The aim of research is to improve the quality of domain dictionaries by expanding the corpus of the documents under study by using short documents. A document model is proposed that allows to define a short document and the need to combine it with other documents to highlight verbose terms. An algorithm for highlighting the substantive part of the document has been developed, since in a short document the heading and closing parts usually contain terms that are not related to the studied domain. A method for preliminary clustering of short documents to highlight verbose terms has been developed. The method is based on highlighting and counting occurrences of nouns (one-word terms) for all analyzed documents. The concept of document proximity is introduced, which is determined by the combination of two criteria: the relative number of matching terms and the relative frequency of occurrence of matching terms. The principle of grouping documents at the customer's site often does not correspond to the principles of grouping necessary for building a dictionary of the domain. In a short document, it is usually impossible to isolate a verbose term because the repetition of terms is very low. A method has been developed for virtual combining of short documents based on the principle of achieving the necessary repeatability of one-word terms. The merged document has the highest possible frequency of terms for the cluster it belongs to. At the same time, the original text of documents is preserved and the ability to associate the selected verbose term with those documents in which it is included. The experiment made it possible to find the best ratio for the elements of the document proximity coefficient and confirm the effectiveness of the proposed preliminary clustering method

Keywords: domain dictionary, short document, clustering, document proximity coefficient, virtual union

Received date 09.09.2020
Accepted date 21.10.2020
Published date 30.10.2020

UDC 004.912
DOI: 10.15587/1729-4061.2020.215190

DEVELOPMENT OF METHODS FOR PRE-CLUSTERING AND VIRTUAL MERGING OF SHORT DOCUMENTS FOR BUILDING DOMAIN DICTIONARIES

O. Kungurtsev
PhD, Professor*
E-mail: akungurtsev19@gmail.com
S. Zinovatna
PhD, Associate Professor*
E-mail: zinovatnaya.svetlana@opu.ua
Ia. Potochniak
PhD, Engineer
Software Development Company "The Product Engine"
Marazlievska str., 7, Odessa, Ukraine, 65078
E-mail: yana.onpu@gmail.com
N. Novikova
Senior Lecturer
Department of Technical cybernetics and information technology named after prof. R. V. Merkt
Odessa National Maritime University
Mechnikova str., 34, Odessa, Ukraine, 65029
E-mail: nataliya.novikova.31@gmail.com
*Department of System Software
Odessa National Polytechnic University
Shevchenka ave., 1, Odessa, Ukraine, 65044

1. Introduction

When designing software products "to order", it becomes necessary to create a dictionary for a narrow domain (DSA). DSA contains the terms of the domain and their interpretation. It allows the customer of the project and the developer of the project to communicate in "the same language", which is necessary at the first stage of the project to determine the requirements for the software product [1]. In the future, such a dictionary is used to create user interfaces, instructions, in the development support process [2]. Automatic construction of domain-specific ontologies is a difficult task [3, 4], since it requires extracting domain-specific terms from the body of documents and assigning them the corresponding labels of domain concepts. On the other hand, the DSA availability allows to solve many problems associated with document processing [5]. Usually, the terminology of the Customer's domain is heterogeneous and specific, which does not allow using existing specialized explanatory dictionaries. The documents that are used in the Customer's organization are examined as a source material for the DSA. Until now, the definition of terms for DSA is usually performed by an expert in "manual mode", which is a very time consuming process. When automating the process of building DSA, there is a problem of highlighting terms in short documents. The low frequency of occurrence of some word or phrase in the document does not allow to consider it a term, however, if the document was processed manually, the expert would probably include it in the DSA.

2. Literature review and problem statement

In [6] it is shown that to represent the domain, it is sufficient to use terms based on nouns. This approach is taken in this article. The analysis of the effectiveness of various approaches to the selection of terms is carried out in [7]. It is shown that the rule-based approach is more accurate, but less versatile and more laborious. A hybrid approach can be a solution to the problem. In [8], using the examples of search engines, it is shown that deep parsing of a text is less effective than primitive analysis and subsequent statistical processing. A similar method of text processing was used in [9] to highlight keywords. However, the author limited himself to one-word terms, which greatly simplified the task, but lowered the quality of the result. The work [10] is devoted to the study of the effectiveness of methods for extracting terms. However, the authors limited themselves to only one-word and two-word terms. At the same time, research has shown that terms can contain up to five or more words [11]. In [12], a general approach to the analysis of texts in natural language was proposed, but the problems of practical implementation of procedures for extracting terms and determining their labor intensity remained unresolved. The most noteworthy is the method that combines statistical processing of texts with partial syntactic analysis [11]. The method is based on forming a possible term from a noun and surrounding words. The implementation of the method is fraught with a number of problems. First, the corpus of documents to be analyzed is heterogeneous in terms of topics. Searching for terms based on the entire corpus can lead to errors both in the selection of terms and in their further interpretation. Secondly, a feature of the array of documents representing a certain organizational structure is the presence of a large number of short documents. In the study [13], the analysis problems associated with the size of text documents are indicated, it is proposed to use the classification method based on weight coefficients. Noteworthy is the proposal to split documents into three types: long, short, and very short, and terms in the thesaurus into unique, rare, and general. However, the work lacks a clear definition of a short document in terms of the number of errors that occur during its analysis. In addition, the corpus of the studied documents was limited to documents from the area of the executive and legislative powers, which made it possible to provide a certain dynamism of the headings, but could not be extended to other domains. The dependence of errors in the selection of verbose terms on the number of words in documents in Slavic languages is determined experimentally in [11] (Table 1). Here, the loss of a term means a one-time appearance of some noun in a document, since in this case there is no way to build a verbose term on its basis. In addition, short documents often have some formalized form [14], the elements of which can significantly affect the selection of terms and their frequency characteristics.

The interpretation of the term depends on the domain where it is used, therefore, to create DSA, it is necessary to cluster the corpus of the documents under consideration. Since the search for verbose terms in short documents is associated with a large number of errors [13], in this study it is proposed to perform preliminary clustering at the first stage using one-word terms. Subsequently, it is required to perform virtual merging of short documents (creating a merged document while preserving its constituent documents), which will allow to highlight verbose terms and, if necessary, carry out clustering based on verbose terms. In [15], clustering of a corpus of short documents is considered. It has been established that "the short length of documents makes it difficult to infer the hidden distribution of topics," therefore, the focus is not on quality, but on the speed of clustering.

Table 1

Dependence of the probability of losing a term on the size of the document

Document size in words	Term loss probability
up to 100	0.95
101÷500	0.90
501÷1000	0.55
1001÷3000	0.22
3001÷5000	0.05
over 10000	0.03

In [16–18], various approaches to document clustering are considered, in particular, the methods of K-means, K*-means, EM algorithm, sGEM, LAR, HSTC and TCFS algorithms, and their efficiency is analyzed. It is shown that all methods are effective in the process of clustering documents, but TCFS performance is slightly better in terms of clustering quality, and the traditional K-means algorithm has a high speed in obtaining clustering results. At the same time, the conducted studies are not extended to short documents. The paper [19] shows the efficiency of the K-means algorithm with four clusters for partitioning into groups of unlabeled documents. A proposal was made on the possible unification of documents. However, the use of the results obtained is limited to a library of dissertations and documents exclusively in pdf format.

Analysis of literary sources allows to conclude that the following unsolved problems exist:

– preliminary processing of short documents;

– preliminary clustering of documents until a full set of their features is obtained;

– virtual merging of short documents.

3. The aim and objectives of research

The aim of research is to present short documents in a form that will allow to highlight verbose terms from them. In this case, the document language can be any of the common European languages for which there is a corresponding analyzer.

Based on the identified problems, the following research objectives are formulated:

– create a mathematical model of a short document;

– develop an algorithm for highlighting the content of the document;

– develop a method for preliminary clustering of short documents;

– develop a method for the virtual combination of short documents;

– carry out an experimental study of the proposed methods.

4. Mathematical model of the document

Let's assume that the corpus of all documents under study is represented by their set D

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

As a result of preliminary analysis of documents, at the stage of which nouns were selected, each document can be represented by a tuple

$$d_i=\langle text_i, nw_i, Mt_i\rangle, \tag{2}$$

where $text_i$ – text of the document; nw_i – document size in words; Mt_i – set of one-word terms of the d_i document.

$$Mt_i=\{<t_q, nn_q>\},\quad q=\overline{1,nn_i}, \tag{3}$$

where t_q – term represented by a noun word; nn_q – the number of occurrences of the term t_q in document d_i ; nn_i – the number of different terms in the document.

5. Algorithm for highlighting the content of the document

Short documents often have a formalized structure (orders, statements, orders, reports, etc.). For example, an order has the following components:

1) the full name of the institution where the order is issued;

2) the title of the document;

3) date;

4) number;

5) the name (title) of the order (what the order is about);

6) main content;

7) signature (position of the head, signature, initials, surname).

Consequently, short documents can contain a number of general terms that define the form, but not the content of the document (name of the organization, address, etc.). Such terms should be excluded from further analysis. In general, formalized documents in an organization can be in the same groups with other documents. If they are collected in separate groups, then this can somewhat simplify the preprocessing process.

It is proposed to present a formalized document in the form of three sections:

– heading part *h*;

– content part *b*;

– the final part (signatures) *f*.

The heading part refers to a term (of one or more words) that defines the type of document. It actually separates the title from the content. The number of document types is small, therefore it was proposed to compose a set of M_b types, in which a separate line corresponds to each type, for example, "Protocol", "Order", "Office memorandum", etc. A distinctive feature of the term representing a type is that it is written as a title in the center of the page. Therefore, the elements of the set are represented by combining the newline character and the text itself, which makes it possible to distinguish the type of document, for example, "Statement" from the word "statement", which may occur in some phrase of the content section of the document.

The final part may include a list of the signatories of the document, indicating the position of each person, perhaps the words "signature", "approve", etc. It is proposed to create a set of M_f terms that could be included in the final part. The upper boundary of the final part of the document should be sought if it was previously established that the document in question belongs to the group of formalized documents. For this, the occurrence of elements of the set M_f in the text of the document is determined, starting from the end of the document.

To reduce the processing time of the document, it is proposed to limit the search for the boundary of the heading part *K* to words, and the final part to *N* words.

The content of the M_b, M_f sets and the values of *K* and *N* are determined on the basis of the accepted office work and are refined with the participation of a domain expert.

In the general case, an element of the sets M_b, M_f is a phrase that includes a different number of words.

The text of the document is checked for the occurrence of word combinations from M_b, M_f according to the following algorithm.

Let n_{stb} be the number of the word with which the content of the document begins d_i; n_{stf} is the number of the word from which the final part of the document begins; l is the number of words in the studied phrase, j is the position in the document in which the studied phrase is located, measured in words.

1. $n_{sb}=0$; $n_{stf}=0$.

2. Define the text ts as a fragment from the beginning of the document with a length of *K* words.

3. $m=1$, w_m – the key word combination under study, $w_m \in M_b$.

4. Determine j as the result of the function of finding the position of the substring w_m in *ts*.

5. If j is non-empty, then $n_{stb}=j$. Go to step 7.

6. $m=m+1$. If $m \le |M_b|$, then go to step 4.

7. Define the text *ts* as a fragment from the end of the document, length *N* words.

8. $m=1$, w_m – the key word combination under study, $w_m \in M_f$.

9. Determine j as the result of the function of finding the position of the substring w_m in *ts*.

10. If j is non-empty, then $n_{stf}=j$. Go to step 12.

11. $m=m+1$. If $m \le |M_f|$, then go to step 9.

12. Mark up the document by marking the text from position 0 to $(n_{stb}+l_b)$ and from n_{stb} to nw_i as comments that are not considered in further analysis of the document d_i.

Since the concept of a "short document" is rather subjective before the frequency of the terms included in it is determined, the extension of the procedure for isolating the content of the document to documents that will not turn out to be short will not lead to any negative consequences.

6. Method of preliminary clustering of short documents

The aim is to get clusters within which it is possible to automatically combine short documents, which will significantly reduce the time of the merging process. Getting into clusters of documents that are not short should not in any way affect the achievement of the goal. Other goals, for example, to use the resulting clusters in the future to search for information in the corpus of documents, are not set, but are a "side" result of the method.

At this stage, the original language of the document actually doesn't matter.

6. 1. Determining the proximity of documents

To include documents d_i and d_j in a certain cluster, let's define set of the same terms in these documents.

$$Mt_{i,j}=\left\{t_k|t_k\in Mt_i\wedge t_k\in Mt_j\right\},\quad k=\overline{1,n_{i,j}}. \tag{4}$$

The number of identical terms $n_{i,j}$ of documents d_i and d_j, referred to the number of all terms in two documents, can serve as a component of the coefficient of closeness of the 1st order of documents d_i and d_j.

$$Ks^1_{i,j}=\frac{n_{i,j}}{\left|Mt_i\right|+\left|Mt_j\right|}.$$

Also, the closeness of documents d_i and d_j is determined by the repetition of the same terms. To do this, let's define a lot of common terms in the documents d_i and d_j, which are repeated at least two times, and their number $n'_{i,j}$.

$$Mt'_{i,j}=\left\{\begin{matrix}t_k|\left(t_k\in Mt_{i,j}\right)\wedge\\ \wedge\left(t_k\in Mt_it_k.nn_k\geq 2\right)\wedge\\ \wedge\left(t_k\in Mt_j\wedge t_k.nn_k\geq 2\right)\end{matrix}\right\},\quad k=\overline{1,n'_{i,j}}.$$

Determine the second component of the 1st order proximity coefficient

$$Ks^2_{i,j}=\frac{n'_{i,j}}{\left|Mt_i\right|+\left|Mt_j\right|}.$$

The ratio of the values $Ks^1_{i,j}$ and $Ks^2_{i,j}$ in the general formula for the proximity coefficient of the 1st order of two documents $Ks_{i,j}$, j is not obvious, therefore let's introduce a certain factor γ, which will be determined experimentally:

$$Ks_{i,j}=Ks^1_{i,j}+\gamma\cdot Ks^2_{i,j}. \tag{5}$$

Let's introduce the concept of the minimum value of the first-order proximity coefficient Ksmin.

If accept Ksmin=0, then all documents will be combined into one cluster. If to take Ksmin=1, then with γ=0 each document will be a separate cluster. A number of experiments were carried out to determine the recommended values of Ksmin and γ.

6. 2. Stages of the method of preliminary clustering of documents

There is an ordered set of documents $D=(d_1, d_2,..., d_i,..., d_{p-1}, d_p\}$ and the value Ksmin is given:

1) f=1;

2) select the first document d_1 from the set D as the current one;

3) create the f-th cluster Mcd_f, which includes d_1 and those documents $d_i\in D$ whose closeness coefficient to d_1 is greater than or equal to Ksmin.

$$Mcd_f=\left\{d_1\right\}\bigcup\left\{d_i\,|\,Ks_{1,i}\geq Ks\min\right\};$$

4) documents included in Mcd_f, are excluded from the set D, that is, $D=D\backslash Mcd_f$;

5) if $\left|D\right|>0$, then $f=f$+1. Go to step 2;

6) the end of the algorithm.

A number of documents that fell into a certain cluster could fall into other clusters as well. It is possible that the clustering result will be better than the initial one.

To improve the distribution of documents across clusters, let's introduce the concept of "cluster core" in the form of a set of documents that can't be included in other clusters. For each core, let's find a document for which the relative total coefficient of proximity with other documents included in the Mcd'_l largest (central document):

$$Ksm_i=\max_{1\leq b\leq\left|Mcd'_l\right|,b\neq i}\left[\frac{\sum Ks_{i,b}}{\left|Mcd_l\right|}\right].$$

Obviously, the last cluster Mcd_f contains only the core. The penultimate cluster $Mcd_{f\text{-}1}$, in addition to the core, possibly contains documents that can be included in the previous cluster $Mcd_{f\text{-}2}$, etc. Some d_h document from the Mcd_j cluster should be transferred to the $Mcd_{j\text{-}1}$cluster if it turns out that $Ks_{j\text{-}1,h}>Ks_{j,h}$, where $Ks_{j\text{-}1,h}$ – the proximity coefficient of the central document of the cluster $Mcd_{j\text{-}1}$ and the document d_h; $Ks_{j,h}$ – the proximity coefficient of the central document of the cluster Mcd_j and the document d_h.

Possible document movements between clusters are shown in Fig. 1.

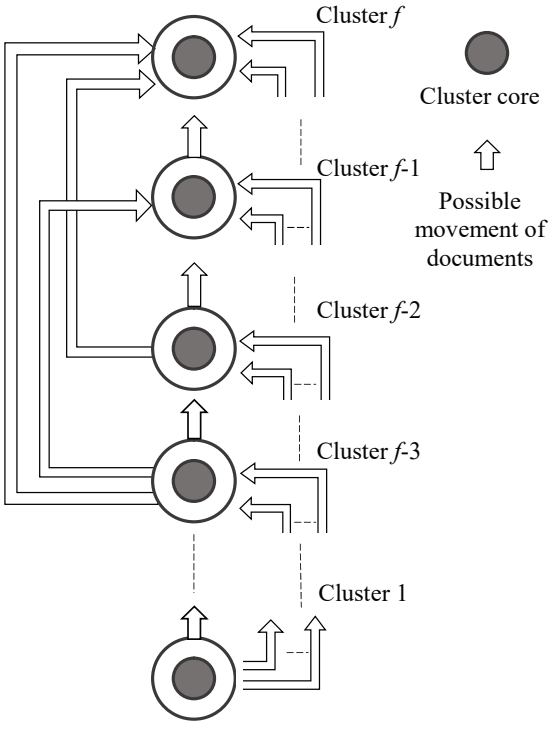

Fig. 1. Possible movement of documents between clusters

Upon completion of the clustering process, cluster f contains only the core. However, in the process of redistributing documents between clusters, cluster f can be supplemented with documents from other clusters.

6. 3. Algorithm for preliminary clustering

1. Perform preliminary processing of documents.
2. Determine Ksmin and .
3. Form clusters using the pre-clustering algorithm.
4. For the last cluster, find the central document.
5. For each next cluster, in descending order of numbers, determine the core by excluding documents that may be included in previous clusters.
6. For all documents of the current cluster that are not included in its core, let's determine the possibility of transferring to the previously considered clusters.
7. Repeat steps 5 and 6 until all clusters have been adjusted.

7. Method of virtual merging of short documents

In the future, let's consider documents that belong to only one cluster, and short documents that have been pre-processed.

From Table 1 it follows that significant errors are possible if the document size is less than 5,000 words. Thus, let's consider a document d short if

$$d.nw \leq 5{,}000, \quad (6)$$

where nw – the number of words in the document. Raising the short document detection threshold does not affect the quality of the merge result, and only leads to an increase in the overall time spent on document processing.

As mentioned earlier, it can be assumed that verbose terms formed on the basis of nouns included in the documents once will be lost. Therefore, it is of interest to form a set of unique nouns for each document.

$$Mt1_i = \left\{ t_{i,k} | t_{i,k} \in Mt_i \wedge nn_{i,k} = 1 \right\},$$

where $nn_{i,k}$ – the number of occurrences of the term $t_{i,k}$ in the text of the document d_i.

Thus, the document will be represented by the following tuple

$$d_i = <text_i, nw_i, Mt_i, Mt1_i>. \quad (7)$$

Based on (7), it is possible to clarify the concept of a short document and select a subset of them from the corpus of all documents.

$$Ds = \left\{ d_i \,|\, d_i \in D \wedge \frac{|Mtq_i|}{|Mt_i|} \geq Kc \right\}. \quad (8)$$

Here Kc defines the boundary value of the concept of a short document and can be established by an expert.

By the virtual union of two documents, let's mean a meta document that is represented by a tuple (7) obtained from two or more documents, if their texts were combined.

To combine the documents, let's introduce the coefficient of closeness of the 2nd order $Ku_{i,j}$ of the document d_i to the document d_j. It will be determined by two components:

$$Ku_{i,j} = Ku^1_{i,j} + \delta \cdot Ku^2_{i,j}. \quad (9)$$

The coefficient $Ku^1_{i,j}$ determines the relative number of terms that, after the documents are combined, will no longer be unique:

$$Ku^1_{i,j} = \frac{|Mt1_i \cap Mt_j|}{|Mt1_i|}. \quad (10)$$

The factor $Ku^2_{i,j}$ determines the increase in the repetition of terms in document d_i after being combined with document d_j.

Let's introduce the concept of common terms of the two documents

$$Mt_{i,j} = Mt_i \cap Mt_j.$$

Then

$$Ku^1_{i,j} = \frac{|Mt_{i,j}|}{|Mt_i|}. \quad (11)$$

To determine the joint proximity of the 2nd order of documents, it is proposed to introduce an integrated coefficient

$$KP_{i,j} = Ku_{i,j} + Ku_{j,i}. \quad (12)$$

Since the coefficient of proximity of the 2nd order of the document d_i to the document d_j differs from the coefficient of proximity of the 2nd order of the document d_i to the document d_j, the formula for $KP_{i,j}$ gives a certain priority to the combination of short documents.

The method contains the following steps:

1) determine the total number of unique terms in all documents

$$n1 = \sum_{i=1}^{n} |Mt1_i|; \quad (13)$$

2) form a variety of unique terms in the corpus of documents

$$Mt = \bigcup_{i=1}^{n} Mt_i;$$

3) transform the set Mt into the set Mt'', replacing each term t_j with a $<t_j, nn_j>$ and setting the values nn_j=0:

$$Mt \Rightarrow Mt';$$

4) perform the procedure Proc1 of combining a set of Mt'' with sets of unique terms of all documents

$$Proc1 = if\left(t_j = t_{i,k}\right) nn_j = nn_j + 1; \quad i = \overline{1,n}, \quad j = \overline{1,|Mt1_i|}, \quad (14)$$

where $t_{i,k}$ – another term from document d_i.

5) form a set of unique terms for a virtual document that unites all documents in the corpus:

$$Mt'v = \left\{ \left\langle t_j, nn_j \right\rangle, nn_j \right\} = 1. \quad (15)$$

The number of such terms

$$n1v = |Mt'v|; \quad (16)$$

6) construct a matrix of documents proximity (Table 2). The rows and columns of the matrix are numbered in accordance with the numbering of documents in the corpus. At the intersection of the i-th row and the j-th column, $KP_{i,j}$ is written – the sum of the coefficients of proximity of the 2nd order of the document d_i to the document d_j and the coefficient of proximity of the 2nd order of the document d_j to the document d_i (12). There should be no empty cells in the table

7) the highest $KP\max_i$ coefficient is determined for each document;

8) the largest coefficient $KP\max$ from $KP\max_i$ is selected, as well as $i_{\max}$ and $j_{\max}$ – the index of the maximum coefficient $KP\max = KP_{i\max,j\max}$;

9) documents $d_{i_{max}}$ and $d_{j_{max}}$ are virtually combined into a document $d_{i,j}$, which is presented in the form:

$$d_{i,j} = <text_{i,j}, nw_{i,j}, Mt_{i,j}, Mt1_{i,j}>,$$

where $nw_{i,j} = nw_{i_{\max}} + nw_{j_{\max}}$; $Mt_{i,j} = Mt_{i_{\max}} \cup Mt_{j_{\max}}$, taking into account the total number of repetitions of terms in $d_{i_{\max}}$ and $d_{j_{\max}}$;

$$Mt1_{i,j} = Mt1_{i_{\max}} \cup Mt1_{j_{\max}} \setminus \left\{ t_k \,|\, t_k \in Mt_{i,j} \wedge nn_k > 1 \right\}.$$

Document $d_{i,j}$ is added to the corpus of documents.

Table 2

Document proximity matrix

No. of document	1	2	...	n	KPmax
1	×	$KP_{1,2}$	...	$KP_{1,n}$	$KP\max_1$
2	$KP_{2,1}$	×	...	$KP_{2,n}$	
...	...	...	...	...	
n	$KP_{n,1}$	$KP_{n,2}$	...	×	$KP\max_n$
		$i_{\max}$	$j_{\max}$		KPmax

The current number of unique terms in the corpus $n1$ is determined, taking into account the union. If $n1 \neq n1v$, then documents $d_{i_{\max}}$ and $d_{j_{\max}}$ are removed from the proximity matrix, document $d_{i,j}$ is added, and the transition to step 6 is performed. Otherwise, the processing of documents is completed.

8. Experimental study of the proposed methods

To test the clustering method, the DocCluster program was created (Fig. 2). The source documents go to the preprocessing block, where the content of the document is highlighted, as well as converted to the format the parser works with.

Cognitive Dwarf (Russian and English, version for non-commercial use), MySteam (Russian, very fast and compact analyzer), Language Tool API (Ukrainian) were used as a parser. For the Cognitive Dwarf and MySteam analyzers, the preprocessing block was used to convert the original text of documents into txt format.

In the block for extracting nouns from the tables obtained at the output of the analyzer, nouns are selected and their occurrence in the document is counted. Thus, the internal form of the document is formed, represented by one-word terms and the number of their occurrences in the document (frequencies). Further processing of documents does not depend on the language of the source documents.

In the pre-clustering block, in accordance with the given algorithm, the documents are distributed into clusters. To control the results, it is possible to see the resulting clusters in the form of groups of source documents.

In the block for virtual merging of documents, documents are grouped within clusters. The resulting groups can later be used to identify verbose terms.

To determine the values of Ksmin and γ, 22 short documents were selected from the university's field of activity (educational process, scientific work, personnel management and repair work. It was assumed that a lot of documents should be grouped into 4 clusters. This was confirmed by the results of the analysis of the work of the noun extraction block At the same time, a number of terms were identified that were included in different supposed clusters (student, teacher, audience, workload, task, etc.). The pre-clustering unit processed these documents 100 times for 10 Ksmin values and 10 γ values. Results clustering were estimated by the relative error $Rer = \frac{nDer}{nD}$, where $nDer$ is the number of documents that did not fall into the "own" cluster; nD is the total number of documents. From the analysis of the graphs in Fig. 3 it follows that the values of Ksmin should be set in the range 0.3–0.4, and the values of γ – in range 0.3–0.6. These recommendations were accepted for further research.

To test the algorithm for the virtual combination of documents, the cluster of documents representing the educational process was expanded to 10 documents. After preliminary processing and identification of one-word terms, 8 documents fell into the category of short ones when determining the value of Kc=0.05. The experiment showed that the second component of the proximity coefficient $\delta \cdot Ku^2_{i,j}$ can improve the repeatability of terms in a combined document (the case was observed once) by combining documents that have more repeated terms than other pairs of documents. However, since priority should be given to the "exclusion" of unique terms, it was proposed to determine δ by a formula $\delta \le 0.5 \frac{Ku^1_{i,j}}{Ku^2_{i,j}}$ for all further research.

After the introduction of the recommended values of Ksmin, γ and δ, the quality of clustering and merging was experimentally tested on another body of documents. In total, the corpus includes 32 documents. As documents, selected reports from conferences on various topics ("Business Management in the Digital Economy" [20], "Strongly Correlated Two-Dimensional Systems: From Theory to Practice [21]," Transport in the Integration Processes of the World Economy " [22]," Digital transformation of education [23], "Ensuring life safety at the present stage of development of society" [24]).

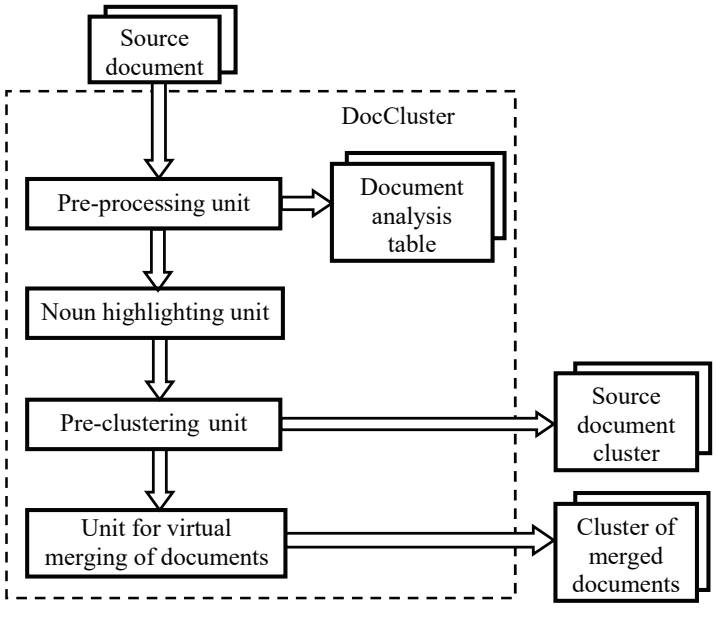

Fig. 2. The structure of the DocCluster program

As a result of clustering, 6 groups of documents were formed (Fig. 4).

The expert confirmed the number of clusters, but considered that one document fell into the wrong cluster. From the point of view of virtual merging of documents, there were no errors in the clusters. As a result of virtual

merging of documents, for example, in cluster cluster_4, the number of unique terms has decreased by 3.8 times.

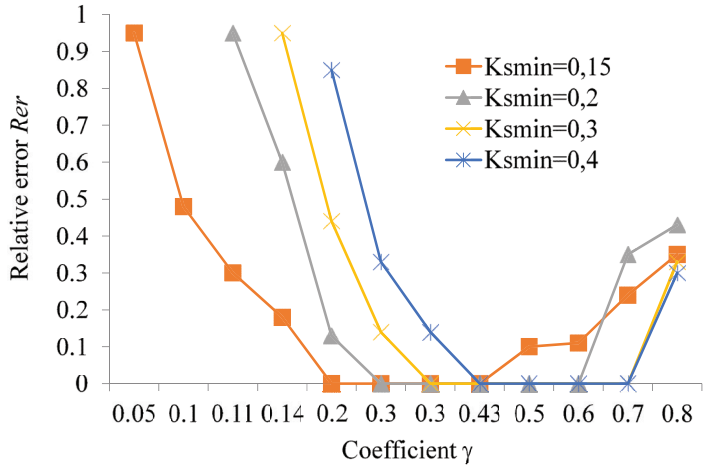

Fig. 3. Influence of the minimum value of the first-order proximity coefficient Ksmin and the coefficient γ on clustering errors

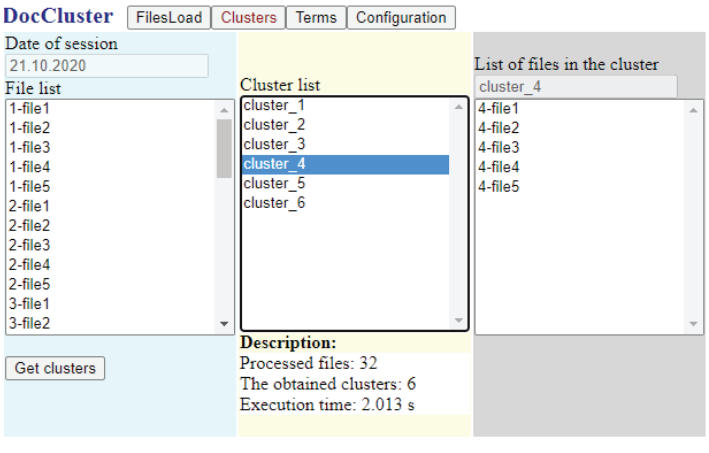

Fig. 4 An example of cluster formation

9. Discussion of the results of clustering and virtual merging of documents

The mathematical model of the document in accordance with the formula (2) allowed in the future to determine the coefficients of the proximity of documents, which is the basis for their clustering.

The results obtained on the clustering of documents became possible, firstly, due to the selection of the content of the documents, which is especially important for short documents. According to the document model, the value of the nw_i (2) component is reduced, that is, words that can't be considered as terms are excluded.

Secondly, for the clustering of documents, a proximity coefficient (5) was proposed, which contains two components. The first was determined by the relative number of overlapping terms, and the second was determined by the repeatability of these terms. Since the ratio of these components (γ) was not obvious, the recommended value was obtained experimentally (Fig. 3). The clustering method provides for a preliminary determination of the number of clusters, and then the refinement of their composition based on the calculation of nuclei for each cluster (Fig. 1).

The virtual merging of documents allows to consider a group of documents as one document (meta-document) in terms of highlighting terms. The merging process is based on the proposed second order proximity factor (9), which provides priority in merging short documents. The virtual combination of documents, built on the principle of maximum "elimination" of unique terms in the combined document, made it possible to achieve the theoretically possible minimum of this indicator in the cluster. This can be verified by comparing the number of unique terms in an imaginary document that combines the entire cluster (16) with the number of unique terms in all combined documents. The proposed method allows to expand the concept of "unique term", supplementing it with cases when the term in one document occurs within a specified number of times.

Some discrepancy between the results of manual and automated clustering observed during the experiments (one document out of 32 fell into a wrong cluster) is determined by the fact that the expert evaluated the result in terms of the semantics of the document, and not the frequency of occurrence of terms. Therefore, the expert's conclusions are only a recommendatory assessment.

The main purpose of the proposed method of preliminary clustering is to reduce the time for building DSA. However, it showed quite good results in solving the clustering problem for further information retrieval. In the future, it is planned to improve the quality of clustering by using verbose terms.

At this stage of the study, preliminary clustering was performed. In the future, it is planned to carry out clustering based on verbose terms, which will reduce the time needed to find the necessary information.

10. Conclusions

1. A mathematical model of the document has been developed, taking into account such characteristics as the number of words, a set of one-word terms, and their frequency. The model is needed for further clustering and possible merging of documents.

2. An algorithm for highlighting the substantive part of the document has been developed, which implies the removal of individual structural components of the document, which obviously do not contain words that can be attributed to terms of a narrow domain. A distinctive feature is the definition of the heading and closing parts of the document according to the introduction of the concept of a document type, determined by a set of keywords. The implementation of the algorithm can significantly reduce the time for preprocessing documents.

3. A method for preliminary clustering of short documents has been developed, which is distinguished by the use of the proximity coefficient, which takes into account both coinciding terms and their relative frequency. To improve the quality of grouping of documents, an iterative process of their distribution among clusters is provided. The method allows to reduce the time and the number of errors when grouping documents manually.

4. A method has been developed for the virtual combination of short documents based on their closest proximity. The method is distinguished by the ability to reduce the number of unique terms in the document corpus to the theoretically achievable level.

5. An experiment was carried out to process 32 documents from 6 different domains. The documents were submitted to the program input in a random order, as a result, the program identified 6 clusters, which confirmed the efficiency of the proposed methods and algorithms. Also, experimental studies made it possible to clarify the ratio of the components of the proximity coefficient; Ksmin in the range 0.3–0.4, γ – in the range 0.3–0.6. The quality of clustering and virtual combining of documents allows using the proposed methods in the technology of creating DSA, and also showed the prospects of their development based on the use of verbose terms.

Acknowledgements

We would like to express our gratitude to students I. Mileiko and A. Androsov for their participation in software development.

References

1. Bourgeois, D., Mortati, J., Wang, S., Smith, J. (2019). Information Systems for Business and Beyond (2019). Information systems, their use in business, and the larger impact they are having on our world. Available at: https://opentextbook.site/exports/ISBB-2019.pdf
2. Kungurtsev, A. B., Potochniak, I. B. (2014). User interface for users communication with information systems in a natural language. Elektrotehnicheskie i komp'yuternye sistemy, 14 (90), 74–81. Available at: http://nbuv.gov.ua/UJRN/etks_2014_14_12
3. Kim, S. N., Cavedon, L. (2011). Classifying Domain-Specific Terms Using a Dictionary. In Proceedings of Australasian Language Technology Association Workshop, 57–65. Available at: https://www.aclweb.org/anthology/U11-1009.pdf
4. Kolle, P., Bhagat, S., Zade, S., Dand, B., Lifna, C. S. (2018). Ontology based Domain Dictionary. 2018 International Conference on Smart City and Emerging Technology (ICSCET). doi: https://doi.org/10.1109/icscet.2018.8537346
5. Deng, Q., Hine, M. J., Ji, S., Sur, S. (2019). Inside the Black Box of Dictionary Building for Text Analytics: A Design Science Approach. Journal of International Technology and Information Management, 27 (3), 119–159. Available at: https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=1376&context=jitim
6. Maynard, D., Bontcheva, K., Augenstein, I. (2016). Natural Language Processing for the Semantic Web. Morgan & Claypool publishers. Available at: https://tianjun.me/static/essay_resources/RelationExtraction/Paper/NaturalLanguageProcessingfortheSemanticWeb.pdf
7. Siddiqi, S., Sharan, A. (2015). Keyword and Keyphrase Extraction Techniques: A Literature Review. International Journal of Computer Applications, 109 (2), 18–23. doi: https://doi.org/10.5120/19161-0607
8. Tamsin Maxwell, K. (2016). Term Selection in Information Retrieval. University of Edinburgh. Available at: https://era.ed.ac.uk/bitstream/handle/1842/20389/Maxwell2016.pdf?sequence=1&isAllowed=y
9. Vivek, S. (2018). Automated Keyword Extraction from Articles using NLP. Available at: https://medium.com/analytics-vidhya/automated-keyword-extraction-from-articles-using-nlp-bfd864f41b34
10. Nokel, M., Loukachevitch, N. (2013). An Experimental Study of Term Extraction for Real Information-Retrieval Thesauri. Proceedings of 10th International Conference on Terminology and Artificial Intelligence, 69–76. Available at: https://istina.msu.ru/publications/article/4964490/
11. Kungurtsev, O., Zinovatnaya, S., Potochniak, I., Kutasevych, M. (2018). Development of information technology of term extraction from documents in natural language. Eastern-European Journal of Enterprise Technologies, 6 (2 (96)), 44–51. doi: https://doi.org/10.15587/1729-4061.2018.147978
12. Vavilenkova, A. I. (2017). Analiz i syntez lohiko-linhvistychnykh modelei rechen pryrodnoi movy. Kyiv, 152. Available at: https://er.nau.edu.ua/bitstream/NAU/42436/1/блок%20в%20печать.pdf
13. Kozlov, P. Yu. (2017). Automated analysis method of short unstructured text documents. Programmnye produkty i sistemy, 30 (1), 100–105.
14. Wahlin, L. (2020). Fundamentals of Engineering Technical Communications. A Resource & Writing Guide for the Fundamentals of Engineering Program. The Ohio State University. Available at: https://ohiostate.pressbooks.pub/feptechcomm/
15. Liang, S., Yilmaz, E., Kanoulas, E. (2016). Dynamic Clustering of Streaming Short Documents. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. doi: https://doi.org/10.1145/2939672.2939748
16. Punitha, S. C., Punithavalli, M. (2011). A Comparative Study To Find A Suitable Method For Text Document Clustering. International Journal of Computer Science and Information Technology, 3 (6), 49–59. doi: https://doi.org/10.5121/ijcsit.2011.3604
17. Hartmann, J., Huppertz, J., Schamp, C., Heitmann, M. (2019). Comparing automated text classification methods. International Journal of Research in Marketing, 36 (1), 20–38. doi: https://doi.org/10.1016/j.ijresmar.2018.09.009

18. Novokhatska, K., Kungurtsev, O. (2016). Application of Clustering Algorithm CLOPE to the Query Grouping Problem in the Field of Materialized View Maintenance. Journal of Computing and Information Technology, 24 (1), 79–89. doi: https://doi.org/10.20532/cit.2016.1002694
19. Fernández, J., Antón-Vargas, J. A., Villuendas-Rey, Y., Cabrera-Venegas, J. F., Chávez, Y., Argüelles-Cruz, A. J. (2016). Clustering Techniques for Document Classification. Research in Computing Science, 118 (1), 115–125. doi: https://doi.org/10.13053/rcs-118-1-11
20. Vtoraya mezhdunarodnaya konferentsiya «Upravlenie biznesom v tsifrovoy ekonomike»: sbornik tezisov vystupleniy (2019). Sankt-Peterburg. Available at: https://events.spbu.ru/eventsContent/events/2019/digital/tez_new.pdf
21. Sil'no korrelirovannye dvumernye sistemy: ot teorii k praktike: tezisy dokladov Vserossiyskoy konferentsii s mezhdunarodnym uchastiem (2018). Yakutsk: Izdatel'skiy dom SVFU. Available at: https://www.s-vfu.ru/universitet/rukovodstvo-i-struktura/instituty/fti/kres/conference/Сборник%20тезисов%20конференции/2D%20systems%20abstracts.pdf
22. Transport v integratsionnyh protsessah mirovoy ekonomiki (2020). Materialy Mezhdunarodnoy nauchno-prakticheskoy onlayn-konferentsii. Gomel'. Available at: https://www.bsut.by/images/MainMenuFiles/NauchnyeIssledovaniya/Konferencii/materialy/2020/transport_febt_2020.pdf
23. Tsifrovaya transformatsiya obrazovaniya (2018). Nauchno-prakticheskaya konferentsiya. Minsk. Available at: http://dtconf.unibel.by/doc/Conference.pdf
24. Obespechenie bezopasnosti zhiznedeyatel'nosti na sovremennom etape razvitiya obshchestva (2019). Materialy respublikanskoy studencheskoy nauchno-prakticheskoy konferentsii. Gorki, 69. Available at: https://baa.by/upload/science/conferencii/snk-bzd-19.pdf