Search Query Extension Semantics

Olga Ataeva[1[0000-0003-0367-5575]], Vladimir Serebryakov[2[0000-0003-1423-621X]], Natalia Tuchkova[3[0000-0001-6518-5817]]

[1,2,3]Dorodnicyn Computing Center FRC CSC of RAS, Vavilov str., 40, 11933, Moscow, Russia
[1]oli@ultimeta.ru, [2]serebr@ultimeta.ru, [3]natalia_tuchkova@mail.ru

Abstract. The problems of extracting the most complete information from the semantic library by accounting for related documents are considered. Expert knowledge encrypted in the subject area can be made available when the user obtains additional information from linked documents. A feature of the approach is the use of a shallow neural network algorithm to expand the search query in mathematical subject areas, where expert knowledge is available with a significant scientific background of users. The solution to this problem can be achieved by means of semantic analysis in the knowledge space using machine learning algorithms. The paper investigates the construction of a vector representation of documents based on paragraphs in relation to the data array of the digital semantic library LibMeta. Each piece of text is labelled. Both the whole document and its separate parts can be marked. Since the problem of enriching user queries with synonyms was solved, when building a search model in conjunction with word2vec algorithms, an approach of "indexing first, then training" was used to cover more information and give more accurate results.

Keywords: Search Model, Word2vec, Synonyms, Query, Query Extension.

1 Introduction

The history of research the problems of expanding the request for the most complete coverage of information is quite long [1-8]. The problem itself is directly related to the understanding of the subject of the search, that is, the level of competence of the user and the capabilities of the information retrieval system to use expert knowledge. Ideally, the use of query enhancement and refinement functionality assumes the presence of the actual data and knowledge base and the ability to reformulate the original query in order to improve the search result.

Many approaches have been developed with the advent of artificial intelligence algorithms and corresponding programming tools in this area [9]. The first expert system using query refinement technique, Dendral [10, 11] was developed in 1965 for the analysis of chemical compounds. An example of another system based on medical expertise was MYCIN [12] presented to the scientific community in 1972. During the dialogue, MYCIN offered options for the diagnosis and further investigation of the

patient. Using about 500 inference rules, MYCIN performed at about the same level of competence as blood infection specialists and better than general practitioners.

The next stage of introducing artificial intelligence into knowledge systems is due to the use of neural network algorithms [13]. Despite the fact that the ideas of creating mathematical models based on the functioning of biological neural networks have been developing since 1943 [14], their practical implementation has gained popularity with the accumulation of digitized data, that is, already in the 21st century. Some researchers have noted this as a new era in the "partially forgotten" for the time of artificial intelligence. Search algorithms began to learn [15] on the accumulated queries, accumulate the most frequent of them, as well as the corresponding answers. All this contributed to an increase in the reaction speed of the search service, the development of targeted offers and user tips.

More complex links and structures are embedded in scientific libraries, which is dictated by the logic of subject areas and requires more careful processing of links to provide users with advanced query capabilities [16]. One such subject area is mathematics. It is of interest to study and replenish the mathematical encyclopedia, to identify unaccounted for semantic relationships of concepts and formulas.

This work is devoted to the use of shallow neural network algorithms [17] to expand the search query in mathematical subject areas based on the LibMeta [18] library, presented in the form of an ontology, and is a continuation of the authors' research in this direction [19-24]. The description of the subject area is terminologically limited to the terms of the mathematical encyclopedia [25]. As a corpus of texts, many mathematical articles are considered, which are partially supplied with codes of thematic classifiers MSC (https://msc2020.org/) and UDC (https://teacode.com/online/udc/) and correspond to a certain structure.

The LibMeta resources include a thesaurus on ordinary differential equations (ODE), dictionaries for special functions of equations of mathematical physics. All dictionaries are semantically linked to a mathematical encyclopedia [25]. These resources are used to analyze semantic relationships.

This paper presents a search model (part 2), outlines a technique based on the use of algorithms for vector representation of texts [26-29] (part 3); shows the application of the search model to add synonyms to a search query (part 4); and examples of search query extension (part 5), which demonstrate the application of the model to improve search results, also provides estimates of the completeness and accuracy of the algorithm, and also shows the process of ranking documents.

2 Search Model

The construction of the search model in LibMeta is based on three main key points, namely:

- converting documents to searchable format;
- requests are presented in a format that allows expressing the user's information needs;
- the assessment of the compliance of the document with the request.

In our case, for the preparation of documents, preprocessing of full texts was carried out to remove the publisher's markup and highlight the main parts of the text. Then a *full-text document index* was created, which allows you to efficiently load and store data and provide quick access to it. Queries written in natural language are used, which can be enriched with synonyms by the system. The assessment of the compliance of a document with a request is subjective and depends on the method used.

One of the most commonly used document and query presentation models is the *vector space model* [26-29]. In this model, one of the models based on artificial neural networks, both the request and the document are represented by a vector and the distance between them is measured, which estimates the degree of closeness of the document and the request.

In vector notation, each word is associated with a weight, which can be calculated in different ways. One of the most commonly used algorithms is the *TF-IDF* [30] algorithm, the main idea of which is that the more often a word appears in one document, the more important it is. And at the same time, the more common a word is in a corpus of documents, the less important it is. Another common model is the probabilistic model, which is based on an estimate of the likelihood that a document is relevant to a particular query. One of the popular scoring algorithms in this model is *Okapi BM25* [30, 31].

The main problem of any search model is to provide relevant results in relation to the user's information needs: from query analysis to ranking search results. This work is devoted to options for resolving this problem. One of the modern approaches is to use neural networks for text processing, since text is an example of data that can be parsed into smaller structures such as paragraphs, sentences, words, etc. depending on the text. This approach to text processing allows you to capture the semantics of the text, since closely related words or fragments of text occur in the same context and lie side by side in vector space. The search model used in this work is based on the vector representation of words and documents built using the *word2vec* [27-29] neural network algorithm [17].

Integration of neural network and index can be done in the following ways:

- first training on the corpus of texts, then indexing the texts and share them 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.

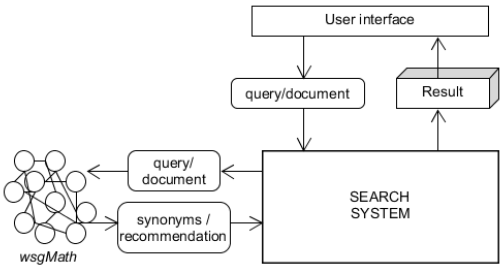

Fig. 1. Joint use of a search engine and a neural network model built on the basis of an index using word2vec algorithms.

Since we were solving the problem of enriching user queries with synonyms, in the LibMeta system we used the *"*indexing first, then training*"* approach to provide more results and more accurate results, based on extended queries on the one hand. On the other hand, using the extended version of *word2vec* in conjunction with the LibMeta search engine, it becomes possible to give users smarter recommendations based on the documents found. This approach to sharing the index and search engine and neural network allows for relevant models and ranking functions that adapt well to the underlying data. The version of the model built on the LibMeta search index using *word2vec* algorithms, hereinafter we will be abbreviated as *wsgMath*.

Figure 1 schematically illustrates the operation of a search based on a neural network, which receives a query string as input, then returns synonyms to the query using the model built by *word2vec*. In another case, a document on a vector representation can be submitted to the input, which, using the constructed model, gives recommendations in the form of a list of documents similar to it.

3 Vector Representation of Documents

Studies show [26-29] that vector representations of text are well suited for taking into account the semantics of words, but the meaning and deep semantics of text documents depend not only on the meaning of individual words. For this purpose, you need to study the semantics of phrases and longer text fragments.

For convenience, we will use the term "paragraph" to denote a paragraph, as such, but also for fragments of a paragraph or several phrases from the text. As applied to our field and the specifics of the structure of a mathematical text, these can also be theorems, lemmas, etc.

Note that the term "important" fragment will also be used. In scientific texts, this is an abstract, introduction, conclusion, theorem, etc. This term is defined since the specified elements of a scientific publication will be used as defining for documents belonging to a certain subject area.

Content for research is the resources of the LibMeta digital library [32], where, along with the accumulated original thesauri and dictionaries (for special functions,

ordinary differential equations, mixed equations of mathematical physics), a mathematical library is integrated [25].

Therefore, to construct *wsgMath*, taking into account the context for paragraphs, we used a version of the word2vec algorithm which is a generalization (extension) of the original *doc2vec* algorithm [29, 33]. For this, during training, one more component is added to the vector. Thus, when training "vectors of the word *w*", the "document vector *d*" is also trained, and upon completion of training, we obtain a vector representation of the document. As a result of the processing of the original content, the presentation of documents as a set of "related contents" was obtained. "Related content" is a semantically similar article related to articles from a mathematical encyclopedia and thesauri.

The procedure for highlighting such content will be used to offer the user semantically related documents. It is essential that without the application of the algorithm for highlighting related content, such documents will not be displayed in the search results by request, since they may not contain keywords from the query or not directly related to a certain subject area in other terms.

The peculiarity of common search models, such as the vector space model with *TF-IDF*, is that they only take into account individual terms. This approach does not always lead to optimal results because contextual information is discarded. The word context is understood as N words in the text before the word for which the vector is constructed, and N words after this word. In contrast to the *TF-IDF* model, the individual elements of the vector are not interpretable, but the distance between the vectors is investigated, which is interpreted as the semantic proximity of words.

Based on the vector representation, the proximity of the texts is measured. Using the search index and vector document representation together leverages the ability of these views to capture the semantics of text when building search models that are well adapted to the data.

The main metrics for measuring the proximity of texts are cosine distance and Euclidean distance, which are used to capture semantically similar words, sentences, paragraphs, etc.

4 Revealing Synonyms

The analysis of mathematical texts is conventionally considered as the analysis of the actual mathematical text as a whole, the analysis of formulas as a "separate language" for the representation of mathematical knowledge and the establishment of semantic links between the text and formulas. Further, only the analysis of the mathematical text as a whole is considered.

To extract synonyms for query terms from the constructed model, lexical and grammatical templates were used, which are one of the recognized methods for extracting links from text [34-38]. Based on the idea of using such patterns, we investigated the task of extracting synonyms of concepts and extracting / constructing simple patterns from them to identify relationships.

The implementation of the model consists in the application of an iterative research algorithm, which will be called *iraWsgMath* below. We list its main stages:

- *allocation of synonyms of terms*

As an example, we will demonstrate the query *"Cauchy problem"* (For the convenience of the reader, the examples have been translated into English, but the work was done for texts in Russian. In Russian the considering term is *"задача Коши"*), which consists of two words, *"problem"* and *"Cauchy"*, each of which has its own synonyms, which are presented in Table 1.

Table 1. Synonyms for each words of query *"Cauchy problem"* (задача Коши)

problem (задача)	*Cauchy* (Коши)	*Cauchy problem* (задача Коши)
equation (уравнение)	Riemann (Риман)	to define (определять)
inequality (неравенство)	boundary (краевой)	boundary (краевой)
boundary (краевой)		

The third column presents the query context as one unit *Cauchy problem* (задача Коши). Extracting its synonyms, it is clear that the list consists of words where the adjective *boundary* falls, which also occurs in the synonyms of individual words in the first two columns.

In this case, the term *"Cauchy problem"* itself has the following synonyms: *"Cauchy equation"*, *"Cauchy inequality"*, which were determined on the basis of high estimates of the proximity of the following pairs of synonyms, for example, for a pair *(problem, equation)*, the proximity estimate is *0.84*.

Note that when constructing synonymous terms, synonyms of the word *Cauchy* were not used, since it was defined as a named entity *Cauchy* based on a dictionary that includes a list of persons mentioned in the mathematical encyclopedia. But at the same time, we note that *Riemann* got into the synonyms of *Cauchy*.

- *determination of classes of synonyms by parts of speech*

Lexical and grammatical templates were used to extract synonyms for query terms from the constructed model. They are one of the recognized methods for extracting links from text [34, 35]. Based on the idea of using such patterns, we investigated the task of extracting synonyms of concepts and extracting / constructing simple patterns from them to identify relationships. Consider a link extraction pattern based on a simple adjective <term> pattern that most often indicates generic links. The original term is a generic concept, and the combination corresponding to the pattern is a specific concept [36, 37, 38].

Each word was considered separately, the synonyms are filtered by parts of speech and a possible synonym (candidate) of the term is formed from them. After that, a sentence was formed for the term and its synonyms based on the selected templates.

Based on these synonyms, the following sentences were formed, which were obtained in accordance with the pattern *"adjective <request term> <synonyms of the request term>": [Cauchy boundary value problem, Cauchy boundary equation, Cauchy boundary inequality]* (In Russian was used word *"краевой"*).

When compiling an extended query, it was also proposed to use the adjective *boundary* as a synonym, therefore additional queries were used: *[Cauchy boundary value problem, Cauchy boundary equation, Cauchy boundary inequality]*. (In Russian was used word *"краевой"*).

- *selection of patterns of "capture" of links*

The <term> verb pattern was considered to analyze and construct more complex relationship patterns in the <term> verb <term> thesaurus. Using this pattern to fill the links with it requires a separate analysis and is beyond the scope of the word2vec-based algorithm considered in this article.

In the process of training the model, several verbs were defined to identify patterns using the considered algorithm. When analyzing context-sensitive synonyms, the list of emerging verbs was rather limited, which is not surprising due to the specifics of the subject area. The list of these verbs is limited to such as: *apply, use, apply, base, prove, consider, consider, define, depend, be, embody*. Also often used are verbal nouns formed from the listed verbs: *application, use, application, basis, definition*.

- *improving the "quality of terms" and checking them*

To improve the search for extended domain terms matching templates for verbose terms in the thesaurus, possible spellings of terms were considered, for example, for "ordinary differential equation", the possible options are "ODE", "ordinary DE", etc. All possible spellings were explored as separate terms. Since there are few such terms in the studied subset of terms, they did not have a significant meaning on the results.

Validation of the model and the links it retrieves was performed based on the thesaurus ODE.

The problem of synonyms and their extraction using *word2vec* with search index is covered in more detail in [24].

5 Examples

The combined use of the full-text index [40, 41] and the search model *wsgMath* makes it possible to extend the original query with synonyms. Extending queries with synonyms without *wsgMath* requires pre-compiled synonym dictionaries. You can use resources such as WordNet (https://wordnet.princeton.edu/) or RuWordNet (https://ruwordnet.ru/ru), but the main problem is that synonyms from pre-compiled dictionaries are not tied to the data being indexed and their use does not improve the results.

Figure 2 shows the main steps of forming a model *wsgMath* for generating query synonyms in LibMeta content. The query string coming from the full-text search in-

terface goes through the *Analyzer*. *Analyzer* is a functional part of the model, where the basic operations for interacting with the *wsgMath* model are performed. All operations described in points in the previous section refer to its main functionality.

The *Analyzer* splits a string into words, analyzes and transforms them. From the *wsgMath* model, synonyms for words are extracted and filtered, an extended query is formed, with the help of which the corresponding documents are extracted from the full-text index.

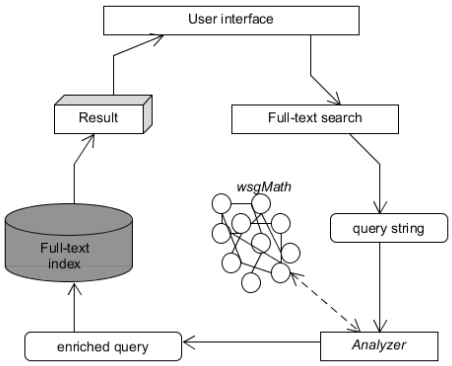

Fig. 2. Joint use of a search engine and a neural network model built on the basis of an index using word2vec algorithms to generate an extended query with synonyms.

A user's information need is defined as a chain of requests that leads him to the information he needs. Each subsequent request in this chain is a refinement of the previous one.

A real information request, as a rule, consists of an initial request and clarifications. Let's consider an example, when the primary query leads to excessive information noise, and the refinement allows you to get a more pertinent answer, and compare the search results using the *wsgMath* model and without it. For comparison, statistical characteristics are calculated (denote score) obtained using the *TF-IDF* algorithm.

The example below demonstrates three lists (List 1-3) with different scores depending on how the query is expanded. For example, when searching for the test query *"Cauchy problem"*, the user enters the qualifying query *"Cauchy boundary value problem"* and finds the information of interest. Based on the fact that the search index contains 3654 scientific articles, of which only 637 contain a mention of the *"Cauchy problem"*. Of these, 59 pieces were selected for the user, since the query words were found in significant parts of the document *(title and annotation)*. With this approach, the document of interest to the user is in 18th place. Part of the list is shown below, "score" value shows how well a document matches the request and is calculated based on statistical characteristics such as *TF-IDF*.

List 1:

1. The Cauchy problem for the system of equations of the theory of elasticity and thermoelasticity in space
score = 0.65376675
2. The Cauchy problem for the system of thermoelasticity equations in space
score = 0.64415324
…….
18. On the Well-Posedness of a Boundary Value Problem on the Line for Three Analytic Functions
score = 0.5233538

With the refinement query *"Cauchy boundary value problem"*, the list of results looks as shown below, and the document of interest is moved to the fifth position, while the number of documents satisfying the query text is reduced to 338, while the user is recommended only 20 of them.

List 2:

1. Projection procedures for non-local improvement of linearly controlled processes
score = 0.8902895
2. On one method of constructing parametric synthesis for a linear-quadratic optimal control problem
score = 0.8708762
……
5. On the Well-Posedness of a Boundary Value Problem on the Line for Three Analytic Functions
score = 0.85024154

Let us consider the situation when the query *"Cauchy problem"* is extended by synonyms and is transformed into the form *"boundary", "problem or equation or inequality Cauchy"* (in Russian: *"краевая или граничная", "задача или уравнение или неравенство Коши"*) using the *wsgMath* model. In parentheses in an extended query, synonyms are listed, connected by a logical operation *OR*. The presence of at least one of these synonyms is required. This approach insignificantly increases the completeness of the answer, and the accuracy also increases, therefore, the degree of satisfaction of the user's need increases. The list of results obtained is displayed below and the searched document is in second place. The number of documents corresponding to the request is 395 and the user will receive the desired answer already in the first positions, while the size of the issue by the system is 65.

List 3:

1. On a positive radially symmetric solution of the Dirichlet problem for one nonlinear equation and a numerical method for obtaining it

score = 0.9809638

2. On the Well-Posedness of a Boundary Value Problem on the Line for Three Analytic Functions

score = 0.9587569

3. On a positive radially symmetric solution of the Dirichlet problem for one nonlinear equation and a numerical method for obtaining it

score = 0.9512307

This example illustrates the effect of this approach already at the level of extending queries with synonyms based on indexed documents. With this approach, all suggested synonyms are found in the search engine index and the query extension is guaranteed to offer answers to the user's queries.

The use of the extended version of *word2vec* (*doc2vec* or *paragraph2vec*, in different sources in different ways) [29, 33] allows you to introduce an additional element, such as a label for a text fragment or the entire document, and based on the vectors of these labels, select similar documents not only by the exact match of keywords or terms, but based on the context of individual fragments or the entire document. As an illustration, Fig. 3 shows the main steps of this approach. This feature is used to issue documents that are close in meaning, which do not appear in the search results, but may be of interest to the user.

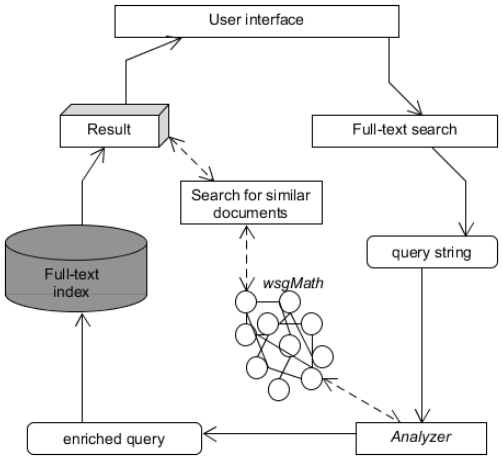

Fig. 3. Joint use of a search engine and a neural network model built on the basis of an index using *word2vec* algorithms to generate an extended query with synonyms and refine search results based on a selection of similar documents.

Let's take a closer look at 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 labels of the nearest documents are returned, the cosine distance of which exceeds a certain threshold, determined experimentally as 0.6. Below is the result of the work on the example

of the document, which in the previous example was the desired one. As the closest to it, 9 documents were found whose cosine distance exceeded 0.6.

List 4:

1. Some classes of singular integral equations solvable in closed form
cosineSimilarity = 0.8136491179466248
2. Riemann's boundary value problem for a half-plane with a coefficient exponentially decreasing at infinity
cosineSimilarity = 0.8028532266616821
3. Algorithm for constructing a quasiregular asymptotic representation of the solution of singularly perturbed linear multipoint boundary value problems with fast and slow variables
cosineSimilarity = 0.7246567010879517
4. Solution in closed form of an integral equation of convolution type in the hyperelliptic case
cosineSimilarity = 0.6468908786773682
5. On biorthogonal systems generated by some involutive operators
cosineSimilarity = 0.6454607248306274
6. On linear periodic systems in the plane having matrices of the required form
cosineSimilarity = 0.6165973544120789
7. On integral equations for the Riemann function
cosineSimilarity = 0.6134763956069946
8. Gakhov's equation for an exterior mixed inverse boundary value problem with respect to a parameter ...
cosineSimilarity = 0.6059825420379639
9. On a nonlinear integral equation of the first kind
cosineSimilarity = 0.6017340421676636

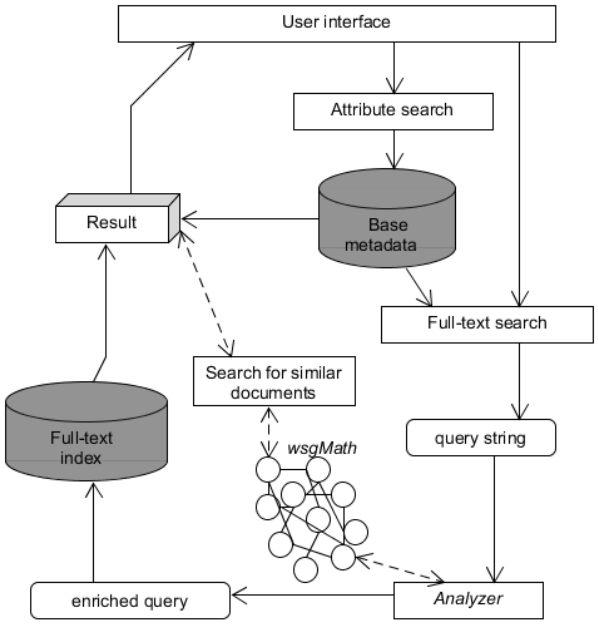

Fig. 4. Joint use of a search engine and a neural network model built on the basis of an index using *word2vec* algorithms using attribute search.

In Fig. 4 adds steps that include attribute search and how it interacts with the previously described search components. Attribute-based search delineates the boundaries in which documents are searched (by author, by year, etc.), then a transition to full-text search can be performed on them, and/or its results can also be refined based on the similarity of documents.

6 Conclusion

Vector representation of documents is proposed to expand the search query, increase the coverage of information on demand. It is shown that the quality of an answer to a request is improved by taking into account semantically close text fragments.

The model proposed in the work was tested on primary data, namely, arrays of articles not systematized by subject matter. Note that the technology of processing and thematic classification of primary data using machine learning methods has been tested. This technology can be used for the subject classification of the texts of scientific articles in Russian and the comparison of selected subjects with the English-language classification by comparing the MSC and UDC classifiers.

Integration of neural network and search indexes makes possible to give users smarter results based on the identified relations among documents.

Also, the considered search model can be used for thematic processing, both primary texts of scientific articles, and already systematized, provided with keywords and links to classifiers. In the second case, this can help to identify interdisciplinary research, as well as erroneous assignments of the subject area, since not only secondary documents, but also the texts of articles (primary documents) are taken as the basis for thematic analysis.

Aknowledgement. The work is presented in the framework of the implementation of the theme of the state assignment "Mathematical methods of data analysis and forecasting" FRC CSC of RAS and partially supported by grant #20-07-00324 of the Russian Foundation of Basic Research.

References

1. Furnas, G.W., Landauer, T.K., Gomez, L.M., and Dumais, S.T.: The vocabulary problem in human-system communication. Commun. ACM, 30(11), 964–971 (1987).
2. Biswas, G., Bezdek, J., and Oakman, R.L.: A knowledge-based approach to online document retrieval system design. In Proc. ACM SIGART Int. Symp. Methodol. pp. 112–120. Intell. Syst. (1986).
3. Voorhees, E.M.: Query expansion using lexical-semantic relations. 17th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Dublin, Ireland (1994).
4. Buckley, C., Salton, G., Allan, J., and Singhal, A.: Automatic query expansion using SMART: TREC 3, presented at the 3rd Text Retr. Conf. (TREC) (1995).
5. Efthimiadis, E.N.: Query expansion. Annu. Rev. Inf. Sci. Technol., 31(5), 121-187 (1996).
6. Guarino, N.: OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, May-June , pp. 70-80 (1999).
7. Bhogal, J., MacFarlane, A., and Smith, P.: A review of ontology based query expansion, Inf. Process. Manage., 43(4), 866-886 (2007).
8. Qui, Y., Frei, H.: Concept based query expansion. SIGIR '93 Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval Pittsburgh, Pennsylvania, USA June 27 – July 01, 1993. ACM New York, NY, USA, pp. 160–169 (1993) https://doi.org/10.1145/160688.160713.
9. Berk, A.A.: LISP: the Language of Artificial Intelligence. New York: Van Nostrand Reinhold Company, 1-25 (1985).
10. Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A., and Lederberg, J.: DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. Artificial Intelligence, 61 (2), 209-261 (1993).
11. Lederberg, J.: An Instrumentation Crisis in Biology. Stanford University Medical School. Palo Alto (1963).
12. Copeland, B.J.: "MYCIN". Encyclopedia Britannica, 21 Nov. 2018, https://www.britannica.com/technology/MYCIN, last accessed 2021/07/27.
13. Gurney, K.: An Introduction to Neural Networks. CRC Press. London and New York (1997).
14. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics 5, 115–133 (1943). https://doi.org/10.1007/BF02478259.
15. MachineLearning.ru, http://www.machinelearning.ru/, last accessed 2021/07/27.

16. Gavrilova, T.A., Horoshevskij, V.F.: Bazy znanij intellektualnyh sistem. SPb. Piter (2000).
17. Aggarwal, C.C.: Machine Learning with Shallow Neural Networks. In: Neural Networks and Deep Learning. Springer, Cham. (2018) https://doi.org/10.1007/978-3-319-94463-0_2.
18. Sererbryakov, V.A., Ataeva, O.M.: Ontology based approach to modeling of the subject domain "Mathematics" in the digital library. Lobachevskij Journal of Mathematics. 42(8), (2021). pp. 1920–1934.
19. Ataeva, O., Serebryakov, V., Tuchkova, N.: Ontological Approach: Knowledge Representation and Knowledge Extraction. Lobachevskii Journal of Mathematics. 41(10), 1938–1948 (2020) https://doi.org/10.1134/S1995080220100030 ISSN 19950802.
20. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P.: Mathematical Physics Branches: Identifying Mixed Type Equations. Lobachevskij Journal of Mathematics. 40(7), 876–886 (2019) https://doi.org/10.1134/S1995080219070047.
21. Ataeva, O.M., Sererbryakov, V.A., Tuchkova, N.P.: Mathematical Physics Problems: Thesaurus and Ontology. Selected Papers of the XXI International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019) Kazan, Russia, October 15-18. Vol-2523, pp. 158-168, (2019) http://ceur-ws.org/Vol-2523/paper16.pdf.
22. Muromskij, A.A., Tuchkova, N.P.: Predstavlenie matematicheskih ponyatij v ontologii nauchnyh znanij. Ontologiya proektirovaniy. 9(1), (31), 50-69 (2019) https://doi.org/0.18287/2223-9537-2019-9-1-50-69.
23. Ataeva, O.M., Sererbryakov, V.A., Tuchkova, N.P.: Query Expansion Method Application for Searching in Mathematical Subject Domains, 38-48 (2020) http://ceur-ws.org/Vol-2543/rpaper04.pdf, last accessed 2021/04/27.
24. Ataeva, O.M., Sererbryakov, V.A., Tuchkova, N.P.: Using Applied Ontology to Saturate Semantic Relations. Lobachevskij Journal of Mathematics. 42(8), 1776–1785 (2021).
25. Vinogradov I.M.: Mathematical Encyclopedia, Vol. 1-5, Soviet Encyclopedia, Moscow, (1982).
26. Gonçalves, A., Zhu J., Song D., Uren, V., Pacheco, R.: LRD: Latent Relation Discovery for Vector Space Expansion and Information Retrieval. Technical Report KMI-06-09. Conference: Advances in Web-Age Information Management, 7th International Conference,WAIM 2006, Hong Kong, China, June 17-19, 2006, Proceedings (2006). DOI:10.1007/11775300_11.
27. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR (2013).
28. Mikolov, T., Yih, W.T., Zweig, C.: Linguistic Regularities in Continuous Space Word Representations. Proceedings of NAACL HLT (2013).
29. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Document. International Conference on Machine Learning, pp. 1188-1196 (2014).
30. Manning, C.D., Raghavan, P., and Schütze, H.: Introduction to Information. Retrieval. Cambridge Univ. Press, Cambridge (2008).
31. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval. 3(4), 333-389 (2009). DOI:10.1561/1500000019.
32. Ataeva, O.M., Serebryakov, V.A.: Ontologiya cifrovoj semanticheskoj biblioteki LibMeta. Informatics and Applications. 12(1), 2-10 (2018).
33. Lu, Y., Zhai, Y., Luo, J., Chen, Y.: MLPV: Text Representation of Scientific Papers Based on Structural Information and Doc2vec, American Journal of Information Science and Technology. 3(3), 62-71 (2019) https://doi.org/10.11648/j.ajist.20190303.12.

34. Bullinaria, J.A., Levy, J.P.: Extracting Semantic Representations from Word Co-occurrence Statistics: A Computational Study, Behavior Research Methods, vol. 39, pp. 510–526 (2007).
35. Klaussner, C., Zhekova, D.: Lexico-syntactic patterns for automatic ontology building, Proceedings of the Second Student Research Workshop associated with RANLP, 109–114 (2011).
36. Raza, M.A., Mokhtar, R., Ahmad, N., Pasha, M., and Pasha, U.: A Taxonomy and Survey of Semantic Approaches for Query Expansion, in IEEE Access, vol. 7, pp. 17823-17833, (2019) https://doi.org/10.1109/ACCESS.2019.2894679.
37. Wang, C., Cao, L., Zhou, B.: Medical synonym extraction with concept space models https://arxiv.org/abs/1506.00528. (2015), last accessed 2021/07/27.
38. Mitchell, J., and Lapata, M.: Vector-based Models of Semantic Composition (2008).
39. Polozov I.K., Volkova I.A.: Applying word2vec technology to shifter extraction task. International research journal 4-1 (94) (2020).
40. Makinen, V.: Compact suffix array — a space-efficient full-text index. Fundamenta Informaticae 56(1–2), 191–210 (2003).
41. Makinen, V., and Navarro, G.: Compressed full-text indexes ACM Computing Surveys 39, (1), 1–79 (2007) https://doi.org/10.1145/1216370.1216372.