DOI: 10.17323/2587-814X.2021.2.60.74

The problem of loss of solutions in the task of searching similar documents: Applying terminology in the construction of a corpus vector model

Fedor V. Krasnov
E-mail: fkrasnov@naumen.ru

Irina S. Smaznevich
E-mail: ismaznevich@naumen.ru

Elena N. Baskakova
E-mail: enbaskakova@naumen.ru

NAUMEN R&D
Address: 49A, Tatishcheva Street, Ekaterinburg 620028, Russia

Abstract

This article considers the problem of finding text documents similar in meaning in the corpus. We investigate a problem arising when developing applied intelligent information systems that is non-detection of a part of solutions by the TF-IDF algorithm: one can lose some document pairs that are similar according to human assessment, but receive a low similarity assessment from the program. A modification of the algorithm, with the replacement of the complete vocabulary with a vocabulary of specific terms is proposed. The addition of thesauri when building a corpus vector model based on a ranking function has not been previously investigated; the use of thesauri has so far been studied only to improve topic models. The purpose of this work is to improve the quality of the solution by minimizing the loss of its significant part and not adding "false similar" pairs of documents. The improvement is provided by the use of a vocabulary of specific terms extracted from the text of the analyzed documents when calculating the TF-IDF values for corpus vector representation. The experiment was carried out on two corpora of structured normative and technical documents united by a subject: state standards related to information technology and to the field of railways.

The glossary of specific terms was compiled by automatic analysis of the text of the documents under consideration, and rule-based NER methods were used. It was demonstrated that the calculation of TF-IDF based on the terminology vocabulary gives more relevant results for the problem under study, which confirmed the hypothesis put forward. The proposed method is less dependent on the shortcomings of the text layer (such as recognition errors) than the calculation of the documents' proximity using the complete vocabulary of the corpus. We determined the factors that can affect the quality of the decision: the way of compiling a terminology vocabulary, the choice of the range of *n*-grams for the vocabulary, the correctness of the wording of specific terms and the validity of their inclusion in the glossary of the document. The findings can be used to solve applied problems related to the search for documents that are close in meaning, such as semantic search, taking into account the subject area, corporate search in multi-user mode, detection of hidden plagiarism, identification of contradictions in a collection of documents, determination of novelty in documents when building a knowledge base.

Key words: similarity of documents; semantic proximity; thesauri application; corpus vector model; applied intelligent information systems; algorithm explainability; similarity evaluation; text mining.

Citation: Krasnov F.V., Smaznevich I.S., Baskakova E.N. (2021) The problem of loss of solutions in the task of searching similar documents: Applying terminology in the construction of a corpus vector model. *Business Informatics*, vol. 15, no 2, pp. 60–74. DOI: 10.17323/2587-814X.2021.2.60.74

Introduction

Among the tasks that distinguish intelligent applied information systems from systems for automating business processes, there is the task of discovering insights in a large amount of information, in particular, in a company's text documents. One of the goals is to find documents that are "close in meaning." To solve this problem, a semantic model of a text corpus is built within which the similarity of documents is defined as the distance between the vector representation of documents.

One of the problems is the possible partial loss of pairs of documents that are similar in human opinion but do not satisfy the condition of exceeding the similarity threshold value set in the applied intelligent information system. This leads to the task of detecting that part of the results that do not show a sufficient degree of similarity with the existing methods, but must be taken into account from the point of view of an expert who uses the system ("true similar" pairs of documents).

When calculating the vector model of the corpus of text documents, different vocabularies are used, and their characteristics and limitations affect the quality of the solution. In particular, focus of the vocabulary on the subject area of the corpus, the proportion of frequently used and rare words, the choice of the *n*-gram range and other parameters have an impact.

It should be noted that if the vocabulary is expanded too much or the similarity threshold is reduced too much in order to include the specified "true similar" pairs of documents in the set of solutions, then the result again deteriorates. This happens because together with the return of the lost part the solutions also include unnecessary, "false similar" pairs that reveal proximity due to an insignificant part of the vocabulary (words with low weight for semantics within the given corpus). One of the methods that allows us to achieve a balance

between the inclusion of unnecessary pairs of documents in the solution and the loss of a significant part of the results is the use of thesauri of the subject area when constructing a text corpus model.

Many applied problems require calculation of similarity indicators between text samples and their constituent parts – paragraphs or sentences. The most obvious example is when a user searches for information in the system and the search engine compares the query text with the texts of previously saved documents in order to find the most relevant document. The user's query is a short text, and the system displays the documents that are most similar to the query using the ranking function.

In addition to the semantic search components, the content similarity of documents is used in the following IT solutions:

- ✦ a recommendation system for authors that determines the most suitable journal for publication;
- ✦ incoming requests routing system that selects an expert in accordance with the documents previously processed by him;
- ✦ software for building project teams for a specific technical task;
- ✦ EDMS for determining an approval route for the document based on its content.

To construct a similarity matrix, it is necessary to use a vector representation of documents, which can be created using the TF-IDF statistical measure, which is often used as a ranking function. To calculate the TF-IDF values, the entire vocabulary of the corpus is taken into account; therefore, common vocabulary may have a dominant influence on the document similarity, while the industry specificity of documents may be lost. So there arises an additional task to identify documents that are close in meaning due to specific terminology. Modification of the method can be made by changing the set of terms which are used to determine the similarity of documents. In particular, the quantity and accuracy of the results obtained can be improved by adding the domain thesaurus to the algorithm for calculating the ranking function.

This modification of the algorithm improves the quality of recommendations in information systems and accelerates user decision-making. This is done by increasing the explainability of the algorithm, which is, reducing the time that the user needs to understand why the system recommends to him some pair of documents as similar [1]. Ultimately, this improvement contributes to the growth of user confidence in the system's recommendations and simplifies his analytical work with text documents, e.g. searching for information in corporate sources, checking for duplicates in the database, detecting intersections and inconsistencies. All of this together leads to a reduction in time spent by an employee to process large volumes of text data.

The aim of this work is to improve the model based on ranking function by using a vocabulary of domain-specific terms as a thesaurus. The authors focused on the corpora of structured text documents in a specific subject area. Such documents, for example, include the regulatory and technical base of organizations [2], income and expense contracts, CVs of candidates, state standards and many others. The applied problem of finding documents with similar meaning in the text corpus was considered.

The problem of the loss of a part of the solution was investigated when calculating the proximity of documents using a ranking function. The core of this problem is that some pairs of documents are similar according to human assessment, but the program does not define them as such, since they show a low degree of similarity according to the TF-IDF algorithm, even taking into account the optimization of the vocabulary by removing the most frequent and rare words. In practice, such a problem arises in the development of applied intelligent information systems: there is a task to find in

the corpus all documents that are close enough in meaning, i.e. those whose degree of similarity exceeds a certain threshold and, therefore, is significant for the user. This paper proposes a modification of the existing methods for calculating the proximity of documents using ranking functions, thereby avoiding the loss of the specified part of the results.

The hypothesis of the research is the following: when searching in a text corpus for pairs of documents similar in meaning and using the TF-IDF ranking function for calculating proximity, the part of the solution that is lost if vector representation is made with the complete corpus vocabulary can be found when constructing a vector model based on the vocabulary of specific terms from subject area. *Figure 1* shows the hypothesis schematically.

This article includes an overview of the available research in the field of thesauri applicability and the problems of their construction, a description of research methods and experimental confirmation of the research hypothesis, as well as an analysis of the results obtained.

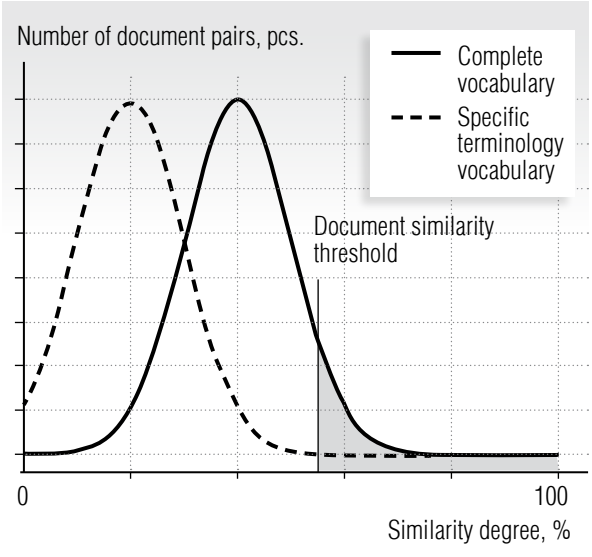

Fig. 1. Expected shift in the distribution of the degree of document similarity when replacing the complete vocabulary in the TF–IDF algorithm with a vocabulary of domain–specific terms

1. Using thesauri

For decades, computational linguistics investigated the possibility of using thesauri in text analysis to identify semantic relationships in the corpus, since the dictionaries of definitions of terms and concepts created by experts have high semantic accuracy.

The first methods proposed in the 1970s and 1980s were aimed at creating semantic language models through the analysis of computer dictionaries, which were considered as sources of taxonomic relations "hyponym – hyperonym." To extract relations from dictionaries, rule-based algorithms were used [3, 4]. The advantage of these methods was that they extracted semantic relations from reliable sources, which can be considered already partially structured, since dictionaries work as "implicit taxonomies." However, these methods inherited the problems of lexicographic data associated both with the unreliability of data due to incorrect updating of sources, and with incompleteness of dictionaries, since even today they are not always corpus.

Many studies have shown interest in finding co-hyponymic relationships, i.e. identifying groups of words that are defined using the same hyperonym, such as types of systems, equipment, nets [5, 6]. Such words are considered paradigmatically related. This means that they tend to repeat in similar syntagmatic contexts and presumably have common semantic features.

Another strategy for identifying semantic relations in the corpus is to create a language model based on the statistical analysis of texts without using any previously obtained knowledge about the dictionary relations. In [7], in order to identify pairs of hyponyms and hyperonyms, a directed graph of the coincidence of terms is constructed, and if terms are too rare, a bias is introduced into the frequency distribution: an analogy relation between hyponyms is introduced based on the composition and mor-

phological features of terms. For example, if it is statistically found that Asperger's syndrome and Carpenter's syndrome (hyponyms) are diseases (hyperonyms), then the term "Meretoya's syndrome" has the same hyperonym "disease"; the word "arthritis" can also be referred to it by analogy with "gastritis." The authors of [8] combined both strategies to create hypernymic chains using approaches based on distributive semantics.

A potential limitation of quantitative methods is the lack of accuracy, but this disadvantage can be overcome by increasing the volume of linguistic data. Therefore, this approach is considered effective and is quite popular. In addition, being language independent, it can be easily replicated and used to create multilingual resources.

Within the framework of the statistical approach, to solve many applied problems vector models of the document corpus are constructed based on ranking functions. The frequently used ranking function is TF-IDF, which is based on two concepts: the frequency of occurrence of a word (term) in a document (term frequency, TF) and the importance of this word for the entire set of documents – the inverse frequency of occurrence of a word in all documents of the corpus (inverse document frequency, IDF). This method was proposed in the 1970s [9–12] and is still widely used to analyze the similarity of texts, since it allows them to preserve certain features when projecting text data onto a numerical space. Similarity measurement can be done by calculating the proximity between TF-IDF vectors, for example in a cosine measure. The principle of comparison using TF-IDF can be called "word-by-word," since the number of components in the sparse TF-IDF vector corresponds to the number of terms in the vocabulary. However, the set of terms that determine the similarity of a document pair is formed by frequency discrimination, but not by the content analysis of these documents. TF-IDF ranking is used in various applications such as document clustering [13–16] and topic modeling [17, 18].

The use of thesauri in the text corpus analysis by the distributive semantics methods has been considered in a number of studies. In particular, various authors investigated the problem of working with the thesaurus at one of the steps of the text search algorithm. For example, the article [19] describes an approach to automatic indexing of abstracts using a subject area thesaurus. The authors of [20] presented a method for automatically extracting key phrases using semantic information about terms and phrases collected from a domain-specific thesaurus. In [21], a polythematic (i.e. covering many subjects) thesaurus is used for the purposes of semantic comparison of concepts.

A noticeable number of studies are devoted to improving the parameters of topic vector models of the corpus, in which the vector representation of documents is made on a set of topics identified as a result of text analysis on the entire corpus. For example, in [22], the improvement of the topic model is performed by artificially increasing the coincidence of synonyms, and the authors of [23] introduce information about synonyms into the prior Dirichlet distribution in order to enhance the coherence of the topics. The work [24] proposed such a concept as Thesaurus-Based Topic Model and compared various topic models. All of the above studies are united by the general result of the experiments: the elimination of hyponyms and an increase in the frequency of phrases improves human evaluation of the quality of the topics obtained, since this method gives more weight to more specific words and phrases that are better defined.

At the same time, the possibility of using thesauri to improve the semantic model of the text corpus, in which a proximity matrix is built on the basis of a ranking function (such as TF-IDF), has not yet been sufficiently studied.

2. Building thesauri

As defined by the International Organization for Standardization (ISO), a thesaurus is a vocabulary formally organized in order to establish explicit a priori relationships between concepts[1]. The elements of the thesaurus are lexical units and semantic relations (connections) between them. Thesaurus relations (genus – species, part – whole, complex – element, cause – effect) are imposed on the taxonomy structure, i.e. they are identified as the main taxonomies of the subject area.

The methodology for creating thesauri is defined in the Russian state standard "GOST R ISO 704-2010 Terminological work. Principles and methods" [25]. Historically, thesauri were created to manually index documents and were not meant for automatic indexing. The difficulty of constructing a thesaurus corresponding to the entire topic variety of indexed information makes the thesaurus a self-sufficient information product, but at the same time it is the main reason for unpopularity of thesauri in modern information systems.

As an outstanding example of specialized thesauri, the Russian thesaurus in the field of agriculture created by the Central Scientific Agricultural Library of Russia should be noted [26]. The thesaurus is built as an extension of the Russian GRNTI dictionary-classifier (state rubricator of scientific and technical information).

Labor intensity of the manual thesaurus compilation and the resulting problems can be seen on the example of such regulatory and technical documents as state standards (which are named particularly in Russia as "GOSTs"). Most of them are accompanied by a thesaurus – a section describing terms and definitions. Organizations responsible for the development of industry standards are faced with the fact that in different documents there are contradictions such as conflicting definitions of the same terms, different names for the same concept, or conflicts between the normative values of indicators. The reason for this phenomenon is that the GOST database consists of hundreds of thousands of documents and it has been accumulating over many decades. At the same time, due to the limited search capabilities, some statements of the standards were often duplicated, turning over time into contradictions due to the updating of documents without taking into account their topic links with others.

Example 1. Consider the various definitions and names of the concept of "acoustic (noise protection) screen" available in different regulatory and technical documents[2]:

1. GOST R 51943-2002. Acoustic screens for protection against traffic noise. Methods for experimental evaluation of efficiency (2002): "**Acoustic screen, screen:** Barrier (limited obstacle) installed in the path of the propagation of noise from a real source to the object to be protected from noise".

2. GOST 32957-2014. Automobile roads for general use. Acoustic screens. Technical requirements (2014): "**Acoustic screen**: An artificial barrier installed in the path of noise propagation from road transport to the object to be protected from noise. A typical acoustic baffle is a prefabricated structure consisting of the following main parts: a foundation (if provided for by the design documentation), a supporting structure (in particular, support posts) and panels. Seals, transverse profiled beams, fasteners, acoustic interconnections, canopies, wickets, gates, frames of breaks, etc. are used as additional elements".

3. GOST 33329-2015. Acoustic screens for railway transport. Technical requirements (2015): "**Acoustic screen**: An extended artificial barrier installed in the path of noise propagation from a real source (railway transport) to an object protected from noise".

[1] https://www.iso.org/standard/53657.html
[2] Translation from Russian

4. SP 338.1325800.2018. Noise protection for high speed rail lines. Design and construction rules (2018): "**Soundproof (acoustic) screen (screen)**; SS: Structures in the form of vertical or inclined extended artificial barriers of various designs, earth embankments, excavations, galleries, etc., installed along the railways on the path of traffic noise propagation to the protected object in order to reduce noise".

5. ODM 218.8.011-2018. Methodological recommendations for determining the characteristics and selection of noise protection structures for highways (2018): "**Noise protection screen (NPS)**: An extended artificial barrier installed on the path of noise propagation from road transport to an object protected from noise, the width (or thickness) of which is much less than its height consisting of a foundation and a soundproofing sheet fixed on it".

To avoid the influence of the indicated disadvantages of manually compiled thesauri on the results of searching for similar documents in applied information systems, it is necessary to use automatic methods of forming terminological dictionaries for the corpus of documents.

Existing pattern-based approaches to the extraction of semantic relations, similar to those described in the previous paragraph, are used in the field of computational terminology. For example, in [27], extraction methods are described that are focused on identifying the relations hyponym – hyperonym (particular and generalization), meronym – holonym (a part and a whole), synonyms and cause-and-effect relationships, which are collectively used to define terms and their relationships in within the data fusion pipeline approach. However, computational terminology focuses mainly on the study of the patterns of semantic relations themselves, their description, interpretation and formalization of their linguistic properties, as well as on the analysis of patterns beyond the limits of their detection. But for the problem under consideration, it is sufficient to obtain a set of specific terms that are used in the investigated corpus of documents without taking into account the semantic relationships of these terms.

3. Methods

To test the formulated hypothesis, an experiment was carried out to compare the results of solving the problem of searching for similar documents in the corpus obtained by two methods: the basic TF-IDF and its modification using the thesaurus.

In this study we considered a set of N_D structured text documents in Russian united by a subject.

As a thesaurus (as the most accessible its option), we used a dictionary T which is a vocabulary of specific terms of the subject area, obtained as a result of automatic texts analysis of the documents. To do this, all documents from the set were processed by the rule-based NER algorithm, which extracted terms from the special section of each document, where specific terms used and their descriptions are listed. Then all the obtained terms were aggregated in the domain terms vocabulary specific to this corpus.

The vocabulary included only terms consisting of one or two words. Words were brought to normal form, information about morphological features was added, numbers and English symbols were excluded.

The documents were converted to the following text format: a set of documents is presented as an array of strings M, in which one line corresponds to one document and the sequence of words was preserved.

To obtain the $\mathbf{S}_{TFIDF}$ matrix containing degrees of "word-by-word" similarity between text documents of the M corpus, the vector transformation TF-IDF was used in several variants, differing in vocabulary:

✦ All words from the corpus of documents were used as a vocabulary. Several options for constructing a complete vocabulary were con-

sidered, depending on the threshold frequency of word use within the entire corpus: if a word was encountered less than a certain number of times (parameter df_{min}), then it was excluded from the vocabulary;

✦ The vocabulary consisted of the set of specific terms *T*.

As a result, for each array, five variants of the $\mathbf{A}_{TFIDF}$ matrix of the "documents – terms" type were obtained with the dimension $N_D \times N_T$ (the number of words in the vocabulary). The vector proximity matrix (containing the degree of document similarity) $\mathbf{S}_{TFIDF}$ was calculated by the formula:

$$\mathbf{S}_{TFIDF} = \mathbf{A}_{TFIDF} \times \mathbf{A}^{T}_{TFIDF}.$$

Using the $\mathbf{A}_{TFIDF}$ matrix formed by a set of specific terms, a knowledge graph $\mathbf{G}_{TFIDF}$ for the text document corpus was built. The vertices of the graph (documents and terms) were connected by edges if the TF-IDF value for the corresponding document – term pair was higher than the threshold (the edge weight is equal to the corresponding value in the $\mathbf{A}_{TFIDF}$ matrix).

4. Experiment results

For the experiment, a set of normative documents from the GOST library was selected: $N_{D\_IT} = 667$ documents in docx format related to the field of information technology. After processing the documents in accordance with the selected methodology, an array of M_{IT} text strings was obtained. $T_{IT} = 4417$ terms were extracted from the text of the documents.

The calculation results for information technology documents are shown in *Figure 2*.

Table 1 shows the values of the mathematical expectation and standard deviation of the similarity degree of documents of pairs from M_{IT}.

The experiment confirmed the hypothesis put forward: replacing the general vocabulary with a set of specific terms in the algorithm based on the ranking function increases the number of solutions found. It can be seen from the graphs presented and from the data in the table that a model built on specific terms determines a larger number of similar documents pairs, and exclusion of words rarely found within the corpus do

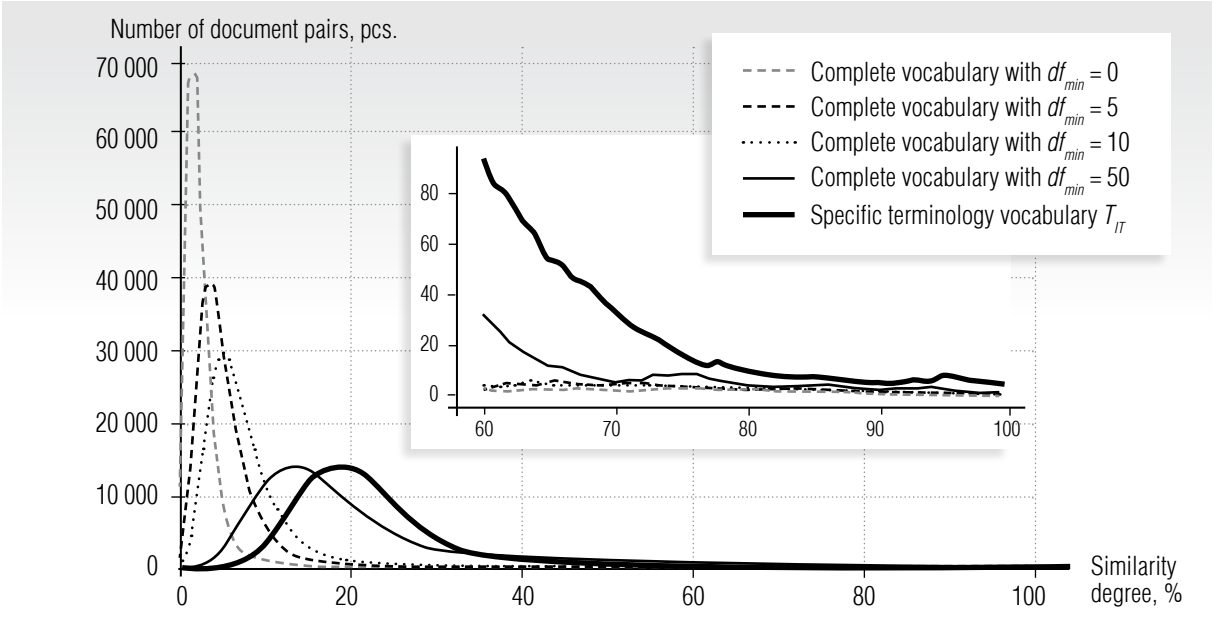

Fig. 2. The number of M_{IT} document pairs of different similarity degrees with different options of the vocabulary, incl. number of M_{IT} document pairs with similarity in the range 60–100% with different vocabulary options

not make significant differences. Therefore, for further comparison of the methods, including the analysis of the documents proximity distribution, from the first group of algorithms we chose the one with the calculation of the similarity by the complete vocabulary at $df_{min} = 5$.

For subject analysis, we consider a similarity range from 60% to 100%. The quantitative characteristics of the obtained result are presented in *Table 2*.

Example 2. *Table 3* shows examples of pairs of documents on information technology that have a high degree of TF-IDF similarity based on the vocabulary of specific terms and a low one by the complete vocabulary of the corpus.

Example 3. *Figure 3* shows one of the options for the graph of links between documents and $\mathbf{G}_{TFIDF}$ terms, namely, its fragment with the documents GOST 34.971-91 and GOST R ISO / IEC 9066-1-93 discussed above. For each of the documents, links with the five most "weighty" (significant) terms are shown.

The experiment has been repeated on another set of documents: $N_{D\_RW} = 218$ documents in docx format related to the field of railways. For these documents an array of M_{RW} text strings was generated and a set of T_{RW} terms was selected.

The results of the second experiment confirmed the conclusion that, despite the shift in the distribution curve of the number of document pairs by the similarity degree, the revealed tendency is preserved, namely: when calculating the proximity by the TF-IDF method based

Table 1.

Expected value and standard deviation of the similarity degree of M_{IT} document pairs with different vocabularies

Conditions	Expected value, %	Standard deviation, %
Complete vocabulary with $df_{min} = 0$	2.9	2.3
Complete vocabulary with $df_{min} = 5$	5.8	3.9
Complete vocabulary with $df_{min} = 10$	7.9	4.8
Complete vocabulary with $df_{min} = 50$	17.4	8.0
Specific terminology vocabulary	21.9	8.0

Table 2.

The number of pairs of M_{IT} documents in different intervals of similarity found by the TF-IDF algorithm with different vocabularies

The M_{IT} similarity degree	The number of similar document pairs determined by the complete vocabulary with $df_{min} = 5$	The number of similar document pairs determined by the specific terminology vocabulary
60% – 70%	42	459
70% – 80%	26	133
80% – 90%	25	55
90% – 100%	7	56

Table 3.

Examples of pairs of M_{IT} documents with high TF-IDF similarity based on specific terminology vocabulary and low similarity on the complete vocabulary of the corpus

No of pairs	Documents	The absolute value of TF-IDF similarity according to the specific terminology vocabulary	The absolute value of TF-IDF similarity according to the complete vocabulary
1	GOST 34.971–91 (ISO 8822–88) Information technology (IT). Open Systems Interconnection. Definition of connection–oriented presentation layer services	0.6944633	0.1744794
	GOST R ISO/IEC 9066–1–93 Information processing systems. Text transmission. Reliable transmission. Part 1. Service model and definition		
2	GOST 28147–89. Information processing systems. Cryptographic protection. Cryptographic Transformation Algorithm	0.6806692	0.17162557
	GOST R 34.13–2015 Information technology (IT). Cryptographic information protection. Modes of operation of block ciphers (with amendment)		
3	GOST R 34.964–92 (ISO 8602–87) Information technology (IT). Open Systems Interconnection. Connectionless transport protocol	0.7068537	0.1946876
	GOST R ISO/ IEC 10025–3–94 Information technology (IT). Data transfer and information exchange between systems. Connection–Mode Transport Layer Qualification Testing Using Connection–Mode Network Layer Services. Part 3. Test Management Protocol Specification		

on the specific terms vocabulary, more similar documents are found in the set than when considering the complete vocabulary of the corpus.

The difference between the document similarity matrix based on specific terms and the complete vocabulary similarity matrix is shown in *Figure 4*.

5. Discussion

The proposed algorithm gives a significant difference when searching for similar documents in the range of similarity of 60–80%. With a higher degree of document similarity, the distribution of proximity for specific terms does not differ much from the distribution of proximity for a complete vocabulary, since the concentration of terms in documents approaches the concentration of commonly used words.

From the graph in *Figure 2*, it can be seen that for a given set of documents, high proximity based on the complete vocabulary means also high proximity based on specific terminology. However, in the opposite case, searching

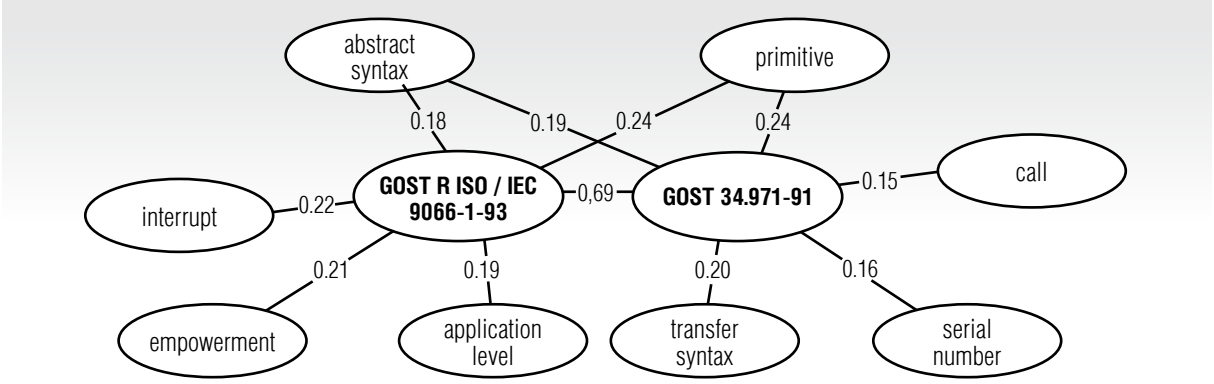

Fig. 3. Graph of a pair of documents GOST 34.971–91 and GOST R ISO/IEC 9066–1–93 and the five most significant terms for each of them

for similar documents by specific terms gives a higher degree of similarity.

The examples of documents given in *Table 3* demonstrate that the described method of calculating the similarity by the ranking function based on specific terms vocabulary can reduce the risk of losing that part of the solution where the similarity is below the threshold, but the documents are similar from the system user's point of view.

It should be noted that in the general case switching from the complete vocabulary to terminology when searching for similarity can in theory lead to the loss of that part of the solution where documents are similar according to human assessment. However, in real information systems, this loss is insignificant, since the user is primarily focused on identifying similarities within a certain topic, that is, similarities based on vocabulary specific to the subject area. Therefore, when moving from a complete vocabulary to specific terms, it is possible to lose only those pairs of similar documents that are close due to the common vocabulary. Quantitative indicators of such losses depend on the proportion of specific terms in the complete vocabulary, which varies from corpus to corpus. Various metrics can be used to calculate the contribution of common words to the overall similarity score, but their development is beyond the scope of this study.

In the experiment, no documents were found in which the similarity in the complete vocabulary was greater than the similarity in specific terms. On the graph showing the difference between the similarity based on the specific terms and on the complete vocabulary (*Figure 4*) the number of negative values is insignificant, and the difference between the similarity for such pairs of documents is hundredths of a percent.

Comparison of the results of two experiments (on the M_{IT} and M_{RW} arrays) made it possible to draw additional conclusions and highlight the factors affecting the quality of the result when searching for similar documents using vector decomposition in a vocabulary of specific terms, including one created automatically based on the same set of documents.

If documents were originally obtained as image files, then the data processing suffers from the uncertainty that appears at the stage of text recognition. This uncertainty is a key factor in the analysis of the corpus of documents. Insufficient recognition quality when part of the text is lost reduces the average similarity according to TF-IDF when using a complete vocabulary; however, according to the specific terms vocabulary such similarity of documents is still

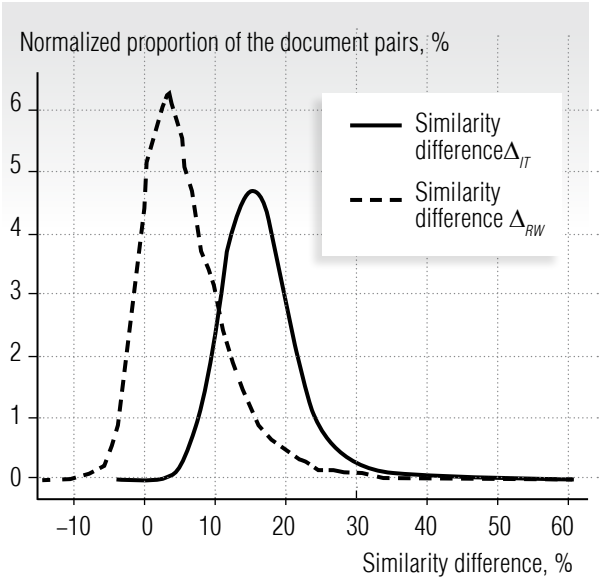

Fig. 4. Distribution of the difference in the similarity degree of document pairs calculated using two vocabularies: all words of the corpus without rare ones (complete vocabulary) and a vocabulary of specific terms (Δ_{IT} – difference in the M_{IT} corpus, Δ_{RW} – difference in the M_{RW})

detected. Thus, the proposed algorithm avoids the cost of preparing text for analysis – preliminary cleaning of the text layer and structuring the document to search for key sections.

The result also depends on the choice of the method for automatic compilation of the glossary of specific terms. The experiment showed that if these terms are separated from the text and listed in a special section of the analyzed documents, then the results are affected by the presence of the described terms in the content of the document: if the terms are only listed in the section, but are not used further, then the meaning of the similarity between the documents is distorted.

With automatic compilation of a vocabulary of specific terms, the similarity also depends on the number of errors in the terms description, since in this case part of the specific vocabulary may be lost: terms are used in the text of the document, but are not included in the vocabulary (which is used for vector representation of documents).

In the general case, the similarity of documents by specific terms depends on the choice of the range for *n*-grams when compiling a vocabulary: long terms consisting of several words probably will never be found in the text of the document entirely, when all words of the *n*-gram go in the right order and stand in a row.

Example 4. In the document from the M_{RW} set "GOST 33798.4-2016 (IEC 60077-4: 2003). Electrical equipment of railway rolling stock. Part 4. Automatic switches for alternating current. General technical conditions" the following terms are introduced among others:

- "**Manual On / Off Lever:** Lever for manually placing the circuit breaker on or off." The specified term is never used in the content of the text of this document.
- "**Indoor switch:** A switch designed for installation and use only with protection from adverse operating conditions (wind, rain, snow, increased dirt deposits, unusual environmental conditions, ice and frost)". This term is not found in the text in its full form, it is used only once and in a modified form: "Switches are classified: ... according to the type of construction, that is, switches for outdoor or indoor installation." This means that this trigram does not appear anywhere in the document, except for the very description of the term. Thus, in this case, TF value (the importance of the term for this document) changes. Therefore, when calculating the similarity of documents according to TF-IDF based on the specific terms vocabulary, it is advisable to include only unigrams and bigrams in it.

Finally, if there is an inversion in the term description (which is allowed in Russian and some other languages, for example: "protective shield" – "shield protective"), then there is also high probability that this term cannot be found in the text as the original *n*-gram.

Example 5. In the document GOST 33798.4-2016 discussed above, the term **"semiconductor switch"** is defined as "switch semiconduc-

tor." In accordance with the Russian language rules, collocations in their inversed form are often used in formal language, while in a coherent text the words of this term will most likely be reversed: the adjective will be placed before the noun. Indeed, in the main part of the document, this term is found only once, without inversion: "For each switch [there should be] indicated: the type of device (for example, an air switch, a vacuum switch, ..., a semiconductor switch ...".

The described experiment has been repeated on two text corpora of the same type. If it is carried out on any other set of documents, the result may be less expressive. This can be facilitated by both the above factors and other features of documents: their structure, stylistics of presentation, industry specificity. In addition, in the general case, the result can be influenced by the subject of the corpus, as well as etymological features of terminology, such as the use of common vocabulary as highly specialized terms. However, preliminary theoretical studies allow us to conclude that when considering any other corpus of documents united by a topic the tendency will remain the same: the number of similar documents found increases when their vector representation is based on the vocabulary of specific terms.

The approach proposed in this work has been implemented in practice within the framework of the Naumen LegalTech information system (the product is being developed with grant support of the Russian Fund for the Development of Information Technologies). It is a platform for analytical processing of a large flow of legal documents focused on users whose interests are far from the field of information technology and natural sciences: lawyers, methodologists, developers of corporate regulatory documents, authors of legislative initiatives, standardization specialists, representatives of legal expertise departments. The specificity of their work is that any mistake of the system in document processing can have legal consequences, but at the same time intelligent algorithms help a lot when applied in an advisory mode. For such users, it is especially important to trust the recommendations from the information system without plunging into the technical details of its implementation, and the explainability of decisions is one of the key factors.

Conclusion

In this work we consider the possibility of using a vocabulary of specific terms as a thesaurus in the vector representation of a text documents corpus using a ranking function.

We study the problem of a partial loss of the solution while searching in the text corpus for pairs of elements that are close in meaning, which is understood as "word-by-word" similarity of texts, taking into account the significance of terms.

The paper describes an experiment carried out on two sets of normative and technical documents. The results of the experiment showed that vector representation based on the vocabulary of specific terms allows us to reduce the loss of that part of solutions that do not meet the given condition of exceeding the threshold value of the similarity degree between two documents, but should be still recognized as significant according to human assessment.

We determined possible factors influencing the solution quality for the selected method of searching for similar documents in the corpus: the way of constructing a vocabulary of specific terms, limiting the length of an n-gram in the vocabulary, grammatical features of terms.

Document corpus models based on a specific terminology vocabulary are applicable in various IT solutions that take into account the similarity of texts: in recommendation systems, in semantic search components, when routing incoming requests. The proposed algorithm increases the accuracy and explainability of recommendations in information systems, which speeds up decision-making by users and increases the efficiency of their work with a large volume of text documents. ■

References

1. Krasnov F.V., Smaznevich I.S. (2020) The explicability factor of the algorithm in the problems of searching for the similarity of text documents. *Computational Technologies*, vol. 25, no 5, pp. 107–123 (in Russian). DOI: 10.25743/ICT.2020.25.5.009.
2. Otradnov K.K., Zhukov D.O., Novikova O.A. (2017) Clustering model of low-structured text data. *Modern Information Technologies and IT-Education*, vol. 13, no 3. Available at: http://sitito.cs.msu.ru/index.php/SITITO/article/view/295 (in Russian). DOI: 10.25559/SITITO.2017.3.439.
3. Calzolari N. (1977) An empirical approach to circularity in dictionary definitions. *Cahiers de Lexicologie*, vol. 31, no 2, pp. 118–128.
4. Amsler R.A. (1980) *The structure of the Merriam-Webster pocket dictionary*. Austin, TX: The University of Texas.
5. Grefenstette G. (1994) *Explorations in automatic thesaurus discovery*. New York: Springer Science & Business Media. DOI: 10.1007/978-1-4615-2710-7.
6. Bullinaria J.A. (2008) Semantic categorization using simple word co-occurrence statistics. Proceedings of the *ESSLLI Workshop on Distributional Lexical Semantics (ESSLLI 2008), Hamburg, Germany, 4–9 August 2008*, pp. 1–8.
7. Nazar R., Vivaldi J., Wanner L. (2012) Co-occurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del Lenguaje Natural*, no 49, pp. 67–74.
8. Santus E., Lenci A., Lu Q., Walde S.S. (2014) Chasing hypernyms in vector spaces with entropy. Proceedings of the *14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, 26–30 April 2014*, pp. 38–42.
9. Jones K.S. (1972) A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, vol. 28, no 1, pp. 11–21.
10. Jones K.S. (1973) Index term weighting. *Information Storage and Retrieval*, vol. 9, no 11, pp. 619–633.
11. Salton G., Yang C.S. (1973) *On the specification of term values in automatic indexing*. Ithaca, NY: Cornell University.
12. Salton G., Wong A., Yang C.S. (1975) A vector space model for automatic indexing. *Communications of the ACM*, vol. 18, no 11, pp. 613–620.
13. Otradnov K.K., Raev V.K. (2018) Experimental study of text documents vectorization techniques and their clustering algorithms efficiency. *Vestnik of Ryazan State Radio Engineering University*, no 64, pp. 73–84 (in Russian). DOI: 10.21667/1995-4565-2018-64-2-73-84.
14. Bafna P., Pramod D., Vaidya A. (2016) Document clustering: TF-IDF approach. Proceedings of the *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT 2016), Chennai, India, 3–5 Mach 2016*, pp. 61–66. DOI: 10.1109/ICEEOT.2016.7754750.
15. Fomin S.A., Belousov R.L. (2017) Detecting semantic duplicates in short news items. *Business Informatics*, no 2, pp. 47–56. DOI: 10.17323/1998-0663.2017.2.47.56.
16. Qaiser S., Ali R. (2018) Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, vol. 181, no 1, pp. 25–29. DOI: 10.5120/ijca2018917395.
17. Krasnov F., Ushmaev O. (2018) Exploration of hidden research directions in oil and gas industry via full text analysis of OnePetro digital library. *International Journal of Open Information Technologies*, vol. 6, no 5, pp. 7–14.
18. Kim S.-W., Gil J.-M. (2019) Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, no 9, article no 30. DOI: 10.1186/s13673-019-0192-7.
19. Evans D.A., Hersh W.R., Monarch I.A., Lefferts R.G., Handerson S.K. (1991) Automatic indexing of abstracts via natural-language processing using a simple thesaurus. *Medical Decision Making*, vol. 11, no 4 (suppl.), pp. 108–115.

20. Medelyan O., Witten I.H. (2006) Thesaurus based automatic keyphrase indexing. Proceedings of the *6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06), Chapel Hill, NC, USA, 11–15 June 2006*, pp. 296–297. DOI: 10.1145/1141753.1141819.

21. Golitsyna O.L., Maksimov N.V., Fyodorova V.A. (2016) On the definition of semantic proximity based on the links of the combined thesaurus. *Scientific and Technical Information. Series 2: Information Processes and Systems*, no 6, pp. 30–44 (in Russian).

22. Loukachevitch N., Nokel M., Ivanov K. (2017) Combining thesaurus knowledge and probabilistic topic models. Proceedings of the *6th International Conference on Analysis of Images, Social Networks, and Texts (AIST 2017), Moscow, Russia, 27–29 July 2017*, pp. 59–71.

23. Andrzejewski D., Zhu X., Craven M. (2009) Incorporating domain knowledge into topic modeling via Dirichlet forest priors. Proceedings of the *26th Annual International Conference on Machine Learning (ICML 2009), Montreal, Canada, 14–18 June 2009*, pp. 25–32. DOI: 10.1145/1553374.1553378.

24. Loukachevitch N., Ivanov K. (2018) Evaluating thesaurus-based topic models. Proceedings of the *23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018), Paris, France, 13–15 June 2018*. P. 364–376. DOI: 10.1007/978-3-319-91947-8_38.

25. Standardinform (2012) *GOST R ISO 704-2010. Terminology work. Principles and methods*. Moscow: Standardinform (in Russian).

26. Bunin M.S., Pirumova L.N. (2020) Information and search thesaurus on agriculture and food of the Central Scientific Agricultural Library, *Russian Agricultural Science*, no 5, pp. 72–75 (in Russian). DOI: 10.31857/S2500262720050178.

27. Aubin S., Hamon T. (2006) Improving term extraction with terminological resources. Proceedings of the *5th International Conference on Natural Language Processing (FinTAL 2006), Turku, Finland, 23–25 August 2006*, p. 380–387.

About the authors

Fedor V. Krasnov

Cand. Sci. (Tech.);

Expert, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail: fkrasnov@naumen.ru

ORCID: 0000-0002-9881-7371

Irina S. Smaznevich

Business Analyst, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail ismaznevich@naumen.ru

ORCID: 0000-0002-5996-4635

Elena N. Baskakova

Leading System Analyst, Department of Management Information Systems, NAUMEN R&D, 620028, 49A, Tatishcheva Street, Ekaterinburg 620028, Russia;

E-mail enbaskakova@naumen.ru

ORCID: 0000-0002-7071-8961