phrases include prepositions, connections, pronouns, and so forth. Normalization is essential, for example, in different definitions, the word forms are used with some modifications depending on the noun cases: the words "network", "on network" or "to network" are semantically same terms, and therefore considered as different components when calculating the volume of the relation strength among the terms. For this reason, the nouns and adjectives should be reduced to the nominative case of noun.

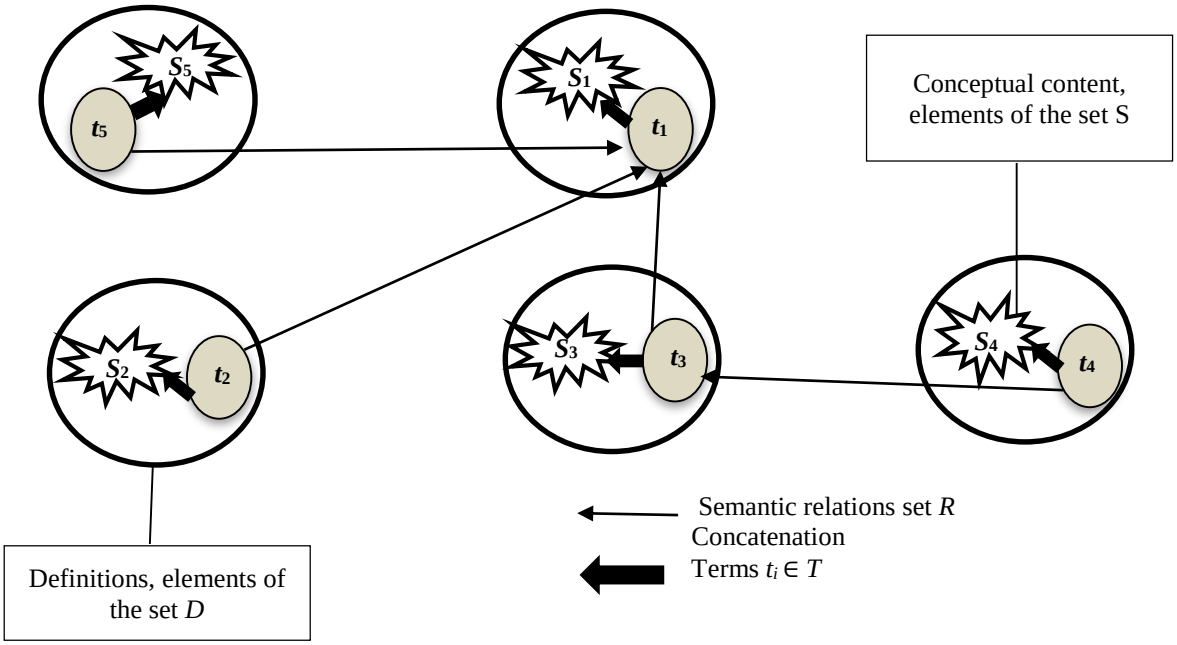

Figure 2. Semantic domain fragment of terminological dictionary

A terminological database is created for the application of the method [12]. The structure of the database is presented in Figure 3. As it is described in Figure 3, the Terminological Database (TDB) consists of three tables. The first table (Terms) contains all the terms that cover the subject area. The second table (Cat) contains all the components involved in definition of the terms in the table Terms. The third table is a relations table, and it contains only the components of the terms from the table Cat that corresponds to the term from the table Terms.

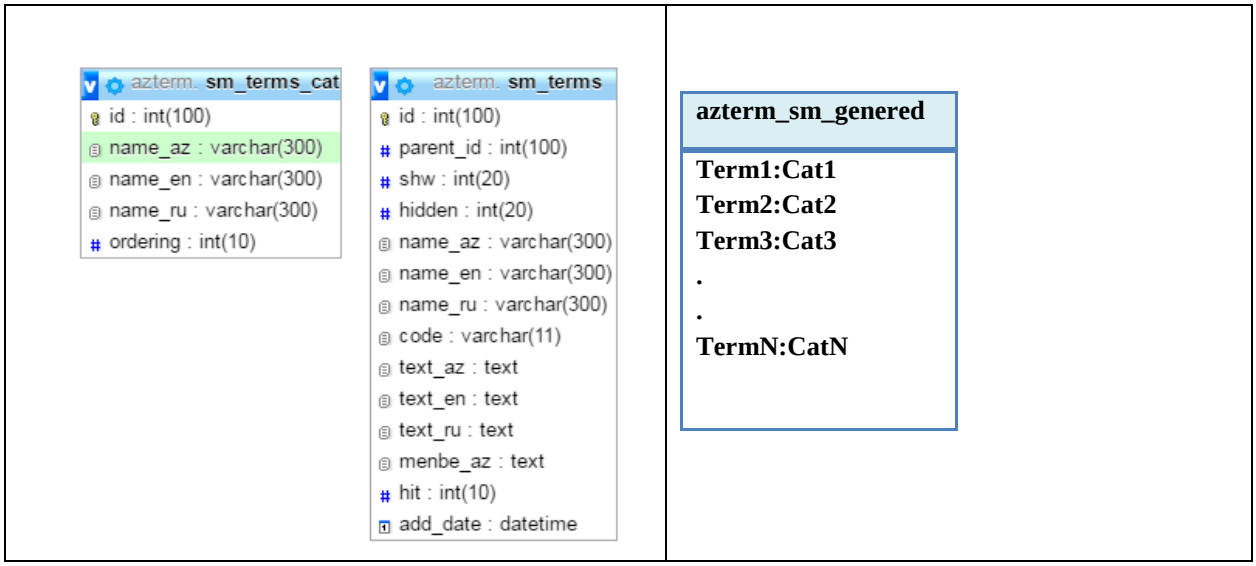

Figure 3. Structural scheme of TDB

Establishing query generation function in the database management system enables the processing of data on the TVB. With the help of complex queries in TVB, the comparison of components of two term components on formula (1) and their semantic relationships are calculated. The software has been developed to calculate the power of semantic relationships between terms. The terminological database, based on terminological dictionaries, is used here as the initial data. The program provides a semantic chart table description (Table 1). Semantic

related terms are shown in the first and third columns of the table, consisting of six columns. The second (M) and fourth (N) columns show the total number of components that form the meanings of the terms that are being compared. The fifth (C) column shows the overlapping components of the terms that are compared. In the Sixth (V) column, the value of the semantic relationship force calculated by the formula (1) is displayed.

Table 1

Semantic domain description

First term	M	Second term	N	C	V
Science	8	Knowledge	6	1	0,142856
Scientific-research activity	5	Science	8	2	0,307692
Scientific research Institute	6	Scientific-research activity	5	2	0, 363636
Scientist	8	Science	8	1	0,125
Science	8	Scientific research Institute	6	1	0,142856
Knowledge	6	Scientist	8	1	0,142856

Conclusion

Semantically related term domain was presented in the fragment of the terminology dictionary included in NTIS. The software was designed to calculate the strength of the semantic relationships among the terms, and the obtained results were presented in Table 1. The method described above allows the automation of the semantic relationships for the formation of terms and notions based on the rules of the natural language. This facilitated the explicit description of the structure of the concept of the subject field. The presentation of the Concept of the Subject Field as a semantic domain enhances the quality of textbooks and paves way for new opportunities.

It should be noted that NTIS has more capabilities. For example, it can be used to expand keywords in search.

The automation of the semantic domain of terms can be used to develop a corporate knowledge base.

This work was financially supported by the Science Development Fund under the President of the Republic of Azerbaijan - Grant # EIF-2014-9 (24) -KETPL-14/02/1

References

1. Naykhanova L.V. The main types of semantic relations among terms of the subject domain // Technical sciences. Computer science, computer facilities and management, 2008, No 1, pp.62-71.
2. Eduardo Blanco, Hakki C. Cankaya, Dan Moldovan. Composition of Semantic Relations: Model and Applications. Human Language Technology Research Institute, The University of Texas at Dallas, Coling 2010: Poster Volume, Beijing, August 2010, pp.72-80.
3. Tikhonov A.N., Ivannikov A.D., Tsvetkov V.Ya. Terminological relations // Scientific and theoretical journal "Fundamental Research", Moscow, 2009, No5, pp.146-148.
4. Kandelaki T.L. Semantics and motivation of terms. Moscow: Nauka Publishers, 1977.
5. Fedorchenko L.A. Peculiarities of constructing the linguistic ontology of educational-methodological material // Vestnik of the International Slavic University. Series "Technical Sciences", 2008, No. 1, pp. 34-44.
6. Gavrilova T.A. Knowledge bases of the intellectual systems. St. Petersburg: Peter, 2001, 384 p.
7. Gladun A.Ya. Ontologies in corporate systems. Part 2. Corporate systems, 2006, No1, www.management.com.ua/ims/ims116.html
8. Mamedova M.H., Skorokhodko E.F. Automated-System of Terminological Dictionary Analysis // Nauchno-tekhnicheskaya informasiya, seriya 2 - Informatsionnye protsessy i sistemy, 1981, pp.14-18.

9. Mammadova M.H. Creating a terminological data bank of the Azerbaijani language // Turkology, 1990, No2, pp. 84-89.
10. Skorohodko E.F. Semantic networks and automatic text processing. Kiev: Naukova Dumka, 1983, 212 p.
11. Fedorchenko L.A. Formalized presentation of the fragments of the text of educational and methodological material // Bulletin of the International Slavic University, Series "Technical Sciences", 2007, No. 1, pp.44-52.
12. Henning Agt, Ralf-Detlef Kutsche, Automated Construction of a Large Semantic Network for Domain-Specific Modeling. Database Systems and Information Management Group DIMA / CAİSE'13 Proceedings of the 25th International Conference on Advanced Information Systems Engineering, Valencia, Spain, June 17-21, 2013, pp.610-625.
13. Fedorchenko L.A., Khayrova N.F., Dovnar A.I., Khayrova S.O. The method of automated construction of the semantic network of terms of the academic discipline // Radio-electronic and computer systems, 2011, No 4 (56), pp.115-120.
14. www.azterm.az
15. Alguliyev R.M., Shukurlu S.F., Kazimova S.I. Basic terms used in scientific activity, Baku: "Information Technologies", 2009, 201 p.