УДК 681.3

METHOD OF COMPLEX INFORMATION AND PSYCHOLOGICAL DOCUMENT ANALYSIS

S. A. Sidchenko, Ph.D; ***T. V. Saprykina***

I. Kozhedub Kharkiv Air Force University

sidserg@list.ru

Method of complex information-psychological analysis of the document. The technique is based on statistical and semantic approaches to text analysis in combination with the methods of phonetic analysis. On the one hand allows to select the text document completed sections of text that meet certain topics , and a summary of the document. On the other hand suggestively to determine the orientation of the text on the subconscious mind of man. The methodology more attention is paid to the phonetic analysis of the text and its connection with the linguistic (semantic) analysis of the structure of the document. In terms of features method meets the current requirements of corporate electronic document.

Keywords: document analysis; information-psychological analysis; content-analysis.

Запропоновано методику комплексного інформаційно-психологічного аналізу документу. Методика заснована на статистичних та семантичних підходах до аналізу текстів у поєднанні з методами фонетичного аналізу. Це, з одного боку, дозволяє виділити в тексті документу закінчені відрізки тексту, що відповідають визначеним тематикам, та створити реферат документу. А, з іншого боку, дає змогу визначити сугестивну спрямованість тексту на підсвідомість людини. У методиці більшу увагу приділено фонетичному аналізу текстів і зв'язку його з лінгвістичним (смисловим) аналізом структури документу. За своїми можливостями методика відповідає сучасним вимогам корпоративного електронного документообігу.

Ключові слова: аналіз документу; інформаційно-психологічний аналіз; контент-аналіз.

Formulation of the problem

Search methods, analysis, drafting and formation of texts (information) are presented very widely. However, the system that implements these methods are not always suitable for conducting information-psychological confrontation.

Lot of implemented systems used to analyze the formation and texts from specified parameters impact which oriented on Russian texts are narrow. Another drawback of most of these systems is the lack of descriptions of their mathematical basis and the governments of most states ban on the export of the products in full.

That is why it is necessary to create their own information and analysis systems. Any system is good only when they have good mathematics and software and hardware from which it is composed, and the staff that it serves.

Analysis of recent researchs and publications

There are several approaches to the analysis of text in a document or document library [1–4]. They discussed approaches to the implementation of an automated process of quasireferencing of electronic documents and forming a plurality of keywords in information-search systems based on using analysis of the text semantic structure and its logical segmentation.

Software catalog and cesource Analysis and linhvistic word processing on the Internet is presented in [5]. The most common system VAAL [6] and TextAnalyst.

The system VAAL (http://www.vaal.ru) allows predicting the effect of unconscious influence of texts to a mass audience, analyzing texts from the point of view of the impact, to make the text with a given vector of influence and identify individual psychological qualities lyricist.

The TextAnalyst system (http://www.analyst.ru) allows us to build a semantic network of concepts selected in the treated text with links to context. Present function of semantic search text fragments based on latent semantic relationships in the text of your search terms. It allows us to analyze the text by constructing a hierarchical tree of topics / subtopics, which affect the text. Alsothere is a of abstracting the text.

Problem definition

Suggest an integrated document analysis method, which allows selecting in text documents complete sections of text which correspond determined topics and select key components allocation suggestive of directional text.

Statement of the main material

The document as a whole is seen as a sequence of words that can be grouped into sentences, paragraphs and sections. Sometimes it is possible to identify the main parts: the header (title), annotation, analysis of literature's office, the purpose, the majority (consisting of sections), references and others.

There are several approaches to the analysis of text documents.

The using only a statistical approach to analyze the contents of the document allows you to create information about the structure of the text only by analyzing the frequency of words occurrence in text. The "central" word domain, which is found in the text at least a specified number of times, will be referred to the set of keywords.

In contrast, the methods of semantic orientations convertible at identifying the content (subject matter) of the text, its thematic focus and the determination of relationships between the individual elements of the text and the text as a whole. In this case, if the structure of two sentences (paragraphs) have the same keywords or words with the same meaning then these sentences (paragraphs) will be considered semantically coherent.

Linguistic approaches (based on the syntactic and morphological methods) allow to bring the text document to word forms dictionary.

Phonetic analysis methods allow you defined orientation suggestive lyrics.

Content analysis reveals the frequency of text occurrence specified characteristics which interest the researcher, and allow him to make some conclusions about the intentions of the creator of the text or the possible reactions of the recipient.

Using a set of these methods allow to select in text document completed sections of text that meet certain topics and to provide them with the release of the key components of orientation of suggestive text as a whole and its main part.

Algorithm for complex document analysis is shown in Figure. Consider the its common stages.

First, selection of document analysis and open it. The document is converted into a convenient form. At first the document choice for the analysis and opening. The document is converted into a convenient form.

The statistical analysis of the document — a quantitative count of words, letters, sentences, lines, paragraphs, pages, sections, figures, tables, list of references, and so on.

After a preliminary analysis of documents are conducted downloading dictionaries.

For the analysis of texts is desirable (and in some cases also necessary) using a set of specially trained dictionaries. To ensure the best quality of analysis should provide an opportunity for self-adjustment of the subject area. Setting dictionaries allow filtered layered not the interesting information in the text and opposite, to provide an important part of it.

Each dictionary should contain three main sections:

– deleted words — contains words that are simply removed from the text in the analysis. In most words that are removed include prepositions, adverbs, numerals, and some adjectives that generally do not carry useful information. In addition, their removal does not affect the grammatical structure of sentences;

– all-used words — contains words that are not removed in the analysis as separate concepts. Basically these are the words of a total value not represent self-interest for the analysis of the material text, the low-informative;

– special words — words containing domain. In this section, it is desirable to identify words that are of greatest interest to the user.

During using vocabulary from the analysis of text documents filters deleted and all-used words, and all other special finds words researched text, which are formed on the basis of key concepts.

As key concepts selects:

– domain specific words and their ratios ended, including all-used words encountered in the text at least a given number of times;

– word-benefits that met in the text at least once and their completed connections, including special and all-used words that met in the text at least a given number of times.

Limit of the occurrences frequency that used for the selection of concepts is determined by the amount of the processed text.

To bring the various forms to the total (normalization) can be used in two ways: automatic allocation root base and full transfer of word forms. In the first, simpler case, rather to present in the dictionary one (arbitrary) form of words. In an analysis of all the other forms that have the same root are considered equivalent. In the second, more difficult, but sometimes necessary case if all word forms are given in the form of a list, which shall be deemed equivalent to the analysis of text.

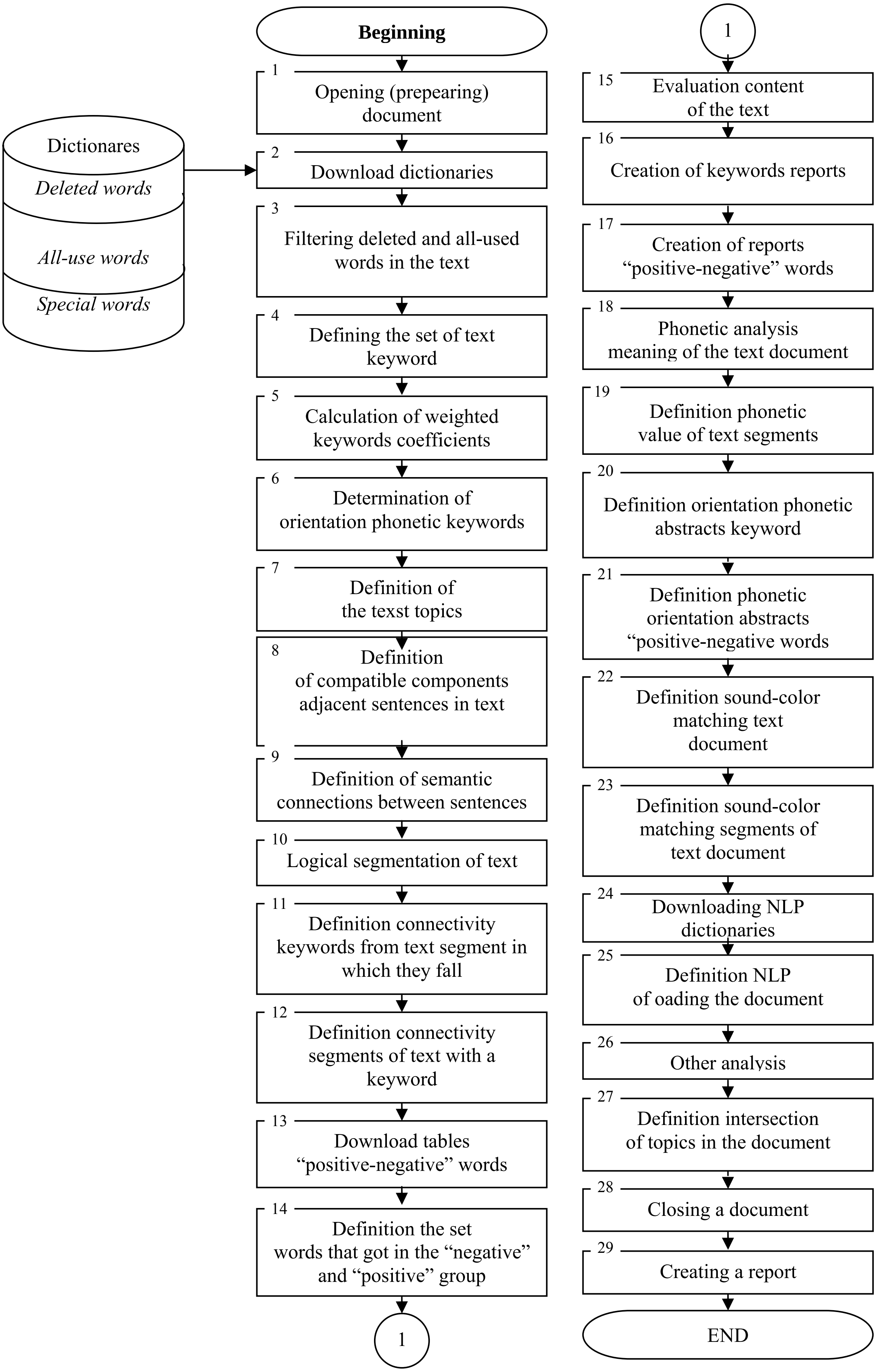
Then calculates the weighted coefficients of keywords. Weighting factor ϖ_t of keyword $t \in S$ is calculated as the ratio of sentences of text documents in which the word (concept) occurs, to the total number of text sentences:

$$\varpi_t = \frac{N_t}{c}, \tag{1}$$

where N_t — number of sentences of the text document, which meets keyword (term) t; c — total number of text sentences.

Determination of orientation phonetic keyword conducted to determine the load keywords suggestive on mans perceptions.

Defining the subject of the text. Subject text of the document can be determined by comparing the keywords $t \in S$ of text document (including their weights ϖ_t) with the keywords that can be assigned to one of the subjects.


Beginning
Dictionares
Deleted words
All-use words
Special words
1 Opening (prepearing) document
2 Download dictionaries
3 Filtering deleted and all-used words in the text
4 Defining the set of text keyword
5 Calculation of weighted keywords coefficients
6 Determination of orientation phonetic keywords
7 Definition of the texst topics
8 Definition of compatible components adjacent sentences in text
9 Definition of semantic connections between sentences
10 Logical segmentation of text
11 Definition connectivity keywords from text segment in which they fall
12 Definition connectivity segments of text with a keyword
13 Download tables “positive-negative” words
14 Definition the set words that got in the “negative” and “positive” group
1
1
15 Evaluation content of the text
16 Creation of keywords reports
17 Creation of reports “positive-negative” words
18 Phonetic analysis meaning of the text document
19 Definition phonetic value of text segments
20 Definition orientation phonetic abstracts keyword
21 Definition phonetic orientation abstracts “positive-negative words
22 Definition sound-color matching text document
23 Definition sound-color matching segments of text document
24 Downloading NLP dictionaries
25 Definition NLP of oading the document
26 Other analysis
27 Definition intersection of topics in the document
28 Closing a document
29 Creating a report
END

Algorithm for complex document analysis

After selecting a topic it may specify by determining the semantic relation keywords of this subject, keywords of text document and name of the document (if available).

Determination of compatible components adjacent sentences in text. Common elements of two adjacent sets of sentences are defined as follows:

$$S_{ij} = P_i \cap P_j,\ j = i+1,\ i = \overline{1,c}, \qquad (2)$$

where P_i, P_j — set of special words of sentences i, j-accordingly; c — total number of text sentences; S_{ij} — set consisting of the same elements P_i and P_j of the set.

Definition of semantic connections between sentences. Degree of semantic relation sentence with the following sentence j equals the number of elements of the set S_{ij}.

Logical segmentation of text. When the text refers to logically segment without crossing segments of text (sentences, paragraphs or plural), each of which it goes about any characteristic of that information.

To determine the degree of "saturation" of logical segments of text information, which is typical for them, we introduce a linear coefficient U_i — parameter that characterizes the average degree of semantic coherence of i-sentence with n sentences that lie ahead:

$$U_i = \frac{1}{n}\sum_{w=1}^{n}\left|S_{i-w,j-w}\right|,\ j = i+1,\ i = \overline{1,c}, \qquad (3)$$

where n — number of sentences of the text segment preceding-th sentence; $\left|S_{i-w,j-w}\right|$ — degree of semantic connection $i-w$ sentence with $j-w$ sentence.

The beginning of a new segment $f_k \in F$ in the text, and subsequently the boundary between segments, we assume a reduction of linear factor of the current sentence, compared with a linear coefficient U_i of the preceding sentence i, provided that between the current sentence i and the following sentence j is no semantic relationship:

$$f_k : U_i < U_{i-1},\ \left|S_{ij}\right| = 0. \qquad (4)$$

Otherwise, the segment continues to sentences in which semantic link with the previous sentence is missing.

Definition connectivity keywords with segments of text in which they fall is determined calculating the semantic weight L_t of word $t \in S_k$ that describes the number of segments sentences semantically related word t:

$$L_t = \sum_{i=1}^{t}\left|P_i \cap t\right| - 1,\ t \in S_k, \qquad (5)$$

where t — element of the set S_k of keywords text document.

Definition connectivity segment text from a single keyword. Degree of semantic communication segments of text with a keyword-level equals number of elements of the set S_k of common components (words) sets specific segments $F_k \in F$ of words with one keyword, which is defined by the following way:

$$S_k = F_i \cap \ldots \cap F_j,\ i \neq j,\ k = \overline{1,l}, \qquad (6)$$

where F_i, F_j — set of special words segments i, j respectively, which contain the same keyword; l — number of proposals for all segments of a keyword; S_k — set consisting of the same elements F_i and F_j of the set.

Downloading table with "positive-negative" words. Tables "positive-negative" words are using to identify areas of the text with the same set of words belonging to the same group of terms. Name the tables are provisional and subject to change.

Definition set of words that got in the "negative" and "positive" group. Determine the set Y consisting of words, what are caught in a "positive" group and set Z — with words, what are caught in the "negative" group.

Evaluation content of the text is defined as the ratio of sentences of text documents in which there are the words from "negative" ("positive") sets the total number of sentences of text

$$v = \frac{H}{c}, \qquad (7)$$

where H — number of sentences of text documents in which the word is found with a "positive" Y ("negative" Z) sets; c — total number of text sentences.

Creating abstracts of keywords. Performed for each keyword (concept) by selecting all the text of the document sentences in which it occurs.

In some cases, the abstract may be included and whole paragraphs (segments) of the text, if this keyword is connected.

Creation of reports with "positive-negative" words. Is based on the selection of sentences (paragraphs segments) of text documents, which contain the same key components of the sets of "positive" ("negative") words corresponding tables (dictionaries).

Analysis of phonetic meaning of the text document. Phonetic analysis of texts based on the analysis of alphabetic writing based on soft consonants. The text as it is presented in the voice-letter form.

Definition phonetic value of the text segments is carried out to tracing the dynamics of change of loading suggestive on the human perception and definition of phonetic stress on different parts of the text. It is held at logic with obtained by logical segmenting of text segments or by splitting the text into document-level segments (usually at least 10).

Definition phonetic orientated abstracts keyword conducted to tracing suggestive loading on the human perception text segments with similar keywords.

Definition phonetic orientated reports with "positive-negative" words is held to tracing suggestive loading on the human perception of text segments into pieces that are constructed with the use of the words "positive" and "negative" groups.

Definition sound-color matching text. A peculiar aspect of the symbolism of speech sounds is the sound color matching. They have purely sinestic basis and apply only to certain sounds.

The most clear sound color matching can be traced mainly to Russian vowels: э, о, ы, у, и, а.

We see that most clearly in solid colors painted three sounds: а — bright red, и — blue, о — light yellow. These three colors in the spectrum are essential in the sense that by mixing them in different proportions you can get all the other colors and shades.

Definition sound color matching text document is conducted by counting the vowel letters. The text of the document takes coloring dominant vowel letters. Dominance is determined by the excess number of vowel letters above threshold. Typically, the threshold is defined as 80–90 % of maximum.

Definition sound color matching segments text document is held to tracing the dynamics of change in coloring of the text and identify sound-color stress on different parts of the text. It is held at logic obtained by segmenting of text segments or by splitting the text into a document-level segments (usually at least 10).

Download dictionaries with neuron-linguistic programming (NLP). To determine the orientation of NLP (load) of text designed (or developed) dictionary words, which are typical for people with a variety developed perception. As such dictionaries are the words to the dictionaries audials, visual, kinystetiks.

Definition NLP load on the text of the document is made by counting the number of words that fall into different groups. View NLP loading is matched group which hit the maximum number of words.

In addition for visual determination of loading carried counting the number of graphics (pictures). Their quality is measured visually or by using special techniques. Other analysis. For example, conducting motivational analysis "Aspiration K" — "Getting away from".

Definition of topics intersection in the text of the document is made possible by identifying additional subjects excluding keywords that relate only to the main topic of the document that has been defined before. After completing the analysis of the document text is made to bring it in its original form or closing. The analysis results are entered into a database and can be used for further work and / or corrections medium (normal) estimates attribute scales and dictionaries for all types of analysis.

The report is created for all types of analysis in graphical and text form.

Conclusions

The methods of complex analysis document, which allows us to select text in a document completed sections of text that meet certain topics and to provide them with the release of the key components of orientation of suggestive text is offered. In terms of features meet modern requirements of corporate electronic document and can be widely used in various fields.

REFERENCES

1. *Gerasimov B. M.* Extracting information from primary phrases electronic documents in an information retrieval system / B. M. Gerasimov, O. Sergeyev, I. Y. Subach // USiM. — 2006. — Number 1. — Pp. 26–29.

2. *Rybakov F. I.* Automatic indexing of natural language / F. I. Rybakov, E. A. Rudnev, V. A. Petukhov. — Moscow : Energiya. — 1980. — 160 p.

3. *Skorokhodko E. F.* Linguistic Foundations of automated information retrieval / E. F. Skorokhodko. — Higher School. — 1970. — 242 p.

4. *Salton G. A.* Automatic processing, storage and retrieval of information / G. A. Salton. — Moscow : Sov. radio, 1973. — 560 p.

5. *Program* analysis and linguistic text processing. — [Electron resource]. — Mode of access: http://www.rvb.ru/soft/catalogue/index.html.

6. *Competitive* Intelligence in the Internet / V. V. Dudihin, O. V. Dudihina. — Moscow: OOO "Publisher AST": Publisher "NT Press", 2004. — 229 p.

Стаття надійшла до редакції 12.02.2014