Розглянуто обґрунтування особливостей застосування технології Web Mining для визначення ключових слів. Web Mining дозволяє використати переваги контент-моніторінгу тексту на основі Стеммера Портера для визначення ключових слів. Запропоновано формальний підхід реалізації стемінгу українномовного тексту. Отримано експериментальні результати запропонованого методу для визначення ключових слів в слов'янськомовних наукових текстах технічного профілю

Ключові слова: Web Mining, NLP, контент, контент-моніторінг, ключові слова, контент-аналіз, Стеммер Портера, лінгвістичний аналіз

Рассмотрены обоснования особенностей применения технологии Web Mining для определения ключевых слов. Web Mining позволяет использовать преимущества контент-мониторинга текста на основе Стеммера Портера для определения ключевых слов. Предложен формальный подход реализации стемминг украиноязычного текста. Получены экспериментальные результаты предложенного метода для определения ключевых слов в славянскоязычных научных текстах технического профиля

Ключевые слова: Web Mining, NLP, контент, контент-мониторинг, ключевые слова, контент-анализ, Стеммер Портера, лингвистический анализ

UDC 004.89
DOI: 10.15587/1729-4061.2017.98750

DEVELOPMENT OF A METHOD FOR DETERMINING THE KEYWORDS IN THE SLAVIC LANGUAGE TEXTS BASED ON THE TECHNOLOGY OF WEB MINING

V. Lytvyn
Doctor of Technical Sciences, Professor*
E-mail: yevhen.v.burov@lpnu.ua
V. Vysotska
PhD, Associate Professor*
E-mail: victoria.a.vysotska@lpnu.ua
P. Pukach
Doctor of Technical Sciences, Associate Professor**
E-mail: petro.y.pukach@lpnu.ua
O. Brodyak
PhD, Associate Professor**
E-mail: brodyakoksana@mail.ru
D. Ugryn
PhD, Associate Professor
Department of Information Systems
Chernivtsi Department National Technical University
"Kharkiv Polytechnic Institute"
Holovna str., 203 A, Chernivtsi, Ukraine, 58000
E-mail: ugrind@mail.ru
*Department of Information Systems and Networks***
Department of Mathematics*
***Lviv Polytechnic National University
S. Bandery str., 12, Lviv, Ukraine, 79013

1. Introduction

Web Mining technology provides obtaining valuable knowledge from a text of information resources from the Internet sources. One of the up-to-date tasks of research is to determine effectiveness of applying the methods of machine learning and data mining for obtaining useful knowledge from the text content, as well as the structure, use, and purpose of the web resources. One of the relevant problems of today is the application of Web Mining methods for the optimization of Web-resources and promotion them in the Internet [1].

A promising task in the promotion of Web-resources in search engines is to provide not only high quality and unique (and not always unique) content for potential and/or permanent target audience. First of all, the content should be relevant and competitive. For this purpose, it is necessary to provide the relevance of keywords in the text content of Web-resources to the keywords, applied by the search engines users. High-quality and effective methods of determining the keywords of a text enable us to optimize the Web-resources and enhance the results of their promotion in the Internet. A set of keywords categorizes the text content. A set of the categorized text content defines the subject area of a Web-resource. However, it is only from the standpoint of the owner and/or the moderator of this Web-resource. It usually does not provide the relevance of the content and, as a result, promotion of this Web-resource. Web Mining allows us to take into account the opinion of potential users and the target audience in general for the formation of Web-resource content, its optimization and further promotion engaging the potential audience. To determine the keywords, the known Zipf law or the TF*IDF method for the identification of statistical patterns of words occurrence

in texts are commonly used. For this purpose, not the words themselves, but rather their stems (without inflexions, prefixes, and suffixes) are analyzed considering parts of speech and lemmatization. The only downside is automatic determining a word stem for the Slavic language texts due to the complexity of morphological analysis of this group of languages. It is not problematic to determine a word stem for the group of the Germanic languages. For this process, it is best to use the Porter stemmer for the group of the Germanic languages and the modified Porter stemmer for the group of the Slavic languages. The Porter stemmer is based on the morphological analysis of words. It is not sufficiently developed for the group of the Slavic languages. There are many methods of automated morphological analysis, but each of them is rather difficult for the implementation of optimal isolation of word stems for the Slavic texts.

2. Literature review and problem statement

Web Mining is a modern direction of obtaining relevant knowledge from Internet sources [2]. A set of measures of Web Mining methods essentially affects the optimization and promotion of a Web-resource (Table 1). Nevertheless, in the framework of present work, we will focus our attention on the problems of Web Content Mining, the main objectives of which are [3]:

– measurement of similarity between two or more text contents;

– formation of the vector input of text content, which facilitates the computation of similarity with the use of space-vector operations.

Table 1

Classification of Web Mining methods

Type	Formation	Application
Web Content Mining	Sets of relevant knowledge from Web-resource content	– Content clustering and categorization – Determining a subject-matter – Concept opening – Scanning – Content personalization – Content monitoring
Web Usage Mining	Models of interaction of a user with Web-resources	– Modeling of user's behavior – Optimization of Web-Resource – Management of interactions with users – Web marketing – Targeted advertising – Referral system
Web Structure Mining	Models of structure of Web-resource hyperlinks	– Document search and ranking – Opening of «hubs» and «influencers» – Opening of Web communities – Analysis of social networks

Classification of the application of Web Content Mining [3]:

– content monitoring: measurement of similarity between a request in the form of a vector and indexed vectors of text content for returning the ranged list of relevant content;

– clustering: a set of content based on similarity or difference (distance) between them;

– categorization: measurement of similarity of a new document which is categorized with the notions of existing categories (for example, middle vector that sets a group of content vectors);

– personalization: content recommendation based on its similarity to the input of a user's profile (using either the term vector that represents concepts, or terms based on user's interests).

For the successful Web-resource promotion, it is necessary to use a combination of Web Content Mining methods [4]:

– content-monitoring (for registering the categories of relevant content of both transitions from search engines and of monitoring the resource itself, as well as the registration of a set of keyword requests from users, registration of popular topics, etc);

– clustering (formation of a group of relevant content by the subject-matter, by author, by a target audience group, etc.);

– categorization (for further determining the popularity of subject-matter content);

– personalization (registration of results of users' activity regarding proposed recommendations for analysis of popularity of both the subject-matter and the text content itself, considering the results of analysis for further content formation of the resource).

Construction of systems of automated processing of the native language content based on the appropriate methods for Natural Language Processing (NLP) and the formalization of respective processes of linguistic analysis/synthesis is considered to be the main problem of the IT intellectualization. Rapid and impetuous development of the Internet dramatically accelerated creation of various information linguistic resources. It also boosted current research, aimed at the development and implementation of information linguistic systems, mathematical methods and NLP software and content analysis of text data arrays [5]. To automate the stages of analysis/synthesis of the native language texts, different models of NLP processes are created, effective algorithms and structures of the representation of natural language data arrays are substantiated [6]. A linguistic analysis of arrays of natural language texts is presented as a sequence of processes in the morphological, syntactic and semantic analysis/synthesis [7]. For each process, appropriate models, methods and algorithms were created:

– oriented at specific groups of languages (morph-lexical analysis);

– systems of Holiday grammars, grammars by N. Chomsky;

– trees of subordination and systems of components by Gladky, expanded navigation networks (sentence syntax);

– semantic networks and frame models by Minsky (text semantics).

The need for the automation of NLP processes contributed to the occurrence of relevant formal and mathematical linguistic models and methods of their analysis/synthesis [8]. The most complex problems of NLP are predetermined by the phenomena of polysemy, homonymy, homonymy that characterize the ambiguity of a language and complicate the process of identifying the correct representation of a semantic-syntactical structure of text into a formal representation through logical interpretation [9]. This is solved within a framework of semantic analysis [10]. However, the application of resource-consuming production rules of a logical-semantic analysis complicates and slows down the NLP programs. When understanding a text, they do not

often apply logic, but rather carry out the associative search for a semantic concept that is closest to the desired word and is contextually close to its own environment. Therefore, the associative search is a promising method for interpreting the natural language data arrays [11]. To implement a syntactic analysis of the text content with the aim of finding keywords and reducing the number of text processing steps, it is necessary [12]:

I. To isolate a verb group from the nominal group (only the words from the nominal group may be keywords) in the analyzed term chain (a sentence in Ukrainian). This is realized through the results of stemming. We analyze inflexions and work only with those words, the inflexions of which correspond to adjectives and nouns (in Ukrainian, adjectives and nouns do not belong to the verb group) [13].

II. In the nominal group, after finding the first set of keywords, the words, adjacent to the keywords, are analyzed. Here, by keywords we imply the words used in the text with certain frequency within the limits, set by a moderator. However, these words are only adjectives in the nominative case of masculine, nouns in the nominative case or abbreviation. We look for keyword combinations, that is, define terms $\mathrm{Noun} \in \mathrm{U}_{\mathrm{K1}}$ as word combinations of a noun or an adjective with a noun among the set of words of content, in particular [14]:

1. If a keyword is an adjective (inflexion of a word is *-ий* – nominative case masculine), then all the words, which are used to the right of this adjective in any case (search by the stem of this adjective), are found in the text. After that, the frequency dictionary is built for them. The word combinations that are used beyond a certain limit (sometimes used less often than an adjective) are new keywords. The limit is determined by the moderator.

2. If a keyword is a noun (inflexion of the word is not *-ий*), all the words to the right and to the left of it are analyzed.

2. 1. First, all the words on the left of it are checked for inflexions. Frequency dictionary is constructed. A set of words, which are found beyond the limit, specified by the moderator, is the new keywords.

2. 2. Then all the words on the right are analyzed – they all have to be without an inflexion. Similarly, we construct the frequency dictionary, by which a set of keywords is defined.

3. The aim and tasks of research

The aim of present work is to develop a method for determining the keywords in the Slavic language texts based on the Web Mining technology.

To achieve the aim, the following tasks were set:

– to develop a lexical analysis of the Slavic language texts and an algorithm of syntactic analyzer of the text content;

– to develop an algorithm for determining the keywords out of the text content based on the linguistic analysis of the text content;

– to develop software for the content-monitoring to determine the keywords in the Slavic language texts based on Web Mining;

– to obtain and analyze results of experimental verification of the proposed content-monitoring method for determining the keywords in the Slavic language scientific technical texts.

4. Peculiarities of lexical analysis of the Slavic language texts

We will describe the process of deriving a terminal chain in Ukrainian, which is characterized by the free word order in a sentence [6]. However, it does not deny the existence of a fixed order of arranging separate language elements [7]. For a simple complete sentence with direct word order, the structural scheme will be fixed, noun and verb groups will be the main syntactic categories of such sentence [14]. The unlimited grammar, built on the same principles as in the previous examples, has no application due to its complexity [15]. To form the context-dependent grammar, we will introduce certain restrictions, above all, on the structure of standard sentences. Based on the rules for constructing the sentences with a direct word order in the Ukrainian language, we consider a nominal group $\tilde{\mathrm{N}}$ of the structural schemes $\tilde{\mathrm{N}} = \{\mathrm{AN}\}$ or $\tilde{\mathrm{N}} = \mathrm{N}^{\mathrm{p}}$. Examples of sentences with direct word order is the case when an adjective is in preposition to a noun; the elements of the nominal group are grouped around a noun, etc. An adjective and a noun in the nominal group agree with each other in case, number and gender. These grammatical categories are also grammatical categories of the pronoun. We will consider the nominal group of structural scheme $\tilde{\mathrm{R}} = \mathrm{R}\tilde{\mathrm{N}}$ or $\tilde{\mathrm{R}} = \tilde{\mathrm{N}}\mathrm{R}$. Given grammatical characteristics of the verb in the English language, verbal and nominal groups agree in number, gender and person (Table 2).

Table 2

Grammatical categories of nominal and verb groups in the Ukrainian language

Type	Description
Nominative group/Ñ	adjective/A, noun/N, pronoun/$\mathrm{N}^{\mathrm{pro}}$
Verb group/ $\tilde{\mathrm{R}}$	verb/R, within Ñ adjective /A, noun /N
Number/NB	singular/sin, plural/pl
Gender/GR	masculine/m, feminine/f, neuter/n
Case/CS	nominative/n, genitive/g, dative/d, accusative/a, instrumental/i, prepositional/p, calling/c
Person/PR	1^{st} /1, 2^{nd} /2, 3^{rd} /3
Time/TM	present/pr, past/p, future/f

We will consider sentences with the nominal group in the third person and the verb group in present tense. Abridged designation of the nominal group is $\tilde{\mathrm{N}}_{\mathrm{GR,NB,CS,PR}}$, and of its components $\mathrm{A}_{\mathrm{GR,NB,CS}}$, $\mathrm{N}_{\mathrm{GR,NB,CS,PR}}$, $\mathrm{N}^{\mathrm{pro}}_{\mathrm{GR,NB,CS,PR}}$. If it is necessary to emphasize the use of different meanings of grammatical categories, we will use the following designations: two nominal groups with different meaning of a category, for example, gender, will be designated as: $\tilde{\mathrm{N}}_{\mathrm{GR',NB,CS,PR}}$, $\tilde{\mathrm{N}}_{\mathrm{GR'',NB,CS,PR}}$. Abridged designation of the verb group is $\tilde{\mathrm{R}}_{\mathrm{GR,NB,TM,PR}}$, of the verb – $\mathrm{R}_{\mathrm{GR,NB,TM,PR}}$. Implementation of the rules and regularities of the Ukrainian language affects the input of transformations. For example, it is known that most frequently the nominal group is expressed by a noun or a pronoun in the nominative case, and forms of the verb in the present tense coincide for all genera in singular (*він/вона/воно летить*, *Ukr.*). Consideration of such regularities is represented respectively in designations of the nominative and the verb groups – $\tilde{\mathrm{N}}_{\mathrm{GR,NB,n,PR}}$ and $\tilde{\mathrm{R}}_{\mathrm{NB,pr,PR}}$. The way of introducing the context-dependent grammar, which represents sentences of the introduced structural scheme

(taking into account certain regularities of the Ukrainian language) is illustrated by this sentence:

У своїй найбільш важливій роботі він показує барвистий світ українського села в його неповторній привабливості (Ukr.) In his most important work he shows a colorful world of Ukrainian village in its unique appeal.

Consider the grammar G_3=(V, T, S, P), where designations of syntactic categories will be for convenience represented without indices:

1. V=(S, Ñ, R̃, A, N, R, E, N^{pron}, #, у, свій, найбільш, важливий, робота, він, показувати, барвистий, світ, український, село, в, неповторний, привабливість) – alphabet;

2. T=(#, у, свій, найбільш, важливий, робота, він, показувати, барвистий, світ, український, село, в, неповторний, привабливість) – term meanings;

3. # is the symbol of the sentence limit;

4. S is the initial symbol.

Every step of representation is a convolution of one of the characters of the previous chain or replacement it with another one (other characters are rewritten without changes) [6]. The intermediate chain contains exactly one secondary character in the last place (the sentence is formed from left to right) [7].

5. Method of determining the keywords in text content

As functional-semantic-structural unity, a text has the rules of construction, reveals the patterns of content and formal connection of the constituent units. Coherence is manifested through external structural indicators and formal dependence of the text components, while integrity – through thematic, conceptual and modal dependence. Integrity leads to the notional and communicative text organization, while coherence leads to the form and structural organization. The operator of detection of keywords of commercial content $\alpha:(X,U,T)\to C$ is representation of commercial content C_2 to a new state, which is different from the previous state by existence of a set of keywords that generally describe its content. When analyzed, the multi-level content structure is explored: a linear sequence of characters; a linear sequence of morphological structures; a linear sequence of sentences; a network of interconnected unities (alg. 1).

Algorithm 1. Linguistic analysis of text content.

Stage 1. Grammatical analysis of text content X.

Step 1. Division of text content X into sentences and paragraphs.

Step 2. Division of the chain of characters of content X into words.

Step 3. Distinguishing figures, numbers, dates, set expressions and abbreviations in X.

Step 4. Removal of non-text characters of content X.

Step 5. Formation and analysis of a linear sequence of words with auxiliary designations for content X.

Stage 2. Morphological analysis of text content X.

Step 1. Obtaining stems (word forms with cut-off endings).

Step 2. For each word form, grammatical category is formed (collection of grammatical notions: gender, case, declension, etc.).

Step 3. Formation of a linear sequence of morphological structures.

Stage 3. Syntactic analysis $\alpha:(X,U,T)\to C$ of content X (alg. 2).

Stage 4. Semantic analysis of text content C.

Step 1. The words are matched against semantic classes from the dictionary.

Step 2. The selection of morpho-semantic alternatives for this sentence.

Step 3. Linking words in a uniform structure.

Step 4. Formation of an orderly set of records of superposition's of basic lexical functions and semantic classes. The accuracy of the result is determined by completeness/correctness of the dictionary.

Stage 5. Reference analysis for the formation of inter-phrase unities.

Step 1. Contextual analysis of text content C. With its help, the resolution of local references (this, which, his) and selection of statement, the unity nucleus, are implemented.

Step 2. Thematic analysis. Division of statements into topics distinguishes thematic structure, which is used in reviewing.

Step 3. We determine regular repeatability, synonymization and repeated nomination of keywords; reference identity (correspondence of words to the object of image; existence of implication on situational links).

Stage 6. Structural analysis of text content C. Preconditions for the usage is a high degree of coincidence of terms of unity, elementary discourse unit, sentences in semantic language and statements.

Step 1. Identification of the basic set of rhetorical links between unities.

Step 2. Construction of nonlinear network of unities. Openness of a set of links allows its expansion and adaptation for analysis of structure C.

A text implements a structurally displayed activity, which implies a subject and an object, a process, a goal, means and a result, which are represented in notional-structural, functional, and communicative indicators. Units of the internal organization of the text structure include the alphabet, vocabulary (paradigmatics), grammar (syntagmatics), paradigms, paradigmatic relations, syntagmatic relations, identification rules, expressions, inter-phrase unity and fragments-blocks. At the compositional level, we distinguish sentences, paragraphs, sections, chapters, sub-chapters, pages, etc., which, except for the sentence, are indirectly related to the internal structure, therefore, are not considered. Using a database (databases of terms/morphemes and auxiliary parts of speech) and the determined rules of text analysis, a term is searched.

Syntactic analyzers work at two stages: they identify notional lexemes and create an analysis tree (alg. 2).

Algorithm 2. Syntactic analyzer of text content.

Stage 1. Identification of notional lexemes $U_1 \in U$ for content X.

Step 1. Determining a term chain in the form of a sentence.

Step 2. Identification of nominal group with the help of the stem dictionary.

Step 3. Identification of the verb group with the help of the stem dictionary.

Stage 2. Creation of the analysis tree from left to right. The tree output involves the convolution of one of the characters of the previous sequence chain of linguistic variables, or its replacement with another one; the other characters are rewritten without changes. During convolution, replaced/rewritten characters (ancestors) are linked directly to characters that come out as a result of convolution, replacement

or rewriting (descendants), and a tree of components, or a syntax structure for the content meaning is obtained.

Step 1. Deployment of the nominal group. Deployment of the verb group.

Step 2. Implementation of syntactic categories with word forms.

Stage 3. Determining a set of keywords $\alpha:(X,U,T)\rightarrow C$ for X.

Step 1. Determining the terms $Noun\in U_1$ – nouns, combinations of a noun or an adjective with a noun among the sets of content words.

Step 2. Computation of unicity Unicity for terms

$$Noun\in U_1.$$

Step 3. Computation $NumbSymb\in U_3$ (number of signs without spaces) for $Noun\in U_1$ at $Unicity\geq 80$.

Step 4. Computation

$$UseFrequency\in U_2$$

of the frequency of occurrence of content keywords (*minimal word weight, % within [a,b])). For terms with

$$NumbSymb\leq 2000,$$

the frequency UseFrequency is within $(6;8]$ %, with

$$NumbSymb\geq 3000-[2;4)\ \%,$$

with

$$2000>NumbSymb<3000\ -\ [4;6]\ \%.$$

Step 5. Calculation of frequency of keywords occurrence BUseFrequency at the beginning, IUseFrequency in the middle and at the end EUseFrequency of the content.

Step 6. Comparison BUseFrequency, IUseFrequency and EUseFrequency for priority setting. Keywords with larger values BUseFrequency have greater priority than with EUseFrequency.

Step 7. Sorting the keywords according to their priorities.

Stage 4. Filling in the base of the search images C, (attributes $KeyWords\in U_4$).

Relying on the rules of generative grammar, term correction is carried out according to the rules of its use in context (Fig. 1).

Sentences set limits of action of punctuation marks, anaphoric and cataphoric references. Text semantics is determined by the communicative task of information transfer. The structure of the text is determined by the internal organization of text units and regularities of their relationships. During syntactic analysis, the text is arranged in the data structure – a tree that corresponds to the syntactic structure of the input sequence and is best suited for further processing. After analysis of the text fragment and the term, a new term is synthesized as a keyword of the content subject-matter, using a database of terms and their morphemes (Fig. 1). Further, the terms for forming a new keyword are synthesized, using the base of auxiliary parts of speech. The principle of identifying keywords by their meaning (terms) is based on the Zipf law. It is reduced to the choice of words with an average frequency of occurrence (the most used words are ignored due to "stop-dictionaries", and rare words of a text are not taken into account).

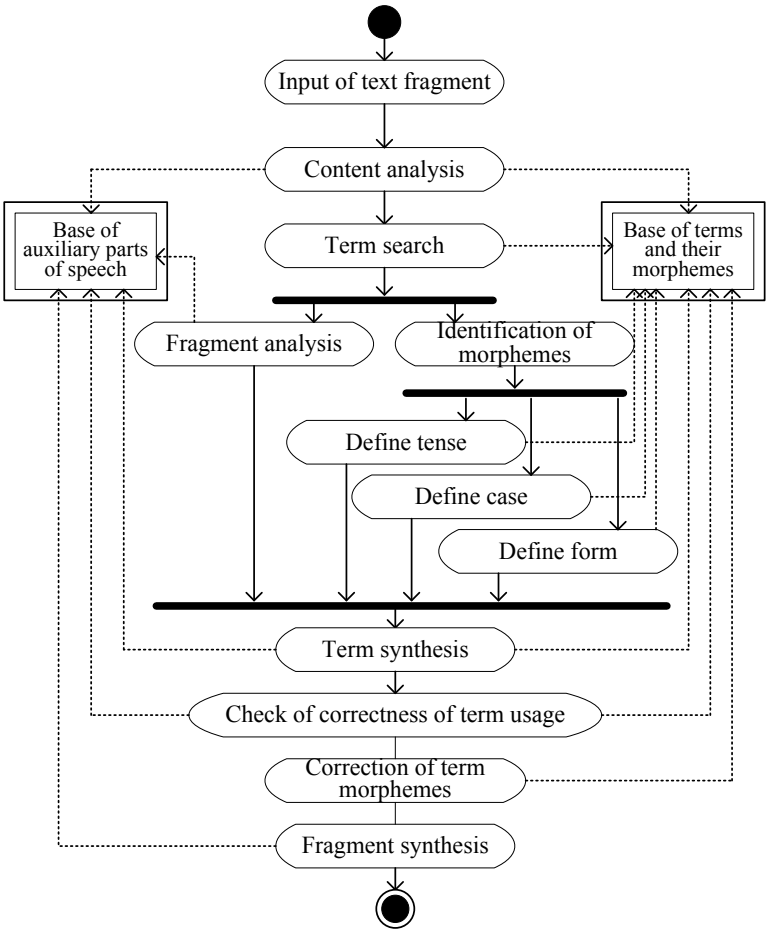

Fig. 1. Diagram of activity for the process of identifying the content keywords

6. Results of examining the keywords of text content in the Slavic language texts based on Web Mining technology

To achieve the goal of the research, the system with the possibility of selecting the language/languages of the analyzed content was developed. The process of finding a set of key words with regard to stems of thematic words was implemented on Web-resource Victana [16]. An analysis of statistics of functioning of the system of keywords set detection from 100 scientific articles of technical area was conducted at two stages:

1. To analyze all the articles with the check of general blocked words and thematic vocabulary (before learning of system).

2. To analyze all the articles with the check of specified blocked words and refined subject dictionary after system learning on the same content set. This need for research is explained by the fact that with a larger number of system launches, an additional set of unknown words (missing both in thematic dictionary and in a set of blocked words) is formed.

At each stage of the system operation, the check was performed at two steps for each article: analysis of the entire article and its abridged version. The latter version did not consider the title, authors, UDC, abstracts in two languages, author's keywords in two languages, work place of authors, and list of literature. This approach was used to determine the error of forming a set of keywords for different modifications of the proposed method. To evaluate the selection of keywords using TF*IDF model, the module that implements this algorithm was developed. The purpose of the experiment is the algorithm assessment. As input examples, a collection of 100 scientific articles from 2 issues (783 and 805) of Lviv Polytechnic National University Bulletin of the Information Systems and Networks series was used [17]. For each content, we built a vector model, 20 words that gained most weight were selected as keywords. Each document from the collection underwent expert assessment gaining from 0 to 10 points (0 – none of the words may be a keyword, 10 – all words are keyword for this content). The data on each collection were averaged. The experiment had two stages. At the first stage, each content of the collection was preliminarily processed:

– Lemmatization – bringing a word to normal form (conducted using the parser and Porter stemmer).

– Removal of stop words (conjunctions, connections, some adverbs, single letters and figures).

– The second stage included additional measures. The list of stop words was extended with some words that do not carry the notional load (for example: verbs *be*, *have, may*), and were not included in the initial list. Some part of words was separated in accordance with the Zipf laws. For each content, the vector for word occurrence statistics was built and low-score words were collected. Abbreviation parameters were chosen empirically and made approximately 5 %. High-score words were not removed, as the words that do not carry the meaning load, but are often found, in most cases were separated at the stage of stop words removal. The obtained results show that the method as a whole copes completely with separation of keywords. However, very often the highest position is occupied by the words that are not keywords for the content. Content transformation and noise reduction in it led to enhancement of quality of keywords selection.

An analysis of statistics was performed based on comparison of these magnitudes for each article, retrieved at different stages of research:

– set of the author's keywords (defined and written in the article by the authors of these works);

– set of keywords defined at the first and second stages with different weights of words UseFrequency$\in U_2$ (but more than the one defined in option *Min. word weight, % within [1;5]).

Research was held among scientific papers at an average arithmetical value of the author's key word combinations/words of about 5 (4.77), which on average are formed from 10 (9.82) words. The word weight UseFrequency is calculated as relative frequency of the word stem appearance throughout the text. Table 3 includes such designations, as:

– A (total number of keywords formed at the assigned word weight UseFrequency within [1, 5]);

– B (notional words from the list of the formed words, i. e. without unknown abbreviations, verbs, auxiliary words, etc.);

– C (corresponding of words to those, defined by the author of the article);

– D (accuracy of correspondence of the found keywords to the author's keywords);

– E (additional keywords, defined by the system, but not defined by the author).

Table 3

Statistical data on the examined text content of articles

Title	Word weight	Stage 1					Stage 2				
		A	B	C	D	E	A	B	C	D	E
Step 1	≥1	5.46	3.92	2.51	2.08	1.74	7.43	7.03	3.27	3	4.18
	≥2	1.08	0.88	0.63	0.59	0.26	2.67	2.64	1.65	1.54	1.12
	≥3	0.41	0.38	0.22	0.21	0.16	1.21	1.2	0.85	0.79	0.41
	≥4	0.15	0.13	0.09	0.09	0.04	0.46	0.45	0.33	0.31	0.15
	≥5	0	0	0	0	0	0	0	0	0	0
Step 2	≥1	6.51	5.02	2.68	2.23	2.37	8.35	7.78	3.25	2.91	4.99
	≥2	1.34	1.11	0.74	0.72	0.39	3.12	3.07	1.81	1.67	1.43
	≥3	0.51	0.45	0.29	0.27	0.17	1.42	1.4	0.93	0.85	0.54
	≥4	0.19	0.17	0.12	0.12	0.05	0.73	0.72	0.45	0.42	0.31
	≥5	0.11	0.1	0.06	0.06	0.04	0.33	0.32	0.25	0.23	0.1

Fig. 2, *a* shows the chart of analysis of statistics of the formation by the system of the sets of all potential keywords compared to the set defined by the authors of the articles. Correspondent value of each column in Fig. 2, *a* means arithmetic mean value of (Table 4):

– keywords, defined by the author (A_1);

– words that make up these author's keywords (A_2);

– potential keywords, defined systematically at stage 1, step 1 (A_3);

– potential keywords, defined systematically at stage 1, step 2 (A_4);

– potential keywords, defined systematically at stage 1, step 1 (A_5);

– potential keywords, defined systematically at stage 1, step 2 (A_6).

Table 4

Results of comparison of values [A_1, A_6]

Value	A_1 (4.77)		A_2(9.82)		A_3(5.46)		A_4(6.51)		A_5(7.43)		A_6(8.35)	
	–	/	–	/	–	/	–	/	–	/	–	/
A_1(4.77)	0	1	5.05	2.0587	0.69	1.144654	1.74	1.36478	2.66	1.557652	3.58	1.750524
A_2(9.82)	5.05	2.0587	0	1	4.36	1.7985	3.31	1.5084	2.39	1.3217	1.47	1.176
A_3(5.46)	0.69	1.144654	4.36	1.7985	0	1	1.05	1.1923	1.97	1.36	2.89	1.5293
A_4(6.51)	1.74	1.36478	3.31	1.5084	1.05	1.1923	0	1	0.92	1.1413	1.84	1.2826
A_5(7.43)	2.66	1.557652	2.39	1.3217	1.97	1.36	0.92	1.1413	0	1	0.92	1.1238
A_6(8.35)	3.58	1.750524	1.47	1.176	2.89	1.5293	1.84	1.2826	0.92	1.1238	0	1

An author of an article usually defines fewer keywords than are actually present in this work. Setting parameters of the system increases the number of defined keywords almost by 2 times. The total value increment, obtained by the system depending on moderation of dictionaries makes respectively for A_3 14.46541; A_4 – 36.47799; A_5 – 55.7652; A_6 – 75.05241.

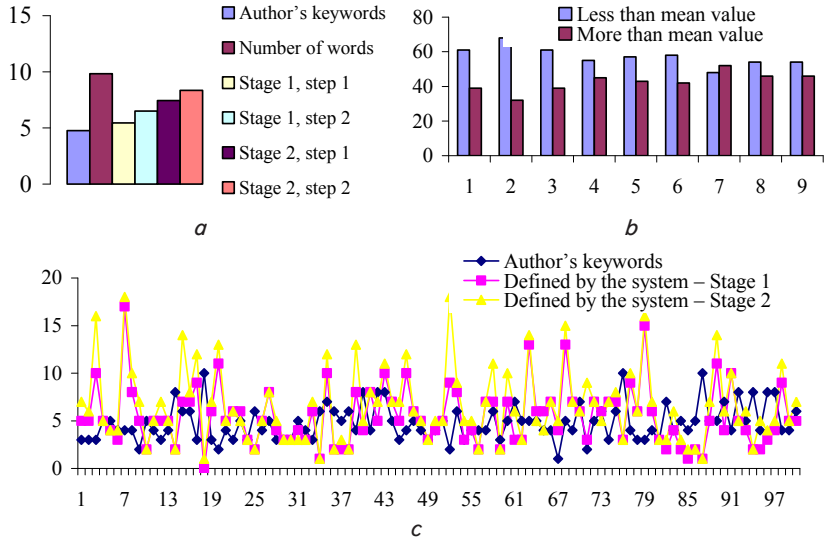

Fig. 2. Results of checking 100 articles: *a* — chart of analysis of statistics of forming the sets of all keywords by the system; *b* — chart of analysis of statistics of the text density distribution in analyzed articles; *c* — chart of distribution of forming the sets of keywords for each article by the system

Fig. 2, *b* shows the chart of analysis of statistics of the text density distribution in the analyzed articles, where respectively there is a result of number analysis (less/more than average value):

1 – pages of articles;
2 – passages in the article;
3 – lines with text;
4 – words;
5 – signs;
6 – signs and spaces;
7 – words on the page;
8 – characters on the page;
9 – characters and spaces on the page.

Fig. 2, *c* shows the chart of distribution of formation by the system of sets of all potential keywords for each article compared to the set, defined by the authors of the articles. Accuracy of defining keyword increases during dictionary moderation.

The difference between the number of keywords, defined by the author, and those, defined by the system at stage 1, step 1 is 44.39919 % (a percentage difference). Accuracy improves at stage 1, step 2 – 33.70672 %, significantly improves at step 2, step 1 – 24.33809 %, and at stage 2, step 2 is already 14.96945 %. Table 5 presents the results of analysis of statistics of formation by the system of sets of all potential keywords for each article compared to the set, defined by the authors, where respectively to the column:

– A – for the author's keywords;
– B – for keywords, defined by the system at stage 1 (step 1);
– C – for keywords, defined by the system at stage 1 (step 2);
– D – for keywords, defined by the system at stage 2 (step 1);
– E – for keywords, defined by the system at stage 2 (step 2).

Fig. 3 shows correspondent histograms for groups A–E according to statistical data of articles analysis in the formation of keywords sets.

An author of a scientific article usually arbitrarily chooses the number of keywords in the range from 2 to 8 words (most often – 3–5 keywords). The system defines a different number of words, depending on the particular author's style of writing (there are such articles, in which the system does not find by the Zipf law a single keyword). For group B, most often the system defined the number of 5, 7, and 3 (more than 10), although distribution of found keywords was in the range of 1 to 18 words (besides 17). For group C, the system most often defined the keywords number of 5, 7 and 3. Although distribution of the found keywords was in the range of [1, 18] words (besides 17), the number of found words increased and reached the highest reliability indicator. For group D, the system most often defined the keywords number of 7, 6, 5, 10 and 8, although distribution of found keywords was in the range from 2 to 14 words (the range considerably narrowed).

Table 5

Statistical data on the formation of keywords for the examined articles

Value	A	B	C	D	E
Mean	4.808081	5.515152	6.565657	7.505051	8.434343
Standard error	0.180859	0.310393	0.39035	0.301297	0.324611
Median	4	5	6	7	8
Mode	4	5	5	7	8
Standard deviation	1.799528	3.088371	3.883932	2.997869	3.229841
Dispersion of sample	3.238301	9.538033	15.08493	8.987219	10.43187
Excess	0.652815	1.705273	0.748643	-0.45645	-0.50438
Asymmetry	0.947939	1.125305	1.065716	0.537598	0.517047
Interval	8	16	17	12	13
Minimum	2	1	1	2	3
Maximum	10	17	18	14	16
Total	476	546	650	743	835
Score	99	99	99	99	99
Most (1)	10	17	18	14	16
Least (1)	2	1	1	2	3
Reliability level (95.0 %)	0.35891	0.615965	0.774637	0.597914	0.64418

For group E, the system most often defined the keywords number of 8, 5, 7 and 10, although distribution of found keywords was in the range from 3 to 16 (accuracy improved).

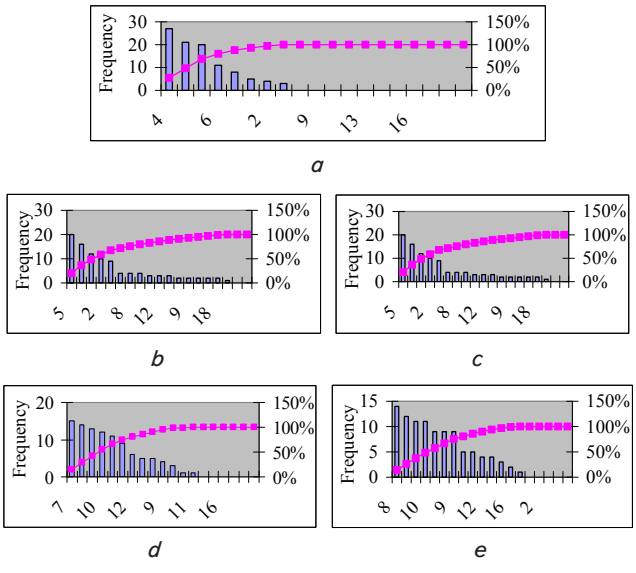

Fig. 3. Histogram for sample A–E for defined keywords: *a* – by authors; *b* – by system at stage 1 (step 1); *c* – by system at stage 1 (step 2); *d* – by system at stage 2 (step 1); *e* – by system at stage 2 (step 2)

7. Discussion of the results of exploring the Web Mining approach for determining the keywords in the Slavic language texts

Fig. 4 shows a comparative diagram of % of using the keywords, found by the system in the filtered text (abridged version) Per_f and the original author's text Per_0. The results were obtained without refining of the thematic dictionary by the moderator through addition of blocked words. The obtained mean values for 100 texts $\overline{Per_f} = 0{,}28$ and $\overline{Per_f} = 0{,}28$ show that such filtering of research articles improves the keywords density by 1.48 times or by 47.83 %.

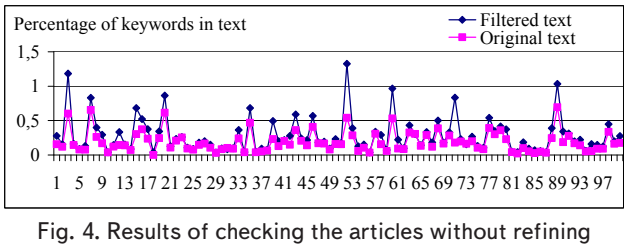

Fig. 4. Results of checking the articles without refining thematic dictionary

Fig. 5 shows a comparative diagram of % of using the keywords, found by the system in the filtered text (abridged version) Per_f^v and the original author's text Per_0^v. The results were obtained with regard to refining the thematic dictionary by the moderator through addition of blocked words. The obtained mean values for 100 texts show that filtering with simultaneous moderation of the thematic dictionary improves the keywords density by 1.35 times or by 35.44 %.

In Fig. 6 shows a comparison diagram of % of using the keywords, found by the system in the author's text without/with refining by the moderator of the dictionary through addition of the blocked words (Per_0 and Per_0^v, respectively). Comparison of values $\overline{Per_0} = 0{,}19$ and $\overline{Per_0} = 0{,}19$ demonstrates effectiveness of moderation of thematic vocabulary in an original text – keywords density increases by 1.34 times or by 34.33 %.

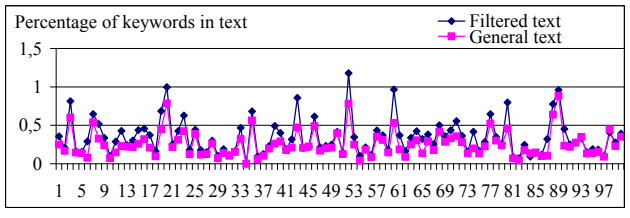

Fig. 5. Results of checking with regard to refining the dictionary

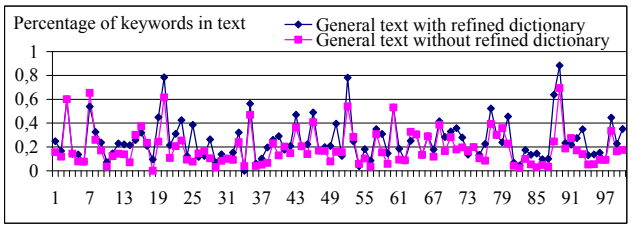

Fig. 6. Results of checking the authors' articles by different dictionaries

Fig. 7 shows the comparative diagram of % of using the keywords, found by the system in the filtered author's text without/with refining by the moderator of thematic dictionary through addition of blocked words (Per_f and Per_f^v, respectively). Comparison of values $\overline{Per_f} = 0{,}28$ and $Per_f^v = 0{,}34$ demonstrates effectiveness of the moderation of thematic topic dictionary in the filtered text – keywords density increases by 1.23 times or by 23.14 %.

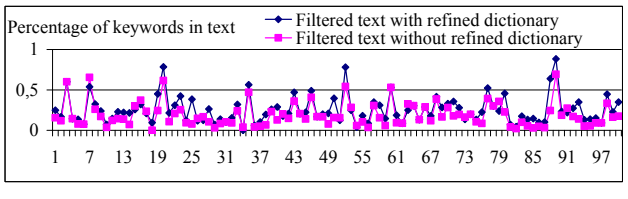

Fig. 7. Results of checking the filtered articles by different dictionaries

In the Internet space, information SEO-resources that define keywords within $[100 \div 1000]$ words in the text, are usually available, for example:

– http://msurf.ru/tools/keygeneratortext/;
– http://syn1.ru/tools/keygeneratortext/;
– http://webmasta.org/tools/keygeneratorurl/;
– http://labs.translated.net/terminology-extraction/;
– http://www.keywordstext.therealist.ru/.

The shortcomings of such SEO-resources include inaccuracy and incorrectness of processing texts in Ukrainian in the absence of correctly constructed morphological dictionaries, stem dictionaries and blocked words dictionaries. Another drawback of the majority of such SEO-resources is limited processing of volumes of text data arrays. For example, an article in the Ukrainian language, which has more than 800 words, was syntactically analyzed by a number of SEO-resources. A certain part of the above mentioned SEO-resources process incorrectly a large volume of information or does not process it at all.

One of the best SEO-resources is http://advego.ru/text/seo/, which works best with Ukrainian texts. It provides semantic analysis of a text online and the SEO-analysis of a text. The result is the closest to the result, obtained by the developed system. Nevertheless, there are some shortcomings. This SEO-resource does not define a set of keywords, but only frequency of using words, word combinations or parts of words (which are not necessarily parts of words as stems). It generally does not work with stems. For this SEO-resource, the words *ключових* and *ключові* are different. The developed SEO-resource http://victana.lviv.ua/kliuchovi-slova works with word stems, focused on texts in Ukrainian, Russian, and English, as well as of the mixed type. Taking this article as an example, SEO-resource defined the following set of keywords {word, key, content, analysis, chomsky, system}. Repeatability of words (times): word – 120; key – 49; content – 46; analysis – 39; chomsky – 37; system – 37. The authors defined the following keywords: text, Ukrainian, algorithm, content monitoring, keywords, linguistic analysis, syntactic analysis, generative grammars, structural scheme of the sentence, information linguistic system. Coincidence of lists of defined keywords with the lists of authors' keywords without considering extra words (repeatability>30 for the text volume of over 4800 words), makes according to such SEO resources:

– http://syn1.ru/tools/keygeneratortext/ – approximately 35 %;

– http://labs.translated.net/terminology-extraction/ – approximately 57 %;

– http://advego.ru/text/seo/ – approximately 83 %;

– http://victana.lviv.ua/kliuchovi-slova – approximately 90 %.

Authors typically define more keywords compared to the real situation in accordance with regularities of words frequency distribution by the Zipf law. The author of the content typically defines a set of keywords with capacity in the range of [2; 10] (most often [3; 5]). The system [16] defines a different number of keywords in the analyzed content depending on the style of its author (typically within [0; 7]).

8. Conclusions

1. We developed the algorithm of lexical analysis of the Slavic language texts based on decomposition of the content-monitoring method into interrelated components of content-analysis of text information and determining a set of keywords. Its features are adaptation of morphological and syntactic analysis of lexical units to peculiarities of the structures of the Ukrainian words/texts.

2. The approach to development of the Web Mining information system for determining the keywords and of the text content was proposed. Theoretical and experimental substantiation of the content-monitoring method for a text in Ukrainian was presented. The method is aimed at the automated detection of notional keywords of a Ukrainian text due to the proposed formal approach to the implementation of stemming of the Ukrainian language content.

3. The algorithm of the syntactic analyzer the text content was developed. Its features include deployment of the nominal/verb groups and construction of appropriate tree for analysis of each sentence taking into account specific features of their structures as elements of the Slavic language texts.

4. The algorithm support of the main structural components of the proposed method, based on the Porter's Stemmer algorithm, adapted to the Ukrainian language, was developed. The ways of enhancing efficiency of keywords search, in particular keywords density in the text, were found. They are based not on an analysis of words themselves (nouns, sets of nouns, adjectives with nouns, other parts of speech are ignored), but rather of word stems in the Slavic language texts. In the rules of stems separation in the texts, not only inflexion separations are considered, but also suffix separations as well as letters alteration at noun and adjective declension.

5. 100 scientific publications of two issues (783 and 805) of the Lviv Polytechnic National University Bulletin of the Information Systems and Networks series were examined [17]. The obtained results demonstrated a positive impact of filtering the text of an article and moderation of the thematic dictionary for keywords definition. It was found that for technical scientific texts of the experimental base, best results were attained by the method of analysis of an article without its beginning (title, authors, UDC, abstracts in two languages, author's keywords in two languages, work place of authors) and without the list of literature with the check of specified blocked words and refined thematic dictionary – for it the average keywords density in the text achieves $\mathrm{Per}_{\mathrm{f}}^{\mathrm{v}} = 0{,}34$, which is by 81 % higher than the correspondent value of density of the original text. With numerical data of statistical analysis, it was proved that setting the parameters of the system increases the number of defined keywords almost by 2 times, without decreasing indicators of accuracy and reliability. Testing of the proposed method for determining the keywords from other categories of texts, such as scientific humanitarian, fiction, and journalistic require further experimental research.

References

1. Mobasher B. Data mining for web personalization [Text] / B. Mobasher // The adaptive web. – Springer, Berlin, Heidelberg, 2007. – P. 90–135.
2. Dinuca, C. E. Web Content Mining [Text] / C. E. Dinuca, D. Ciobanu, D. Ciobanu // Annals of the University of Petrosani, Economics. – 2012. – Vol. 12, Issue 1. – P. 85–92.
3. Xu, G. Web content mining [Text] / G. Xu, Y. Zhang, L. Li // Web Mining and Social Networking. – 2010. – P. 71–87. doi: 10.1007/978-1-4419-7735-9_4
4. Bolshakova, Y. Avtomaticheskaya obrabotka tekstov na yestestvennom yazyke i komp'yuternaya lingvistika [Text] / Y. Bolshakova, E. Klyshinskiy, D. Lande, A. Noskov, O. Peskova, Y. Yagunova. – Moscow: MIEM, 2011. – 272 p.

5. Lytvyn, V. The method of formation of the status of personality understanding based on the content analysis [Text] / V. Lytvyn, P. Pukach, I. Bobyk, V. Vysotska // Eastern-European Journal of Enterprise Technologies. – 2016. – Vol. 5, Issue 2 (83). – P. 4–12. doi: 10.15587/1729-4061.2016.77174
6. Khomytska, I. The Method of Statistical Analysis of the Scientific, Colloquial, Belles-Lettres and Newspaper Styles on the Phonological Level [Text] / I. Khomytska, V. Teslyuk // Advances in Intelligent Systems and Computing. – 2016. – P. 149–163. doi: 10.1007/978-3-319-45991-2_10
7. Khomytska, I. Specifics of phonostatistical structure of the scientific style in English style system [Text] / I. Khomytska, V. Teslyuk // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589887
8. Vysotska, V. Information technology of processing information resources in electronic content commerce systems [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589909
9. Vysotska, V. The commercial content digest formation and distributional process [Text] / V. Vysotska, L. Chyrun, L. Chyrun // 2016 XIth International Scientific and Technical Conference Computer Sciences and Information Technologies (CSIT). – 2016. doi: 10.1109/stc-csit.2016.7589902
10. Lytvyn, V. Classification Methods of Text Documents Using Ontology Based Approach [Text] / V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H. Rishnyak // Advances in Intelligent Systems and Computing. – 2016. – P. 229–240. doi: 10.1007/978-3-319-45991-2_15
11. Jivani, G. A. A Comparative Study of Stemming Algorithms [Text] / G. A. Jivani // Int. J. Comp. Tech. Appl. – 2011. – Vol. 2, Issue 6. – P. 1930–1938.
12. Mishler, A. Using Structural Topic Modeling to Detect Events and Cluster Twitter Users in the Ukrainian Crisis [Text] / A. Mishler, E. S. Crabb, S. Paletz, B. Hefright, E. Golonka // Communications in Computer and Information Science. – 2015. – P. 639–644. doi: 10.1007/978-3-319-21380-4_108
13. Vysotska, V. Linguistic analysis of textual commercial content for information resources processing [Text] / V. Vysotska // 2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET). – 2016. doi: 10.1109/tcset.2016.7452160
14. Kowalska, K. Sentiment Analysis of Polish Texts [Text] / K. Kowalska, D. Cai, S. Wade // International Journal of Computer and Communication Engineering. – 2012. – P. 39–42. doi: 10.7763/ijcce.2012.v1.12
15. Kotsyba, N. The current state of work on the Polish-Ukrainian Parallel Corpus (PolUKR) [Text] / N. Kotsyba // Organization and Development of Digital Lexical Resources. – 2009. – P. 55–60.
16. Victana [Electronic resource]. – Available at: http://victana.lviv.ua/index.php/kliuchovi-slova