Bull. Nov. Comp. Center, Comp. Science, 26 (2007), 133–146
© 2007 NCC Publisher

Ontology-based approach to text analysis[*]

Elena Sidorova, Yury Zagorulko

Abstract. The paper presents an approach to document analysis adaptable to a certain genre of the document and subject domain. The approach is intended for updating a database of an information system by means of information obtained by automated analysis of electronic documents. This approach implies using linguistic knowledge about the document language and formal structure of document texts, as well as expert knowledge on a subject domain of the information system.

1. Introduction

The need for automatic processing of non-structured or semi-structured text resources is vital for many information systems. The necessity of this service is primarily conditioned by a great volume of “raw material” (electronic business letters, Internet pages, text databases) which cannot be processed manually. The correctness, completeness and topicality of information provided to a user depends, in the final analysis, on the speed and quality of material processing.

The process of analysis of a natural language text is a hardly-formalizable problem. The existing universal methods of text analysis do not allow us to fully automatize the process of filling the information system with new data.

Other approaches imply strict orientation of methods of analysis to limited subject domains and (sub)languages. This allows applying special knowledge to solving the problem of semantic ambiguity which arises in text analysis. The main disadvantage of these approaches is impossibility of reusing the implemented modules of text processing for other subject domains and systems. A solution to this problem can consist in using a flexible, easily extensible and adjustable knowledge base.

Recently, for the purpose of a more effective use of linguistic knowledge in text analysis, various technologies for building and using thesaurus [1], lexical [2, 3] and linguistic [4] ontologies have been developed. (Hereinafter, we will use the term “linguistic ontology” to mean the set of linguistic and expert knowledge necessary for the natural language text analysis.) In particular, knowledge presented in the form of ontology is used for information extraction (IE), that is a method for analyzing texts which represent facts in

[*] The authors are grateful to the Russian Foundation for the Humanities (grant 07-04-12149) for financial support of this work.

a natural language and extracting relevant pieces of information from these texts [5]. Extracting information from texts requires knowledge of grammars describing a specific syntax of the texts to be analyzed, as well as semantic and ontological knowledge.

In this paper, the ontology-based approach to text analysis is considered. It is characterized by using the linguistic knowledge about document language and formal document text structure, as well as the expert knowledge on a subject domain of an information system. The use of such knowledge presented in a form of ontology makes the information system easily adjustable to any subject domain and any type of information resources.

2. Ontology in information systems

The peculiarity of modern information systems (IS) is in the fact that they actively use knowledge about both subject and problem domains, and the knowledge is explicitly described by ontology [6]. The use of ontologies allows formalizing and unifying the operations of information processing for improvement of quality of various information services. One of the most demanded services related to the topic of this paper is providing the information content to the IS based on automatic analysis of text resources.

The information systems based on ontologies possess the following specific features:

- **They use ontological descriptions of objects stored in their databases.** The classes of objects of the information system corresponding to the concepts of subject and problem domains of IS are described in the ontology. Structures of the classes and possible connections (relations) between them are also presented here. The entities of the subject and problem domain of IS and various information resources (text documents, images, video files, etc.) can be considered as objects of IS (hereinafter we will name them the information objects). Thus, any information stored in IS is represented by a set of information objects (IO) and their relations. Note that none of information objects or resources can be stored in a system if it is not described by some class (concept) of ontology.
- **The content of each information resource is presented as a set of IOs stored in a database of IS**. It should be noted that such a description of an information resource is not its semantic copy as the ontology of IS represents only those aspects which are important for solution of tasks of the information system.

The use of ontology makes the knowledge system of IS easily extensible and adjustable: both a new knowledge (e.g., new concepts and relations within a subject domain) and new types of information resources (e.g., new genres of documents) can be integrated into IS.

Data are presented as a set of IOs of various types. Each IO is an instance of some element of the ontology (class or relation) and has a structure with a fixed set of attributes specified by the expert.

Any IO may be considered as having three different aspects — structure, content, and context. The structure of IO is characterized by a set of its attributes. The context specifies possible environment of IO and is defined by a set of relations with other IOs. The format of IO structure and its context are defined by the ontology.

For example, the context of IO can be formed by the following relations of the ontology:

- *Part (Publication, Collection)* — the relation that connects a part to the whole (e.g., an article and a collection of articles);
- *Author (Person, Document)* — the relation that connects a document and its creator;
- *Publisher (Organization, Collection)* — the relation that connects the book with the organization that issues it.

The content of IO is described by a network of objects of the subject domain.

The content of IO that is a text resource (document) can be obtained as a result of the automatic analysis of its text. The purpose of the analysis is extracting information about the objects of a subject domain.

Thus, the technology of text analysis uses the following components of the knowledge system of IS:

1) the ontology that includes the concepts and relations of the subject domain; from the point of view of the text analysis, the ontology describes data to be extracted from texts and put in the database of the system;
2) ontological concepts (classes) required for representing text resources;
3) the information content of the system, or its database.

3. Linguistic ontology

A linguistic ontology includes all kinds of knowledge that are necessary for the text analysis:

1. Vocabulary or the list of minimal units of the language that are used for the description of information significant for IS.

2. The linguistic knowledge specific for a given language: morphological classes, rules of collocations of terms, etc. T his knowledge can be specified with a different degree of details (it depends on requirements to IS and possibilities of its developers).
3. The knowledge about the coordination of the linguistic knowledge with the subject domain knowledge defined in the ontology of IS. To define this coordination, the terms are supplied with semantic types, which secure its correspondence with the ontology elements either directly, or according to a certain scheme.
4. Descriptions of the genre structures of texts that are specified for all types of text resources of IS.

Formally, the linguistic ontology is a union of five elements $\langle$O, T, F, D, R$_{DF}\rangle$, where:

– O is the ontology of the subject and problem domains of IS,

– T is the vocabulary of terms,

– F is a set of ordered lists of schemes of facts (the order in the list defines the sequence of application of schemes of facts during the analysis),

– D is a set of document models (text resources); a proper set of schemes of facts is defined for each model,

– R$_{DF} \subseteq$ D$\times$F is an incidence relation defined between a set of models D and the set of ordered lists of schemes of facts F.

Let us consider the components of the linguistic ontology in detail.

3.1. The vocabulary of terms

The vocabulary of terms is T = $\langle$E, M, S, G$\rangle$, where

– E is a set of terms,

– M = $\{$M$_1$, ..., M$_m\}$ is a set of morphological types,

– S is a set of semantic types,

– G = $\{$G$_1$, ..., G$_n\}$, where G$_i \subseteq$ E is a set of subsets of the terms forming groups of synonyms.

The morphological information allows us to carry out the morphological analysis of word-forms of the input text and to extract terms-lexemes and terms-collocations. The following concepts are used:

1) a morphological attribute;

2) a part of speech;

3) a type of a paradigm;

4) a morphological type (class).

The morphological attribute $a_i = \langle l_i, V_i \rangle$ is represented by its name l_i and the set of values of this attribute V_i (for example, l_i = gender, V_i = {feminine, masculine, neuter}). A part of speech is also an attribute that should always be present in the description of a morphological type.

The type of a paradigm P defines the order and the description of possible endings in changeable lexemes. All paradigms (sets of the endings) of one type have the same length, which allows us to store endings in one table. The endings in paradigms are ordered so that each sequence number in a paradigm (a table column) corresponds to the endings of lexemes with the identical set of values of morphological attributes. For example, inflectional attributes of a regular adjective are a case, number, and gender.

The ordering of paradigms allows us to use a compact form of a record in the form of tree structures. A node of such a structure is a subset of a set of values of attributes $\langle a_i, V_i‘ \rangle$, where $V_i‘ \in V_i$. A pair of functions f: $n \rightarrow v_{i1} \times \ldots \times v_{ik}$, g: $v_{i1} \times \ldots \times v_{ik} \rightarrow n$ for any tree structure provides a transformation of the number of endings in the paradigm n to a set of values of attributes $\{v_{i1}, \ldots, v_{ik}\}$, and conversely. Thus, each changeable lexeme is linked to some paradigm from the table of paradigms, and each paradigm is associated with some type of a paradigm describing its structure.

A morphological type $M_i = \langle a_{ps}, AV_i, P \rangle$ includes a part of speech a_{ps}, a set of morphological attributes of the lexeme $AV_i = \{\langle a_1, v_{1j} \rangle, \ldots, \langle a_k, v_{kj} \rangle\}$ and a type of the paradigm P describing the attributes of word-forms of the lexeme.

Based on the morphological attributes, agreement of the terms occurring in the text can be performed. The syntactic component, which uses a rich store of knowledge about the language, solves this problem. In this paper, we do not consider this component in detail.

The set of semantic attributes $\{S_1, \ldots, S_m\}$, $S_j \in S$ can be associated with the term x_i of the dictionary. Each semantic type is linked to some element of the ontology (attributes, concepts or relations). The association of the term with the semantic type makes it possible to discover which element of the ontology is described by the term in the text.

The group of synonyms allows one semantic value of a subject domain (for example, the domain value) to be linked to the set of alternative terms which can be used for representation of this value in the text.

3.2. Facts description

When considering the problem of coordination of a linguistic knowledge with the ontology of IS, we note that an ontology element is seldom unambiguously described by one term, collocation or other dictionary unit. Such a description occurs very often throughout the text.

To understand which element of the ontology is presented in the text, it is necessary to find the fact that obviously defines the class or attribute of the concept/relation. On the other hand, the fact is a natural language expression with a certain syntactic structure consisting of the dictionary units.

One of the ways of solving the problem of the text analysis is a declarative description of the structure of significant facts and their connections with the concepts of the ontology. Such a description is named *the scheme of the fact*.

The scheme of the fact F_k is a triad ⟨Arg, Res, C⟩, where

- Arg is a set of arguments of the fact, where an argument can be a semantic type of the vocabulary, a concept/relation of the ontology, a type of the fact or an information object of the document whose text is being analyzed.
- Res = ⟨t, op (t), P⟩ is a result of application of the scheme of the fact, where
  - t is a type of element (the class of the resulting object);
  - op (t) is a type of the operation (creation or editing of the object) that is applied when all restrictions C defined in the scheme of the fact are complied.
  - P is a set of rules for generation of the values of attributes of the resulting object. Each rule links an attribute of the resulting object with one of the following elements:
    - the value of the attribute of one of the arguments,
    - a default value,
    - an expert value given in the table of semantic restrictions.

  To apply the scheme of the fact, its arguments should satisfy the set of restrictions on compatibility of arguments. Semantic and structural restrictions are distinguished:
- C = ⟨Sem, St⟩, where Sem is a table of semantic restrictions, St is a set of structural restrictions. (These restrictions are described in detail below.)

Let us give some examples of the schemes of facts:
F1: Research-Object (monument) + Locality (Western Sahara) =>
creation Object-is-found-in (monument, Western Sahara);
F2: Activity (work) + Object (construction project) =>
creation Function (work, construction project, Activity: construction);
F3: Sender (Organization) + Function. Activity =>
editing Document (Activity: Function. Activity).

Semantic restrictions

A semantic restriction is imposed on the semantic characteristics of arguments of a fact. In the proposed approach, semantic restrictions are defined by a table where each line represents compatibility of the semantic characteristics of arguments of a fact and contains additional semantic information which is used further for generation of information objects corresponding to the facts found in a text.

Let us consider a common scheme of the table of semantic restrictions for binary facts (see Table 1).

Table 1. The table of semantic restrictions for binary facts

Characteristics of compatibility						**Characteristics of refinement**								
1st argument			2nd argument			1st argument			2nd argument			Result		
S_1	…			…	S_k	S'_{k+1}	…			…			…	S'_n

The table of semantic restrictions Sem defines n-ary relation on k sets of semantic characteristics S_1, S_2, ..., S_k and (n–k) sets of additional characteristics S'_{k+1}, S'_{k+2}, ..., S'_n. For each column i | 1$\leqslant$ i $\leqslant$ k (combinatory characteristics), a comparison operation Θi: $S_i \times S_i \rightarrow$\{true, false\} is defined. It allows us to determine whether the value of the characteristic specified in the table corresponds to the value of the characteristic of an argument of the fact. This information, in particular, allows us to use hierarchical relations for comparison of the characteristics which are semantic classes.

Thus, the characteristics of compatibility contain the conditions which should satisfy the parameters of arguments of the fact. Characteristics of refinement contain the values allowing us to refine the objects forming the fact or to generate the information object corresponding to the fact found in the text.

Below we can see a little fragment of the table of semantic combinations:

Work (class) + Construction_project (class) => Work: construction.

"Development" (term) + Natural_resources (class) => Work: nature management.

"Development" (term) + Document (class) => Work: document creation.

Structural restrictions

In addition to the semantic restrictions, restrictions of other language levels, such as syntactic and genre restrictions, should be considered.

For each scheme of the fact, additional conditions on its arguments should be given:

- the condition on a segment, i.e. what type of a segment the arguments should be situated in;
- the position of arguments in the text (contact position, pre- and post-position, priority of positions);
- syntactic conditions (valencies of terms, prepositional phrases, etc.);
- rules of combining (homogeneity, projectivity or maximal connectivity rules).

Verification of syntactic compatibility may involve a simple comparison of syntactic features of terms or construction of a local syntactic dependency tree.

Consider an example of a scheme of the fact with structural restrictions:
Fact (a1:Work, a2:Object)
– condition on a Sentence segment;
– to check valencies of terms of Work class;
– to check syntactic compatibility;
– search for homogeneous terms;
– to conform to the projectivity rule;
– to give priority to the postposition of Object terms relative to Work terms.

If this scheme will be applied to the sentence

"It takes about 2 months to complete the installation ⟨1⟩ of equipment ⟨2⟩ and systems of automatics ⟨3⟩ in view of the necessary field change ⟨4⟩, carrying out of production tests ⟨5⟩ a nd preparations for shipment ⟨6⟩ of the 2-nd diesel power station ⟨7⟩".

then the following facts will be extracted:
⟨1⟩ [installation] – ⟨2⟩ [equipment]
⟨1⟩ [installation] – ⟨3⟩ [systems of automatics at]
⟨4⟩ [field change] – ⟨2⟩ [equipment]
⟨4⟩ [field change] – ⟨3⟩ [systems of automatics]
⟨5⟩ [production tests] – ⟨2⟩ [equipment]
⟨5⟩ [production tests] – ⟨3⟩ [systems of automatics]
⟨5⟩ [production tests] – ⟨7⟩ [power station]

3.3. Formal structure of the text

To solve the problem of comprehensive analysis of documents, the system should possess a function of segmentation of the text. It will provide an expert with the opportunity to work with the text structure inhered in the document.

Each text resource (document), depending on its type or genre, has a certain structure of the text that can be formally represented by the hierarchy of segments.

The text has, at least, three levels of the formal structure — physical, logical and genre. The first one depends on format of the document (text, html, etc.) and provides a representation of the text on a page, for example, using tags or tables of styles. The second level can include "polygraphic" elements, such as paragraph, line, and sentence. The third level is presented by decomposition of the text into genre parts. For example, the text of a business letter has the following genre parts: heading (the sender, addressee, resume and treatment), the basic part (the text of the letter, notes and enclosures) and the signature. The genre parts of the text are characterized by a certain lexicon and have a certain structural organization (the structure and position with respect to other parts), and they are realized within the limits of formal segments of other levels.

Segments are described by markers. A marker is a list of alternative elements $\mathrm{m}_i = \{\mathrm{m}_{i1}, \ldots, \mathrm{m}_{in}\}$, where an element m_{ij} can be a symbol, a dictionary object (a term) or a segment of another type. Markers can set the beginning and the end of a segment either explicitly or implicitly — using boundaries of another formal segment. Joint use of markers of the second and third type allows us to determine a condition on the presence of specific objects inside a segment.

The construction of a segment is based on the following restrictions:

- *single* — the segment should not intersect other segments of the same type; a special case of this restriction is the absence of insertions;
- *min* — choosing the minimal possible segment;
- *max* — choosing the maximal possible segment.

Other types of restrictions are also possible.

Thus, the segment s is a triad $\langle \mathrm{N}_S, \mathrm{Mark}, \mathrm{pr}\rangle$, where N_S is a unique name of the segment, $\mathrm{Mark} = \{\mathrm{m}_1, \ldots, \mathrm{m}_k\}$ is a set of markers, $\mathrm{pr}\in\{$*single*, *min*, *max*$\}$ is a restriction.

The model of the document is a triad $\mathrm{D}_i = \langle \mathrm{Seg}, \mathrm{R}_S, \mathrm{R}_I\rangle$, where

- $\mathrm{Seg} = \{\mathrm{s}_1, \ldots, \mathrm{s}_m\}$ is a set of segments,
- $\mathrm{R}_S \subseteq \mathrm{Seg}\times\mathrm{Seg}$ is the antisymmetric, transitive, irreflexive relation of consequence setting of the partial order on set the S,

– $R_I \subseteq Seg \times Seg$ is the antisymmetric, transitive, irreflexive inclusion relation setting of the partial order on the set S.

The set of models of documents developed by experts can be used for automatic identification of the genre of a document.

4. Technology of the text analysis

The proposed system of knowledge provides a transparent scheme of the text analysis, when a consecutive stage-by-stage analysis of the text is performed, however each subsequent stage can eliminate ambiguity of the results of the previous stages.

The system architecture (see Figure 1) includes a set of editors, dictionaries, executing modules and application programming interface (API).

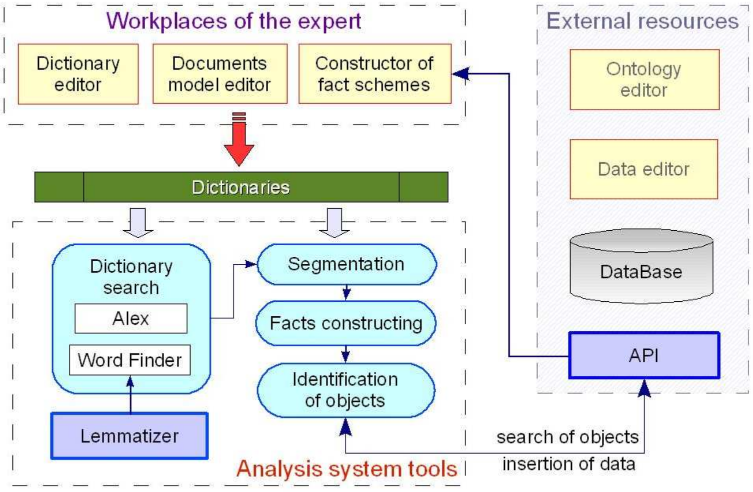

Figure 1. The architecture of the system of analysis

The editors allow us to describe knowledge and to form the linguistic ontology. The ontology and data editors are external editors. The system also has and its own (internal) editors: the dictionary editor, document model editor and the constructor of fact schemes. Editors hold knowledge either in the dictionaries or DB. Dictionaries are used by the executing modules which perform the step-by-step analysis of the text.

The important requirement for implementation of this technology is its independence of a concrete DB (it mainly concerns the module of identification of objects and the constructor of fact schemes). For this purpose the program interface (API) to a DB was developed. It is implemented as a separate module that provides data access in terms of ontology.

4.1. Segmentation

There are two kinds of text segmentation — primary and genre ones.

During the primary segmentation, the linear representation of a text is splitted into an ordered list of the string objects which are used for forming segments. The genre segmentation is performed after the lexical analysis. It is based on lexical objects that mark different genre segments.

The mechanism of segmentation is realized by the Alex system [7] included in the dictionary component of technology.

4.2. The lexical analysis

The lexical analysis performs extraction of lexical objects from the set of ordered string objects obtained by the primary segmentation of the text. A lexical object is either a lexical pattern described in the Alex system, or a word/phrase represented in the dictionary.

The tasks of the given stage are the following:

- application of lexical patterns;
- execution of the morphological analysis and phrase search;
- identification of genre segments.

The process results in an ordered list of objects with the following set of parameters: the name (the canonical form of a word or collocation, or the name of a pattern), position in the text, value (the main word in a synonymic group or the numerical value), grammatical class (morphosyntactic information about the word form), semantical class, statistical characteristics.

4.3. Facts constructing

The mechanism of facts constructing is based on a preliminary planning which is performed for each class of documents on the basis of the prespecified schemes of facts.

The tasks of planning are as follows:

- Generation of executed rules on the basis of schemes of facts.
- Organization of a queue of rules to be executed.
- Maintenance of correctness and convergence of the process of constructing fact.

When the queue is organized, the two aspects are taken into account:

- interdependence between the schemes of facts and the order of creation of objects;
- the order of segments and their nesting level (the analysis is carried out starting with the smallest segment in the nesting hierarchy and proceeded up to the largest one).

During the document processing, the rules are successively taken from the queue and applied. This process goes on until the queue becomes empty. For each rule the data are grouped around the segments specified in the condition of a rule. Extraction of the facts is limited by the frameworks of one segment.

The table of semantic combinations and syntactic rules for checking the compatibility of grammatical characteristics of terms and controlling the coordination, projectivity and connectivity is also used for fact constructing.

For a list of lexical objects obtained after lexical analysis, the appropriate combinations are selected from the table of semantic combinations. These combinations are further considered as separate schemes of facts (however, the syntactic rules are to be applied as well).

The closely adjacent objects of the same class are combined in one group. After that the contact groups are checked for semantic and syntactic compatibility.

All the methods use the same approach to disambiguation based on the use of weight of terms and objects. The weight depends on such factors as:

- whether the term is a part of a collocation,
- whether the term is a constituent of a fact,
- statistical characteristics,
- compatibility of adjacent terms, etc.

4.4. Identification of information objects

Further processing consists in forming the content of the document. For this purpose it is necessary to identify the obtained objects and provide their correct insertion into the information space of the system.

The tasks at this stage are as follows:

- Reconstruction of objects with complex structural names by means of a “part-whole” hierarchy determined in a database;
- Reference resolution (identical objects are integrated);
- Search in a database of IS for the objects that are similar to objects obtained from the text of the document;

- Disambiguation, in case when the database includes several objects the description (content) of which corresponds to the obtained object.

The object is considered as identified if its class and a set of its key attributes are defined. This property allows us to distinguish the obtained object from other objects, i.e. uniqueness of objects in a database of the system is ensured.

The set of unambiguously identified objects forms the content of the document. Uniqueness of objects in the content provides its correct insertion into the database of IS.

5. Conclusion

The proposed approach is substantially based on the ideas presented in [4]. In particular, we exploited the idea of a collaborative use of the subject domain ontology and thesaurus, as well as the methods of semantically oriented analysis of a text. In the course of practical implementation of the proposed approach, the methods and algorithms developed for an experimental system for extraction of information from weather forecast telegrams [8] and industrial intelligent document management system InDoc [9] were also used.

Our immediate goals are to complete the development of technology based on the proposed approach and to apply it to solution of a laborious problem concerned with providing the knowledge portal on computational linguistics with new knowledge and data [10].

References

[1] Ageev M., Dobrov B., Loukachevitch N. Sociopolitical Thesaurus in Concept-Based Information Retrieval: Ad-hoc and Domain Specific Tasks // Proc. of the 6th Workshop of the Cross-language Evaluation Forum (CLEF'2005), Vienna, Austria, September, 2005.
`http://www.clef-campaign.org/2005/working_notes/workingnotes2005/ageev05.pdf`

[2] Guarino, N. Some Ontological Principles for Designing Upper Level Lexical Resources // Proc. of the First Internat. Conf. on Lexical Resources and Evaluation, Granada, Spain, May 1998.

[3] WordNet. An electronic lexical database / Ed. by Ch. Fellbaum. — MIT Press, 1998. — 445 p.

[4] Narin'yani A.S. TEON-2: from Thesaurus to Ontology and Backwards // Proc. of Internat. Workshop on Computational Linguistics and Intellectual Technologies (Dialogue'2002). — Moscow: Science, 2002. — Vol. 1. — P. 307–313. (In Russian)

[5] Nédellec C., Nazarenko A. Ontology and Information Extraction: A Necessary Symbiosis // Ontology Learning from Text: Methods, Evaluation and Applications / Ed. by P. Buitelaar, P. Cimiano, B. Magnini. — IOS Press Publication, July 2005.

[6] Guarino N. Formal Ontology and Information Systems // Formal Ontology in Information Systems / Ed. by N. Guarino. — Proc. of FOIS'98, Trento, Italy, 6-8 June 1998. — Amsterdam: IOS Press. — P. 3–15.

[7] Zhigalov V., Zhukov A., Kononenko I. et al. ALEX — A System for Multi-Purpose Automatized Text Processing // Proc. of Internat. Workshop on Computational Linguistics and Intellectual Technologies (Dialogue'2002). — Moscow: Science, 2002. — Vol. 2. — P.192–208. (In Russian)

[8] Kononenko I., Kononenko S., Popov I. et al. Information Extraction from Non-Segmented Text (on the material of weather forecast telegrams) // Proc. of Internat. Conf. on Content-Based Multimedia Information Access (RIAO'2000). — 2000. — Vol. 2. — P. 1069–1088.

[9] Zagorulko Yu., Kononenko I., Sidorova E. A Knowledge-Based Approach to Intelligent Document Management // Proc. of 7th Internat. Workshop on Computer Science and Information Technologies (CSIT'2005). — Ufa-Assy, 2005. — Vol. 1. — P. 33–38.

[10] Borovikova O., Bulgakov S., Zagorulko Y. et al. Ontology-Based Approach to Development of Adjustable Knowledge Internet Portal for Support of Research Activity // Bull. of NCC. Ser. Computer Science. — 2005. — Iss. 23. — P. 45–56.