8. Recognition based segmentation of connected characters in text based CAPTCHAs / Hussain R., Gao H., Shaikh R. A., Soomro S. P. // 2016 8th IEEE International Conference on Communication Software and Networks (ICCSN). 2016. doi: https://doi.org/10.1109/iccsn.2016.7586608
9. Abdullah Hasan W. K. A Survey of Current Research on CAPTCHA // International Journal of Computer Science & Engineering Survey. 2016. Vol. 7, Issue 3. P. 1–21. doi: https://doi.org/10.5121/ijcses.2016.7301
10. Anti-captcha. URL: https://anti-captcha.com/mainpage/
11. Myroniv I. Development of the character recognition software on the base cellular authomata // VI-th International Conference of Students, PhD-Students and Young Scientists “Engineer of XXI Century”. 2016. P. 229–240.
12. OpenCV library. URL: https://opencv.org/
13. Leonenkov A. V. Samouchitel’ UML. Sankt-Peterburg: BHV Peterburg, 2004. 576 p.
14. Fake Captcha is the #1 free fake captcha maker! URL: https://fakecaptcha.com/

UDC 004.4’413
DOI: 10.15587/1729-4061.2018.147978

DEVELOPMENT OF INFORMATION TECHNOLOGY OF TERM EXTRACTION FROM DOCUMENTS IN NATURAL LANGUAGE

O. Kungurtsev
PhD, Associate Professor*
E-mail: abkun@te.net.ua

S. Zinovatnaya
PhD, Associate Professor*
E-mail: zinovatnaya.svetlana@opu.ua

Ia. Potochniak
Postgraduate student*
E-mail: yana.onpu@gmail.com

M. Kutasevych*
E-mail: masteryoda290@gmail.com

*Department of System Software
Odessa National
Polytechnic University
Shevchenka ave., 1,
Odessa, Ukraine, 65044

Показано, що словники предметних областей широко використовуються на різних етапах створення і експлуатації програмних продуктів. Процес створення словника, особливо виділення термінів, досить трудомісткий та вимагає високої кваліфікації експерта. Проведено дослідження по виявленню найбільш важливих характеристик багатослівних термінів, таких як: ймовірності присутності в документі термінів, що містять різну кількість слів; розташування іменників в багатослівних термінах; можливу кількість іменників в багатослівних термінах. Проаналізовано контекст використання термінів та визначено можливі межі термінів в тексті. Запропоновано процедуру попереднього групування документів, що дозволяє уникнути «втрати» термінів, що входять в короткі документи. Визначено залежність помилок при виділенні термінів від розміру аналізованого документа.

Запропоновано математичну модель представлення терміна, що заснована на визначенні безлічі ланцюжків слів, згрупованих близько опорного слова – іменника. Фільтрація ланцюжків виробляється в залежності від частоти їх входження в текст на основі зіставлення нормалізованих уявлень багатослівних термінів.

Розроблено механізми заповнення словника предметної області новими записами і коригування існуючих у міру аналізу вхідного документа. Запропоновано рішення щодо коригування частоти появи термінів на основі виявлення міжфразових зв'язків. Всі процеси і моделі об'єднані в єдину інформаційну технологію створення словника предметної області. Проблема визначення тлумачень термінів в даній роботі не розглядається, оскільки вимагає окремого рішення. Розроблено програмний продукт, що дозволяє в значній мірі автоматизувати процес виділення термінів з текстових документів. Результати апробації запропонованих рішень показали відсутність «загублених термінів» і, як результат, скорочення часу виділення термінів з текстів обсягом в 10000 слів на 1.5 години за рахунок звільнення експерта від аналізу вихідного документа. Результати дослідження можуть бути використані на різних етапах створення і експлуатації програмних продуктів

Ключові слова: словник предметної області, багатослівний термін, морфологічний розбір, математична модель терміна, текстовий документ

1. Introduction

Domain dictionaries (DD) are widely used in software design [1]. In particular, when determining the roles of members of the development team [2]; when constructing data dictionaries [3, 4]; in the problems of selection and clustering of materialized database representations [5, 6]. Based on DD, job descriptions and many other documents are created. To construct a DD, an analysis of various texts used in a specific knowledge domain is made. These can be orders, reports, contracts, instructions and other documents sufficiently reflecting the activities of a particular organizational system. The main stage of DD construction is an extraction of terms from texts. In this case, the term means not only individual words, but also set phrases or multi-word terms (MWT) in a specific knowledge domain. Manual extraction of MWT requires long work of a highly qualified specialist [7, 8]. Therefore, research aimed at automating the process of MWT identification for DD construction is relevant.

2. Literature review and problem statement

In [9], the LEXTER software package for term extraction is proposed. It is of interest that terms are formed on the basis of allocation of nouns. Related words are determined by empirical rules, which limits the package scope to French. The statistical method of term extraction considered in [10] is applicable to Slavic languages. However, the authors solve the problem of term extraction in the context of document clustering and search for contrasting terms. This leads to a large number of false terms. In [11], the task to extract not only individual keywords, but also phrases united by frequency characteristics was set. However, the solution is proposed for the construction of hierarchical document clustering, when not all terms are to be defined and extracted terms contain no more than two words. It is of interest to study the extraction of key phrases [12], which can be used in the interpretation of terms in DD. However, from the point of view of term extraction, the key phrase requires further analysis. In [7], the method of automated preliminary text grouping and term extraction by frequency characteristics and simplified syntax rules is proposed. This allowed extracting terms from two and partially three words. However, term formation according to the "noun + adjectives" scheme did not allow extracting all multi-word terms, and a double pass through documents reduced productivity. In [13], the deep syntactic and semantic analysis of documents in natural languages is given. However, the proposed models are not brought to such a degree of formalization, which allows using them in applied problems. The interesting solution to reduce the complexity of keyword extraction by organizing parallel computing is proposed in [14]. However, the proposed algorithm is applicable only for extraction of single-word terms and loses effectiveness with a small number of documents, which is typical for narrow knowledge domains.

Thus, the task of term extraction for DD construction has a number of unsolved problems, namely:

– the study of characteristics of terms that allowing to formulate requirements for extracting them from the text (the number of words, arrangement and type of head-words, limits);

– preliminary text grouping, ensuring term detection in small documents;

– development of the technology providing the extraction of terms containing an arbitrary number of words;

– development of a software product implementing the proposed technology and allowing to approve the decisions made.

3. The aim and objectives of the study

The aim of the study is to reduce the time and improve the quality of term extraction from documents in a narrow knowledge domain.

To achieve the aim, the following objectives were formulated:

– to determine the characteristics of terms affecting the technology of their extraction from the text;

– to develop an information technology of term extraction from the text, including preliminary document grouping;

– to develop a software product and estimate the quality of term extraction.

4. Characterization of multi-word terms

To automate the process of MWT extraction, it was necessary to identify a number of MWT characteristics affecting the technology of this process. The defined characteristics use the concept of "head-word" – a noun included in the MWT. To work with MWT, the following characteristics were required:

– possible number of words included in the term;

– arrangement of "head-words" in the MWT;

– possible number of "head-words" in the MWT;

– definition of words and punctuation marks that limit the MWT.

The domain dictionary is used for the design and maintenance of the software product. Therefore, text documents from various fields of technology and applied sciences in Russian, Ukrainian and Belarusian languages were chosen for the study. 200 terms were extracted for each knowledge domain.

Fig. 1 shows the average probability distribution of occurrence of a certain number of words in the multi-word term. The scatter of values determined by a particular knowledge domain did not exceed 1–2 %.

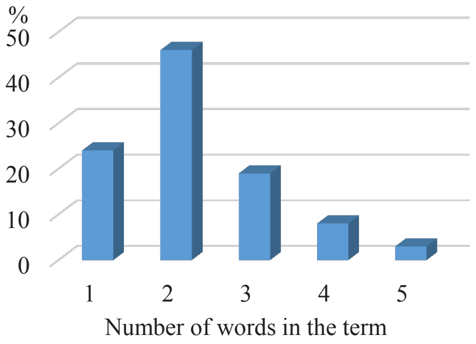


Fig. 1. Probabilities of occurrence of the term containing one or more words

Fig. 2 shows the results of the analysis of the arrangement of the head-word in the multi-word term. Nouns were chosen as the head-word, for example, for the term "information system" the head-word is "system". If the term contains more than one noun, as, for example, in the term "relational databases", each of them was assigned to the corresponding category ("bases" – in the middle, "data" – on the right).

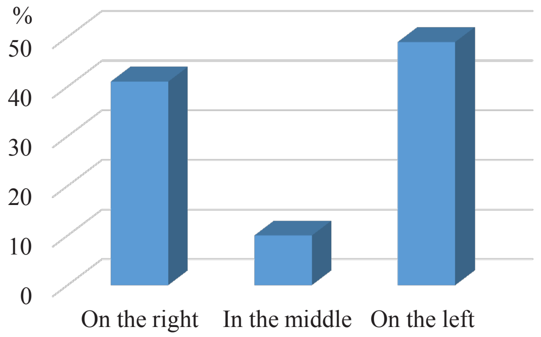


Fig. 2. Probabilities of arrangement of the head-word in the multi-word term

Fig. 3 shows the probability of occurrence of several head-words (nouns) in the MWT. From the above data, it follows that the probability of occurrence of several nouns (head-words) in the term is high. Therefore, the method of term extraction by a noun and related adjective [7] leads to large errors, and a deeper analysis of the connection of words requires complex syntax analysis. This confirms the need to look for a more efficient method of MWT extraction.

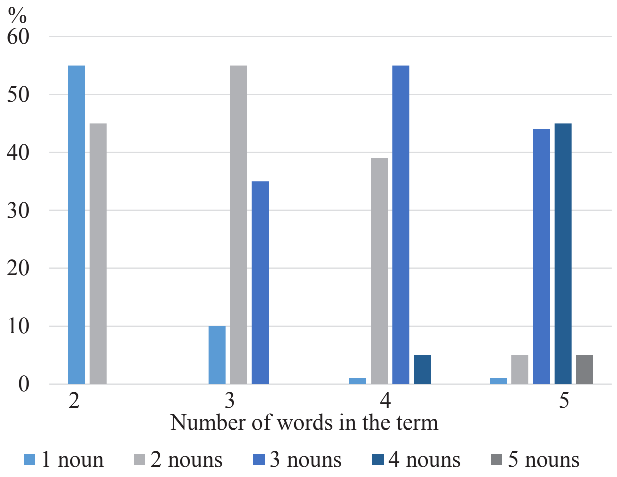


Fig. 3. Probability of occurrence of nouns in the MWT

Table 1 shows the results of determining possible MWT limits.

Table 1

Possible limits of MWT inclusion in the text

No.	Left limit	MWT inclusion	Right limit
1	Space	Included	Space
2	, space	Included	,
3	– space	Included	Space –
4	: space	Not included?	:
5	; space	Not included	;
6	. space	Not included	.
7	? space	Not included	?
8	! space	Not included	!
9	) space	Not included	Space (
10	» space	Not included	Space «
11	Pronoun space	Not included	Space pronoun

The case when the comma is included in the MWT (No. 2) turned out to be the only one out of 1,000 analyzed MWT (“decision makers”).

In accordance with the results of the study, the following conclusions were made:

– a multi-word term can be represented by no more than five words;

– the arrangement and number of head-words in the multi-word term can be any;

– the sequence of words included in the multi-word term should be limited at the left and right by punctuation marks or pronouns.

5. Technology of term extraction from the text

The technology involves a number of stages:

– selection and grouping of documents reflecting the knowledge domain;

– document format conversion;

– morphological analysis of the analyzed text, extraction of nouns;

– determination of possible MWT based on head-words;

– identification of inter-phrase relations and replacement of references with terms;

– calculation of the number of MWT occurrences in the text;

– dictionary updating by searching for the occurrence of some terms in another.

The proposed technology is applicable to the most common European languages. For Slavic languages, due to case declension, gender agreement and rather complex rules of plural formation, the mechanism for using the normalized form of their representation is additionally introduced for comparison of terms.

5. 1. Preliminary document grouping

In the process of knowledge domain (KD) analysis, the system analyst has to deal with a variety of documents in order to determine the requirements for the developed software product. These documents may represent various aspects of the organization’s activities. Term extraction from the entire set of documents as a whole can lead to an underestimation of those terms that are concentrated in separate small documents.

The processing of each document separately with a small amount of some of them may not provide for the accumulation of statistics. To determine the influence of the document size on the quality of term extraction, the study of a set of documents of different volume was conducted. It was believed that if some phrase occurred in the document once, then it would not be identified as a term in the automated method of term extraction. If the expert considers this phrase a term, this indicates a potential error in the automated term search. Based on the analysis of 100 documents with different number of words, the dependence of the probability of the term definition error on the size of the document was obtained (Fig. 4).

In [7], it is proposed to group documents on the basis of the normalized distance between them, which provides for a partial morphological analysis. In this paper, we introduce the concept of the volume v_i of some document D_i. If it turned out that $v_i<5,000$, then in accordance with Fig. 4, we assume that the document D_i should be combined with other documents into some integrated document T_x in order to achieve the volume of at least 5,000 words in the group:

$$T_x = \{D_i\} \mid \sum_{i=1}^{n} v_i \geq 5000.$$

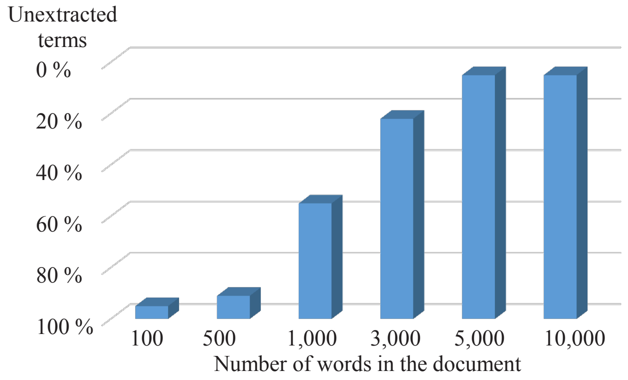


Fig. 4. Dependence of the term extraction error on the document size in words

The problem of document selection for the group can be assigned to an expert in the knowledge domain or systems analyst.

5. 2. Document format conversion

Known text analyzers [15] accept documents in the .txt format at the input, therefore the corresponding conversion is required:

$$T_x \Rightarrow T_{txt}.$$

Such conversion is performed in any text editor.

5. 3. Mathematical model of multi-word terms

The proposed model considers the single-word term as a special case of a multi-word term. As a result of processing of the text T_{txt}, a list of terms should be obtained. We will call this list a dictionary, because later on, when term interpretations are added to the list, it becomes a KD dictionary. At the stage of term extraction, the dictionary will be represented as a set of records

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

Each record has the following form:

$$r = < tm, lsn, nf, q >, \tag{2}$$

where *tm* is the set of term representations, *lsn* is the list of head-words (nouns) included in the term in the normalized form, *nf* is the normalized representation of the term, *q* is the number of occurrences of the term in the document.

Introduction of the normalized form of term representation is necessary only for Slavic languages. As an example, three sentences in Russian, English and French are given.

Мы работаем с ***реляционными базами данных****.*

Нами внесены изменения в ***реляционную базу данных****.*

Реляционная база данных *содержит множество таблиц.*

We work with ***relational databases****.*

We made changes to a ***relational database****.*

The ***relational database*** *contains a set of tables.*

Nous travaillons avec les ***bases de données relationnelles****.*

Nous faisons les changements dans la ***base de données relationnelle***

La ***base de données relationnelle*** *contient la multitude de tableaux.*

All sentences contain the term "relational database". However, the options of its representation in Russian significantly differ one from another, whereas in English and French these differences are minimum. Therefore, for comparing terms from texts in non-Slavic languages, fuzzy string matching can be successfully applied (for example, using the Levenshtein distance), while for texts in Slavic languages it is suggested to use the normalized form of term representation.

Representation of one term by a set of *tm* options is also used for Slavic languages, since it allows choosing the correct representation of the multi-word term containing several head-words at the end of the analysis. Each element of the set *c* consists of the same sequence of words. The elements differ in cases and number of the corresponding words. Table 2 illustrates the use of a set of representation options of one term in Slavic languages.

Table 2

Representation options of the multi-word term

Option number	Term		
	Belarusian language	Ukrainian language	Russian language
1	рэляцыйнымі базамі дадзеных	реляційними базами даних	Реляционными базами данных
2	рэляцыйную базу дадзеных	реляційну базу даних	реляционную базу данных
3	рэляцыйная база дадзеных	реляційна база даних	Реляционная база данных
	Normalized form of term representation		
	Реляцыйныя база дадзенае	реляційний база дане	реляционный база данное

The normalized form of representation *nf* is the same for all representation options of the term. In [13], it is proposed to compare the content of texts by means of a special linguistic processor. Using the normalized form allows comparing multi-word terms using a very simple procedure of non-fuzzy string matching, which significantly reduces text processing time.

For non-Slavic languages, the set *tm* from (2) will contain one element (term), and *nf* – the same term.

The list of the head-words of the term *lst* at the completion of forming the dictionary *D* will allow choosing the most appropriate representation of the term *tm*. According to the diagram in Fig. 1, a multi-word term can include up to 5 words. According to the diagram in Fig. 2, the head-word can take any position in the multi-word term. Therefore, it is proposed to form all possible groups of words relative to the head-word. In order to reduce the number of possible groups, the set of types of left and right limits of MWT is defined in accordance with Table 1:

$$B = \{":", ";", ".", "?", "!", "(", ")", "<<", ">>", \backslash pron\}. \tag{3}$$

It is proposed to form possible terms as sequences of 5, 4, 3, 2 and one word containing at least one head-word. First assume that the sequence will include only one head-word.

We represent a piece of the text S as a sequence of elements:

$$e_1,...,e_l,..,e_m. \quad (4)$$

An element can be a single word or a punctuation mark. Each word is represented by a sequence of letters W (directly from the text), the set A, the normalized form of representation nf (the analyzer result):

$$e = < W, A, nf >. \quad (5)$$

We define the attributes that will be needed to determine the MWT limits, as well as to take into account the inter-phrase relations [16]. Let $A1$ represent a part of speech, $A2$ – number, $A3$ – gender, $A4$ – person, $A5$ – case.

Punctuation marks are represented only by their spelling $e = < W, \varnothing, * >$.

Let some element be the head-word e_0=<W, A, q>, where $A1$=*noun* (noun), q – the number of occurrences of the term in the text S.

We formulate the rules for composing sequences of words:

– the sequence is formed of nearby elements;

– the head-word must be included in the sequence;

– the number of elements in the sequence should not be more than 5 and less than 1 (punctuation marks included in the sequence are not taken into account);

– the sequence can be limited to the left or right of the head-word, if some element of the sentence e_i, provided that $e_j \in B$.

Let some text contain a sequence of elements:

$$e_{-5}e_{-4}e_{-3}e_{-2}e_{-1}e_0e_1e_2e_3e_4e_5,$$

where e_0 is the head-word.

Then the possible sequences of words (without limits) will be as follows:

$$\begin{bmatrix} e_{-4}e_{-3}e_{-2}e_{-1}e_0 \\ e_{-3}e_{-2}e_{-1}e_0e_1 \\ e_{-2}e_{-1}e_0e_1e_2 \\ e_{-1}e_0e_1e_2e\ e_3 \\ e_0e_1e_2e_3e_4 \\ e_{-3}e_{-2}e_{-1}e_0 \\ ...,e_{-1}e_0 \\ e_0e_1e_2e_3e_4 \\ ...,e_0e_1 \end{bmatrix}. \quad (6)$$

The formula for determining the number of possible combinations is proposed:

$$K = \sum_{i=0}^{i \le 5-2} (5-i) = 14. \quad (7)$$

We consider possible limits for combinations. Let some element $e_j \in B$. Then all combinations including elements with indices $i \le j$ are excluded from further analysis. The formula for determining the number of possible combinations under the left limit is:

$$K_l = 14 - \sum_{i=1}^{5-j} (5 - j + i - 1). \quad (8)$$

Let us consider a more general case, when a group may include more than one head-word. Let some text have a sequence of elements:

$$e_i....e_j^*...e_k^*...e_l,$$

where e_j^* and e_k^* are head-words. Then, provided that:

$$k - j \ge 5. \quad (9)$$

Formed word sequences will contain one head-word as terms. If k–j<5, then, using the previously described method of forming word sequences separately for the head-word e_j^* and for the head-word e_k^*, we obtain repeated sequences. For example, for the sentence fragment:

$$e_{=4}e_{-3}e_{-2}e_{-1}e_0^*e_1e_2e_3^*e_4e_5e_6e_7.$$

On the basis of e_0^*, we obtain the following sequences with two head-words:

$$\begin{bmatrix} e_{-1}e_0^*e_1e_2e_3^* \\ e_0^*e_1e_2e_3^*e_4 \\ e_0^*e_1e_2e_3^* \end{bmatrix}.$$

And on the basis of e_3^*, we obtain the same sequence with two head-words:

$$\begin{bmatrix} e_{-1}e_0^*e_1e_2e_3^* \\ e_0^*e_1e_2e_3^*e_4 \\ e_0^*e_1e_2e_3^* \end{bmatrix}.$$

It will be shown below how to eliminate repeated word sequences in the dictionary.

The number of possible word sequences in the presence of several head-words in a sequence depends on the number of head-words, but cannot exceed K from (6) per one head-word.

5. 4. Inclusion of the word sequence in the dictionary

Each sequence of words E, obtained after the accounting of limits, should be represented by the record (2) in the dictionary (1). For this purpose, we define its normalized form E_{nf}. We introduce the notation for the belonging of some word sequence to the dictionary $E \in_e D$. If:

$$r_i \mid r_i \in D \wedge r_i.nf = E_{nf},$$

then the combination of words is already present in the dictionary. In this case, we increase the number of occurrences by 1 ($r_i.q$: $r_i.q$+1) and check the occurrence of E in $r_i.tm$. If $E \in r_i.tm$, then we add a new version of the term to the set of options $r_i.tm = r_i.tm \cup \{E\}$.

If $E \notin_e D$, then we form a new record in the dictionary:

$$r = < E, lsn, E_{nf}, 1 >,$$

where *lsn* will contain all nouns (head-words, selected at the stage of building sequences) from *E* in the normalized form.

5. 5. Accounting of inter-phrase relations

The main criterion of term selection is the frequency of occurrence in the analyzed text. Inter-phrase relations occur if a term in the subsequent sentences is replaced with a pronoun, ordinal number, etc. For example, in the sentence *"Hard drive is the main data storage device for the majority of personal computers."*, the phrase *"Hard drive"* can be defined as a term. In the next sentence, *"Usually it is characterized by capacity and speed."*, the term *"Hard drive"* is replaced with the pronoun "it". If the relation between the sentences is not found, then only one occurrence of the term *"Hard drive"* will be defined. In the present study, we used the results obtained in [16], where algorithms for identifying inter-phrase relations are presented. Let some element of the sentence e_i be an anaphor (replacement or reference) of the previously found term $e_i \rightarrow r_i.t$, then the number of occurrences of $r_i.t$ in the text should be increased:

$$r_i.q := r_i.q + 1.$$

Here the sign «:=» means assignment.

5. 6. Updating of the dictionary

For each term, it is necessary to introduce the lower limit *Be* of the number of term occurrences in *G*:

$$\forall r \in G \,|\, r_i.m \geq Be.$$

The minimum value of the lower limit is *Be*=2. With this value of *Be*, some sequence of words, extracted in accordance with (8), repeatedly occurred in the text. For large texts, the value of *Be* can be increased. It is recommended to entrust this operation to an expert in the knowledge domain. Consistently increasing the *Be* value, the moment should be fixed when all important terms for the given knowledge domain still remain in the dictionary.

As a result of the analysis of the document, the terms that are included in other terms may occur. The question of keeping such terms in the dictionary or excluding them from the dictionary depends on their independent use in the text. The procedure of dictionary updating provides a comparison of records. If:

$$r_i.nf \in_t r_j.nf \wedge r_i.q = r_j.q,$$

the record r_i is excluded from the dictionary.

If:

$$r_i.nf \in_t r_j.nf \wedge r_i.q > r_j.q,$$

then it is necessary to analyze $\Delta = r_i.q - r_j.q$. If $\Delta \geq Be$, then the record r_i is not excluded from the dictionary.

After determining the terms to be included in the dictionary, it is necessary to choose one of the representation options of each term in the set *tm*. For this purpose, we introduce the concept of "main word" in the term. There are a number of signs that distinguish it from other words:

– it must be a noun (mandatory);

– it usually ranks first among other nouns in the term;

– its spelling options (case and number variations) usually define various options of term representation in *tm*.

Thus, the process of choosing an option of term representation involves the following sequence of actions.

We determine the number of elements of the set *tm*.

If $tm|=1$, then there is only one option of term representation in the dictionary.

If $|tm|=k \wedge k>1$, then the number of head-words in the list *lsn* is determined.

If $|lsn|=1$, then there is only one head-word $w_1 \in lsn$ in the term. Its position *j* in options of term representation is determined based on the position of this word in *r..nf*:

$$w_1 = w_j \,|\, w_j \in r..nf. \tag{10}$$

Then, the word $w_{i,j}$ in the position *j* is selected from each representation option of the term $tm_i \in tm$ and compared with the normalized representation.

If:

$$w_{i,j} = w_1, \tag{11}$$

then all elements except tm_i are removed from the set *tm*, that is, $tm=\{tm_i\}$.

If the condition (11) is not met, then:

$$tm = \{tm_1\}, \tag{12}$$

and the problem of formulating the term definition is solved by an expert.

If $|lsn|=l \wedge l>1$, then there are several head-words in the definition of the term. For each head-word $w_p \in lsn|p=1, l$, its position *j* is determined in options of term representation in accordance with (10).

Further, from each representation option of the term $tm_i \in tm$, the word $w_{i,j}$ in the position *j* is selected and compared with the normalized representation.

If $w_{i,j}=w_p$,then all the elements except tm_i are removed from the set *tm*, that is, $tm_i \in tm$ and the cycle of searching for the best option of term representation is completed. Otherwise, $p=p+1$ and the cycle continues. If the best representation was not found, then the decision is made in accordance with (12) and the expert solves the problem of formulating the term definition.

6. Development of the software product and assessment of the quality of term extraction

To implement the proposed technology and models, the TermsSelect software product was developed. The scheme of document processing is presented in Fig. 5.

Fig. 6 presents the window allowing the expert to edit the list of terms found in the text. The terms were obtained as a result of the analysis of 15 texts on the subject "Materials and technologies of ceramics production" with a total volume of about 20,000 words [17]. The arrangement of terms is determined by the first noun. The content of the first column "Term" is subject to editing. In addition, the expert can remove a row from the table or enter a new term.

The purpose of testing the software product was a comparative assessment of new and previously existing technologies by time characteristics and quality of term extraction. Quality was understood as the percentage of errors of the first kind (“excess” terms) and the second kind (“lost” terms) of the total number of the terms found.

To test the proposed technology and software product, texts from various fields of science and technology were used. As a result of the experiments, it was found that when using TermsSelect, the average time of term extraction from the document of 10,000 words was 15.6 seconds. The timing of the expert’s work on the extraction of terms and their frequency characteristics “manually” gave the result of about 10 hours. The simplified task – term extraction only was performed by the expert within 1.5 hours. During the term extraction with the program, “excess terms” were found. They made up about 5 % of the extracted terms. At the same time, “lost terms” were not found. It should be noted that the removal of “excess terms” does not require a special procedure, since in all cases the list of extracted terms is viewed by the expert.

For comparison, testing of the DictionaryCreator software product, proposed in [7] was carried out. Here, time of term extraction was 12.4 seconds for the document of 10,000 words. However, the number of “lost terms” was 22 % (mostly terms of three or more words). Definition of “lost terms” is a very labor-intensive procedure that can only be performed manually. Thus, with an insignificant increase in the time of text processing, it was possible to obtain a significant improvement in the quality of extraction of multi-word terms.

7. Discussion of the results of research on the speed and quality of term extraction

A significant reduction in the number of “lost terms” at a high speed of processing of source texts is explained by two main solutions:

– the proposed method of forming potential terms as all admissible chains of words located near the head-words;

– preliminary grouping of short documents.

Representation of a term as a set of word chains allows defining terms as a subset of chains that are repeated in the text. Such a principle can be used for the majority of natural languages and requires only morphological analysis. Grouping of short documents for the period of analysis allows finding terms that occur in a document once. The proposed solution requires the expert to only edit the term included in the dictionary. The existing solutions for determining the frequencies of single words are characterized by high speed, but leave a lot of work to the expert related to the analysis of source texts. The methods of term extraction as a noun with related adjectives do not cover the whole variety of terms. The speed of such method is commensurate with the one proposed in this work, however, a large number of “lost terms” also require the expert to work with the source text.

The studies were limited to Slavic and most common European languages, for which the concept of the head-word can be introduced. They can not be applied, for example, to Vietnamese and other languages like Chinese.

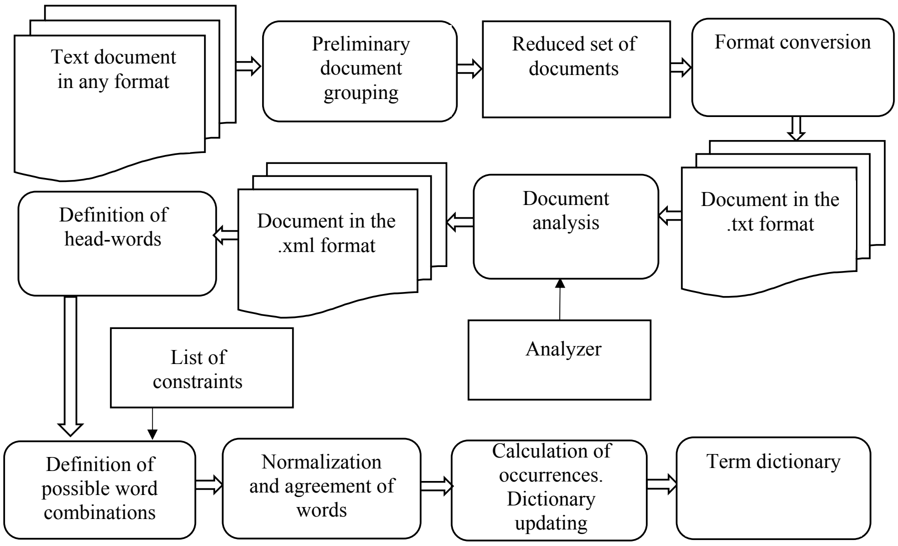
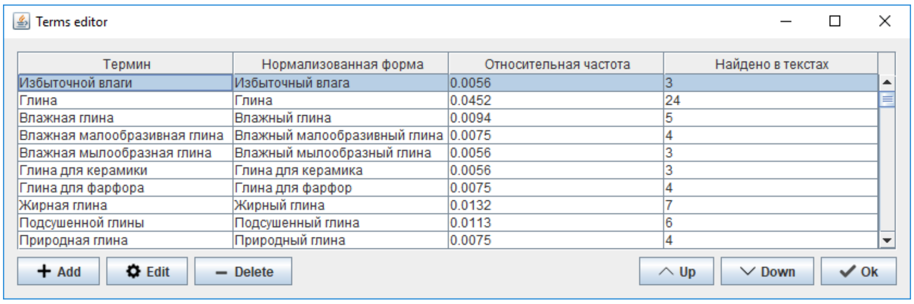

Text document in any format
Preliminary document grouping
Reduced set of documents
Format conversion
Document in the .txt format
Document analysis
Document in the .xml format
Definition of head-words
Analyzer
List of constraints
Definition of possible word combinations
Normalization and agreement of words
Calculation of occurrences. Dictionary updating
Term dictionary

Fig. 5. Functional diagram of the TermsSelect program



Термин	Нормализованная форма	Относительная частота	Найдено в текстах
Избыточной влаги	Избыточный влага	0.0056	3
Глина	Глина	0.0452	24
Влажная глина	Влажный глина	0.0094	5
Влажная малообразивная глина	Влажный малообразивный глина	0.0075	4
Влажная мылообразная глина	Влажный мылообразный глина	0.0056	3
Глина для керамики	Глина для керамика	0.0056	3
Глина для фарфора	Глина для фарфор	0.0075	4
Жирная глина	Жирный глина	0.0132	7
Подсушенной глины	Подсушенный глина	0.0113	6
Природная глина	Природный глина	0.0075	4

Fig. 6. Editing of the term dictionary

The disadvantages of the study include the representation of the extracted term in the form, which in some cases requires editing by an expert. Attempts to present a multi-word term in the final form without the use of known labor-intensive methods of text generation have so far failed.

In addition to the generally accepted concept of a term in the texts representing a narrow knowledge domain, specific abbreviations and names (programs, processes, machines, etc.) can be used, which can also be attributed to terms. Extraction of such terms requires the formalization of the concepts of “abbreviation”, “name” and is a continuation of this study. The domain dictionary should contain an interpretation of terms, which is currently performed manually. Automation of this process involves the search for relevant sources of information and selection of suitable pieces of text. This problem also requires further research.

8. Conclusions

1. Such parameters of terms as the possible number of words

included, the possible number and arrangement of nouns in the term, as well as possible limiters of the chain of words included in the term are determined. The results of the study are needed to construct a mathematical model of the term.

2. The information technology of term extraction from text documents, containing document grouping; mathematical model of the term, allowing to extract it from the sentence; adjustment of the frequency of terms based on the identification of inter-phrase relations and occurrence of some terms in others is developed. The technology allows term extraction without a detailed syntax analysis of the sentence, which significantly reduces the processing time of the document.

3. The TermsSelect software product, implementing the proposed technology is developed. Text documents in any standard formats were submitted to the input. To allocate parts of speech and obtain the normalized form of word representation, freely available plug-in text analyzers were used. The maximum length of the word chain was set equal to five. The expert's task was only the editing of terms. The analogue was the earlier developed DictionaryCreator software product, which extracts terms as nouns and syntactically related adjectives. Comparative tests of the products on the same texts showed that with almost the same time spent on text processing, TermsSelect found all the terms, and DictionaryCreator found 78 % of the terms. The search for "lost terms" was estimated at 1.5 hours of work of the expert work. Thus, the achieved improvement in the quality of term extraction significantly reduced the total time of term extraction.

References

1. Izbachkov Yu. S., Petrov V. N. Informacionnye sistemy: ucheb. Piter, 2011. 544 p.
2. Liubchenko V., Sulimova I. Examining the attributes of transitions between team roles in the software development projects // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 1, Issue 3 (85). P. 12–17. doi: https://doi.org/10.15587/1729-4061.2017.91597
3. Best Practices for Data Dictionary Definitions and Usage Version 1.1. 2006. URL: https://s3.us-west-2.amazonaws.com/org-pnamp-assets/prod/best_practices_for_data_dictionary_definitions_and_usage_version_1.1_2006-11-14.pdf
4. 10 Ways Data Dictionary Increases Software Developers Productivity. URL: https://dataedo.com/blog/ways-data-dictionary-increases-software-developers-productivity
5. Novokhatska K., Kungurtsev O. Application of Clustering Algorithm CLOPE to the Query Grouping Problem in the Field of Materialized View Maintenance // Journal of Computing and Information Technology. 2016. Vol. 24, Issue 1. P. 79–89. doi: https://doi.org/10.20532/cit.2016.1002694
6. Novokhatska K., Kungurtsev O. Developing methodology of selection of materialized views in relational databases // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 3, Issue 2 (81). P. 9–14. doi: https://doi.org/10.15587/1729-4061.2016.68737
7. Kungurcev A. B., Potochnyak Ya. V., Silyaev D. A. Method of automated construction of explanatory dictionary of subject area // Technology audit and production reserves. 2015. Vol. 2, Issue 2 (22). P. 58–63. doi: https://doi.org/10.15587/2312-8372.2015.40895
8. Califf M., Mooney R. J. Bottom-up relational learning of pattern matching rules for information extraction // Journal of Machine Learning Research. 2003. Vol. 4. P. 177–210.
9. Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases // COLING '92 Proceedings of the 14th conference on Computational linguistics. 1992. P. 977–981. DOI: https://doi.org/10.3115/993079.993111
10. Method of rare term contrastive extraction from natural language texts / Bessmertny I. A., Nugumanova A. B., Mansurova M. Y., Baiburin Y. M. // Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2017. Vol. 17, Issue 1. P. 81–91. doi: https://doi.org/10.17586/2226-1494-2017-17-1-81-91
11. Popova S. V., Hodyrev I. A. Izvlechenie klyuchevyh slovosochetaniy // Nauchno-tekhnicheskiy vestnik Sankt-Peterburgskogo gosudarstvennogo universiteta informacionnyh tekhnologiy, mekhaniki i optiki. 2012. Issue 1 (77). P. 67–71.
12. Hasan K. S., Ng V. Automatic keyphrase extraction: a survey of the state of the art // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014. P. 1262–1273. doi: https://doi.org/10.3115/v1/p14-1119
13. Vavilenkova A. Methods of identifying logical connections between parts of text documents // Bulletin of the National Technical University «KhPI» Series: New solutions in modern technologies. 2017. Issue 7 (1229). P. 118–122. doi: https://doi.org/10.20998/2413-4295.2017.07.16
14. Realizaciya algoritma izvlecheniya klyuchevyh slov iz tekstov predmetnoy oblasti na osnove modeli MapReduce / Bessmertniy I. A., Karimov A. T., Novoselov A. O., Nugumanov A. B. // Trudy VIII Mezhdunarodnoy nauchno-prakticheskoy konferencii «Sovremennye informacionnye tekhnologii i IT-obrazovanie». 2013. P. 617–624.
15. Programmniy paket sintaksicheskogo razbora i mashinnogo perevoda. URL: https://www.cognitive.ru/
16. Uchet mezhfrazovyh svyazey pri avtomatizirovannom postroenii tolkovogo slovarya predmetnoy oblasti / Kungurcev A. B., Gavrilova A. I., Leongard A. S., Potochnyak Ya. V. // Informatika i matematicheskie metody v modelirovanii. 2016. Issue 2. P. 173–183.
17. Materialy i tekhnologiya izgotovleniya keramicheskih izdeliy. URL: http://art-con.ru/node/233