Advances in Economics, Business and Management Research, volume 81
1st International Scientific Conference "Modern Management Trends and the Digital Economy: from Regional Development to Global Economic Growth" (MTDE 2019)

Information search model based on the use of percolation properties of semantic networks of texts

Alyoshkin A.S.
MIREA - Russian Technological University
Moscow, Russia
Antony@testor.ru

Zhukov D.O.
MIREA - Russian Technological University
Moscow, Russia
ZhukovDm@yandex.ru

***Abstract*— The paper examines the models of information search in the texts presented in the multidimensional vector space. Describes approaches to semantic representation of a text document. The concept of information conductivity of the document which can be used for the information search task is discussed. The developed model of construction of semantic multiconnected document and offered algorithm of construction such network for the document from the collection of texts.**

The second part of the article describes the application of percolation theory for the description of information conductivity of multiconnected semantic networks. Since the percolation threshold determines the loss of meaning and the separation of the text into separate unrelated fragments, its value can be taken in determining the relevance of documents for solving problems of information search. Shown the correlation of percolation threshold and "semantic force" of the document.

The last part of article describes the "greedy" algorithm of clustering documents using the value of percolation threshold as measures of information conductivity. In conclusion there is given approaches for performing practical calculations based on the described theoretical approaches.

Keywords — percolation theory; informational search; document clustering; greedly algorithm

I. INTRODUCTION

When solving problems of search information and clustering of texts by semantic groups, their vector representation and definition of relevance and the cosine angles between them used are most often at present days.

Elements of vectors are calculated on the basis of lexical or semantic methods of text analysis, mainly based on the sets and frequency of occurrence in them of words, terms, objects of associative semantic classes, etc. It should be noticed that the analysis of their semantic network is increasingly used to determine the elements of the vectors of texts. However, the parameters that characterize the topology of the semantic network do not apply.

In our opinion, the semantic difference between documents should be determined not only by the set and frequency occurrence in documents of words, terms, associative units, etc. (possibly obtained by analyzing their semantic networks), but also by the topological properties of their network structures, which must be taken into account when solving tasks of informational search.

A semantic network of text is a graph in which concepts (specific and abstract entities) are stored in nodes, and arcs (links or network edges) convey relationships between concepts. Nodes of a network can have different weight (that corresponds to the status of concepts in a patrimonial hierarchy), and arcs-length (that corresponds of force of connections).

A kind of semantic networks of texts representation is are conceptual graphs — universal model of representation of texts meaning, developed by John Sova [1]. This model combines formal possibilities of logic and expressiveness Natural language. A conceptual graph is a finite, cohesive bipartite graph with nodes of two types: one type corresponds on concepts, and the other is a conceptual relation. The structure of the conceptual graph has the following limitations: Arcs correspond to binary relationships, nodes-relationships can be associated with any number of nodes-concepts.

Another model for describing the texts meaning is their representation in the form of a semantic field, uniting the words of different parts of speech, the meanings of which have one common semantic trait [2], which is called an integral scheme.

Unlike other models, the semantic field is a simpler way of presenting knowledge and does not have a rich set of tools and methods for working with data. The disadvantages of this model include:

-A common semantic characteristic is required for data binding.

-There are no logical links.

The concept of using semantic networks has found its application for the presentation of information in Internet. We can mention the model of formal-structured semantics (RDF).

The RDF model has an abstract syntax that reflects graph-based data, and formal semantics (with detailed concepts of the relationship investigation) provide a basis for reliable logical reasoning over RDF data [3].

All expressions in RDF are based on triplet collections, each of which consists of a subject, a predicate, and an object. The set of such triplets is called the RDF graph.

One of the problems with RDF documents is that the SPARQL query language, which is proposed by the W3C consortium and used to extract knowledge from cohesive structures, has several significant shortcomings in the grammar and query semantics.

The second disadvantage is heterogeneity ontologies of RDF documents, which makes it impossible to transfer information from one ontology to another or transfer data to another system. An attempt to solve these problems was the development of the OWL language, which was designed to solve the problems of data transmission and bringing to a single type of ontology.

Another disadvantage is the lack of diverse ontologies, the one major project in this direction "DBpedia" is the presentation of Wikipedia data in the form of coherent data or ontologies, suitable for the RDF format. It is worth mentioning also the lack of ontologies for the Russian language in DBpedia.

This brief overview of research state in this domain shows that the parameters that characterize the topologies of the semantic network overall, do not apply to solving the problems of information search. In our opinion, this is very important, and this issue requires deep study.

II. INFORMATION CONDUCTANCE OF SEMANTIC NETWORKS

Before we describe developed by our team the new approach, we will discuss the concept of their information conductivity, which we use as part of our model for information retrieval.

Consider the semantic network of texts, nodes (semantic units) which will be the vertices of the graph, and the edges connecting vertices-this is multiple logical relationships between semantic units of text. If all nodes are not excluded from the network, then there is at least one (and in fact set) path between any far-spaced (not neighboring) randomly selected nodes (communication that defines the meaning of this document). And that parameter can be defined as the information conductance of the text. If you start to randomly exclude nodes from the network, then there will be a situation where between two sufficiently far apart, randomly selected nodes are not a single communicative path on the remaining nodes. Such a situation can be defined as the loss of information conductivity, i.e. the text splits into separate, unrelated fragments, and the so-called threshold of percoation (leakage/blockage) is reached. The percoation threshold is the minimum value of the proportion of the connecting nodes of the given network (for example, it can be semantic units of the text) which should be excluded (delete, block, or something more) to the network has disintegrated on unrelated areas (was in generally lost the meaning of the text).

A. Building the structure of a multi-layer semantic network of text

In order to determine the threshold of the semantic multi-network of a document, it is necessary to develop a model and algorithm of its construction. The essence of our proposed methodology (see Figure 1) is as follows:

1) The text is processed and the Sentence Boundary disambiguation-SBD is defined with using token sets and flags.

2) Must be selected semantic units (elements) and relations between them in the sentences of the text and between proposals (based on the analysis of syntactic structures ("parsing") and their meaningful analysis).

3) Next step is the construction of semantic network, in which the found semantic units (elements) within sentences and between sentences are connected by the identified logical links.

Figure 1 shows a graphic example in which the text is represented as an ordered set of lines of different lengths. The length of each line conditionally shows the length of the sentence corresponding to it. Oval objects indicate the semantic units included in the sentences. Since the semantic units can be not only words, but also complex terms, the oval objects in Figure 1 has of different sizes (depending on the structure and size of the semantic element). Semantic units can be including integral Semas and tripletes. The lines connecting the oval objects in Figure 1 show the logical relationships between the semantic elements of the text.

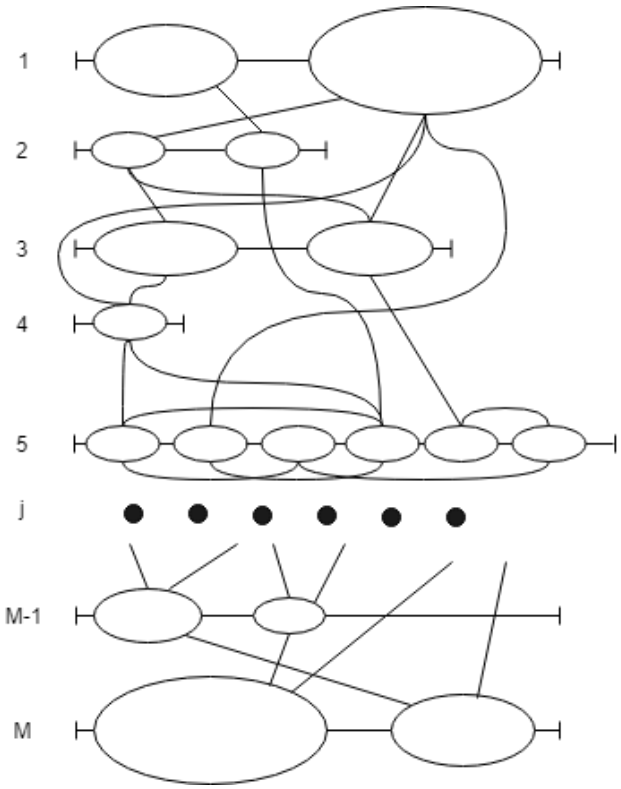

Figure 1. Graphic illustration of the semantic network of the document

In the example presented in Figure 1, the text consists of M-sentences. Proposals 1–3, (M-1) and M contain two semantic units, each of which can have its own structure, and the sentence with number 4 – one unit, the 5th – six, etc. The semantic units themselves can have both bidirectional or unidirectional direct links within the proposal and the relationships between the proposals (see Figure 1).

The proposed model of semantic network construction of the document has much in common with the model of conceptual graphs. But the significant difference is that nodes are the semantic units, which can be any, and not only relationships and concepts.

Let's describe the algorithm of selection semantic units from the text. This algorithm has the following steps:

1. After the whole text is processed and prepared, the first sentence is analyzed, the semantic units and the links between them are defined.
2. Analysis of the second sentence is carried out, it defines the semantic units and the links between them. Further the analysis of presence and establishment of links between semantic units of the given sentence and the previous one is carried out.
3. The analysis of the third sentence is carried out, it defines the semantic units and the links between them. Further the analysis of presence and establishment of links between semantic units of the given sentence and the previous one is carried out. In addition, there is an analysis of the existence and establishment of links between the semantic units of the sentence and the first.

4. The analysis of the fourth sentence is carried out, it defines the semantic units and the links between them. Further the analysis of presence and establishment of links between semantic units of the given sentence and the previous one is carried out. In addition, there is an analysis of the existence and establishment of links between the semantic units of the proposal and the first one, second one, etc.

5. Given that the volume of calculations and their complexity in establishing links between semantic units of text will be nonlinear increase with the increase in the number of analyzed sentences, the execution of step №4 can be stopped on some sentence with a number n (for example n= 7). And when you move to a sentence with a number (n+1) and up to the value (2n+1) conduct an analysis of the presence and establishment of links between the semantic units of this sentence and only the second sentence. When a quotation is reached with a number (2n+1) and up to (3n+1) conduct an analysis of the existence and establishment of links between the semantic units of this sentence and only the third sentence, etc.

The result is a semantic network of the document, the topologies of which will be different for various texts, and in general, it will be a random multinetwork (there are a lot of connecting paths between randomly selected nodes).

B. Application of the percolation theory for the description of information conductivity of multiconnected semantic networks

Different texts will have a different structure of semantic networks (for example, in such networks there will be different number of links per one node, i.e. various network density), which will have random character. The study of the conductive properties of various network structures (including those with a random structure) is engaged by the percolation theory (conductivity) which can be considered as a theory of probability on graphs and it studies the formation of related objects in ordered and unordered structures.

Percolation (*Перколяция* - Rus) – leakage, conductivity or flow. The percolation processes are widely studied in various fields of science: Mathematics, Physics, Chemistry, Biology and many others. The application of methods of the theory of percolation is described in many monographs and reviews [4-10].

In a percolation theory studies the solution of the nodes problems and the communications problems using of networks with different structures (both regular and random). To explain the basic provisions of percolation theory we consider an infinite square grid. When we have a deal with communications problem, we must see what part of links that need to be broken so that the network is split into at least two parts. In a node problem, we define the percentage of blocked nodes with which network is split into unrelated clusters (see Figure 2), within which the links persist (or vice versa, the proportion of conductive nodes when conductivity occurs). The percentage of non-blocked nodes (in a nodes problem) or connected links (in a communications problem) that results in a conductivity between two randomly selected nodes in a network is called a percoation (leak) threshold.

The use of the notion of the shares of a particular node or relationship is essentially equivalent to the notion of the likelihood of a randomly selected node or relationship being in a non-blocked state. Thus, the value of the threshold of percoation determines the probability of transmission of information through the whole network.

Using of the notion of shares a particular node or relationship is essentially equivalent to the notion of likelihood a randomly selected node or relationship being is in a non-blocked state. Thus, the value of the threshold of percoation determines the probability of transmission of information through the whole network.

In addition to square networks it is possible to consider the percolation in regular (triangular, hexagonal, Kaylee trees, three-dimensional lattices (for example, hexagonal, cubic, etc.)) and random networks.

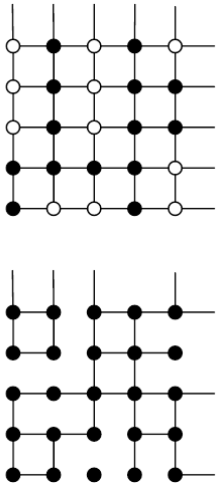

Figure 2 - The type of network structures for the square grid in the node task (up) and the link task (down)

However, if the regular structures have analytical dependencies of the percoation thresholds or cluster sizes from the average number of connections per node (network density), then in random networks for research it is necessary to use Numerical methods and computer modelling [8-10].

The application of the methods of percolation theory to the research of network structures and processes in them allows us to answer for following questions:

1) Find the distribution of clusters of blocked nodes in whole network structured by size, at the given probability of their blocking (or on the contrary, being in active state).

2) Find the statistical characteristics of the clusters, for example, the average size of the blocked nodes cluster in network.

3) How the network structure threshold value depends on its density (the average number of connections per node) and several other issues.

The dependence of the network structure threshold value from its density (the average number of connections per node) is very important in terms of solving the problems of information search. Since the percoation threshold determines the loss of meaning and the separation of text into unrelated fragments, its value must be considered for determining the relevance of the documents.

In relation to the semantic network of text, the value of the percoation threshold can be called "Semantic force".

In Works [8-10] the study of random networks density and percolation threshold was conducted. In these studies, it was shown that the logarithm of the probability of the *impossibility* of transmitting information (Ln$P(x)$) linearly depends on the inverse value of the network density $1/x$. (Where the x is the

value of the random network density). This equation (1) was obtained to simulate the task of breaking links:

$$lnP(x) = -\frac{6{,}581}{x} - 0{,}203 \quad (1)$$

The equation (1) shows the value of the correlation coefficient with experimental data of numerical modeling and linear dependence equal to 0,992.

Equation to simulate a network node exclusion task Has the appearance:

$$lnP(x) = \frac{4{,}39}{x} - 2{,}41 \quad (2)$$

The correlation coefficient of data obtained by numerical modelling for equation (2) is 0.95.

Knowing the structure of the semantic network *i*-Text document can easily determine its density (the average number of links per node), and then determine the likelihood of the *inability* to transfer information (P_i).

Since the percolation threshold is, by definition, the probability of information transmission, and P_i - is the probability of the impossibility of transmitting the information [8-10], its value (ψ_i) for *i*-text can be defined by the formula:

$$\psi_i = 1 - P_i \quad (3)$$

"*Semantic force*"- ψ of texts or the value of percolation threshold of their network is in fact a probability of preservation of meaning. When removed from the semantic network *i*-document of the share words more than the value of "*Semantic force*" (ψ_i), the source text turns into unrelated fragments. Thus, the amount of (3) can be considered as the probability of loss of meaning.

When comparing the text vectors between each other the "semantic force" or the value of networks percolation threshold can be used as "weights" of vectors.

III. THE ALGORITHM OF "GREEDY" CLUSTERING, BASED ON THE ACCOUNTING OF THE PERCOLATIONAL PROPERTIES OF SEMANTIC NETWORKS OF TEXTS

At the heart of the work of any "greedy" algorithms is the assumption that the process, which at each of his stages in terms of selected parameters were optimality or rationality is also optimal/rational.

The suggested algorithm of "greedy" clustering uses a vector space of text documents, which can be obtained with any methods (lexical, semantic, etc.), and consists of the following steps:

1) Conduct processing of some texts: define the bounds of paragraphs and sentences (Sentence Boundary disambiguation - SBD) using sets of tokens and flags. Based on the analysis of syntactic structures and their informative analysis allocating semantic units (elements) and relations between them in sentences of text and between sentences. Carrying out the construction of semantic network.

It is worth to consider the case when semantic network of a document is obtained initially divided into several parts, between which there is no connectedness. In this case we must generate separate document, that falls into the document collection, for each individual area.

2) For each text we build a matrix of links between nodes of semantic network, based on that matrix we calculate its density. Then we compute the equations for the logarithm dependency of the impossibility transferring information (Ln$P(x)$) from the inverse value of the network density $1/x$. Then we calculate two parameters for each document from collection: the probability of impossibility information transmitting, and second one - the value of percoation threshold (ψ_i), "semantic force".

3) We process the texts and vectorizing them. From the collection of n-text documents, we create a matrix that has n-columns and M-strings.

$$L = \begin{vmatrix} l_{1,1} & l_{1,2} & \cdot & l_{1,i} & \cdot & l_{1,N} \\ l_{2,1} & l_{2,2} & \cdot & l_{2,i} & \cdot & l_{2,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ l_{j,1} & l_{i,2} & \cdot & l_{k,i} & \cdot & l_{i,N} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ l_{M,1} & l_{M,2} & \cdot & l_{M,i} & \cdot & l_{M,N} \end{vmatrix} \quad (4)$$

Where $L_{j,i}$ is a numerical characteristic corresponding to the vector element obtained using lexical, semantic or other methods. For example, associative-vector models, where each vector contains a set of linguistic units that correspond to the associative context of a natural language document. Each *i*-document of the collection N will correspond to the vector L_i:

$$L_i = \begin{vmatrix} l_{1,i} \\ l_{2,i} \\ \cdot \\ l_{k,i} \\ \cdot \\ l_{M,i} \end{vmatrix} \quad (5)$$

4) to calculate the difference between vectors, you need to use Euclidean metric (for example). If we consider the "semantic force" (the value of percolation threshold) of semantic network of texts as the "weight" of its vector, we can say that close to zero values of percolation thresholds reduce the length of the vectors, and close to the zero practically do not change. It is quite logical, because at a small "semantic force" for the text to split into unrelated fragments it is necessary to remove a very small fraction of semantic units (nodes from network). Also note that the Euclidean metric depends on the length of the vectors, and the value of the cosine of the angle between them - no.

Thus, for the ranking of texts and the solution of the task of information retrieval, it is necessary for each vector from collection, calculate regarding its "semantic force" (ψ_i) (in the multidimensional space of traits ($\boldsymbol{R^M}$)) - Euclidean distance between this vector and each of the other vectors:

$$d_i(l_{j,i}, l_{j,k}) = \sqrt{\sum_{j=1}^{M} \{\psi_i \cdot l_{j,i} - \psi_k \, l_{j,k}\}^2} \quad (6)$$

When solving the problems of information search rather than clustering by semantic groups, the value of the cosine $d_i(l_{j,i}, l_{j,k})$ will determine the relevancy values that can be used for returning search results (*SERP, search engine return paging*).

5) For a next step we create a matrix that has n-columns and n-lines, and the elements of which are Euclidean distances

$d_i(l_{j,i}, l_{j,k})$. Note that this matrix is symmetrical relative to the main diagonal.

6) Find the sum of the Elements $d_i(l_{j,i}, l_{j,k})$ on each line, and select the line with the lowest amount. The corresponding number of the given line document has the maximal closeness with all others (the less sum, the nearer this vector to all others - this is an optimality criterion for any stage of process "greedy" clustering).

7) Take the vector selected at step №4 and find for it those vectors for which, Euclidean distances do not exceed a certain value of ξ. All the vectors that are in this group, unite in the first cluster. When selecting the ξ value, consider its specification: the number of clusters (as it is used in K-means) or the rules for assigning a vector to a selected class, such as those used in Affinity Propagation, Dbscan, BIRCH, and others.

8) Remove from the matrix created in step №3 all columns and rows corresponding to the vectors in the first cluster.

9) Next step. We find the sum of the elements $d_i(l_{j,i}, l_{j,k})$ for each line of matrix obtained in step №6, and select the line with the lowest amount. We take the chosen vector and find for it those vectors for which Euclidean distances do not exceed the value of ξ. All the vectors that are in this group are merged into the next cluster.

10) Than we repeat the algorithm steps until all the vectors are exhausted in the matrix.

IV. EXPERIMENTAL TESTING OF THE QUALITY OF THE "GREEDY" CLUSTERING ALGORITHM BASED ON THE USE OF THE VALUES OF SEMANTIC NETWORKS PERCOLATION THRESHOLDS

For experimental testing of the proposed algorithm the corpus of texts with the volume of 1130650 documents was used, collected from the news resource «echo Moscow» (http://echo.msk.ru) for a period from 2000-08-17 to 2018-01-31.

When conducting a study of the accuracy of clustering of the developed algorithm, a measurement is performed on the same base of texts compared to the following most commonly used clustering algorithms: K-means, Affinity Propagation, DBSCAN, BIRCH and entropy estimates. We are using following clustering models:

- TF-IDF, the elements of which were normalized frequencies of terms (words and n – grams) in the documents.
- TF-IDF, the elements of which were obtained by applying the text representation model as associative-vector space.

Acknowledgements

This work was supported by the Ministry of Education and Science of the Russian Federation [grant number 28.2635.2017/ПЧ, under project title “Development of stochastic dynamics models with memory and self-organization for weakly structured information, to forecast news events based on natural language textual arrays”].

References

[1] Sowa J. F. Semantics of conceptual graphs. Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics, 1979, California, P. 39-44., DOI: 10.3115/982163.982175.

[2] Semantic field [electronic resource]. URL: https://ru.wikipedia.org/wiki/Семантическое_поле (getting date: 20.08.2018).

[3] Resource Description Framework (RDF): Concepts и Abstract Syntax [electronic resource] The World Wide Web Consortium (W3C). 2004. 10th february. URL: https://www.w3.org/2007/03/rdf_concepts_ru (getting date: 20.09.2018).

[4] Grimmet G. Percolation. — Berlin: Springer-Verlag, — 1989 (2nd ed., 1999).

[5] Sahimi M. Applications of Percolation Theory. — London: Tailor & Francis, — 1992.

[6] Stauffer D., Aharony A. Introduction to Percolation Theory. — London: Tailor & Francis, — 1992.

[7] Kirkpatrik, Scott, и др. Percolation in Dense Storage Arrays. — 2002, — Physica A.

[8] Dmitry Zhukov, Tatiana Khvatova, Sergey Lesko, Anastasia Zaltsman. Managing social networks: applying the Percolation theory methodology to understand individuals’ attitudes and moods. Technological Forecasting and Social Change, Volumes 123, pp. 234–245.

[9] D.O. Zhukov, T.Yu. Khvatova, S.A. Lesko, A.D. Zaltsman. The influence of the connections density on clusterisation and percolation threshold during information distribution in social networks. Informatika i ee Primeneniya, (Informatics and its applications). 2018, 2018, volume 12, issue 2, pp. 90-97.

[10] Khvatova, T.Yu., Zaltsman, A.D., Zhukov, D.O. Information processes in social networks: Percolation and stochastic dynamics. CEUR Workshop Proceedings 2nd International Scientific Conference "Convergent Cognitive Information Technologies", Convergent 2017; Volume 2064, 2017, pp. 277-288.