Monitoring Online Publications about Public Authorities Activity by means of Ontological Semantics

A. Dobrov
St. Petersburg State University, ITMO University, AIIRE company
Saint-Petersburg, 199226
Korablestroiteley st. 16/2
+7 960 232 15 03
adobrov@aiire.org

A. Dobrova
AIIRE company
Saint-Petersburg, 199226
Korablestroiteley st. 16/2
+7 931 224 07 17
adobrova@aiire.org

N. Soms
AIIRE company
Saint-Petersburg, 199226
Korablestroiteley st. 16/2
+7 911 251 44 90
nsoms@aiire.org

ABSTRACT
In this paper, an attempt is made to describe an on-going research, aimed at creating a tool of content-analysis and opinion-mining, based on monitoring online publications about Russian federal pubic authorities activity by means of ontological semantics. This research continues a series of studies focused on the 'agenda' that is formed in the media on topics related to the development of e-government and online services. The problem of linguistic ambiguity is partially solved by semantical restrictions imposed by conceptual relations specified in the ontology, thus increasing precision of the analysis. Recall is also increased by means of conceptual hierarchies and synonymy. These methods have allowed to perform automatic monitoring of online publications that refer anyhow to specific public authorities, to choose a set of publications that contain evaluations of their activity, and to make a marked-up corpus from this collection. Nevertheless, the currently developed corpus shows that, in most cases, the evaluations of actions of public authorities are expressed by very sophisticated linguistic techniques, and a significant development of existing technologies of computer linguistic analysis is needed.

Categories and Subject Descriptors
I.2.4 [**Ontologies**]

General Terms
Algorithms, Human Factors, Languages.

Keywords
Ontological semantics, NLU, conceptual relations, lexical disambiguation.

1.INTRODUCTION
The AIIRE Ltd., in close collaboration with ITMO University Center for Electronic Government, are developing machine understanding technologies to analyze news items about different kinds of Russian public authorities activity with help of ontological semantics. AIIRE is a free open source natural language processor, developed by a team of researchers in Saint-Petersburg, Russia. The team was formed in 2003-2005, in the Laboratory for Informational Linguistic Technologies of the Institute of Linguistic Studies in Saint-Petersburg, Russia, and continues its work as AIIRE company. The 'AIIRE' acronym stands for 'Artificial Intelligence-based Information Retrieval Engine', which was the first production-level application of the developed NLU-kernel. This article describes the research that continues a series of studies focused on the `agenda' that is formed in the media on topics related to the development of e-government and online services, as well as on identifying indicators that reflect the level of discussion of implementing e-government services on the Internet, including social media and blogosphere [1], [2]. During the research it was found that the traditional methods of content analysis, based on key words, may be insufficient [3], [4] to identify the above-mentioned indicators. In order to solve this problem, an attempt was made to use the AIIRE system, which allowed to apply methods of ontological semantics to content analysis tasks. The pilot study has proven that a substantial increase of precision and recall of content analysis can be made by means of ontological semantics, but it needs a profound elaboration of ontology, including creation of different concepts for specific public authorities, their areas of public administration, and for specific administrative units of Russian Federation. This paper describes a part of this research, which aims at implementing machine-driven monitoring of online publications about Russian federal public authorities on the basis of ontological semantics.

2.The task of monitoring the news about the activity of public authorities
The practical need for analytical tools that help to get information about the public reaction to the different actions of public authorities has significantly increased in recent years. Social surveys show less and less part of the real picture and make it impossible to identify the sources of the observed changes, while their scale and speed are becoming less predictable. The amount of politicized texts on the Internet has considerably grown. Today, automation of public opinion studies, based on analysis of online publications, do not provide the required complete and accurate results. This study is aimed at improving the existing methods and tools for analyzing the dynamics of media discourse. Like many other developers the AIIRE team tries to resolve many of the news flow monitoring problems with help of the team's original solutions:

1. Algorithms of ontological semantics are used instead of statistical machine-learning algorithms for lexical disambiguation, which allow to improve the recall and

Conference'10, Month 1–2, 2010, City, State, Country.

precision of the analysis [3], [5]. In order to evaluate textual semantics and to map it to the ontology, a complex linguistic analysis of online publications is performed, including full-scale syntax parsing and evaluation of semantic graphs, as described in [7];

2. Online publications are filtered from information noise (advertisement, links to other publications, irrelevant textual inclusions, etc.), as it is possible, and that is why each text can be accurately processed further. Thus, the text can have the same formatting as the list of the most discussed topics that do not have any relation to the text. If the monitor does not distinguish between the text and the list then the publication can be attributed to incorrect topic [3];
3. Every news item has its own "weight" and the number "coverage factors", which are presented on the original source page. These `coverage factors' are values that can reflect the extent to which a publication covers the online audience (amount of views, comments, reposts, likes or dislikes, retweets, etc.) Each coverage factor has its own weight, which is evaluated empirically in order to calculate the weighted harmonic mean of the coverage factors for each publication [3].
4. Due to use of ontologies, our monitoring system recognizes such relations as synonymy, hyponymy meronymy etc. These relations are used to increase recall of monitoring, e.g., to collect publications about all subclasses of a particular class of public authorities, and also to recognize various names of the same authorities that can be used in different texts. The way that semantic and ontological relations are represented in AIIRE ontology is described in [6].

3.Course of Work

3.1.The Ontology: lexical disambiguation for improving precision

Ambiguity is a major problem, due to which any kind of analysis of text may be inaccurate. The main means for disambiguation in AIIRE is its ontology. Grammatical restrictions sometimes help to reduce ambiguities on the level of morphology, but grammar rarely helps to get rid of homonymy (unfortunately, it does not help when it comes to specific vocabulary related to the public authorities), and never helps to choose between word meanings. E.g., Russian word *Кремль* (Kremlin) often means Russian government, although its main meaning is 'fortress'. Neither morphology, nor syntax helps machine to choose between these two meanings. Attempts to remove lexical ambiguity by analyzing statistics of compatibility of different meanings in textual corpora, for the time being, unfortunately, do not increase precision of the analysis because of their own errors. The main problem of these methods, when it comes to NLU tasks, is that they never guarantee absence of false-negative results (correct hypotheses can be rejected due to statistics), which are inadmissible, because they lead to loss of whole syntactic structures or semantic graphs. The same is true, e.g. for punctuation checkers (cf. [9]).

In order to make precise inference, e.g., that the word *Kremlin* means the government, machine must take the whole context into account, and find those specific concepts in the context, that logically contradict with some attributes of the concept of 'fortress'. E.g., in the phrase *Кремль решил…* ('Kremlin decided to…') such inference can be made due to inability of inanimate objects (including fortresses) to make their own decisions, but this information can be taken only from the ontology.

It should be clarified that the term 'ontology' here refers to the database that contains concepts that stand for lexical meanings, and their attributes, which are formed by relations to other concepts. It is not a semantic dictionary, as in [10], because ontology is not principally language-dependent (concepts can be shared by different languages), and because it contains many concepts that are not meanings of any particular word (e.g., the concept of 'inanimate object'). It is also not a thesaurus like WordNet (cf.[11], [12]), because thesauri contain only word-to-word (or meaning-to-meaning) relations, but not, e.g., object-to-action relations like 'can perform action', which are driven by extralinguistic knowledge (e.g., grammar does not help us to know that governments can make decisions) and are crucial for lexical disambiguation.

Ontology-driven disambiguation has already substantially increased precision of automatic classification of news items [5], which was proved by a series of experiments with independent auditors performing manual classification of the same texts. It must be stressed here, that these experiments have proved that traditional way of evaluating precision of text classification has proved to be substantially inaccurate because of great dispersion in human estimations. For this particular reason, each text was classified (in average) by 10 different auditors (of different genders, occupations, and ages), according to requirements of a psycholinguistic experiment [8]. It was also shown (cf. [5]), that the dispersion in human estimations significantly decreases when it comes to attributing texts to objective subject domains (like 'sports') or to specific objects (like 'government'), in comparison with less definite topics like 'in the world' or 'society', which may apply to almost any publication. Thus, one of the main assumptions in this on-going research is that ontology will be able to provide close to 100% precision.

3.2.The Hierarchy of Public Authorities

Firstly, in order to prepare our ontology to work with a specific topic, the «list of Federal Public Authorities of Russian Federation» was made. This list was gathered from free sources on the Internet, including government sites, and then incorporated into the AIIRE ontology as a set of its concepts. New types of ontological relations were introduced in order to reflect particular relationships between different public authorities and their activities. The hierarchy of ministries and their subordinate agencies was created and completely elaborated using the "to be an immediate part of organization" relation and the reverse relation in some cases ('to immediately contain the organization'). This hierarchy was included into AIIRE ontology, so that each concept of the hierarchy was involved into various ontological relations with other "common" concepts [6].

AIIRE ontology has more relations, and of course it includes equivalence (synonymy) and inheritance (hypernymy, subclass relations). For example, the concept of 'ministry' has such interim subclasses as 'ministry of a specific competence' and 'ministry of a specific country'. These subclasses were required due to the presence of two different bases for the classification of ministries (competence and country in this case). Subclasses of the `Ministry of a specific country' concept have been created on a geographical basis — these are 'Ministry of America', 'Ministry of Australia', 'Ministry of Eurasia' and 'Ministry of Oceania'. The geographical position of the country (mainland) is known for these classes, but the area of public administration, which falls within the competence of the specific Ministry, remains unknown. These subclasses are further classified down to the lower classes and to the instances with both area of public administration, and territorial unit known, such as: 'Ministry of the Caucasus' — `Ministry of the Nagorno-Karabakh Republic' — `The Ministry of

Health of the Nagorno-Karabakh Republic'. In addition, the concept of 'Ministry of a specific country' is connected with 'the government of a specific country' with 'to be an immediate part of organization' relation, which means that the parallel classification is carried out on the basis of countries and regions for the concept of 'government'.

Classifications in the AIIRE ontology often reproduce the structure of each other, in accordance with the underlying rules of inference, and can be `multiplied' when there are several bases of classification (this phenomenon is described in detail in [6]). The classification of public authorities proved to be parallel both to the hierarchy of territorial units (regions), and to the hierarchy of public administration areas. As a result, intermediate classes, like `Ministry of Health of a specific country' or 'Russian Ministry of a specific competence' had to be created. The latter has two superclasses — 'Ministry of Russia' and 'Ministry of a specific country and specific competence'. Among the subclasses of this class, seventeen concepts for specific ministries can be seen, that were created and specified according to the requirements of the ontology. Thus, the concept of 'Ministry of Justice of Russia' is an instance of several classes at the same time: 'Ministry of Justice of an Asian Country', 'Ministry of Justice of a European Country', 'Ministry of a specific competence of Russia' and `Russian law enforcement authority'. This instance inherits relations with various other concepts from these classes, allowing AIIRE Natural Language Processor to build additional hypotheses about the way a text relates to them.

The concept of 'Ministry of Justice' is linked with the concept of 'justice (judicial activities)' with the 'to carry out activities' relation; it is also linked with the concept of 'judicial authorities' by means of the 'to be an immediate part of organization' relation, and with the concepts of 'Federal Penitentiary Service' and 'Federal Bailiffs Service' by means of the reverse relation ('to immediately contain the organization'). In addition, the 'Ministry of Justice of Russia' is `a typical representative' of an abstract 'Ministry of Justice' — this allows AIIRE Natural Language Processor to 'understand' that, in a Russian-language text, in spite of formal logic, the phrase «*the Ministry of Justice*» is likely to mean 'the Ministry of Justice of Russia' and not 'a Ministry Justice of any country'.

It should be clarified that there are certain rules of interpretation of the 'to be an immediate part of organization' its reverse relation ('to immediately contain the organization'). These relations imply that, for example, the Ministry of Justice is included in the judiciary authorities, and the reverse relation means that the Ministry of Justice holds both the Federal Penitentiary Service and the Federal Marshals Service under its jurisdiction. This relation is a particular case (a subclass) of a transitive 'to belong to' relation, which allows the linguistic processor to refer the information to the purview of the Ministry of Justice, even if only some of its subordinate offices is mentioned in the text.

3.3.100% recall?

Hyponymy relation (relation to subclasses) allows to attribute to correct concepts those texts, where only their subspecies are mentioned. E.g., a publication may contain only *МВД* word form ('Ministry of Internal Affairs'), and hyponymy relation allows to attribute this text to the abstract 'ministry' concept. Other systems, when asked to monitor publications about ministries, actually search for forms of the word 'ministry', and therefore do not find particular ministries, when denoted by abbreviations. This leads to significant loss of recall, which is absolutely remedied by ontology. It must be mentioned here, however, that the cost of recall increase in this case is a similar increase of ambiguity, e.g., Russian *MO* abbreviation has a lot of meanings, including 'Moscow district', 'Moscow region', 'Ministry of Defense', and 'Ministry of Education'. However, as it was mentioned above, ontology often helps to remove the ambiguity in these cases.

There is another case when ontology is necessary for improving recall: the same concepts can be denoted by many different natural language expressions (which are called synonyms). E.g., the concept of 'government' can be denoted by Russian words *правительство*, *кабмин*, and *кабинет*, and also by *кабинет министров* expression. These denotations may be treated as referring to different concepts, because they seem to have different significative parts. Nevertheless, they definitely refer to the same object in the real world, and, therefore, have the same denotation. Concepts that share the same denotation are called equivalent, and are linked with each other by 'equivalence' relation. This relation forms so-called 'synsets', like it is done in the WordNet Thesaurus (see above). Each synset is represented by an arbitrary concept, which is called a 'synset root'. During the natural language processing, each concept in the semantic graph is replaced by the synset root, so that different concepts of the same synset have the same representation. Thus, different synonymic natural language expressions are treated uniformly, and concepts are found in texts regardless of specific formulations. Other systems sometimes, but rarely, are able to take some (but not all) synonyms into consideration, but it often leads to 'funny' mistakes (e.g., Google in 2009 was retrieving the words *bush* and *bill* by *~president* query, cf. [13]). These mistakes were made because of inaccuracy of synonymy relations assignment, and, certainly, because of ambiguity: Bush is a surname of two ex-presidents of USA, so, at least, Bush is not a synonym, but a hyponym of 'president'; however, when written in lowercase, it means only 'shrub'. Probably, in order to increase precision, Google has removed these 'synonyms', but now it finds only literal results (*president* word forms). This solution does increase precision, but actually there is no gain in recall, as there are, in fact, some synonyms of *president* (*head*, *chief*, *leader*), that are never retrieved, which is unacceptable in case of monitoring, when the user can not reformulate his or her query.

Thus, in this study an assumption is made that ontology will be able to provide close to 100% recall (exact results, as for precision, will be published after the appropriate series of experiments).

3.4.Monitoring (resources, topics, coverage)

The previously developed system of monitoring and rating everyday audience coverage of topics [3] is used in this study. This system allows to gather publications simultaneously from different web-resources and keep them up-to-date, in a semi-structured form: text is filtered from information noise and is stored separately from metadata, including titles, references to information sources, dates, and the above-mentioned coverage factors. A separate 'monitor' module is made for each website, so that specific features of its current layout can be taken into account and processed individually. The majority of websites provide RSS-feeds of their publications, so, in the majority of cases, the main problem is to reveal the particular way the information is presented to the user on a publication page. The rules of full-text and meta-data extraction are written mostly as XPath-expressions, when possible, but sometimes some extra processing is needed (e.g., to get rid of information noise, or to load some information asynchronously, which is often the case when working with coverage factors). Some websites (especially, forums) do not provide RSS or other feeds or other APIs to get lists of latest publications, so making a monitor module for such websites is harder (these monitors are called HTML feed

monitors, because not only the publications, but also the feeds are actually extracted by means of parsing HTML). However, only one website (http://politikforum.ru) actually needed an HTML feed monitor in this study.

Different ‘sections’ of the same website (e.g., different weblogs that are maintained on the same platform) do not need separate modules to extract publications and their content, but there is a certain set of settings that is necessary for sections.

The list of information resources to monitor was built together with the ITMO University Center for Electronic Government on the basis of the previously created and working set of monitor modules of AIIRE system. The list was extended with a set of resources that are most likely to contain publications with evaluative judgments about different actions or activities of Russian federal public authorities.

In order to perform the pilot study, and to gather a textual collection for further markup, topics were created for each public authority, according to their hierarchy and sets of synonymic denotations registered in the ontology (the list was verified again by theITMO University Center for Electronic Government).

The monitoring system has been working continuously for about three weeks, and have collected 51622 publications, so current results are, at most, preliminary, but they show already, that there is a significant dispersion in coverage of different topics. E.g., the total coverage measure of *Правительство* (government) topic is 10,541.33, whereas *ФСТЭК* (The Federal Service for Technical and Export Regulation) has only 0.05 so far. Current values of total coverage are represented in Table 1.

Table 1. Current total coverage values for different Russian public authorities (top 50 topics)

Authority name (in Russian)	Total audience coverage
правительство	10541,33
Госдума	3727,67
МИД	3280,30
МВД	3020,61
спецслужба	2283,18
разведка	1751,68
ФСБ	1662,90
Правительство	1495,74
Министерство финансов	1020,94
Минфин	922,22
ФСИН	637,83
МЧС	609,69
Кремль	568,67
Правительство РФ	490,97
Минобороны	408,23
Минздрав	400,56
Министерство обороны	352,83
надзор	288,71
Минэнерго	229,08
налоговая служба	224,10
ФСКН	207,70
Министерство иностранных дел	135,31
МЭР	131,05
ФМС	119,43
ФАС	116,07
Минкульт	88,40
Минсельхоз	87,59
Роскомнадзор	81,23
Росстат	70,29
Государственная Дума	66,38
Минтруд	64,54
законодательная власть	62,91
ФНС	52,83
Минюст	48,49
Министерство юстиции	46,69
Роспотребнадзор	38,34
федеральная таможенная служба	32,41
Министерство экономического развития	30,82
Министерство здравоохранения	28,88
Министерство внутренних дел	28,27
миграционная служба	26,88
Минсвязи	26,44
Минтранс	24,01
силовые структуры	22,64
Министерство культуры	22,08
исполнительная власть	20,97
Министерство образования	16,36
федеральная служба безопасности	15,72
Федеральное агентство связи	15,58

Минпромторг	15,20

Current results correspond well with the well-known Zipf law (cf. Figure 1).

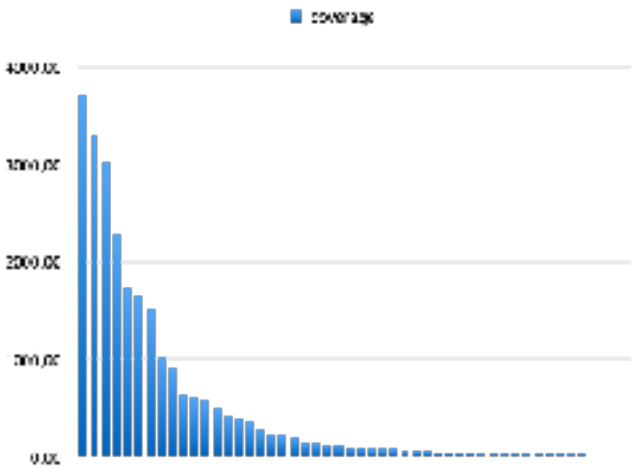

Figure 1. Current distribution of total coverage values

4.Opinion Mining Perspective

Public reaction to the activities of public authorities is primarily manifested in different evaluations of certain actions formulated in on-line publications. The ultimate goal of this work is to create a quality Opinion Mining tool, which should allow to trace emergence of certain evaluative judgments about the activities of public authorities operatively.

Publications that contain aforementioned evaluative judgments were identified in the existing collection of texts (so far, 50 of 51622 texts) by experts of ITMO University Center for Electronic Government, who were working with texts automatically attributed to topics corresponding to specific public authorities by the monitoring system. Special markup has been designed to mark evaluative judgments in the publications. This markup has allowed to mark the public authorities, their actions or activities, and evaluations of these actions separately.

E.g., the statement *Политика МинФина стала агрессивной* (The Policy of the Ministry of Finances became aggressive) is marked up as follows: <es><a> *Политика* <d> *МинФина* </ d> *стала* <e> *агрессивной* </e></es>, where <*es*> tag marks the whole evaluative statement, <a> tag marks the evaluated activity (the whole *policy* in this case), <d> tag stands for department (the *Ministry of Finances*), and <e> tag contains the evaluative word or expression (*aggressive*).

The example above is, however, a very rare case of direct evaluation of a specific kind of activity of a specific public authority. In the great majority of cases, some of these three components are only implied; moreover, it proved to be very hard to find such evaluations, partially because general evaluations of the whole political system are much more frequent, and partially because websites containing such judgments are merely blocked in Russia by local authorities as extremist materials. Online publications are usually written with respect to the possibility of being blocked, so that allegories, metaphors and other linguistic tricks that allow authors to express their evaluations in technically (formally) non-evaluative way, are very widespread.

Only about 4% of the current textual collection contain direct evaluations, the rest of evaluations are indirect and therefore need a special markup.

The marked-up collection (corpus) is now under development. It is planned to use it as a standard to evaluate precision and recall of opinion mining, but now it is used more as a source of examples of different linguistic constructions that denote evaluative judgments. Classification of evaluations that have been found in the texts is being performed now in the ontology. Evaluations are classified by the evaluated attributes when applicable (e.g., rationality, openness, legitimacy, etc.), by author's attitudes (positive or negative), and by their degrees. The latter two bases of classification allow to make rankings of the evaluations (to assign numeric values to them).

The size of the corpus will substantially depend on the dispersions of these rankings: for each public authority, the set of evaluation rankings must be a statistically representative sample, which means that the limiting error of the sample must be considerably low (ideally, no more than 5%). The limiting error is calculated as the value of the Student distribution (which decreases, as the size of the sample grows), multiplied by the square root of the ratio of the sample dispersion to the size of the sample. It means that the more is the dispersion, the more must be the size of the sample. In practice, however, 400 values is usually enough to get less than 5% limiting error, so, the minimal size of the corpus should be about 2000 publications for 50 public authorities.

5.Conclusion

Apparently, a tool for Opinion Mining, which will effectively detect evaluations of public authorities in media discourse, can only be built by a significant development of existing technologies of computer linguistic analysis. In most cases, the evaluations of actions of public authorities are expressed by sophisticated linguistic techniques and can not be detected by searching literal matches. Methods of ontological semantics have proved to be necessary already at the stage of automatic selection of texts relating to specific public authorities.

6.ACKNOWLEDGMENTS

Our thanks to ITMO University Center for Electronic Government for productive co-operation.

7.REFERENCES

1. Bershadskaya, L.A. and Chugunov, A.V. 2014. E-Government Services: an Analysis of Discussions in Social Media. *Interdepartment Information Service, 1* (166). 10-17.
2. Bershadskaya, L.A., Bikkulov, A.S., Bolgova, E.V., Chugunov, A.V., and Yakushev A.V. 2012. Social networks and sociometry research: theoretical basements and practical examples of computerized instruments implementation for virtual communities studies. *Information resources of Russia*. *4*. 19-24.
3. Soms, N.L., Dobrov, A.V. and Dobrova, A.E. 2014. Using Natural Language Processing Tools in the System of Information Resources Monitoring by User Preferences. *Information Society Technologies in science, education and culture: a collection of scientific articles. Proceedings of the XVII All-Russian Joint Conference «Internet and Modern Society»* (St. Petersburg, Russia, November 19-20, 2014) p. 149-158.
4. Dobrov, A.V., Dobrova, A.E., Soms, N.L. and Chugunov, A.V. 2015. Semantic Analysis of News Items on `Electronic Services' Subject Domain: Experience of Applying Methods

of Ontological Semantics. *The state and the citizens in the digital environment: theory and research technologies*. 120-125.

5. Dobrov, A.V. 2014. *Automatic classification of news by means of syntactic semantics*. Doctoral Thesis. Saint-Petersburg State University.
6. Dobrov, A.V. 2014. Semantic and Ontological Relations in AIIRE Natural Language Processor // *Computational Models for Business and Engineering Domains*. Rzeszow-Sofia: ITHEA 147-157.
7. Dobrov, A.V. 2011. A Complex Linguistic Approach to Automatic Classification of News Reports. *Political Linguistics, 3* (37).202-209.
8. Dobrov, A.V. 2012. Back to the issue of the technique to evaluate efficiency of automatic text classification: psycholinguistic aspect. *Psycholinguistics, 9.* 173-178.
9. Petkevič V. 2006. Reliable Morphological Disambiguation of Czech: Rule-Based Approach is Necessary. *Insight into the Slovak and Czech Corpus Linguistics* (Šimková M. ed.). Veda (Publishing House of the Slovak Academy of Sciences & Ludovít Štúr Institute of Linguistics of the Slovak Academy of Sciences), Bratislava, pp. 26–44, ISBN 80-224-0880-8
10. Leontyeva N.N. 2006. Автоматическое понимание текстов. Системы, модели, ресурсы. Москва: Academia, 2006.
11. Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM, 38* (11). 39-41.
12. Fellbaum, Ch. 1998. WordNet: An Electronic Lexical Database.
13. Dobrov, A.V. 2010. Technologies of Intellectual Information Retrieval and Techniques Evaluating Their Effectiveness. *Structural and Applied Linguistics, 8*. 219-233