DOI 10.34185/1562-9945-6-125-2019-08
UDC 004.738.52

Y.R. Kovylin, O.S. Volkovsky

# MATHEMATICAL MODEL FOR AUTOMATIC CREATION THE SEMANTIC THESAURUS FOR THE SCIENTIFIC TEXT

*Annotation. The paper deals with the issues related to the use of the algorithm of constructing the semantic model of the document for the creation the thesaurus of the scientific text terms in the natural language. The purpose of the paper is to develop an approach to create matching between elements of scientific text that do not have a direct syntactic link, but are semantically related to the one field. The relevance of the research is that the system does not use linguistic or vocabulary knowledge during its work, which makes it a universal tool for forming semantic correspondence between terms in a scientific text. Obtained results show that the semantic labels of a document have the highest number of intersections with semantic contours when they contain the largest number of semantically significant stems in their composition, which allows to make assumptions about a direct semantic connection between terms corresponding to such stems.*

*Keywords: Thesaurus, latent-semantic analysis, semantic document model.*

**Formulation of the problem.** A thesaurus is defined as a dictionary with additional information about the links of terms, which uses some types of semantic relations [1]. Unlike the explanatory vocabulary, a thesaurus makes sense not only by definition, but also by correlating words with other concepts and their groups, which can be used to fill the knowledge base of artificial intelligence systems. Creation a thesaurus is an important task of the natural language processing field, as an electronic thesaurus can helps computer systems to solve the complex text analysis problems that require some semantic data on natural language (response generation, plagiarism, etc.). The major problem with the construction of such systems is the need to involve a large amount of linguistic and vocabulary knowledge of the language and its individual parts, which is currently a complex and unsolved mass task for a flexibly-rich Ukrainian language. Therefore, as part of this work, an approach to the automatic integration into semantic clusters of scientific text

terms was created and tested, which has no prior semantic markup in its composition and does not require the use of any linguistic or semantic knowledge of the system.

**Analysis of recent research and publications.** An analysis of existing works has shown that the vector of research of the topic of automatic formation of thesaurus is shifted towards partial automation of the manual labor of the operators who form or verify the thesaurus connections. For example, in [2] an approach to constructing of the thesaurus based on the patterns of nominal phrases is proposed, after which the previously obtained list of terms is manually edited using the utility - thesaurus of subject domain terms. An alternative approach is described in [3], which implements the construction of thesauruses based on a heading, and "makes it easier for the operator to work with distributed information resources". Therefore, in order to reduce the size of participation of the human operator in the process of thesaurus formation, the system presented in this work is based on our developed approach to the formation of the semantic model of the document described in [4, 5].

**Purpose.** To develop a system of automatic formation of the thesaurus of the scientific text on the basis of algorithm for constructing semantic models of the document.

**Main research material.** A feature of the algorithm developed in [4, 5] is the use of latent-semantic analysis, clustering, and artificial intelligence methods to form a semantic model of a document without the involvement of linguistic knowledge, which consists of semantic labels of a scientific text, that combine the words stems from text, and the semantic contours of sentences that may have a connection with semantic labels and have some semantic force of such connection. The hypothesis put forward at this stage implies that the formed semantic labels of a document have the largest number of intersections with semantic contours when they contain the largest number of semantically significant stems in their composition. If so, then the resulting system can automatically compile semantic thesaurus of terms, matching the sets of sentences and the semantic labels of the document, while implementing the scheme of understanding the abstract concept through a set of meaningful words.

Consider an example for processing a technical text whose subject is related to the topic of space, namely cosmic relic radiation (Fig. 1)

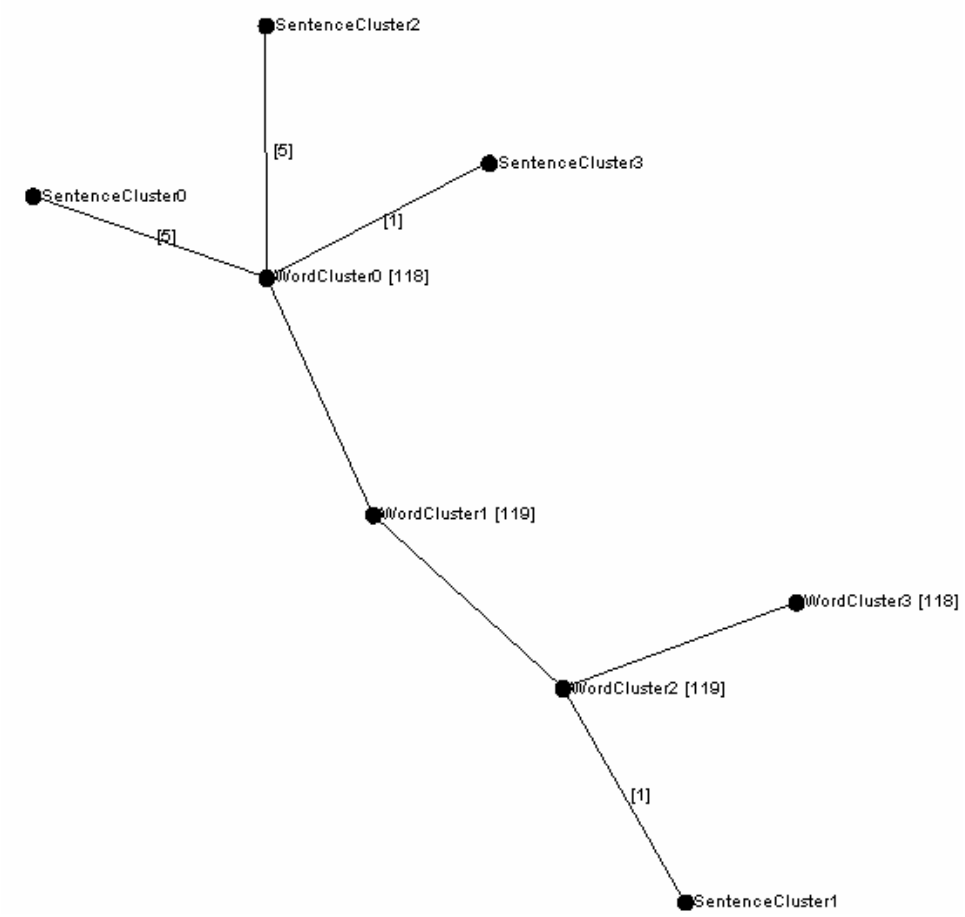


Figure 1 - The Semantic network of the "Relic radiation" text

The semantic text grid shown shows that the semantic label with number 0 (WORDCLUSTER 0) has the greatest total weight of connections to the semantic contours than the others, and the semantic label with number 3 (WORDCLUSTER 3) did not fit into the resulting model at all due to the lack of intersections with semantic contours. Then, according to the hypothesis under test, the null cluster should be the most informative and clearly indicate its content on the subject matter of the text, and the third cluster should have non-informative and semantically ambiguous stems. The contents of the clusters obtained are shown in Table 1. Indeed, comparing clusters, one can see a significant difference - the null cluster contains stems (universe, relict, gravitational, satellite, radiation, space, etc.) that are not only related to the main subject of the text, but also semantically related a sectoral focus. On the other hand, a cluster that does not fall into the semantic network does not represent

any meaning, the whole cluster has only one stems related to the topic of space (telescope), others - are semantically generalized words that do not define any subject area.

Table 1

Examples of semantic labels of the "Relic radiation" text

| **WORDCLUSTER 0** |
|---|
| became predicted by the explosion of Sunyev Universe by polarization must electromagnetic which isotropy experimentally bound ceased plasma stages obtained by Robert time of the line free could ceased cosmological large atom particles takes into account caused by the main process microwave displacement interacted background object satellite energy called very much ranks of propagation thus slowed down due to the expansion of hydrogen helium charged sunyevay density of the wound area sunyayev set sunyayev appear sunyayev ionized space neutral backgrounds constantly again red temperature theory of spectra of photonsdipal hot consists of most environments slowing down existence possessed the cosmic first possibility today's basics of high background fluctuations are radiated by scattering |
| **WORDCLUSTER 3** |
| got maximum estimates to check the impact way The product was less According to the protection of studying According to the maximum of calls Origin of accuracy Is allocated Isolated protection of place of protection Yakov is heterogeneous According to the study of protection the less is less phenomenon of cooling of cooling of calls and other is subject to the same level of product according to the course of studying precision less the number of Origin studying product multiple protection during sign calibration long product widely studying protection origin calibration processing studying several others may be knocked out completely different cooling call yak maximum processing output call telescope precision wide according to other calibrations product telescope recurrent confirm stopped protecting the cooling area less ringing the term |

The described semantic tag evaluation of the semantic labels is not limited to several tests and was conducted for a separate test knowledge base, since it indicates the dependence of the semantic network structure and semantic properties of the document. For this purpose, a test set of the scientific texts was formed (65 texts), that grouped by fields "astronomy", "philosophy", "economy" and "information technology", after which for each document in the set was created an appropriate semantic model of the text. After that, from each constructed semantic model, the clusters with the largest (semantically strong cluster) and smallest (semantically weak cluster) total cross-weight with semantic contours were obtained, for which the value of the sense capacitance $S_V$ was calculated by formula (1):

$$S_V = \frac{N_Q}{N_W}, \quad (1)$$

where $N_W$ is the total number of words in the document required to normalize the results, $N_Q$ is the empirically established number of unique terms in the cluster that are directly related to the field of knowledge to which the text relates (topic of the document). The results of the calculation of the semantic capacity for each thematic set of texts are shown in fig. 2.

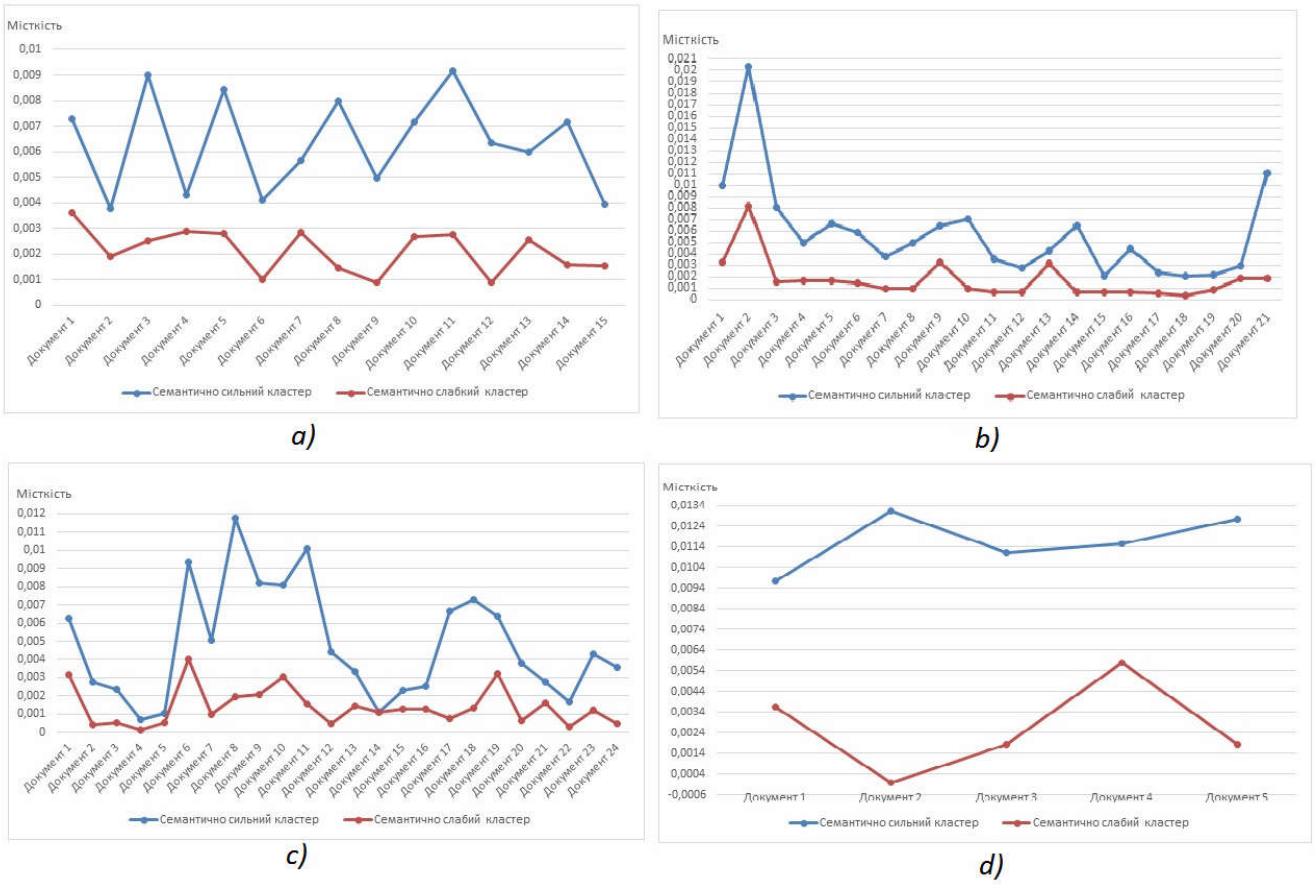


Figure 2 - Distribution of values of sense capacity for topics: a) economics; b) philosophy c) astronomy; d) information technology

**Conclusions.** The obtained graphs show that in the performed tests, the number of semantically meaningful terms in semantically strong clusters significantly exceeds the number of terms in semantically weak clusters regardless of the number, size and thematic orientation of documents - for economics (15 documents) the number of terms in strong clusters 2.98 times higher than the weak, philosophy (21 documents) - 3.375 times, astronomy (24 documents) - 3.473 times, information technology (6 documents) - 4.44

 ISSN 1562-9945

times. Thus, the content of semantic labels with the greatest number of intersections with semantic contours is the thesaurus of a scientific text.

The conducted research does not take into account the weight of the stems or their any syntactic relationship with the text - related topics were evaluated only from the point of view of belonging to the subject of the document, therefore the results indicate a quite adequate formation of the semantic labels of the document - the number of semantically significant terms in the semantic label is directly proportional to the number and the weight of the semantic contours of sentences associated with it in the constructed document models, which proves the dependence of the structure of semantic oh network from the semantic components of the text. Therefore, based on the totality of the tests performed, we can conclude that the algorithm is developed and the semantic model obtained on its basis, despite the absence of semantic markup and semantic dictionaries in its composition, can serve to construct a thesaurus of concepts of one subject area at the text level. The result obtained is limited solely by the degree of formalization of the input text, which should allow making a basic statistical portrait of the document, and does not depend on the subject area of the text.

**REFERENCES**

1. N.M. Bogest. Hierarchical and associative relations in thesaurus on the example of the designer dictionary // Bulletin of the Samara State Aerospace University. - 2012. №2 (33). –p.-228-236.

2. Voloshin P., Svitla S. Automated creation of subject area thesaurus for local search engines // “Knowledge - Dialogue - Solution” International Book Series “information science & computing”, Number 15. - FOI ITHEA Sofia, Bulgaria. - 2009. - p. 24–31.

3. V. Trusov Construction of thesauruses, thematic classifications and rubricators for information retrieval in distributed information systems/ Bulletin of the Novosibirsk State University - 2015. №2 (13). - p. 86-101

4. O.S. Volkovsky, Y. R. Kovylin. Computer System of Building of the Semantic Model of the Document // 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) - p. 322-327 – Lviv, 2018. DOI: 10.1109/DSMP.2018.8478591.

5. O.S. Volkovsky, Y. R. Kovylin. Computer system of intellectual semantic search with the text generation using// Bulletin of the Kherson National University - 2018. №3 (66). -p. 238-245.

Received 10.12.2019.
Accepted 13.12.2019.

***Математична модель для автоматичного створення семантичного тезауруса наукового тексту***

*Розглядаються питання автоматичного створення семантичного тезауруса термінів наукового тексту без попередньої семантичної розмітки і без включення лінгвістичних або словникових знань в систему на основі алгоритму отримання семантичної моделі документа.*

***Mathematical model for automatic creation the semantic thesaurus for the scientific text***

*The questions of the automatic creation of a semantic thesaurus of scientific text terms without preliminary semantic markup and without the inclusion of linguistic or vocabulary knowledge in the system based on the algorithm for obtaining a semantic document model are considered.*

**Волковський Олег Степанович** - кандидат технічних наук, доцент, Дніпропетровський національний університет імені Олеся Гончара.
**Ковилін Єгор Романович** - Дніпропетровський національний університет імені Олеся Гончара, аспірант кафедри автоматизованих систем обробки інформації.

**Волковский Олег Степанович** - кандидат технических наук, доцент, Днепропетровский национальный университет имени Олеся Гончара.
**Ковылин Егор Романович** - Днепропетровский национальный университет имени Олеся Гончара, аспирант кафедры автоматизированных систем обработки информации.

**Volkovskiy Oleg** - candidate of technical science, associate professor, Oles Honchar Dnipropetrovsk National University.
**Kovylin Egor** - Oles Honchar Dnipropetrovsk National University, postgraduate student of department automated information processing systems.