On Synonyms Search Model

Olga M. Ataeva, Vladimir A. Serebryakov, Natalia P. Tuchkova

Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, Moscow, 119333, Russia

Abstract

The problem of finding the most relevant documents as a result of an extended and refined query is considered. To solve it, a search model and a text preprocessing mechanism are proposed. It is proposed to use a search engine and a model based on an index using word2vec algorithms to generate an extended query with synonyms. To refine the search results, the idea of selecting similar documents in the digital semantic library is used.

The paper investigates the construction of a vector representation of documents in relation to the data array of the digital semantic library LibMeta. Each piece of text is labeled. Both the whole document and its separate parts can be marked. Search through the library content, search for new terms and new semantic relationships between terms of the subject area becomes more meaningful and accurate.

The task of enriching user queries with synonyms was solved. When building a search model in conjunction with word2vec algorithms, a "indexing first, then learning" approach is used, which allows obtaining more accurate search results. This work can be considered one of the first stages in the formation of a training data array for the subject area of problems of mathematical physics and the formation of a dictionary of synonyms for this subject area.

The model was trained on the basis of the library's mathematical content. Examples of training, extended query and search quality assessment using training and synonyms are given.

Keywords

Search model, word2vec algorithm, synonyms, information query, query extension

1. Introduction

What is considered "synonyms" is determined not only by general linguistic dictionaries, but also by the subject area. In each subject area, there are well-established expressions that only in a certain context act as synonyms and are not such in general dictionaries of synonyms. In relation to, the selection of synonyms in the mathematical subject area is an independent task. The problem of finding synonyms and "similarity" documents has been studied for a long time [1, 2]. There is such an approach as the Latent Dirichlet Algorithm (*LDA* model) [3], based on the statistical Bayesian model. Algorithms of vector representation of texts like "*tf-idf*" [3] have gained the greatest popularity in their time. The *tf-idf* scheme reduces documents of arbitrary length to lists of fixed length and word count, without reflecting the semantic structure within the document. The *LDA* algorithm uses thematic anchoring of words and thereby facilitates the consideration of semantic relationships between documents and within documents.

The studies presented in [1–3] and other well-known works allow us to say that incorrect information obtained upon request, as a rule, is the result of the use of erroneous semantic connections in databases. This means that at the stage of preliminary data processing, some semantic connections of terms were not taken into account [4, 5]. For scientific papers placed in the search index without including the semantic relationships specific to each subject area, this means that they may not be found by specialists and not cited. In this issue, a special role is played by preliminary data processing and the application of modern approaches to solving the problem of finding reliable scientific information

based on machine learning [6, 7]. The initial data, having acquired a certain structure in the step of processing, can be used already as a source of reliable knowledge [8].

We study the problem of finding documents from the content of the semantic library that are closest to the information request. To select the relevant documents, the procedure was used to find similar documents, which can be obtained by adding the query with synonyms. The aim of the research is to build a search model that will satisfy the conditions of the most complete satisfaction of the user's search needs on the existing set of documents of the sematic library.

The version of the model, built on the LibMeta search index [9] using the *word2vec* algorithms [10–12], will be abbreviated as *wsgMath* in what follows, as is customary in [13]. An approach to the combined use of the *search engine index* and the *neural network* allows obtaining relevant models and *ranking functions* that adapt well to the underlying data.

In this work, the task is to link the search model with the subject area, the boundaries of which are outlined by its thesaurus and classifiers. Thus, the search through the content of the library, the search for new terms and new semantic links between the terms of the subject area becomes more meaningful and accurate.

The structure of the work is as follows: at the beginning we list related researches, the second section outlines the principles of building a search model; the third part describes the construction of extended queries based on the vector representation of texts; the following are examples, conclusion and citation list.

2. Work related research

Research close to the problem of finding synonyms is directly related to the increase in the scope of application of semantic libraries and databases. The IT-community is facing new challenges posed by the thematic diversity of data and sources. In this regard, the information request is processed taking into account the context of the subject area. The report [14] noted that this has led to a new understanding of the processing of an information request as a problem of access to global information. The range of related studies, formulated in [14], includes such sections as:

- User and context sensitive retrieval,
- Multi-lingual and multi-media issues,
- Better target tasks,
- Improved objective evaluations,
- Substantially more labeled data,
- Greater variety of data sources,
- Improved formal models.

The construction of search models includes all the listed sections. In particular, for finding synonyms, the following main directions can be distinguished: adaptation of existing synonyms to a specific subject domain and the formation of a specific dictionary of synonyms separately from the general language dictionary for a specific subject domain. Both of these areas in modern semantic libraries are implemented using machine learning algorithms.

The adaptation of synonyms for the subject area is performed using a more general vocabulary, for example [15–17]. The formation of a specific dictionary of synonyms for the subject area is based on the selection of the "main" terms and semantically related terms to them. This procedure is similar to the construction of the thesauri [18–23] and it is associated with the expansion of the search query in the semantic database [24–27].

The use of machine learning algorithms involves the automated construction of reference samples (training corpus) for training the model and their classification [28–30]. These studies are aimed at developing a scoring system for selected samples. These include searching for candidates (synonyms) and testing candidates based on benchmarks.

When building a model based on machine learning, a dataset has a special role to play, namely: initial data and datasets for testing and training the model. The result of training the model ultimately depends on the quality of the data for testing and the selected samples [29].

In the proposed study, the semantic library focuses on mathematical subject areas such as problems in mathematical physics and equations for these problems. A special dictionary was not used for data analysis, but the algorithms rely on internal sources of the library, a thesaurus for ordinary differential equations, a dictionary of special functions of mathematical physics, a dictionary for mixed equations, as well as classifiers MSC[2], UDC[3] and the mathematical library of I.M. Vinogradov [31].

3. Features of the search model

It should be noted that there are approaches with models built using algorithms trained on publicly available lexical dictionaries and datasets. As a rule, these sets do not include special subject areas and do not include their terminological specifics. To expand search queries [24–27] with synonyms, it is required to use synonym dictionaries. You can use resources such as WordNet[4] or RuWordNet[5], but the main problem is that synonyms from pre-built dictionaries are not linked to the data being indexed, and their use does not improve the results. Therefore, it is necessary to train the model on the subject area and the subject area of this research is "Mathematics". The model of the combined application of the search index and the vector model is built using the *word2vec* algorithm and trained in the mathematical subject area.

To implement this approach, a sequential scheme for working with data was chosen, namely:
- the subject area is determined;
- a vocabulary corresponding to the subject area is determined;
- based on the links between the terms of the dictionary, links are revealed between documents, articles, authors, etc.

Finding and tracking links is done as follows:
- preprocessing of texts is performed;
- machine learning algorithms are used for text processing and analysis;
- vector representations of documents and queries are used to rank search results.

This approach increases the likelihood that the system will more accurately respond to the user's information needs and provide more relevant answers.

In the process of research, the architecture of the search subsystem of the semantic library was determined, which consists of the following parts:
- a text preprocessing component for presenting documents in a searchable format, efficiently loading and storing data and providing quick access to them;
- a component of forming a full-text index of documents;
- a component for constructing a vector model based on an index using word2vec algorithms;
- a component for processing requests and presenting them in a format convenient for expressing the user's information needs in natural language, enriched with synonyms from the subject area;
- a component for generating results based on assessments of document compliance with the request, using the content of the library.

The peculiarity of this approach is the flexible combination of all library tools, such as thesauri, classifiers and encyclopedia for finding synonyms and similar documents, as well as evaluating the results based on them.

In Figure 1 shows the main steps of the formation of search results in the LibMeta library. The query string coming from the full-text search interface goes through the *Analyzer* block. It breaks the string into words, then analyzes and transforms them. Synonyms for words are extracted and filtered from the *wsgMath* model to form an advanced query that retrieves matching documents from the full-text index.

---

[2] https://cran.r-project.org/web/classifications/MSC.html
[3] https://teacode.com/online/udc/
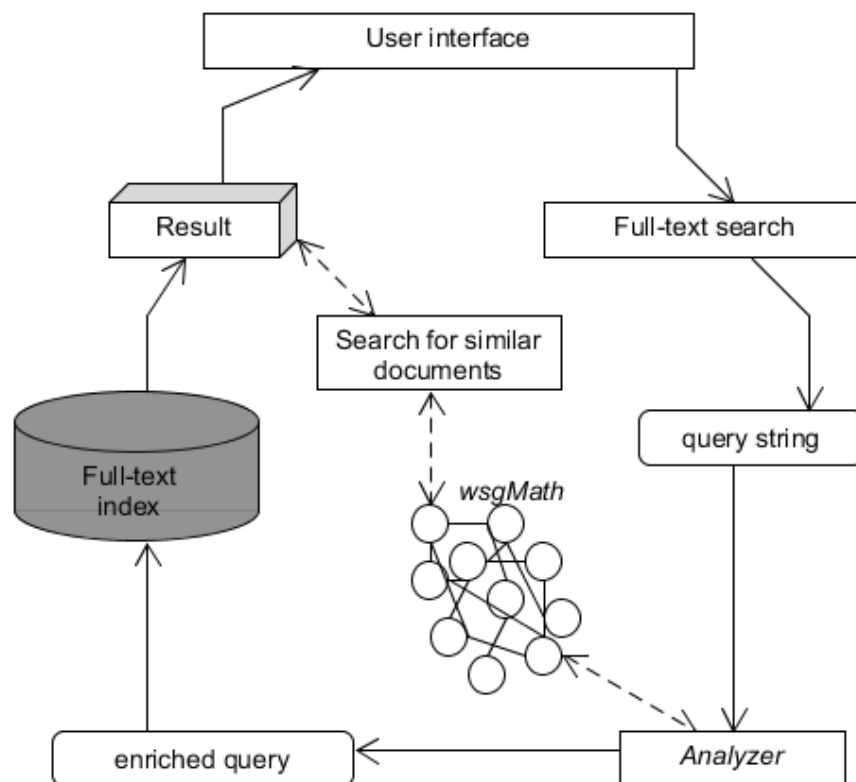[4] https://wordnet.princeton.edu/
[5] https://ruwordnet.ru/ru

Figure 1: Joint use of a search engine and a neural network model built on the basis of an index using *word2vec* algorithms to generate an extended query with synonyms and refine search results based on a selection of similar

An extended version of *word2ve*c (*doc2vec* or *paragraph2vec*, in different sources) allows to introduce an additional element, such as a *text fragment label* or a *document label*. Based on the vectors of these *labels*, we can select similar documents not only by the exact match of keywords or terms, but also *based on the context* of individual fragments or the entire document.

Remark 1. The *text fragment label* is used to display documents that are close in meaning, which do not appear in the search results, but can provide interest to the user.

4. Construction and training of a search system vector space model

4.1. Article preprocessing

One of the necessary stages of preparing data for their loading in certain text formats into an already prepared data infrastructure is the preprocessing and cleaning of this data.

In our case, the data was provided by files in the *TeX* format, decorated with different styles and meta-commands. To begin with, it was necessary to replace all author's tags with standard ones, to clear documents from special characters and unknown tags. At the same time, it was not possible to completely avoid manual processing, but it was possible to reduce it to a minimum.

In Figure 2 shows an example of viewing terms in the LiMeta system, the links of which were formed at the stages of preprocessing and data cleansing.
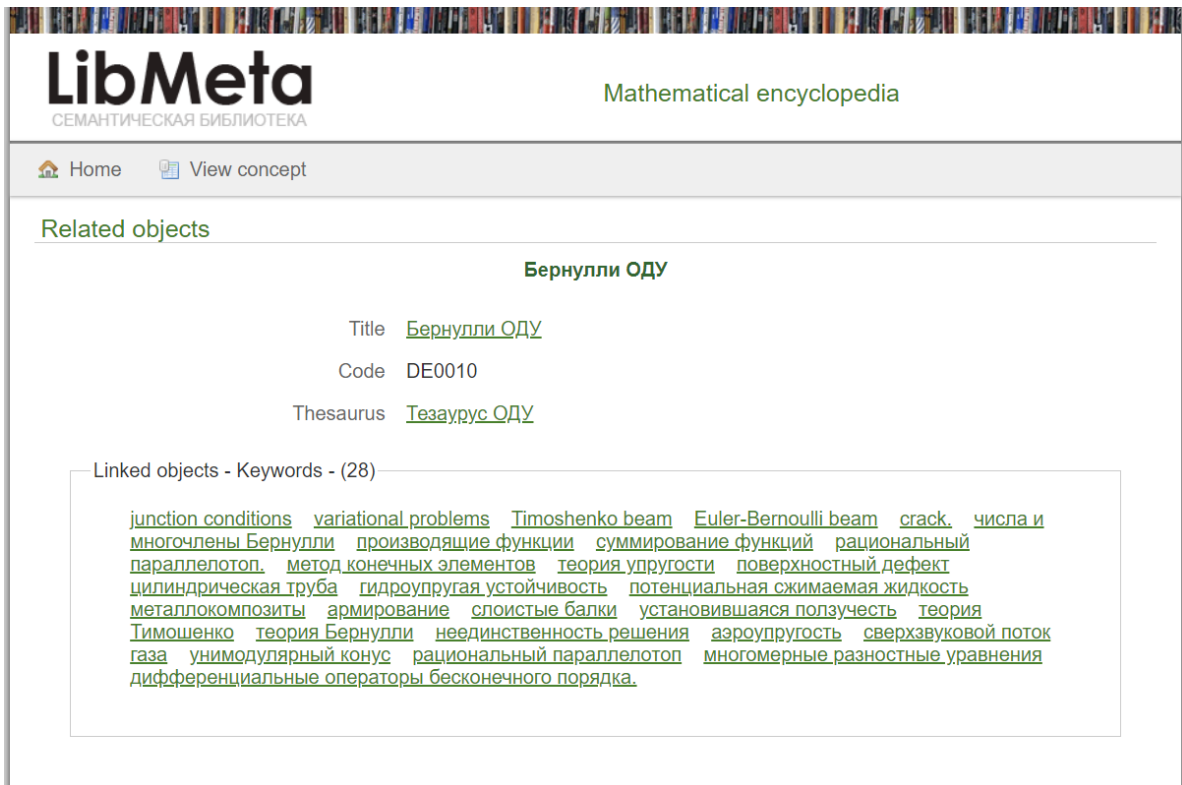
Figure 2: An example of implementation of the preprocessing module in LibMeta

The preprocessing module is made in the *Python* programming language along with the integration of the open-source library *TexSou*p 2015 version and is divided into the following blocks:
- document cleaning;
- converting an article into a tree view;
- processing of all nodes of the tree, recording the revised document.

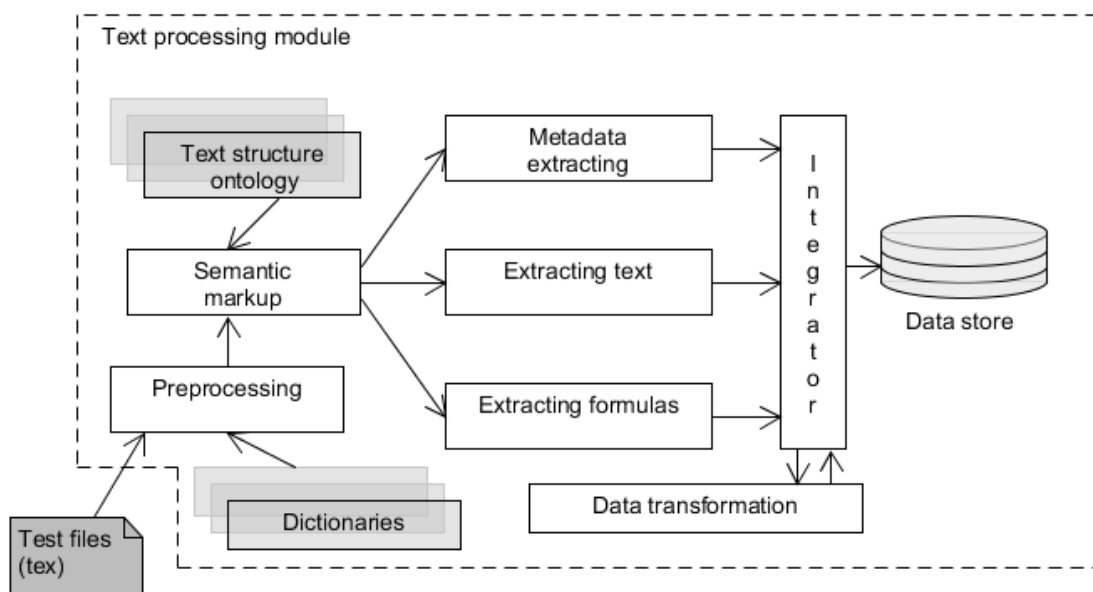In Figure 3 shows the main stages of text preprocessing.

Figure 3: Text preprocessing scheme

4.2. Building an index and training *word2vec*

The goal of training the *wsgMath* model was to obtain synonyms that could expand the search query and get previously unaccounted semantic relations for further retrieval of information relevant to the query.

The search model used in this work implements the integration of a model built on the basis of the *word2vec* neural network and a full-text index. Integration of a neural network and an index can be done in the following ways:

- first training on the text corpus, then the indexing of the texts and the joint use of the trained model and index in the search;
- indexing first, then training on indexed data and sharing in search;
- first training, then extraction / creation of useful resources by the trained network, and then indexing of all resources, both new and original.

The LibMeta library uses an indexing, then learning approach. The problems of how to provide more accurate results based on extended queries [14–16] and how to give users smarter recommendations for further search based on found documents from the subject area in LibMeta were also studied.

Based on an array of preprocessed articles, a full text index was built based on the open source Apache Lucene[6] search library written in Java[7]. This index is used by the library's *full-text search engine* and was also used to train the algorithm and extract contexts.

Words in the context close to the one under consideration are treated as synonyms (*context-sensitive* synonyms, in this case) and analyzed. Their lexical and semantic analysis is carried out, that is, parts of speech, word forms and their own connections are determined, including with dictionaries and thesauri of the subject area. Based on the *wsgMath* model, the proximity of *context-sensitive* synonyms is numerically estimated. With the help of these ratings, candidates are selected, and then the best ones with the highest ratings are selected. For further comparison, classifier codes can be used if the selected words are associated with them.

Table 1 shows examples of words with associations between words (in the first row there is the main word, in the columns below - the identified ones) based on the *wsgMath* model.

Table 1
All connections of the word

пространство (field; space)	*краевой* (boundary value)	*задача* (problem)	*краевой* (boundary)	*напряжение* (stress; tension)	*остаточный* (residual)
оператор (operator)	граничный (boundary)	решение (solution)	интегральный (integral)	деформация (deformation)	концентратор (concentrator)
множество (set)	интегральный (integral)	уравнение (equation)	дифференциальный (differential)	упрочнение (reinforce)	упрочнение (reinforce)
функция (function)		условие (condition)	уравнение (equation)	пластический (plastic)	усталость (fatigue)
		система (system)		(tension)	
		функция (function)		(strain)	
		решение (result)			

---

6 https://lucene.apache.org/
7 https://www.java.com/ru/

5. Examples (query with synonyms)

5.1. Expansion by synonyms

Consider the term "boundary value problem" (*краевая задача*), consisting of two words – "problem" (*задача*) and "boundary value" (*краевая*), each of which has its own synonyms, presented in Table 1.

The context of the term as one unit includes such synonyms as [solution, equation, condition, system, type, function, field, work] (*решение, уравнение, условие, система, тип, функция, область, работа*). In this case, the term "boundary value problem" itself has the following synonyms: "boundary equation", "boundary condition", "boundary function", "integral function" (*граничное уравнение, граничное условие, граничная функция, интегральная функция*), which were determined on the basis of high estimates of the proximity of the following pairs of synonyms and in accordance with the pattern "Adjective + term (noun)" on the *wsgMath* model:

$$sim\ (problem, solution)(\ задача,\ решение\ ) = 0.91$$
$$sim\ (problem, equation)(\ задача,\ уравнение\ ) = 0.86$$
$$sim\ (problem, condition)(\ задача,\ условие\ ) = 0.82$$
$$sim\ (problem, system)(\ задача,\ система\ ) = 0.79$$
$$sim\ (problem, function)(\ задача,\ функция\ ) = 0.73$$

Remark 2. When constructing synonymous terms, synonyms of words defined as a named entity based on a dictionary, which includes a list of persons found in a mathematical encyclopedia, are not used. But at the same time we note that the word "riemann (Riemann)" got into the set of synonyms {cauchy (Cauchy)}, and the word "fourier (Fourier)" got into the set of synonyms {laplace (Laplace)}.

Thus, works with high estimates of the similarity of synonyms were selected.

1.   О положительном радиально-симметрическом решении задачи дирихле для одного нелинейного уравнения и численном методе его получения (On a positive radially symmetric solution of the Dirichlet problem for a nonlinear equation and a numerical method for obtaining it)
     score = 0.90484273

2.   О корректности краевой задачи на прямой для трех аналитических функций (On the correctness of a boundary value problem on the line for three analytic functions)
     score = 0.902505

3.   Проекционные процедуры нелокального улучшения линейно управляемых процессов (Projection procedures for non-local improvement of linearly controlled processes)
     score = 0.8816618

4.   Краевая задача для частного вида уравнения эйлера–дарбу с интегральными условиями и специальными условиями сопряжения на характеристике (A boundary value problem for a particular form of the Euler–Darboux equation with integral conditions and special conjugation conditions on the characteristic)
     score = 0.846388

5.   Теорема валле-пуссена для одного класса функционально-дифференциальных уравнений (Vallee-Poussin theorem for a class of functional differential equations)
     score = 0.84127665

5.2. Search for similar documents

Let's consider an example of using the *text fragment label* element for the process of ranking documents based on the *wsgMath* model when searching for similar documents.

When a document enters the system, its current vector representation is retrieved, a search is performed and the marks of the nearest documents are returned, the cosine distance of which exceeds a certain threshold, determined experimentally as 0.6.

Further, you can also use classifier codes for comparison as one of the options for evaluating similar documents. In this case, various options are possible, associated with the presence or absence of classification codes MSC and UDC in the source documents:

- Documents entering the system are marked with MSC and UDC classifier codes. Documents entering the system are marked with MSC and UDC classifier codes. If UDC codes differ for similar documents, then you can specify them as related subject areas (applications of results, interdisciplinary research, etc.).

- The documents are not provided with codes, but the keywords correspond to the subject area, and there are codes of classifiers in the dictionary (thesaurus, encyclopedia). In this case, the key word codes are compared and the corresponding codes are assigned to the documents.

In Figure 4 shows an example of the correspondence between the classifier codes obtained from the LibMeta content and the procedure for identifying synonyms. In this case, it was revealed that the UDC 515.128 code corresponds to such MSC codes as 54E20, 54E40, 54D65, etc.

Figure 4: An example of correspondence between the codes of the classifiers MSC and UDC

6. Conclusion and future work

In the presented study, the following main results were obtained.

It is shown that the preliminary processing of the input data arrays (texts of scientific articles) allows us to take into account further additional semantic connections and improve the quality of the search.

The use of the mechanism for integrating a neural network and an index makes it possible to implement variants of the search model to obtain relevant documents with a given accuracy.

The combined use of the search engine index and the neural network makes it possible to obtain relevant models and ranking functions that adapt well to the underlying data.

The proposed search model also makes it possible to establish a correspondence between classifier codes for close documents, find synonyms in contextual comparison, and rank documents based on a fragment label.

Problems for further study were identified - the development of a mechanism for assessing the quality of search using various metrics, the use of English and Russian synonyms to enrich the query and improve the quality of search, and assess the learning rate of the model.

This work can be considered one of the first stages in the formation of a training data array for the subject area of problems of mathematical physics and the formation of a dictionary of synonyms for this subject area.

The solution to these problems stems from the research done, which allows us to formulate specific tasks to improve the quality of search. This is the compilation of dictionaries of domain synonyms associated with classifiers and reference documents associated with terms of the domain thesaurus. Such resources can further improve the search quality based on machine learning algorithms.

7. Acknowledgements

8. References

[1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, New York, 1999.

[2] G. Salton, Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[3] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[4] G. W. Furnas, T. K. Landauer, L.M. Gomez, S. T. Dumais, The vocabulary problem in human-system communication, Commun. ACM. 30, 11 (1987) 964–971.

[5] G. Biswas, J. Bezdek, R. L. Oakman, A knowledge-based approach to online document retrieval system design, in: Proceedings of the ACM SIGART Int. Symp. Methodol. Intell. Syst. 1986, pp. 112–120.

[6] U. S. Mak-Kallok, V. Pitts, Logicheskoe ischislenie idej otnosyashchihsya k nervnoj aktivnosti Avtomaty (Perevod anglijskoj stati 1943 g.) Ed. Shennon i Dzh Makkarti, Izd-vo Inostr. Lit., Moscow, 1956.

[7] Professionalnyj informacionno analiticheskij resurs posvyashchennyj mashin nomu obucheniyu raspoznavaniyu obrazov i intellektualnomu analizu dannyh.
URL: http://www.machinelearning.ru/.

[8] T. A. Gavrilova, V. F. Horoshevskij, Bazy znanij intellektualnyh sistem, Piter, SPb, 2000.

[9] O. M. Ataeva, V. A. Serebryakov, Ontologiya cifrovoj semanticheskoj biblioteki LibMeta, Informatika i eyo primeneniya 12, 1 (2018) 2–10.

[10] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, in: Proceedings of Workshop at ICLR, 2013.

[11] T. Mikolov, W. T. Yih, C. Zweig, Linguistic Regularities in Continuous Space Word Representations, in: Proceedings of NAACL HLT, 2013.

[12] Q. Le, T. Mikolov, Distributed Representations of Sentences and Document, in: Proceedings of International Conference on Machine Learning. 2014, pp. 1188–1196.

[13] O. M. Ataeva, V. A. Sererbryakov, N. P. Tuchkova, Using Applied Ontology to Saturate Semantic Relations, Lobachevskii Journal of Mathematics 42, 8 (2021) 1776–1785.

[14] J. Allan, J. Aslam, D. Hiemstra, C. Zhai, et al., Challenges in Information Retrieval and Language Modeling, SIGIR Forum, 37, 1 (2003) 1–17.

URL:http://sigir.org/files/forum/S2003/ir-challenges2.pdf.

[15] D. Turcato, F. Popowich, J. Toole, D. Pass, D. Nicholson, G. Tisher, Adapting a synonym database to specific domains, RANLPIR '00: Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics, V. 11. 2000. P. 1–11. https://doi.org/doi:10.3115/1117755.1117757.

[16] S. Liu, F. Liu, C. Yu, W. Meng, An effective approach to document retrieval via utilizing WordNet and recognizing phrases, SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004. P. 266–272. https://doi.org/doi:10.1145/1008992.1009039.

[17] E. M. Voorhees, Using WordNet for text retrieval. In Fellbaum, 12 (1998) 285–303.

[18] E. I. Moiseev, A. A. Muromskij, N. P. Tuchkova, Tezaurus informacionno-poiskovyj po predmetnoj oblasti: obyknovennye differencial'nye uravneniya, MAKS Press, Moscow, 2005.

[19] E. I. Moiseev, A. A. Muromskij, N. P. Tuchkova, O tezauruse predmetnoj oblasti smeshannye uravneniya matematicheskoj fiziki. CEUR Workshop Proceedings 2260 (2018) 395–405. https://doi.org/10.20948/abrau-2018-43.

[20] ISO 2788:1986 Documentation – Guidelines for the establishment and development of monolingual thesauri. (charge) ISO 5964:1985 Documentation – Guidelines for the establishment and development of multilingual thesauri.
URL: http://www.iso.org/iso/en/ISOOnline.frontpage.

[21] L. Will, Thesaurus principles and practice.
URL: http://www.willpower.demon.co.uk/thesprin.htm.

[22] B. Anne, Thesaurus Management Software.
URL: http://www.fbi.fh-koeln.de/institut/labor/Bir/thesauri_new/thsoften.htm.

[23] R. Gazan, Cataloging for the 21st. Century – Course 3 Controlled Vocabulary & Thesaurus Design, Association for Library Collections & Technical Services Program for Cooperative Cataloging.
URL:https://www.loc.gov/catworkshop/courses/thesaurus/pdf/cont-vocab-thes-trnee-manual.pdf.

[24] E. M. Voorhees, Query expansion using lexical-semantic relations, in: Proceedings of 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland, 1994.

[25] C. Buckley, G. Salton, J. Allan, A. Singhal, Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC), 1995.

[26] E. N. Efthimiadis, Query expansion, Annu. Rev. Inf. Sci. Technol. 31, 5 (1996) 121–187.

[27] J. Xu and W. Croft, Query Expansion Using Local and Global Document Analysis. ACM SIGIR, 1996.

[28] C. Alexander, A Pattern Language. Towns, Buildings, Construction, Oxford University Press, 1977.

[29] V. Lakshmanan, S. Robinson, M. Munn, Machine Learning Design Patterns. Solutions to Common Challenges in Data Preparation, Model Building, and MLOps, O'Reilly Media, Inc. 2020.

[30] E. Freeman, E. Robson, Head First Design Patterns, 2nd Edition. O'Reilly Media, Inc. 2020.

[31] I. M. Vinogradov (red.), Matematicheskaya enciklopediya. Tom 1–5. Sov. enciklopediya, Moscow, 1977. URL: https://dic.academic.ru/contents.nsf/enc_mathematics,
URL: https://encyclopediaofmath.org/wiki/Main_Page.