Method of searching term interpretations for domain dictionaries, used for developing software

[1]O. Kungurtsev, [2]Ia. Potochniak,[3]E. Pilyaev

[1]PhD, Associate Professor,Department of System Software,Odessa National Polytechnic University,Odessa, Ukraine
[1]Ph.D., Professor, Department of System Software of Odessa National Polytechnic University,Odessa, Ukraine
[2]Postgraduate student, Department of System Software of Odessa National Polytechnic University, Odessa, Ukraine
[3]Student, Department of System Software of Odessa National Polytechnic University, Odessa, Ukraine
CorrespondingAuthor: O. Kungurtsev

ABSTRACT: The paper proposes a method for the formation of definitions of terms for the domain dictionary using existing explanatory dictionaries. The mathematical description of procedures of search of interpretations and algorithms which realize them is offered. The method of automated search for the interpretation of terms, which is characterized by the usage of existing dictionaries, the assessment of the quality of interpretation by the presence of terms from the subject area, the synthesis of interpretation in case of a mismatch of terms, what allows to reduce the time of the expert search through a dictionary is offered. Software that allows to significantly speed up the process of searching for interpretations of terms in a domain dictionary has been developed.

KEYWORDS: domain dictionary, one-word term, multiword term, morphological analysis, mathematical model of the term, interpretation definition option, text document.

I. INTRODUCTION

Usually, a domain dictionary (DD) is a specialized explanatory dictionary, which gives definitions to a set of concepts related to the activities of a certain organizational structure [1]. DD is used to solve a variety of tasks related to the creation and maintenance of software products (SP): formulation and coordination of requirements for SP, database design, creation of user interfaces, writing various manuals, etc. [2].

There are two main types of dictionaries according to their content: encyclopedic and linguistic [3]. The object of the description in the encyclopedic dictionary and encyclopedia – various objects, phenomena and concepts; the object of the description in the linguistic dictionary – a unit of language, most often a word.

By the way the material is organized, linguistic dictionaries are divided into alphabetic (most common), family (one dictionary entry interprets not the word, but the entire word-formation family) and semi-family (derived words of a different grammatical category than the headword).

The main type of the linguistic dictionary is the explanatory dictionary. Explanatory dictionaries differ in the volume of the dictionary, technical means of presenting the material. Therefore, before using the dictionary, it is necessary to get acquainted with the "System notes" – conditional abbreviations (usually found in the intro to the dictionary).

According to the functions and purposes of creating, explanatory dictionaries are divided into descriptive and normative.

Descriptive dictionaries are designated for a complete description of the vocabulary of a certain sphere and fixation of all available uses. ("Explanatory dictionary of the common great Russian language" V. I. Dal [4], "Explanatory dictionary of the English language"Oxford Dictionaries Online [5]).

The purpose of the normative dictionary is to show the standard use of the word, eliminating not only the wrong use of words associated with an erroneous understanding of their meanings, but also those uses that do not correspond to the communicative situation.

The dictionary consists of dictionary entries [3]. Dictionary entry is a brief linguistic information about a word. In it's turn, the dictionary entry consists of the following components:

- the headword (usually in bold and capital letters) with the emphasis. Sometimes it contains comments to the pronunciation in square brackets;
- interpretation of the word contains:

a) notes – a brief description the words expressed by the adopted reduction of the corresponding term describing the use of the word (usually in italics);
b) definition (dictionary definition);

- illustrative material as a means of word semantics;

- collocations based on the headword;
- derived words (a sign of the family dictionary).

Compilation of explanatory dictionary for a certain language or a certain branch of knowledge is a very time – consuming and poorly automated process that requires many months of work of highly qualified specialists. Therefore, when compiling the DD, there is a task to automate and speed up the process of dictionary compiling.

II. ANALYSIS OF THE LITERATURE DATA AND FORMULATION OF THE PROBLEM

There are a number of works on creation of DD in English [6], Russian [7], Ukrainian [8, 9] and other languages.

[1]. The paper focuses on the automated selection of terms from the texts in Russian from a given subject area. The interpretation of the terms is entrusted to the expert, who is offered to use a pre-selected set of dictionaries online.

[8]. The paper deals with the automation of the allocation of terms from the texts in the Ukrainian language, but there is no solutions to automate the process of interpretation of terms.

[3, 10]. A number of studies have considered in detail the structure of existing dictionaries and the conditions for the successful search of definitions, however the algorithms for their selection in the automated search is not proposed.

In addition, the dictionary entry of the explanatory dictionary is mainly intended to define one word, while the terms of the subject area, as a rule, contain several words, for example, "operating system", "waybill", "medical history", "schedule of commuter trains".

[11]. There are certain rules for the interpretation of collocations based on the headword, but there are not enough of formalized rules for their allocation.

Therefore, the automation of the retrieval of the interpretations of the terms is highly relevant and largely unresolved challenge.

The problem of interpretations highlighting is the high complexity of their definition.

The proposed solution is to automate the process of determining the definitions of terms for the DD based on the allocation of interpretations from the analyzed document and public dictionaries.

To solve the problem, the following tasks are formulated:

- determination of the conditions for determining definitions directly from the analyzed document;
- allocation of definitions of the term from the dictionary;
- filtering of definitions depends on notes;
- filtering definitions based on the analysis of the entry of terms from the DD;
- selection and layout of definitions for multiword terms.

III. SEARCH OF INTERPRETATIONS COMBINED WITH A SELECTION OF TERMS FROM DOCUMENTS

The analyzed document may introduce some new concepts (terms) or provide a new interpretation of the known ones. Then the definition of the term can be included in the text in close proximity to the term itself. To verify the effectiveness of the interpretation search directly in the text on the basis of which the DD is build, the analysis of text documents from various subject areas with a total volume of 50,000 words was carried out. As a result, 257 terms were allocated. For the 27 terms interpretation were found directly in the analyzed texts. On the basis of the study it was concluded that the development of methods of interpretation search combined with the selection of terms is futile.

IV. THE USE OF EXISTING EXPLANATORY DICTIONARIES

We assume that the search for definitions of terms in all cases will use pre-compiled specialized or broad-profile dictionaries.

At Fig.1 the dictionary entry is presented in a simplified form.

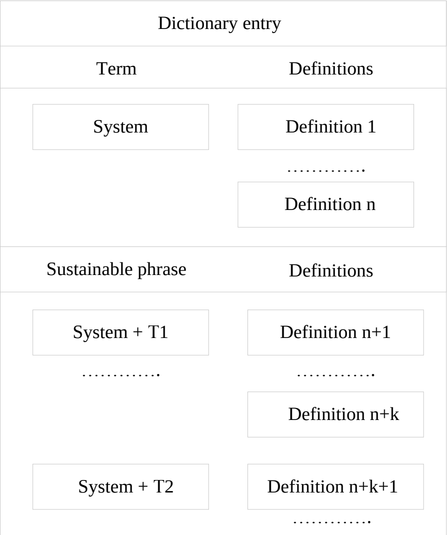


Figure 1. Simplified structure of the dictionary entry

If you want to find an interpretation of a one-word term, you should choose one or more definitions from Definition1 ... Definition n (for example, for the term "System"). If the task is to find a definition for a multiword term (for example," Operating system"), then you should look for a suitable collocation (in Fig.1 is the system + T1, system+T2) and make a choice from the definitions that apply to it.

For the formation of DD, we will use a workpiece for open DD without interpretations (contains only terms and their frequency characteristics), as well as some downloaded DD with online access, organized through some system that allows to extract information from sites such as Interpretatio [12].

Let us present the explanatory DD in the form of:

$$To = \{< to, tt >\} \ , \qquad (1)$$

where to is a term from the subject area;
tt– interpretation of the term (one or more sentences, originally an empty line– tt=«»).
Let's present the explanatory dictionary of a wide profile in the form of:

$$Td = \{< td, ma >\} \ , \qquad (2)$$

where td is a term from the Td dictionary;
ma-multiple interpretations of the term td (dictionary entry).
In general, the term to can consist of several words:

$$to = e_1 ... e_n \qquad (3)$$

Search for interpretations of one-word terms (OT)

We will assume beforehand that the term to is OT (to = e_1). Then the task of interpreting to can be formulated as follows. If td is found, such that td = to, then from a variety of interpretations ma, we need to choose one or more interpretation definition options (IDO), which for some formal features is most suitable for the subject area.

Let's set the task of reducing possible IDO. Typically, dictionaries contain abbreviations that define the scope of the term, for example, area of activity (Mat. – mathematics, mus. - music, comp. sc. – computer science), style (folk-poet – folklore-poetic, dismis. – dismissive), etc. We call such reductions characteristics of the option. Then IDO interpretation of the term (family) can be represented as:

$$ma_i = \{< mc, tx >\} , \qquad (4)$$

where mc is the set of IDO characteristics;
tx – IDO text.
Since not every IDO can contain characteristics, it is possible that mc = Ø.

For a particular domain we will create many invalid mCacharacteristics and lots of options for mCicharacteristics. This allows us to distinguish from the set of ma a subset of ma' interpretations of the term that satisfy the following conditions:

$$\left.\begin{array}{l} \exists c_j \mid c_j \in mc \wedge c_j \in mCi \\ \forall c_j \mid c_j \in mc \wedge c_j \notin mCa \end{array}\right\} \qquad (5)$$

Since not every family has a characteristic and several different families can have the same characteristics, after selecting the IDO according to the conditions (4), there may be several of them. To further reduce the number of IDO, it is proposed to select the most suitable IDO from ma' by counting the number of occurrences of terms from Toin each tx from ma'. To do this, let's introduce ma' as:

$ma' = \{< tx, k >\}$,

where k is the number of term occurrences from Tototx;
tx is represented as a sequence of words (tx = $e_1 \dots e_n$).
We write operation of calculation of the number of occurrences of term from To in tx (preliminary k = 0):

$(if\,((e_i = to_j) e_i \in tx \wedge to_j \in To) k := k+1) i = 1, n; j = 1, q$ (6)

whereq = |To|.
After calculation k all IDO, which are not part of the terms of the subject area:

$\forall tx_j \mid tx_j \in ma' \wedge k_j = 0$,

excluded from ma', and the remaining are sorted in ascending order by k.

$ma' = \{< tx_j, k_j >\} l = 1, m-1; k_l \geq k_{l+1}$,

where m = |ma'|.
Thus, the expert is provided with m IDO to select and edit the term.
The search method of the multiword term interpretations (MT).

There are dictionaries of word combinations [5], but they represent only a small number of commonly used word combinations. In ordinary explanatory dictionaries word combination are used as a IDO of some basic term. This determines the next stages of interpretation search for MT.

A) the allocation of the MT nouns. Since nouns carry the main semantic load, the interpretation of MT suggests the search for interpretation to start with nouns.
We introduce additional information into the MT submission in accordance with (3). To do this, each element (word) is represented as:

$e_i =< w_i, w'_i, a_i >$, (7)

where $w_i$represents one of the words of MT,
w'_i – represents w_i in normalized form,
a_ii - part of speechw_i (noun, adjective, numer, ...).

B) definition of search priorities for nouns. Some nouns that are included in MT can represent OT in the subject area under study. It is recommended to start the search for interpretations with these nouns. This will allow you to partially combine the search of MT and OT. In addition, it is possible to assume that a noun used only as a part of MT has a highly specialized meaning, which will make it difficult to find it in Td. Thus, we form a sequence of words to search for IDO in Td.

$to' = w'_1, \dots, w'_i, \dots, w'_n$, (8)

where their location is subject to the following condition:

$(\forall w'_i \mid (w'_i \in To) \exists w'_{i-1} \in To) \wedge (\forall w'_i \mid (a_i = noun) \exists w'_{i-1} \mid a_{i-1} = noun); i = 2, n$

B) Search of OT in the Td. For each element of (8) satisfying the condition $w'_i \in To$, is searched according to IDO in paragraph 1.
G) Search of IDO for the MT based on noun. Dictionary entry Td can have a number of interpretations of collocations, which include the defined term. The generally accepted word order in such a phrase can be changed. Often defined term (represented as the first letter of the word) takes the first position in the word combination. In accordance with the above, it is proposed to present the explanatory dictionary entry of the Td in the form of:

$ar =< td, \{s, ts\} >$,

where S is a word combination;
ts is IDO, corresponding to the interpretation of the term td in this word combination.
Each phrase will present many of its constituent words, with previous deletion of linking words:

$s = \{ws_i\} i = 1, n$,

Similarly, let's pesent MT:

$to = \{wt_j\} j = 1, k$,

We define a set of ma' IDO for MT that can be provided to an expert. Originally ma'= Ø. Since there are no assurance that in Td will be found the word combination matching with to, it makes sense to memorize the word combination, which partially correspond with to.

The terms of placement ofS inma' is:
$(X = to \cap s) \wedge |X| \geq 2\ ,\ to \subseteq s$

Therefore, if an S containing two or more words from to is found, then such an incomplete interpretation is included in ma'. Also ma ' includes all S that match with to or have additional words.
E) assessment of the relevance of IDO from ma'. If one or more IDO are found, such that to = S then all s ≠ to are excluded from ma'.
If |ma'| ≥ 2, then it is necessary to estimate each IDO entering ma'. To do this, in accordance with the operation (6) the occurrences of terms from a Toin each IDO of the ma' must be calculated. Elements of the ma' set are sorted in descending order by the number of term occurrences. IDO that do not contain terms or have a small number of term occurrences are discarded.
(F) search for IDO based on all the words in the MT. If the search for IDO on the basis of nouns did not give results, in accordance with the previously described procedure, the search for IDO is based on all the words that make up the MT.
G) search for IDO which partially cover MT. If, as a result of the search for IDO in accordance with paragraphs D) and E) in the set of ma' there are no elements such that to = S, it is proposed to use phrases containing only part of the words included in MT.
We assume that the number of words in each element ma' is a measure of its correspondence to MT. To do this, we order the elements of ma' in descending order of their power:
$\forall(s_i \in ma')\exists s_{i-1} \in ma' \,|\, (|s_{i-1}| \geq |s_i|); i = 2, n\,,$
Similarly, to the point E) let's assess the relevance of IDO from ma'. IDO with a small number of words and a small number of occurrences of terms from Toare discarded.
For the remaining IDO of ma' it is necessary to give the interpretation of words that are included in MT, but not included in S and those that are included in S, but not included in MT. We define sets of these words:
$Y1 = to \setminus s$ where $|to| > |s|$,
$Y2 = s \setminus to$ where $|s| > |to|$,
It is necessary to present one IDO as several constituent parts required for the introduction of structuring for the S. Let's present (IDO) as:

$$s = < \{ws_i\}, \{w_l, \{w_j\}\} > , \qquad (9)$$

where $\{ws_i\}$ is the set of words representing the main partial interpretation of MT;
w_l is one of the words in the set Y1 or Y2;
w_j is set of words representing the interpretation of the word w_l.
The representation of IDO in the form (9) allows to provide the expert with the most comprehensive information available on options of the interpretation of MT.

V. METHOD IMPLEMENTATION

In accordance with the proposed method of determining the interpretation, algorithms to find the definition of one-word (Fig. 2) and multiword (Fig. 3) terms were developed.

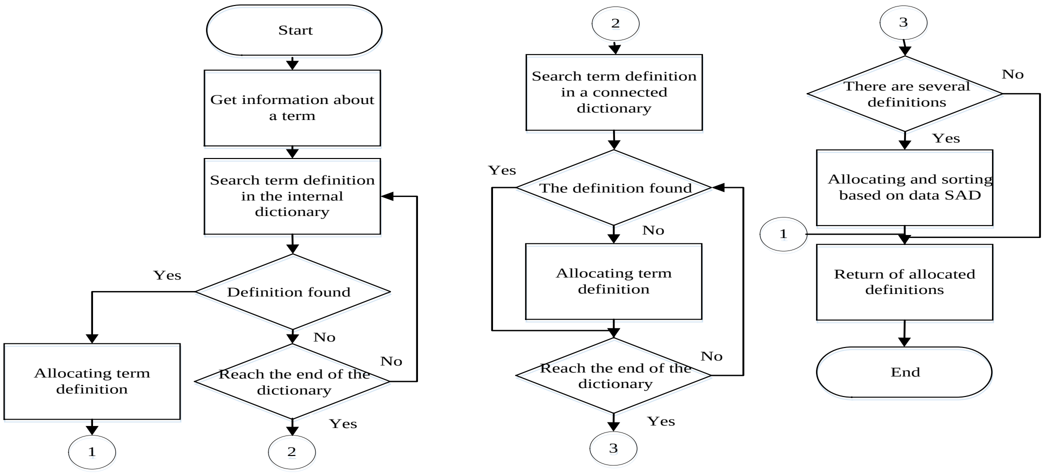

Start
Get information about a term
Search term definition in the internal dictionary
Definition found
Yes
No
Allocating term definition
Reach the end of the dictionary
No
Yes
1
2
2
Search term definition in a connected dictionary
The definition found
Yes
No
Allocating term definition
Reach the end of the dictionary
No
Yes
3
3
There are several definitions
No
Yes
Allocating and sorting based on data SAD
1
Return of allocated definitions
End

Figure 2. Search for the definition of a one-word term

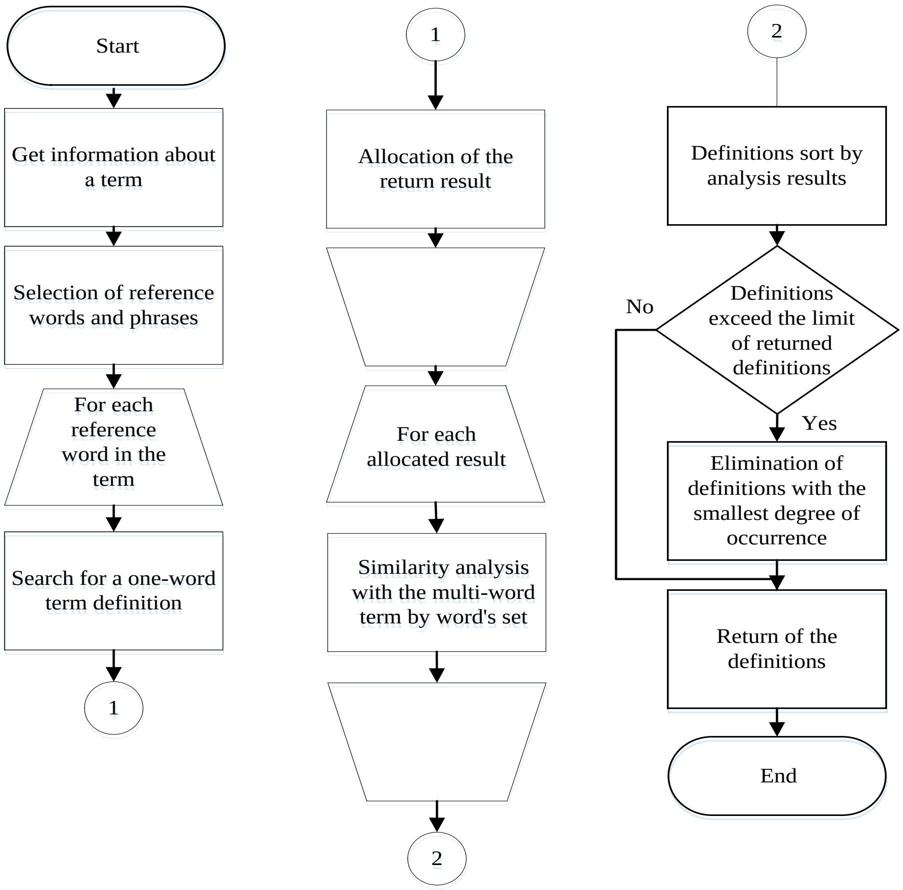
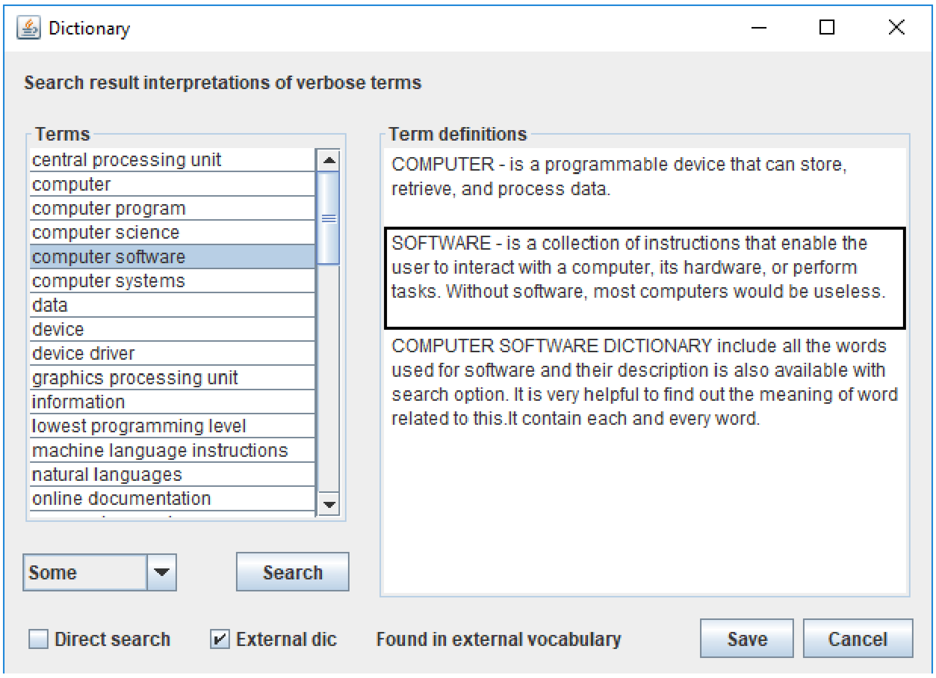

Start
Get information about a term
Selection of reference words and phrases
For each reference word in the term
Search for a one-word term definition
1
1
Allocation of the return result
For each allocated result
Similarity analysis with the multi-word term by word's set
2
2
Definitions sort by analysis results
Definitions exceed the limit of returned definitions
No
Yes
Elimination of definitions with the smallest degree of occurrence
Return of the definitions
End

Figure 3. Searching for interpretation of multiword term

At Fig.4 shown a window of the software product that represents the results of the search for the interpretation of the multiword term “computer software”. The word combination was not found, the result was presented in the form of interpretations for two words –“computer” and “software”.



Search result interpretations of verbose terms

Terms

central processing unit
computer
computer program
computer science
computer software
computer systems
data
device
device driver
graphics processing unit
information
lowest programming level
machine language instructions
natural languages
online documentation

Term definitions

COMPUTER - is a programmable device that can store, retrieve, and process data.

SOFTWARE - is a collection of instructions that enable the user to interact with a computer, its hardware, or perform tasks. Without software, most computers would be useless.

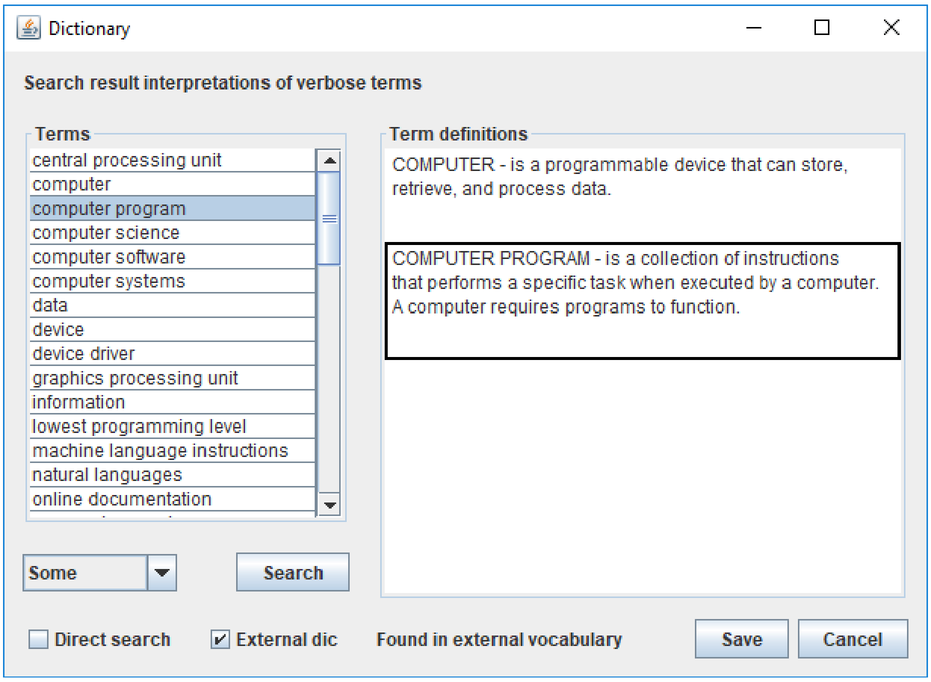
COMPUTER SOFTWARE DICTIONARY include all the words used for software and their description is also available with search option. It is very helpful to find out the meaning of word related to this.It contain each and every word.

Some

☐ Direct search ☑ External dic Found in external vocabulary

Figure 4. Research results for interpretations of the term “computer software”

At Fig. 5 the result of the search for the interpretation of the term “computer programs” is presented. Of the two definitions of “computer” and “computer program”, preference was given to the second one based on the calculation of entries in the definition of other terms from the DD.



Dictionary

Search result interpretations of verbose terms

Terms

- central processing unit
- computer
- computer program
- computer science
- computer software
- computer systems
- data
- device
- device driver
- graphics processing unit
- information
- lowest programming level
- machine language instructions
- natural languages
- online documentation

Term definitions

COMPUTER - is a programmable device that can store, retrieve, and process data.

COMPUTER PROGRAM - is a collection of instructions that performs a specific task when executed by a computer. A computer requires programs to function.

Some

☐ Direct search ☑ External dic Found in external vocabulary

Figure 5. Research results for interpretations of the term “computer program”

VI. EVALUATION OF THE EFFECTIVENESS OF THE METHOD AND SOFTWARE, DEFINITION OF TERMS OF INTERFERENCE

For the experiment with 63400 terms were randomly selected 5 lists of 30 terms from different subject areas. Tests were carried out in 3 modes: in manual mode, when the expert had to find and edit the definition of the term using online dictionaries; in the automated mode, when the search was performed by the program on the built-in and external dictionaries, and the expert edited the results; in the optimized mode, when the search was performed by the program, but the previously obtained terms from the corresponding subject area were taken into account. The results of the experiment are shown in Fig. 6

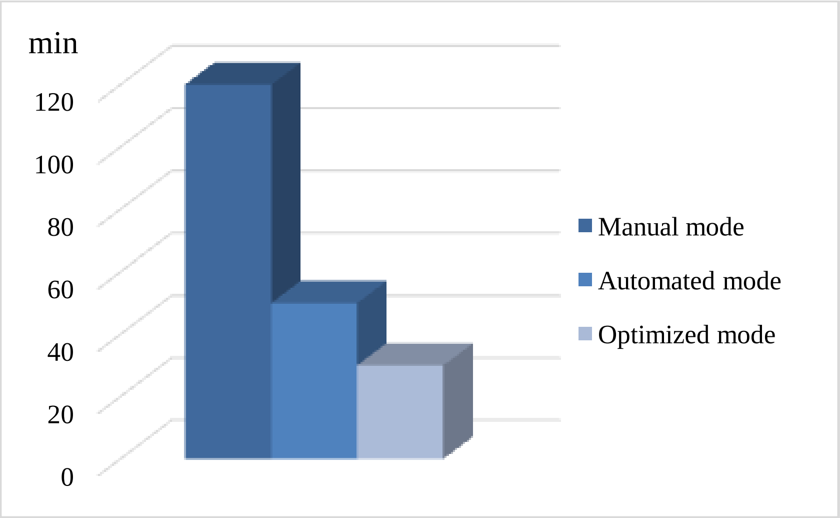


Figure 6. Time to determine the interpretation of terms in different modes

VII. CONCLUSION

The analysis of the existing methods of constructing DD for software projects and the need to reduce the time for the preparation of descriptions of terms are made. The method of definition of interpretations is proposed. This method allows to automate the process of determining the definitions of terms of subject areas. Algorithms and software that implement the proposed method is developed. The experiments have confirmed the effectiveness of the proposed solutions. The results of the study can be used in the development of software products "by order" at various stages of design: collection and analysis of requirements, database development, writing project documentation, etc.

REFERENCES

[1]. Kungurcev, A. B., Potochnyak, Ya. V., Silyaev, D. A., 2015, “Method of automated construction of explanatory dictionary of subject area”, Technology audit and production reserves, vol. 2, issue 2 (22), p. 58–63. doi: https://doi.org/10.15587/2312-8372.2015.40895

[2]. Califf, M. E., Mooney, R. J., 2003,“Bottom-up relational learning of pattern matching rules for information extraction”, Journal of Machine Learning Research, 4, p. 177–210.

[3]. Hartmann, R. R. K., 2003,“Lexicography: Dictionaries, compilers, critics, and users. Routledge”, 412 p.

[4]. Dal, V. I., 2015,“Explanatory Dictionary of Russian language. Part 1. (Russian)”, 636 p.

[5]. Oxford Dictionaries Online, 2018, Oxford University Press. Available at https://www.oxforddictionaries.com/

[6]. Hasan, K.,2014, “Automatic keyphrase extraction: a survey of the state of the art”, Proc. 52nd Annual Meeting of the Association for Computational Linguistics, № 1,p. 1262–1273. doi: https://doi.org/10.3115/v1/p14-1119

[7]. Chainikova, G.R., 2014,“Development of learner’s electronic dictionary of a thesaurus type as means of forming internal foreign-language lexicon”, p. 70-84.

[8]. Kungurtsev, O., Kovalchuk,S., Potochniak, Ia., Shirokostup, M., 2016, “Creating the domain vocabulary on the basis of automated analysis of ukrainian texts”, Technical sciences and technologies, №3 (5), p. 164-174.

[9]. Kungurtsev, O., Zinovatnaya, S., Potochniak, Ia., Kutasevych, M.,2018, “Development of information technology of term extraction from documents in natural language”, Eastern-European Journal of Enterprise Technologies,vol 6, no 2 (96),p. 44-51. doi: https://doi.org/10.15587/1729-4061.2018.147978

[10]. Nielsen, Sandro, 2008,“The Effect of Lexicographical Information Costs on Dictionary Making and Use”, Lexikos.

[11]. Navigli, R., Velardi, P.,2008, “From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions”, IOS Press, pp. 71-87.

[12]. Interpretatio. Available at http://www.softholm.com/download-software-free16427.htm