UDC 004.912

TECHNOLOGY FOR AUTOMATED CONSTRUCTION OF DOMAIN DICTIONARIES WITH SPECIAL PROCESSING OF SHORT DOCUMENTS

Kungurtsev O. B. – PhD, Professor, Professor of the Software Engineering Department, Odessa Polytechnic National University, Odessa, Ukraine.

Mileiko I. I. – Student of the Software Engineering Department, Odessa Polytechnic National University, Odessa, Ukraine.

Novikova N. O. – PhD, Associate Professor of the Department of Technical Cybernetics and Information Technologies named after professor R.V. Merct, Odessa National Maritime University, Odessa, Ukraine.

ABSTRACT

Context. The task of automating the construction of domain dictionaries in the process of implementing software projects based on the analysis of documents, taking into account their size and presentation form.

Objective. The goal of the work is to improve the quality of the dictionary based on the use of new technology, including special processing of short documents.

Method. A model of a short document is proposed, which presents it in the form of three parts: header, content and final. The header and final parts usually contain information not related to the subject area. Therefore, a method for extracting content based on the use of many keywords has been proposed. The size of a short document (its content) does not allow determining the frequency characteristics of words and, therefore, identifying multi-word terms, the share of which reaches 50% of all terms. To make it possible to identify terms in short documents, a method for their clustering is proposed, based on the selection of nouns and the calculation of their frequency characteristics. The resulting clusters are treated as ordinary documents, since their size allows for the selection of multi-word terms. To highlight terms, it is proposed to select sequences of words containing nouns in the text. Analysis of the frequency of repetition of such sequences allows us to identify multi-word terms. To determine the interpretation of terms, a previously developed method of automated search for interpretations in dictionaries was used.

Results. Based on the proposed model and methods, software was created to build a domain dictionary and a number of experiments were conducted to confirm the effectiveness of the developed solutions.

Conclusions. The experiments carried out confirmed the performance of the proposed software and allow us to recommend it for use in practice for creating dictionaries of the subject area of various information systems. Prospects for further research may include the construction of corporate search systems based on dictionaries of terms and document clustering.

KEYWORDS: domain dictionary, information system, term, clustering, information technology, short document.

ABBREVIATIONS

DD is domain dictionary;
IS is information system;
FCA is Formal Concept Analysis;
LDA is Latent Dirichlet Allocation;
OS is organizational system.

NOMENCLATURE

addr is a location of the document;
Cd_i is a i -th cluster of the content of short documents;
d_j is a document included in the cluster;
Ds is a set of short document;
ds_i is a document;
$Ks_{j,p}$ is a coefficient of proximity of the document d_j with the central document of the cluster Cd_p;
kw_i is a tuple of the set of keywords;
mKw is a lists (sets) of keywords;
Mt_i is a set of one-word terms of the document d_i;
Nc is a number of the last words of the document;
Ncd is a number of words in the document;
Ncorp is a number of documents in the corpus;
Nh is a number of the first words of the document;
nm_i is a number of different terms in the document;
nm_q is a number of occurrences of the term t_q in the document d_i;
ns is a quantity of documents;
nw_i is a size of the document in words;
Nww is a number of "erroneous" words;
Qs is a quality of selection;
Qsa is a quality of separating the content for the corpus of documents;
$r1$ is an index of the first word of the content of the document;
$r2$ is an index of the last word of the content of the document;
Tc_i is a concatenation of the texts of the documents included in it;
$text_i$ is a text of the document;
t_q is a term represented by a noun.

INTRODUCTION

The DD is one of the first artifacts that is created in the process of designing an IS. DD allows the Developer and the Customer to determine a common language of communication [1]. With its help, requirements for IS are formulated, user interfaces are created, instructions are written [2]. Such dictionary is recommended to be used in various subject areas [3]. The creation of DD primarily involves the definition of terms. The simplest way to highlight terms is to study the texts of documents [4] that represent the subject area of IS. Manual text analysis is a very time-consuming process that requires special knowledge in the field of linguistics. Therefore, in recent

DOI 10.15588/1607-3274-2023-4-14

years, more and more attention has been paid to the automated selection of terms and their interpretation [5].

The condition for a reasonable selection of a term is its repeated occurrence in the text of the analyzed document. Only in this case it is possible to distinguish stable phrases. The smaller the document size, the lower the probability of correct selection of terms in it.

Fig. 1 shows the existing technology for building DD, which does not provide for special processing of short documents.

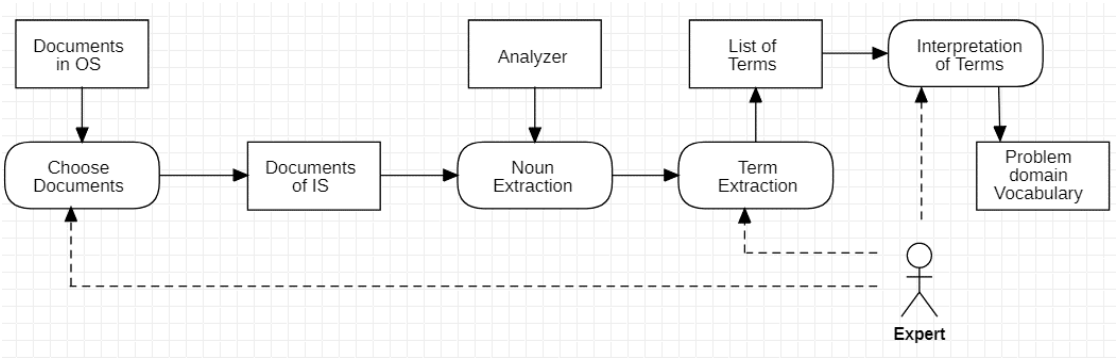

Figure 1 – The existing technology for creating DD

A feature of the corpus of documents used to build DD is many short documents (letters, orders, instructions, etc.) related to different topics, in the structure of which a significant proportion is occupied by the heading and closing parts with terms that do not correspond to the topic of the document. The concept of a short document does not have a clear quantitative definition in terms of isolating terms from its text. Short documents often contain formal header and tail sections that need to be excluded from analysis. To build DD, it is necessary to develop a separate methodology for processing short documents.

1 PROBLEM STATEMENT

Suppose that there are many documents used to develop an automated information system –

$D = Do \cup Ds$, where Do – is a regular document, and Ds – is a short. Since it is impossible to correctly extract a term from a short document, it is necessary to perform a number of transformations on the set Ds:

$Ds \Rightarrow Ds'$, where Ds' – the set of short documents without header and tail parts;

$Ds' \Rightarrow CDs$, where CDs – the set clusters of short documents.

A sufficiently large cluster text can be considered as a regular document. Then $D' = Do \cup CDs$ – a virtual set of documents from which one can get a set of terms – *Term*, that define subject area: $D' \Rightarrow Term$.

2 LITERATURE REVIEW

In [6], the problem of the low frequency of keywords in short documents was noted. An algorithm is proposed that uses the domain ontology to calculate the semantic distance between short documents. The question of ontology formation remained open. In articles [7,8] was proposed a method that uses thematic models that make short text less sparse and more thematically oriented for classification. These methods are difficult to apply to DD, which is created at the first stage of IS design. The task of analyzing short texts is relevant for Web applications such as social networks, forums, and blogs. To solve the problem, an extension of terms, also known as an extension of documents, was proposed [9], based on the classification of texts and their semantic analysis. The paper does not specify the way of obtaining the initial information, but the proposals deserve attention. In [10] a system that can perform automatic summation of several documents using semantic text analysis, clustering, based on the representation of a document as a set of triplets (subject, verb, object) is proposed. The disadvantage of this solution is the rather complicated text analysis and system configuration for a specific user. In [11], dynamic clustering is proposed, which allows you to track the time-varying distribution of topics across documents and words across topics. From the point of view of the formation of DD, the method is difficult to apply due to insufficient time and the number of texts for training. In article [12] is used the concept of weighted similarity scores of terms in Formal Concept Analysis and was explored its effectiveness for expanding terms. It is shown that the weighted measures of similarity of terms, when choosing the appropriate weight value, give a good result. The material is of practical interest and will be partially implemented in this work.

Analysis of the effectiveness of using two approaches to expand terms: weighted measures of similarity, studied in FCA, and a number of measures of correlation, often used in data mining was carried in [13]. It has been empirically shown that the cosine correlation measure is superior to other methods for short documents. The paper [14] describes an experiment comparing short document classifiers based on two methods: Latent Dirichlet Allocation (LDA) and Formal Concept Analysis (FCA). It has been shown that FCA leads to a much higher degree of correlation between terms than LDA, and initially gives a lower classification accuracy. The disadvantage of

DOI 10.15588/1607-3274-2023-4-14

the considered methods is a long and laborious learning process.

In [15], a preliminary clustering method is proposed, which allows one to limit oneself to only a corpus of documents representing the subject area of the designed IS. However, in this work, as well as in [16, 17], there is no clear definition of a short document.

Short documents used to build DD are significantly different from short texts on the Internet [18]. The workflow of most organizational structures is dominated by formalized documents [19] with heading and closing parts. The need to highlight the meaningful part of a short document was noted in [15], but a clear algorithm for solving this problem is not presented.

3 MATERIALS AND METHODS

Short document model. In accordance with [15], we will assume that the corpus of all documents under study is presented in the form

$$D = \{d_i\}, \ i = 1, n. \tag{1}$$

Let's extend the previously used model to represent a single document. Since the documents based on which the DD is built can be located on different computers, disks and directories in the Customer's organization, it is necessary to introduce the concept of an address for a document. To highlight the content of the document, you should limit the search area of the heading and final parts of the document. Thus, document can be represented by a tuple

$$D_i = <addr, text_i, r1, r2, nw_i, Mt_i>, \tag{2}$$

$$Mt_i = \{< t_q, nm_q>\}, \quad q = 1, nm_i. \tag{3}$$

To apply the model, it was necessary to clarify the concept of a "short document". As a result of the experiments (see the Results section), it was proposed to consider a document up to 1400 words as short. Thus, the set of short documents can be represented as

$$Ds = \{d_i \mid d_i \in D \wedge d_i.nw_i \le 1400\}, \ i = 1, ns. \tag{4}$$

A method for highlighting the content of a short document. To highlight the meaningful part, it is necessary to have signs of the heading and closing parts. Such signs are certain "keywords" of these parts of the document. It should be noted that the list of all formalized documents for a certain state has hundreds of names. It is undesirable to use such a list in the algorithm for highlighting the content of a document, since the interpretation of keywords from one subject area may not coincide with their interpretation in another. Therefore, it makes sense to compose sets of keywords for the heading and closing parts in relation to the subject area (perhaps in the narrow sense of the word).

In general, within the framework of one project of *IS*, several lists of *mKw* (sets) of keywords can be created:

$$mKw = \{kw_i\}, \ i = 1, q, \tag{5}$$

where $kw_i = < mKwHead, mKwEnd >$ is a tuple of the set of keywords in the header part of documents; $mKwHead_i = \{ w_{i,j} \}, j = 1, qh_i$ and the corresponding set of keywords of the final part of the documents; $mKwEnd = \{ w_{i,j} \}, j = 1, qe_i$.

For example, for the personnel department of a university, the set *mKw* will look like:

$$mKw = \{\{labor\ contract, order\}, \{signature, date\}, \ldots, \{\}\}.$$

Thus, a short document can be presented in the form

$$Ds_i = <mWhead, mWcontent, mWend>, \tag{6}$$

where the heading is *mWhead* , represented by an ordered set of words

$$mWhead = \{w_1, w_2, \ldots, w_i, \ldots, w_p\},$$

while $\exists\, w_i \mid w_i \in mKwHead \wedge w_i \in mWhead$;
and the final part *mWend* , represented by an ordered set of words

$mWend = \{w_r, w_{r+1}, \ldots, w_j, \ldots, w_z\}$, at the same time $\exists\, w_j \mid w_j \in mKwEnd \wedge w_j \in mWEnd$.

For example, the following heading keywords can be distinguished from personnel documentation: "Agreement", "Order", "Card", "Order", "Time sheet", "Statement", "Act", "Schedule", "Note" and the following the words of the final part: "Signature"/"Signatures", "Acquainted", "Approve".

To highlight the content part, you need to determine its first and last words. To do this, it is proposed to determine the possible boundaries of the heading and closing parts by searching for terms from *mKw*. Terms from *mKwHead* are searched from the beginning to the end of the document, and terms from *mKwEnd* are searched from the end to the beginning of the document. It is proposed to limit the search area of the heading and final parts of the document to reduce the probability of errors in the case when the document does not belong to the category of formalized ones. In addition, limiting the search area reduces the time for document analysis.

It is not possible to analytically determine the boundaries of the header and footer search for many different documents, so experimental studies were carried out on two sets of documents of various formats. 74 documents in Russian and 68 documents in English from the trade organization's workflow were processed. The number of words for highlighting the heading part *Nh* and the closing part *Nc* was set equal to 5, 15, 25, 30, 35 and 50. The quality of highlighting the content part was assessed by an expert depending on the number of "extra" and "missing" words in the content part. Under the quality

DOI 10.15588/1607-3274-2023-4-14

of selection for a separate document, it is proposed to understand *Qs*, calculated by the formula:

$$Qs = \frac{100 \times (Ncd - Nww)}{Ncd}. \quad (7)$$

Quality of separating the content for the corpus of documents for certain values of *Nh* and *Nc* by the formula

$$Qsa = \left| \frac{\sum_{i=1}^{Ncorp} Qs_i}{Ncorp} \right|. \quad (8)$$

It has been experimentally shown that the best value for *Nh* is 35, and for *Nc* – 25 (see the Results section). Let us formulate the steps of the method.

– Find among the first *Nh* words of the document the word $w_i \mid w_i \in mWhead$. If the word is found, then the index of the first word of the substantive part $StartInd = i+1$ is, otherwise $StartInd = 0$.

– Find the word among the last *Nc* words of the document $w_j \mid w_j \in mWend$. If the word is found, then the index of the last word of the substantive part $EndInd = j - 1$ is, otherwise $EndInd = Ncd$.

– Crop the document at the edges – before the index *fInd* and after the index *lInd* .

– Find how many characters need to be further indented after the beginning of the truncated document to remove lines that have less than five words or less than 50 characters.

– Crop the document according to the received data.

Clustering short documents. In accordance with a previous study (Fig. 2), in short documents the average frequency of repetition of nouns is low, which does not allow to qualitatively distinguish verbose terms. Therefore, it is proposed to define terms not within a single document, but within a cluster of short documents. For this purpose, the preliminary clustering method was used [15]. The method allows you to calculate the proximity coefficient $K_{i,j} = K_{1i,j} + \gamma * K_{2i,j}$ of documents d_i and d_j based on the relative number of matching nouns (component $K_{1i,j}$) and the frequency of matching nouns (component $\gamma * K_{2i,j}$). Optimization of the composition of clusters is ensured by adjusting their composition depending on the proximity of the document to the cluster core.

Let us represent the clustering process in the form $Ds \xrightarrow{clustering} \{Cd_i\}$.

The practical use of the short document clustering method [13] showed that after the completion of clustering based on kernels, several clusters Cd_i are formed that contain only one document, that is

$$Cd_i = \{d_j\}, j = 1. \quad (9)$$

Therefore, an additional stage is introduced into the method, at which for each document *dj* a cluster Cd_p is found to which it can be attached.

$$d_j \xrightarrow{add} Cd_p.$$

In this case, the next condition is fulfilled

$$Ks_{j,p} = \max_{q=1,n;q \neq j} Ks_{j,q}.$$

Extracting terms from a cluster

Multiword terms make up a significant part of all terms. The work [20] shows that 29.13% of the terms from the Internet request contain three or more words. In the documents of organizations, terms containing two and three words make up about 50% of all terms. This determines the need to extract multiword terms from short documents.

Let us represent some cluster Cd_i in the form

$$Cd_i = \{d_j\}, j = 1, n_j.$$

Let's form the text of the cluster Cd_i.

$Tc_i = \mathrm{U}_1^{n_j} d_j$, as a concatenation of the texts of the documents included in it.

Let's represent the cluster text as a sequence of elements

$$Tc_i = e_1 e_2 \ldots e_k \ldots e_{q-1} e_q.$$

The text element can be a word or a punctuation mark.

Let us denote the text element corresponding to the noun as eN, and the text of the cluster as

$$Tc_i = e_1 e_2 \ldots e_{k-2}\, e_{k-1}\, eN_k e_{k+1}\, e_{k+2}\, eN_{k+3}\, e_{k+4} \ldots e_{q-1} e_q.$$

To highlight multiword terms in each sentence of the text, sequences of words containing nouns are selected. Let there be some sequence of words, which is bounded on the left and right by punctuation marks:

$$S = e_{k-2}\, e_{k-1}\, eN_k e_{k+1}\, e_{k+2}\, eN_{k+3}\, e_{k+4}.$$

If we take a noun eN_k as a base, standing at the beginning or end of a term, the following sequences of words, which can be terms will be selected from *S*: $e_{k-2}\, e_{k-1}\, eN_k$, $e_{k-1}\, eN_k$, eN_k , $eN_k\, e_{k+1}$, $eN_k\, e_{k+1}\, e_{k+2}$, $eN_k\, e_{k+1}\, e_{k+2}\, eN_{k+3}$, $eN_k\, e_{k+1}\, e_{k+2}\, eN_{k+3}\, e_{k+4}$ At the same time, the term cannot begin and end with a union, a preposition and a numeral, and these parts of speech are not considered "important". In this work, it was decided to limit the length of the term to three "important" words.

Definition of interpretations of the term.

Defining definitions for terms is a long and laborious process [21], which it is desirable to automate [22] . Since such a task is beyond the scope of this study, it is proposed to use a ready-made solution [23], which provides a detailed analysis of a dictionary entry, automated removal of redundant definitions, and a simple procedure for expanding the dictionary bank (software product DictionaryOfInterpretations).

Technology for creating DD with separate processing of short documents. Given the large number of short documents in the corpus of documents to be analyzed for the construction of DDs, it is proposed to introduce additional procedures for processing short documents (Fig. 2).

DOI 10.15588/1607-3274-2023-4-14

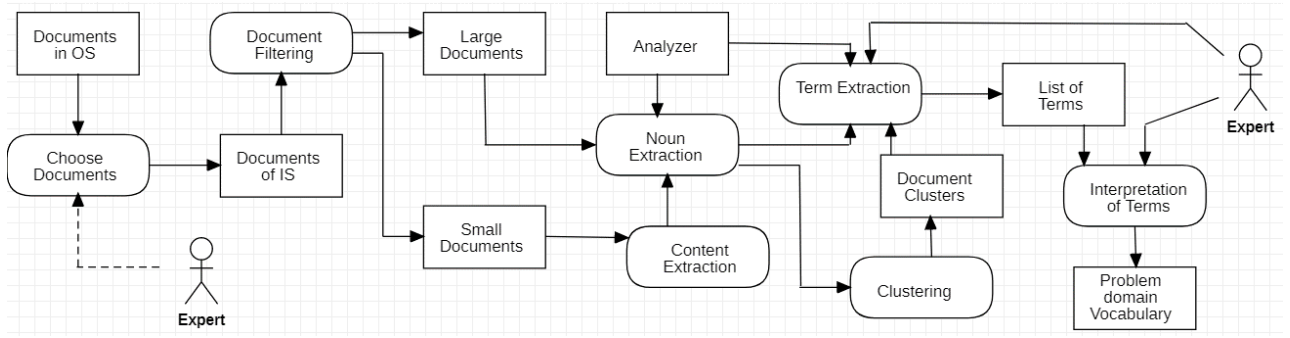

Figure 2 – New technology for building DD

According to Figure 2, the procedure for building an DD is as follows:

– The expert (representative of the customer) selects documents of the organizational system (*OS*) that are of interest to the designed *IS*.

– As a result of filtering, short documents are selected from the entire corpus of documents.

– From short documents, their substantive part stands out.

– Using the analyzer, nouns are highlighted in texts and the number of their occurrences in documents is counted.

– For short documents, clustering is performed, because of which clusters are formed according to the principle of belonging to one topic.

– From documents (large) and clusters of documents (short), terms (generally multiword) are distinguished. The expert analyzes and edits the list of terms in terms of their belonging to the subject area of the projected IS.

– Based on the received list of terms, the user himself or with the help of an external system performs the interpretation of the terms.

4 EXPERIMENTS

Development of a software product. To implement the proposed technology, the software product TerEx was developed , the general class diagram of which is shown in Fig. 3.

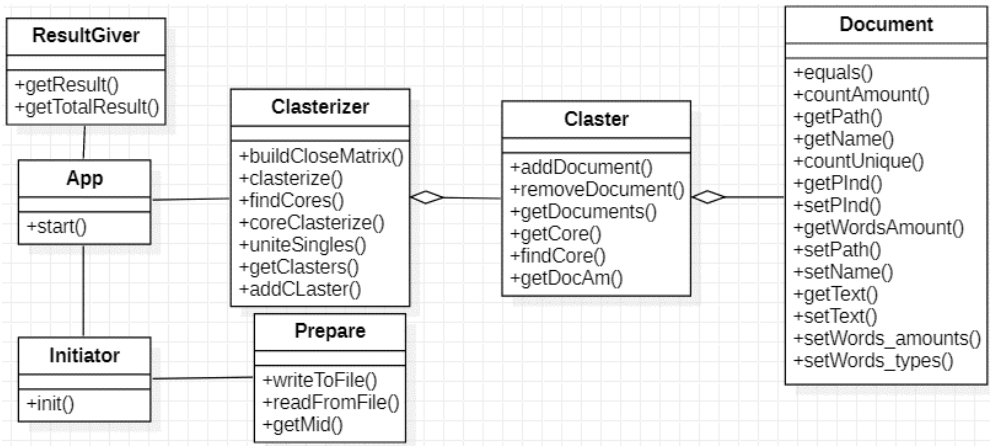

Figure 3 – Class diagram of the TerEx system

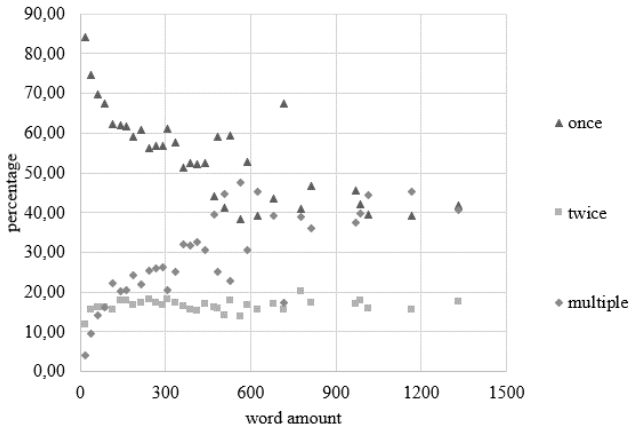

Figure 4 – Graph of the dependence of the frequency of repetition of nouns on the length of the document

DOI 10.15588/1607-3274-2023-4-14

TerEx allows you to perform the following actions on documents:

– highlight individual words, conduct their morphological analysis, count the number of occurrences in texts;

– highlight the content of the document;

– perform clustering of documents;

– highlight single-word and multi-word terms from the text.

5 RESULTS

Definition of a short document. We studied a corpus of 381 documents in Ukrainian, English and Russian, containing from 15 to 1332 nouns. The results of the experiment are presented in the form of a graph (Fig. 4), based on which it was concluded that documents up to 1300–1400 words in size should be considered short. The experiment showed that the subject area and the language of the documents do not have a noticeable effect on the results obtained.

Determining the quality of highlighting the content of a short document.

To determine the dependence of the quality of highlighting the content of the document on the values of *Nh* and *Nc*, 74 short documents were analyzed. The results are shown in Table 1. The best quality is achieved with a limit of 35 words from the beginning of the document and 25 words from the end of the document.

Table 1 – Dependence of the quality of highlighting the meaningful part on the values of *Nh and Nc*

Nh	*Qs*	*Nc*	*Qs*
35	97.14	25	95.64
30	96.16	35	95.41
25	96.06	30	94.99
15	89.16	15	89.79
5	84.16	5	82.98
50	82.37	50	824

After setting these values, the average correctness of extracting the meaningful part of 96.39% was obtained.

Determining the quality of clustering. When requesting clustering, a new folder “ClusterizationResultFolder” is created in the directory selected for displaying the result, in which the file “_ totalTerms . txt ”, containing a general list of terms from all clusters, as well as new folders – Cluster _1 ... Cluster _ N , where N is the resulting number of clusters is created. An example of the contents of the “ClusterizationResultFolder” directory and the structure of the files before they are processed is shown in Figure 5. In the folder of each cluster is the file “_ terms.txt”, containing a list of terms in this cluster, as well as copies of all documents included in this cluster.

104 documents were subject to analysis. Five clusters were found. Analysis of the result showed that the documents in the selected clusters are really close to each other in terms of the nouns they contain. No errors were found.

Determining the quality of highlighting multi-word terms. A comparative analysis of the results of the work of the TerEx product and the well-known online service for text analysis SketchEngine [24] was carried out. A corpus of 100 documents was studied. The results of the experiment are presented in Table 2.

Defective terms are those that contain extraneous characters, cut words, prepositions, or are not words at all.

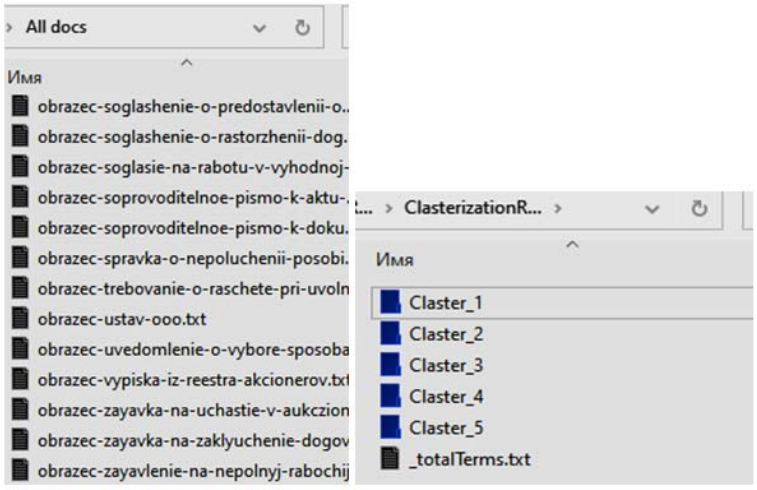

Figure 5 – File structure before (left) and after (right) processing

DOI 10.15588/1607-3274-2023-4-14

Table 2 – Comparative analysis of the quality of term selection

	TerEx	SketchEngine
Number of highlighted terms	11119	15315
Number of defective terms	585	2095
Error Percentage	5.26%	13.68%
Elapsed time	4 minutes 45 seconds	5 minutes 12 seconds
File limit	No limits	100 files
License	Is free	Paid subscription
Distribution of documents by clusters	Performed	Not supported

6 DISCUSSION

The experiments carried out made it possible to quantify a short document in terms of specific requirements for its analysis. The conditions for the implementation of the method of highlighting the content of a short document were identified, and its effectiveness was shown (96.39%). The implemented clustering of short documents can be of independent importance, for example, in information retrieval tasks. However, in this study, it allowed to significantly improve the quality of multiword selection, which was confirmed by the results of comparative tests of TerEx and Sketch Engine (Table 2).

CONCLUSIONS

An information technology for constructing a dictionary of the subject area has been developed, which provides for special processing of short documents.

The scientific novelty of the research lies in:

– improvement of the mathematical model of a short document, by introducing indexing of the beginning and end of the content, as well as the address of the location of the document, which made it possible to further formalize operations for its processing;

– obtaining an experimentally substantiated definition of a short document, which allows you to sort documents quickly and efficiently for further processing.

– development of a method for highlighting the content of a short document that implements the exclusion from the document of the heading and ending parts that contain terms not related to the subject of the document, which made it possible to further improve the quality of *DD*;

– improvement of the method of preliminary clustering of short documents by introducing an additional stage of merging clusters containing 1 document, which made it possible to increase the frequency of terms in clusters

The practical value of the work lies in combining the model and methods into a single technology for creating *DD*.

The conducted experiments confirmed the effectiveness of the theoretical results of the work.

The practical implementation of models and methods can be used to create *DD* in various subject areas.

REFERENCES

1. Larman K. Primenenie UML 2.0 i shablonov proektirovanija. Prakticheskoe rukovodstvo. 3-e izdanie. Moscow, Izdatel'skij dom "Vil'jams", 2013, 736 p. [in Russian].
2. Bourgeois D., Mortati J., Wang S., et al. Information Systems for Business and Beyond. Information systems, their use in business, and the larger impact they are having on our world [Electronic resource]. Access mode: https://opentextbook.site/exports/ ISBB-2019.pdf
3. Artamonov A., Kshnyakov D., Danilova V. et al. Methodology for the Development of Dictionaries for Automated Classification System, *8th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2017). Procedia Computer Science Volume:123*. Moscow, Russia, 1–6 August 2017, pp. 57–62. doi:10.1016/j.procs.2018.01.010
4. Dalglish S. L., Khalid H., McMahon S. A. Document analysis in health policy research: the READ approach, *Health Policy and Planning,* 2020, Vol. 35, Issue 10, pp. 1424–1431.
5. Cheng Y. Huang Y. Research and development of domain dictionary construction system, *Proceedings of the International Conference on Web Intelligence, August* 2017, pp. 1162–1165. https://doi.org/10.1145/3106426.3109046
6. Liang S., Yilmaz E., Kanoulas E. Dynamic Clustering of Streaming Short Documents, *International Conference on Knowledge Discovery and Data Mining, August 2016*, pp. 995–1004.
7. Wang Y., Yang S. Outlier detection from massive short documents using domain ontology. *International Conference on Intelligent Computing and Intelligent Systems, 29–31 Oct. 2010*, Xiamen, China. DOI: 10.1109/ICICISYS.2010.5658426
8. Shi T., Kang K., Choo J. et al. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations, *WWW '18: Proceedings of The Web Conference 2018, April 2018*. Lyon, France. DOI: 10.1145/3178876.3186009
9. Hafeez R., Khan S., Abbas M. et al. Topic based Summarization of Multiple Documents using Semantic Analysis and Clustering, *International Conference on Smart; 8–10 Oct. 2018*. Islamabad, Pakistan. DOI: 10.1109/HONET.2018.8551325
10. Vo D-T., Ock C-Y. Learning to classify short text from scientific documents using topic models with various types of knowledge, *Expert Systems with Applications: An International Journal,* 2015, V. 42, Issue 3, pp. 1684–1698. https://doi.org/10.1016/j.eswa.2014.09.031

DOI 10.15588/1607-3274-2023-4-14

11. Liang S., Yilmaz E., Kanoulas E. Dynamic Clustering of Streaming Short Documents, *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, US, 13 August 2016*, pp. 995–1004. https://doi.org/10.1145/2939672.2939748
12. Seki H., Toriyama S. On Term Similarity Measures for Short Text Classification, *11th International Workshop on Computational Intelligence and Applications (IWCIA), 9–10 Nov.2019*, pp. 53–58, DOI: 10.1109/IWCIA47330.2019.8955045.
13. Seki H., Toriyama S. Using term similarity measures for classifying short document data, *International Journal of Computational Intelligence Studies*, Vol. 10, Issue 2–3, https://doi.org/10.1504/IJCISTUDIES.2021.115430
14. Rogers N., Longo L. A. Comparison on the Classification of Short-text Documents Using Latent Dirichlet Allocation and Formal Concept Analysis, *25th Irish Conference on Artificial Intelligence and Cognitive Science, AICS 2017*. Dublin, Ireland, 7–8 December 2017, V. 2086, pp. 50–62.
15. Kungurtsev O., Zinovatna S., Potochniak I. et al. Development of Methods for Pre-clustering and Virtual Merging of Short Documents for Building Domain Dictionaries, *Eastern-european Journal of Enterprise Technologies*, 2020, Vol. 5, № 2 (107), pp. 39–47. http://doi.org/10.15587/1729-4061.2020.215190
16. Kungurtsev O., Zinovatnaya S., Potochniak Ia. et al. Development of information technology of term extraction from documents in natural language, *Eastern-European Journal of Enterprise Technologies*, 2018, V. 6, №2 (96), pp. 44–51. doi: https://doi.org/10.15587/1729-4061.2018.147978
17. García R. G., Beltrán B., Vilariño D. et al. Comparison of Clustering Algorithms in Text Clustering Tasks, *Computación y Sistemas*, 2021, Vol. 24, № 2. https://doi.org/10.13053/cys-24-2-3369
18. Shevchenko A. Organizacija elektronnogo dokumentoobigu na pidpryjemstvi [Electronic resource]. Access mode: https://uteka.ua/ua/publication/commerce-12-dokumentooborot-2-organizaciya-elektronnogo-dokumentooborota-na-predpriyatii [in Ukrainian].
19. Typova instrukcija z dilovodstva v ministerstvah, inshyh central'nyh ta miscevyh organah vykonavchoi' vlady [Electronic resource]. Access mode: https://borispol-rada.gov.ua/item/39961-typova-instruktsiya-z-dilovodstva-v-ministerstvakh-inshykh-tsentralnykh-ta-mistsevykh-orhanakh-vykonavchoi-vlady.html [in Ukrainian].
20. Lions K. Long-Tail Keywords: What They Are & How to Use Them for SEO. [Electronic resource]. Access mode: https://www.semrush.com/blog/how-to-choose-long-tail-keywords/
21. Borysova, N. V., Kanyshheva O. V., Kanyshheva O. V. The formation of problem domain dictionary, *Eastern-European Journal of Enterprise Technologies*, 2013. Vol. 5, №3(65), pp. 16–19. https://doi.org/10.15587/1729-4061.2013.18
22. Rahoo L. A., Unar M. A. Design and Development of an Automated Library Management System for Mehran University Library, Jamshoro, *Control Theory and Informatics,* 2016, № 6(1), pp. 1–6.
23. Kungurtsev O., Novikova N., Kozhushan M. Automation of Serching for Terms in the Explanatory Dictionary, *Proceedings of Odessa Polytechnic University*, 2020, № 3(62), pp. 91–100. DOI: 10.15276/opu.3.62.2020.11
24. Sketch Engine. [Electronic resource]. Access mode: https://www.sketchengine.eu/

Received 14.09.2023.
Accepted 06.11.2023.

УДК 004.912

ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ПОБУДУВАННЯ СЛОВНИКІВ ПРЕДМЕТНОЇ ГАЛУЗІ ЗІ СПЕЦІАЛЬНОЮ ОБРОБКОЮ КОРОТКИХ ДОКУМЕНТІВ

Кунгурцев О. Б. – канд. техн. наук, професор кафедри Інженерії програмного забезпечення Національного університету «Одеська політехніка», м. Одеса, Україна.

Милейко І. І. – студентка кафедри Інженерії програмного забезпечення Національного університету «Одеська політехніка», м. Одеса, Україна.

Новікова Н. О. – канд. техн. наук, доцент кафедри Технічна кібернетика й інформаційні технології ім. професора Р. В. Меркта Одеського національного морського університету, м. Одеса, Україна.

АНОТАЦІЯ

Актуальність. Розглянуто завдання автоматизації побудови словників предметної галузі у процесі виконання програмних проектів на основі аналізу документів з урахуванням їх розміру та форми подання.

Мета роботи – підвищення якості словника на основі застосування нової технології, що включає спеціальну обробку коротких документів.

Метод. Пропонується модель короткого документа, яка представляє його у вигляді трьох частин: заголовної, змістовної та заключної. У заголовній і заключній частинах зазвичай міститься інформація, що не має відношення до предметної області. Тому запропоновано метод виділення змістовної частини, заснований на використанні множини ключових слів. Розмір короткого документа (його змістовної частини) не дозволяє визначити частотні характеристики слів і виявити багатослівні терміни, частка яких сягає 50% від усіх термінів. Для забезпечення можливості виділення термінів у коротких документах запропоновано метод їх кластеризації, заснований на виділенні іменників та обчисленні їх частотних характеристик. Утворені кластери розглядаються як звичайні документи, оскільки їхній розмір дозволяє виділяти багатослівні терміни. Для виділення термінів запропоновано виділяти в тексті послідовності слів, що містять іменники. Аналіз частот повторення таких послідовностей дозволяє визначити багатослівні терміни. Для визначення тлумачення термінів використано раніше розроблений метод автоматизованого пошуку тлумачень у словниках.

Результати. На основі запропонованої моделі та методів створено програмне забезпечення для побудови словника предметної галузі та проведено низку експериментів, що підтверджують ефективність розроблених рішень.

Висновки. Проведені експерименти підтвердили працездатність запропонованого програмного забезпечення та дозволяють рекомендувати його до використання на практиці для створення словників предметної галузі різних

DOI 10.15588/1607-3274-2023-4-14

інформаційних систем. Перспективи подальших досліджень можуть включати побудову корпоративних пошукових систем на основі словників термінів та кластеризації документів.

КЛЮЧОВІ СЛОВА: словник предметної галузі, інформаційна система, термін, кластеризація, інформаційна технологія, короткий документ.

ЛІТЕРАТУРА

1. Larman K. Primenenie UML 2.0 i shablonov proektirovanija. Prakticheskoe rukovodstvo. 3-e izdanie / K. Larman. – M. : Izdatel'skij dom "Vil'jams", 2013. – 736 p. [in Russian].
2. Information Systems for Business and Beyond. Information systems, their use in business, and the larger impact they are having on our world [Electronic resource] / [D. Bourgeois, J. Mortati, S. Wang, et al.] – Access mode: https://opentextbook.site/exports/ ISBB-2019.pdf
3. Methodology for the Development of Dictionaries for Automated Classification System / [A. Artamonov, D. Kshnyakov, V. Danilova et al.] // 8th Annual International Conference on Biologically Inspired Cognitive Architectures (BICA 2017). Procedia Computer Science Volume:123. Moscow, Russia, 1–6 August 2017. – P. 57–62. doi:10.1016/j.procs.2018.01.010
4. Dalglish S. L. Document analysis in health policy research: the READ approach / S. L. Dalglish, H. Khalid, S. A. McMahon // Health Policy and Planning. – 2020. – V. 35, Issue 10. – P. 1424–1431.
5. Cheng Y. Research and development of domain dictionary construction system / Y. Cheng, Y. Huang // Proceedings of the International Conference on Web Intelligence, August 2017. – P.1162–1165. https://doi.org/10.1145/3106426.3109046
6. Liang S. Dynamic Clustering of Streaming Short Documents / S. Liang, E. Yilmaz, E. Kanoulas // International Conference on Knowledge Discovery and Data Mining. August 2016. – P. 995–1004.
7. Wang Y. Outlier detection from massive short documents using domain ontology / Y. Wang, S. Yang // International Conference on Intelligent Computing and Intelligent Systems, 29–31 Oct. 2010, Xiamen, China. doi: 10.1109/ICICISYS.2010.5658426
8. Short-Text Topic Modeling via Non-negative Matrix Factorization Enriched with Local Word-Context Correlations. / [T. Shi, K. Kang, J. Choo et al.] // WWW '18: Proceedings of The Web Conference 2018, April 2018, Lyon, France. doi: 10.1145/3178876.3186009
9. Topic based Summarization of Multiple Documents using Semantic Analysis and Clustering / [R. Hafeez, S. Khan, M. Abbas et al] // International Conference on Smart; 8–10 Oct. 2018, Islamabad, Pakistan. doi: 10.1109/HONET.2018.8551325
10. Vo D-T. Learning to classify short text from scientific documents using topic models with various types of knowledge / D-T. Vo, C-Y. Ock // Expert Systems with Applications: An International Journal. – 2015. – V. 42, Issue 3. – P. 1684–1698. https://doi.org/10.1016/j.eswa.2014.09.031
11. Liang S. Dynamic Clustering of Streaming Short Documents / S. Liang, E. Yilmaz, E. Kanoulas // 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, US, 13 August 2016. – P. 995–1004. https://doi.org/10.1145/2939672.2939748
12. Seki H. On Term Similarity Measures for Short Text Classification / H. Seki, S. Toriyama // 11th International Workshop on Computational Intelligence and Applications (IWCIA), 9–10 Nov. 2019. – P. 53–58. DOI: 10.1109/IWCIA47330.2019.8955045.
13. Seki H. Using term similarity measures for classifying short document data / H. Seki, S. Toriyama // International Journal of Computational Intelligence Studies. – Vol. 10, Issue 2–3, https://doi.org/10.1504/IJCISTUDIES.2021.115430
14. Rogers N. Comparison on the Classification of Short-text Documents Using Latent Dirichlet Allocation and Formal Concept Analysis / N. Rogers, L. A. Longo // 25th Irish Conference on Artificial Intelligence and Cognitive Science, AICS 2017, Dublin, Ireland, 7–8 December 2017. – V. 2086. – P. 50–62.
15. Development of Methods for Pre-clustering and Virtual Merging of Short Documents for Building Domain Dictionaries / [O. Kungurtsev, S. Zinovatna, I. Potochniak et al.] // Eastern-european Journal of Enterprise Technologies. – 2020. – Vol. 5, № 2 (107). – P. 39–47. http://doi.org/10.15587/1729-4061.2020.215190
16. Development of information technology of term extraction from documents in natural language / [O. Kungurtsev, S. Zinovatnaya, Ia. Potochniak et al.] // Eastern-European Journal of Enterprise Technologies. – 2018. – V. 6, №2 (96), – P. 44–51. doi: https://doi.org/10.15587/1729-4061.2018.147978
17. Comparison of Clustering Algorithms in Text Clustering Tasks. / [R. G. García, B. Beltrán, D. Vilariño et al.] // Computación y Sistemas. – 2021. – Vol. 24, № 2. https://doi.org/10.13053/cys-24-2-3369
18. Шевченко А. Організація електронного документообігу на підприємстві [Electronic resource] / А. Шевченко. – Access mode: https://uteka.ua/ua/publication/commerce-12-dokumentooborot-2-organizaciya-elektronnogo-dokumentooborota-na-predpriyatii
19. Типова інструкція з діловодства в міністерствах, інших центральних та місцевих органах виконавчої влади [Electronic resource]. – Access mode: https://borispol-rada.gov.ua/item/39961-typova-instruktsiya-z-dilovodstva-v-ministerstvakh-inshykh-tsentralnykh-ta-mistsevykh-orhanakh-vykonavchoi-vlady.html
20. Lions K. Long-Tail Keywords: What They Are & How to Use Them for SEO. [Electronic resource] / K. Lions. – Access mode: https://www.semrush.com/blog/how-to-choose-long-tail-keywords/
21. Borysova, N. V. The formation of problem domain dictionary / N. V. Borysova, O. V. Kanyshheva, O. V. Kanyshheva // Eastern-European Journal of Enterprise Technologies. – 2013. – Vol. 5, №3(65). – P. 16–19. https://doi.org/10.15587/1729-4061.2013.18
22. Rahoo L. A. Design and Development of an Automated Library Management System for Mehran University Library, Jamshoro. / L. A. Rahoo, M. A. Unar // Control Theory and Informatics. – 2016. – № 6(1). – P. 1–6.
23. Kungurtsev O. Automation of Serching for Terms in the Explanatory Dictionary / O. Kungurtsev, N. Novikova, M. Kozhushan // Proceedings of Odessa Polytechnic University. – 2020. – № 3(62). – P. 91–100. DOI: 10.15276/opu.3.62.2020.11
24. Sketch Engine. [Electronic resource]. – Access mode: https://www.sketchengine.eu/

DOI 10.15588/1607-3274-2023-4-14