

Semantic Models at Task Monitoring Public Opinions

Aleksandr Dodonov¹, Dmitry Lande^{1[0000-0003-3945-1178]} and Boris Berezin¹

¹ Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine

dodonov.ipri.kiev.ua, dwlande@gmail.com, bberua@ukr.net

Abstract. The paper proposes a method for constructing and usage of the semantic models (SM) for the purpose of continuous monitoring of public opinion, opinion mining (OM) for finding actual subtopics in the Internet message flow. We define semantic model within this work, as the subject domain model, which has the form of a directed graph, vertices of which correspond to concepts of the domain, and edges define relations between them. Semantic models make it possible to use the results of linguistic statistical analysis of texts (Text Mining) and the use of Information Extraction methods for texts from the Internet for opinion mining. While existing public opinion analysis projects are more focused on one-time (static) public opinion analysis on objects and phenomena, this paper proposes a method for automated construction and use of SM based on continuous monitoring of public opinion on the Internet. OM procedure consists of three steps: construction and clustering of the SM; selection of documents and sentiment definition of subtopics; visualization of results. SM construction using compactified horizontal visibility graph algorithm, usage of cluster analysis methods for determining relevant subtopics, estimation of proportion and tonality for individual subtopic in overall topical information flow are shown. As examples, the models of subject areas corresponding to: “One Belt, One Road”, “Nord Stream”, “Genetically Modified Organisms” topics are considered. Obtained results confirm that proposed method can be used for opinion monitoring in various subject areas.

Keywords: Subject Domain Model, Semantic Model, Cluster Analysis, Sentiment Analysis, Content-Monitoring, Opinion Monitoring.

1 Formulation of the problem

We define semantic model, as the subject domain model, which has the form of a directed graph, vertices of which correspond to concepts of the domain, and edges define the relations between them. Concepts can be events, processes, i.e. such a semantic model can be interpreted as a semantic domain map.

Information that is created by Internet users reflects public opinion on various issues and can be collected, analyzed by content monitoring systems and taken into account when planning the activities of companies, organizations, etc. Semantic models make it possible to use the results of linguistic statistical analysis of texts (Text Mining) and the use of Information Extraction methods contained in texts from the

Internet for opinion mining. While the existing public opinion analysis projects are more focused on one-time (static) public opinion research on objects and phenomena, this paper proposes a method for automated construction and use of SM based on continuous monitoring of public opinion on the Internet.

2 Analysis of publications

As part of this work, authors propose a public opinion analysis based on natural language processing methods. This analysis is aimed at determining the attitude of the subject of monitoring public opinion to the chosen topic. One of the main objectives of the analysis of public opinion is classification of emotional coloring of the text (positive, negative or neutral).

Current works devoted to the analysis, extraction of opinions, moods (Sentiment Analysis - SA, Opinion Mining - OM) note that this is a computer study of opinions, people's attitudes to an object, a concept that can represent individuals, events or topics [1,2]. In these works, levels of analysis of opinions are highlighted: the level of the document, the level of the proposal and the aspect level, when an opinion on a certain concept is considered. For example, in product reviews, product itself is usually the concept, and everything related to this product (price, quality, etc.) are aspects of this product. Analysis is often associated with the search not only for general opinions about the concept, but also for finding opinions about aspects. Some approaches use a fixed, predefined list of aspects, while others extract aspects from the analyzed text.

The work [3] analyzes the difference between the public opinion on genetically modified organisms, presented in the Internet resources, and the opinion of experts in scientific publications. For this purpose, the content of websites from Google search results, the headlines of articles from Google News found by thematic query, etc. were considered. For these resources, three semantic networks were built on the basis of word adjacency analysis; words with a repetition rate above the mean were used as concepts. As a result, central words in each network were identified, common words were found in different networks, the tonality of individual network fragments was estimated.

In [4], an analysis of public opinion consisted in identifying a thematic structure in an array of comments when discussing a film on the Youtube channel. At the same time, the results of semantic analysis and thematic modelling are compared. About three thousand comments were collected on the Youtube service server for analysis. In constructing the SM, bigrams were used as vertices. The thematic structure of the discussion was revealed using clustering of the main component of the constructed semantic network. It is concluded that semantic analysis can complement thematic modelling or serve as an alternative.

In [5], public opinion on the US presidential election in 2012 is analyzed on the basis of news articles published on the Internet. With the help of the monitoring system, more than 81,000 English-language articles from 400 news agencies were collected. On the basis of these resources, the "subject-verb-object" triplets were allocated and

with the help of them two semantic graphs were constructed, reflecting the main actors, their election camps, etc. The results of the analysis of the election campaign were obtained by studying the characteristics of semantic graphs.

3 Method for constructing and using semantic models

In this paper, we propose method for constructing and using SM for OM tasks on the Internet, which involves three stages [6]:

- construction and clustering of SM;
- selection of documents and determining the tonality of subtopics;
- visualization of results.

At the first stage:

- selection of an array of documents for the construction of CM;
- finding concepts; definition of SM connections by constructing a compactified graph of horizontal visibility [7];
- graph clustering;
- formation of requests corresponding to clusters (based on the clusters found, experts identify subtopics and formulate requests for the selection of relevant documents).

At the second stage is made:

- selection of documents corresponding to subtopics (subthemes) from the general information flow using queries;
- determine their share in the total flow of documents;
- the tonality of the documents of the relevant subtopics is determined.

At the third stage of the subtopics with the tonality:

- visualized on the map;
- states are recorded in the database (DB) of the monitoring system for subsequent receipt of the dynamics of the results change over time.

Below are considered the main operations performed as part of these three stages.

3.1 Stage of construction and clustering SM

Selection of an array of documents to build a semantic model. On the basis of a given monitoring object and topic, a request is formulated for sampling an array of documents.

Finding concepts. Documents included in the array are pre-processed, service information is removed, as well as stop words that do not carry a semantic load. Stemming can be performed (coercion of words to the base). Then, on the basis of taking into account the frequency of words in an array of documents, or using other well-known metrics, for example TFIDF, the most important, having the greatest weight of concepts are selected from the words of the array of documents [8].

Definition of SM links by constructing a graph of horizontal visibility. The algorithm of compactified horizontal visibility graph (CHVG) [7] is used to determine the

relationships between the concepts and build the semantic model [7], which provides three steps:

1. On the horizontal axis, a number of nodes are marked, each of which corresponds to words in the order of appearance in the text, and weight numerical evaluations are laid on the vertical axis (visually, a set of vertical lines).

2. Build a traditional graph of horizontal visibility. At the same time, there is a connection between the nodes, if they are in “line of sight”, i.e. if they can be connected by a horizontal line that does not cross any other vertical line.

3. The network obtained in the previous step is compactified. All nodes with this word are combined into one node. All connections of such nodes are also combined.

The peculiarity of using the CHVG algorithm in this work is that its first two steps are performed separately for each sentence of the analyzed text. After that, the resulting network is compactified. In the process of developing the proposed method, a study was carried out on the construction of a SM for documents collected on the following topics: One Belt, One Road (OBOR is the initiative of the PRC on the New Silk Road); Nord Stream; GMO et al. (For more detailed analysis of topics and documents, see the results section). A fragment of the graph of the semantic model built for 28 concepts of the OBOR topic using the described algorithm is shown in Fig. 1.

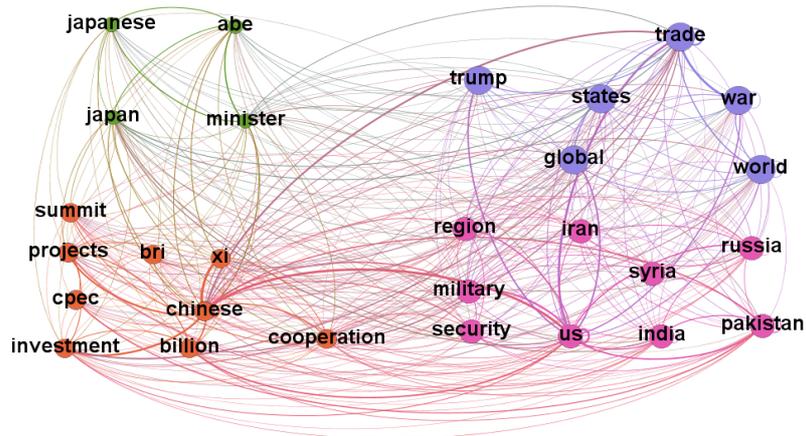


Fig. 1. Fragment of the semantic model graph for 28 OBOR topic concepts.

Clustering graph SM. Given the relevance of the aspect level of the analysis of opinions, after building a semantic model, its network structure is analyzed using graph clustering algorithms—community detection (clustering graph, community detection). In [9], a community is defined as a tightly connected group of nodes that is weakly connected with the rest of the network. Identifying online communities is a complex problem due to the existence of multiple definitions of communities and the complexity of community detection algorithms. In [9], more than a dozen clustering algorithms were considered to identify communities of both disjoint and overlapping (and communities of both types). For the clustering of SM graphs, this paper consi-

dered the use of various known algorithms. The best, most meaningful results were obtained using the Louvain, Leading Eigenvector, and Walktrap algorithms.

Among the well-known community search algorithms in the graph, we can distinguish the Louvain algorithm [10,11], according to which at the beginning of the algorithm each vertex forms a separate community. The step of the algorithm consists of two phases. At the first phase, for each vertex, an attempt is made to find a community, moving to which will give the maximum overall positive change in modularity. You can move a vertex only along adjacent edges, that is, only into those communities that belong to the vertices adjacent to this one. Viewing all vertices continues as long as at least one vertex movement occurs. In the second phase, the graph is compressed: the vertices belonging to the same community form a new super vertex with the corresponding edge transformation. The algorithm stops when the graph stops changing.

The Leading Eigenvector [12,13] algorithm uses the provisions of spectral graph theory. The algorithm is based on maximizing modularity by dividing the graph into two groups of vertices, using the spectrum of the graph. The algorithm proposed in the paper [11] consists in finding the eigenvector corresponding to the first component of the spectrum of the modularity matrix. The partition is determined by the estimate of the leading eigenvector of the modularity matrix.

Walktrap clustering algorithm [14], which allows to find densely connected subgraphs (communities in the graph) on the basis of random walks. The principle of the algorithm is that short random walks, as a rule, remain in the same community. It is argued that transitions from one cluster to another should occur quite rarely. Based on this property, a metric is introduced for the similarity of the vertices.

Fragment of the results of clustering of SM graphs constructed for documents collected on the OBOR topic using the considered algorithms are shown in Table 1.

The first line of Table 1 shows the names of three algorithms that were used to cluster the SM. Columns 1–2 show the results of applying clustering algorithms to the SM based on the concepts selected by frequency of use in the documents, and columns 3–4 show the results of clustering SM based on the concepts selected using the TFIDF indicator. The columns of the fragment of the cluster table show the interrelated sets of concepts of the analyzed topic, which were found using three clustering algorithms.

Formation of queries corresponding to clusters. As a result of the clustering of the semantic model graph, the sets of the most connected graph vertices were found, corresponding to the identified clusters, i.e. sets of similar concepts. On the basis of these concepts in the general stream of documents characterizing the analyzed subject area, subtopics, aspects are highlighted. Experts in this subject area give names to these subtopics and formulate requests for subsequent selection using the information retrieval system of documents corresponding to subtopics from the general flow of documents characterizing the general subject area.

For the OBOR topic under discussion, on the basis of the clusters found, the experts formulated four subtopics and the corresponding queries, given in Table 2.

Table 1. A fragment of the cluster table obtained using different clustering algorithms for the SM graph for OBOR topic documents.

Louvain	Walktrap	Louvain (TFIDF)	Leading eigenvector (TFIDF)
economy military world countries country state global war trade projects economic cooperation security political infrastructure investment development international support national		india military power pakistan country us strategic might russia american political iran russian defense syria turkey	india world relations war united states japan iran defense syria
chinese president people government market xi business million debt including beijing billion summit leaders		sri chinese president people government projects xi debt trump project investment cpec billion port bri	sri chinese president obor people government projects xi cooperation debt project investment beijing summit port bri
india china south asia power region pakistan relations influence us strategic east europe russia american trump japan africa part sea		south region relations indian xinjiang economic cooperation africa development beijing summit pacific taiwan	china's military south power region state global strategic economic security africa uighur pacific taiwan
prime minister foreign policy	prime minister foreign	prime minister foreign japanese abe mahathir malaysia	prime minister foreign japanese abe mahathir russian
china's road belt initiative project silk	road belt initiative silk	china's global obor market research	

Table 2.Subtopics and queries formulated on the basis of the clusters found.

№	Subtopic keywords	Subtopic name
1	xi projects investment	President of the People's Republic of China Xi Jinping about investment projects as part of the Belt and Road Initiative (BRI)
2	india pakistan us	The attitude of India, Pakistan, the United States and other countries to the initiative BRI
3	south region development	The attitude of countries to the development of the southern region in the framework of the initiative BRI
4	japan minister abe	Prime Minister of Japan Shinzō Abe about the initiative BRI

At this, periodically repeated training monitoring system (repetition period from several hours to a day), implemented using the SM construction and clustering algorithms, ends.

3.2 The stage of selection of documents and determining the tonality

Selection of subtopics documents from the general information flow using queries. From the general flow of documents generated by the search query, which characterizes the subject area, the documents of subtopics are selected using search queries, formulated by experts on the basis of clusters identified in the semantic model. For selected documents, each of the subtopics is determined by their share in the total flow of documents.

For the flow of documents generated by the search query to the content monitoring system InfoStream (one-road) & (one-belt) & china, which characterizes the OBOR topic, the names of the subtopics and the corresponding requests are given in Table 2. The shares of the subtopic documents selected using the formulated requests from the total flow, are shown in Table 3.

Table 3.The proportion of documents formulated subtopics in the general flow of the OBOR topic and their tonality.

Subtopic	Share	Negative	Neutral	Positive
xi projects investment	10.1% (101)	2% (2)	0	98% (99)
japan minister abe	9.2% (92)	1% (1)	0	99% (91)
india pakistan us	8% (80)	5% (4)	0	95% (76)
south region development	15.3% (153)	5% (8)	1% (1)	94% (144)

Determining the tonality of the subtopic documents. For the documents of each of the identified subtopics, the tonality is determined - positive, negative, neutral based on the analysis of the words that make up the documents relating to the topics. To determine the tonality, the algorithms proposed in [15] can be used. Under the tonality of the text in this case is meant a positive, negative or neutral emotional coloring of the entire text document as well as its individual parts related to certain concepts, such as persons, organizations, brands, etc. The task of determining the tonality is checked at least three indicators of emotional coloring: positive, negative, neutral and, often, there is also a need to check the combination of these hypotheses (for example, to identify the level of “expressiveness” of the text). The tonality of the subtopic documents selected from the general document flow of the OBOR topic is given in Table 3.

3.3 Stage visualization of the results

At this stage, the visualization of the found subtopics with tonality on the map is performed. Monitoring results are visualized in real time on a geographic map with reference to specific objects. For each subtopic identified in the general flow of documents, a chart is displayed indicating the name of the subtopic and the proportion of documents in the total flow attributable to this subtopic, as well as the proportion of documents in positive, negative and neutral tonality within the subtopics (table 3, figure 2).

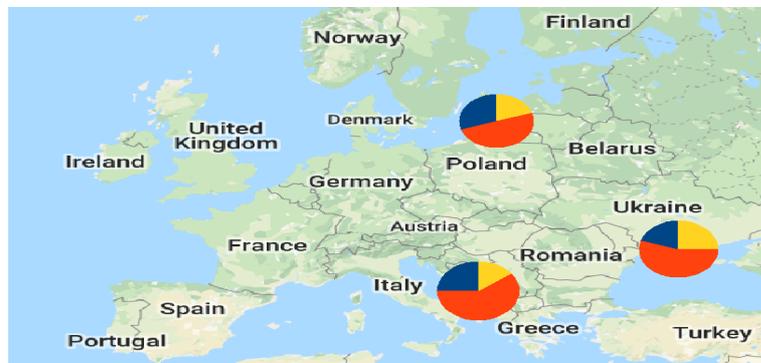


Fig. 2. Visualization of monitoring results on a geographic map. For each of the subtopics found, the map shows the proportion of its documents in the total input stream and the tonality of documents.

The geographic map shows: the subtopics identified in the flow of input documents; share of documents on each subtopic; the tonality of the documents on the subtopics, as well as the dynamics of changes in results over time. The conditions found during the monitoring are recorded in the monitoring database for later obtaining the dynamics of the results change over time. The change in share of documents

and tonality of the formulated subtopics (subthemes) of the main OBOR topic by week is shown in Fig. 3, Fig. 4 and Fig. 5.

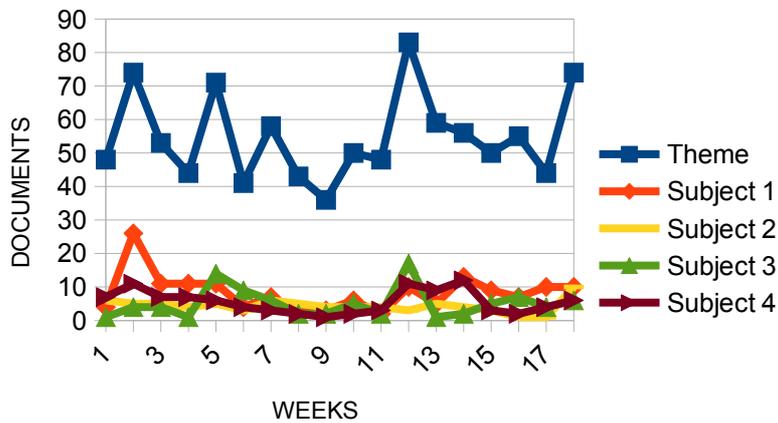


Fig. 3. Changes in the share of documents in formulated subtopics (Subtopic 1 - Subtopic 4, four lower graphs) as part of the total flow of documents on the OBOR topic (Topic, upper graph) by week.

The operations performed in the three stages of the considered method are implemented using the tools of the Gephi software package (<http://gephi.org>), as well as using software tools developed in the programming language for statistical calculations R. The results obtained using the proposed method summarized in the next section.

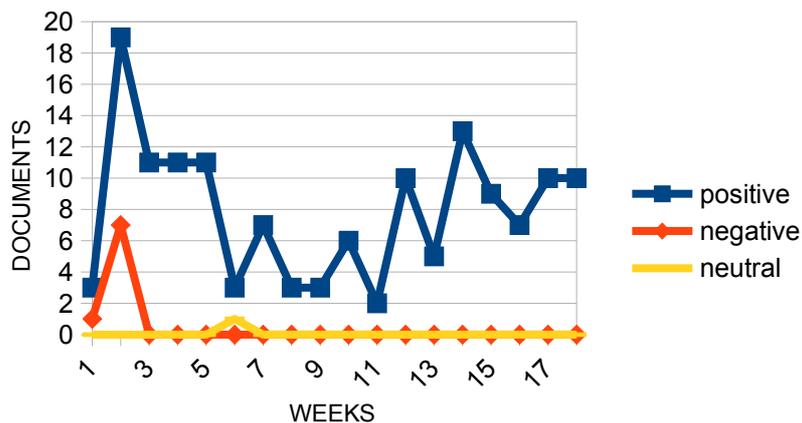


Fig. 4. Changes in the number of documents with a positive, negative and neutral tonality in the “South Region Development” subtopic by week.

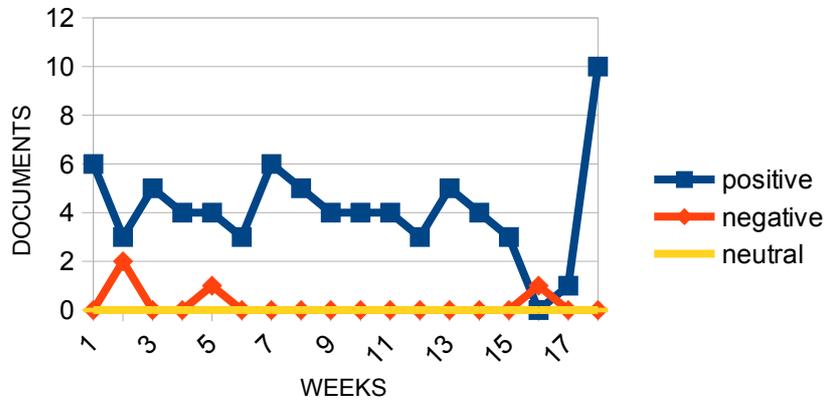


Fig. 5. Changes in the number of documents with a positive, negative and neutral tonality in the “India Pakistan US” subtopic by week.

4 Results

The capabilities of the proposed method of building and using SM for monitoring public opinion were analyzed on several topics:

- One Belt, One Road (OBOR) - an initiative of the People's Republic of China on joint projects of the “Silk Road Economic Belt” and “The 21st Century Sia Silk Road”;
- Nord Stream - a gas pipeline project from Russia to Germany through the Baltic Sea;
- GMO - genetically modified organisms and some other topics.

To monitor public opinion on OBOR topic, an array of 1000 English-language documents (from 30.11.2018 to 07.25.2018) collected using the (one-road) & (one-belt) & china query using the InfoStream system was analyzed.

At the first stage, after sampling an array of documents and its preliminary processing, finding concepts (based on the frequency of use of terms, as well as on the basis of the TFIDF indicator), the corresponding SMs were built (Fig. 1). Clusters of concepts obtained using the Louvain, Leading Eigenvector and Walktrap algorithms based on the constructed SM are shown above in Table 1.

Based on the comparison of the clusters obtained, experts in this subject area can classify document subtopics, as well as requests for a selection of documents relating to the subtopics, determining their share and tonality (Table 2).

At the second stage, the share of documents of the subtopics selected using the formulated requests from the general flow is given in Table 3. Also, the tonality of the documents of the formulated subtopics selected from the general flow of documents of the OBOR topic are given there.

At the third stage, the visualization of the results is performed. A general view of the interface for visualizing the found subtopics, their shares in the general flow and

tonalities is shown in Fig. 2. The dynamics of changes in the share of documents and the tonality of the formulated subtopics by the composition of the OBOR topic documents by week are shown in Fig. 3-5. Fig. 3 shows the change in the proportion of documents in the formulated subtopics (Subtopic 1 - Subtopic 4, four lower graphs) in the composition of the documents of the OBOR topic (Topic, upper graph) by week. Fig. 4 shows graphs of changes in the number of documents with a positive, negative and neutral tonality in the subtopic of the “South Region Development” by week. Fig. 5 shows plots of changes in the number of documents with a positive, negative and neutral tonality in the “India Pakistan US” subtopic by week.

In addition to the OBOR topic, the use of the proposed method of public opinion monitoring on the topics of Nord Stream, GMO and others was considered. For example, for the Nord Stream topic, on the basis of 1000 English-language documents (collected from 02/11/2018 through 08/18/2018) a corresponding SM was built, clusters were found and subtopics were formulated: merkel putin meeting (about Chancellor of Germany Merkel and Putin meeting); gas transit ukraine (on gas transportation through Ukraine); european security energy market (on the security of the European energy market), poland united states (on the relationship of Poland and the United States to the Nord Stream project).

5 Conclusion

A method for constructing and using SM for public opinion monitoring has been proposed, which includes three stages: the construction and clustering of SM; selection of documents and the definition of the tonality of subtopics; visualization of results.

Construction of the SM using the compactified horizontal visibility graph algorithm, use of cluster analysis methods for determining relevant subtopics, estimating the proportion and tonality of individual subtopics within the general topical information flow are shown.

Obtained results confirm the possibility of using the proposed method of monitoring public opinion in various subject areas.

References

1. *Schouten K., Frasinca F.* Survey on aspect-level sentiment analysis // *IEEE Transactions on Knowledge and Data Engineering*, 2016. – Iss. 28(3). – pp. 813-830.
2. *Medhat W., Hassan A., Korashy H.* Sentiment analysis algorithms and applications: A survey // *Ain Shams Engineering Journal*, 2014. – Iss. 5(4). – pp. 1093-1113.
3. *Jiang K., Anderton B., Ronald, P., Barnett G.* Semantic Network Analysis Reveals Opposing Online Representations of the Search Term “GMO” // *Global Challenges*, 2018. – Iss. 2(1). – pp. 1700082. DOI: <https://doi.org/10.1002/gch2.201700082>.
4. *Юдина Д.И., Дудина В.И.* Семантическая сеть на биграмммах как метод валидации результатов тематического моделирования в социологическом исследовании // *Журнал социологии и социальной антропологии*, 2016. – Iss. 19(4). – С. 71-83.
5. *Sudhahar S., Veltri G., Cristianini N.* Automated analysis of the US presidential elections using Big Data and network analysis // *Big Data & Society*, 2015. – Iss. 2(1). – pp. P. 21-49.

6. Додонов А.Г., Ландэ Д.В., Березин Б.А. Построение и использование семантических моделей для мониторинга общественного мнения // Міжнародна науково-практична конференція "Інтелектуальні технології лінгвістичного аналізу": Тези доповідей. - Київ: НАУ, 2018. - С. 6.
7. Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V. The use of horizontal visibility graphs to identify the words that define the informational structure of a text // 12th Mexican International Conference on Artificial Intelligence (MICAI), 2013. – pp. 209-215. DOI: 10.1109/MICAI.2013.33
8. Ландэ Д.В. Підходи до автоматичного визначення термінологічних основ онтологій // Тези доповідей. Міжнародної науково-технічної конференції "Інтелектуальні технології лінгвістичного аналізу". – Київ: НАУ, 2014. – С. 7-9.
9. Harenberg S., Bello G., Gjelteta L., Ranshous S., Harlalka J., Seay R., Samatova N. Community detection in large-scale networks: a survey and empirical evaluation // Wiley Interdisciplinary Reviews: Computational Statistics, 2014. – Iss. 6(6). – pp. 426-439.
10. Blondel V.D., Guillaume J.L., Lambiotte R., Lefebvre E. Fast unfolding of communities in large networks // Journal of Statistical Mechanics: Theory and Experiment, 2008. –Iss. 10. – pp. P10008
11. Louvain, <http://contest.dislab.org/algs>, last accessed 2019/02/06.
12. Newman M.E. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 2006. – Iss. 74(3). – pp. 036104.
13. Ключарёв, П. Г., Басараб М. А. Спектральные методы анализа социальных сетей. *Машиностроение и компьютерные технологии*, 2017, 5. – С. 168-177.
14. Pons, P., Latapy, M. Computing communities in large networks using random walks. In International symposium on computer and information sciences, 2005, pp. 284-293, Springer, Berlin, Heidelberg.
15. Lande D.V. Identification of information tonality based on Bayesian approach and neural networks // E-preprint arXiv: 0806.2738 (2008).