Semantic Analysis of Text Data with Automated System

O. Chernenko[1], O. Gordeeva[1]

[1]*Samara National Research University, 34 Moskovskoe Shosse, 443086, Samara, Russia*

Abstract

This paper describes application of the basic methods of semantic analysis of text data – Porter stemming, frequency semantic analysis, latent semantic analysis and syntactic semantic analysis using an automated system. The system allows analyzing the text using these methods. The characteristics and features of the methods' implementation as well as the obtained results of their applying to texts of small complexity are considered. The research allows to reveal features of usage of the methods according to the text analysis purposes.

Keywords: text analysis; frequency semantic analysis; Porter stemmer; latent semantic analysis; core words

1. Introduction

At present days, it is difficult to imagine an effective work with text data without using computer processing. One of the most relevant and ever-evolving types of text processing is the semantic analysis. Depending on the criteria which are set in the automated system, the most appropriate type of semantic analysis can be selected. For example, in the case of search audit of a site, the criteria for choosing a method of semantic analysis are the speed of processing and the minimal dictionary volume. In the case of literal piece of art with complex speech turns, the main criterion of selecting an analysis method is the quality of processing. Consequently, the algorithm of semantic analysis should achieve the results that are as close to natural human as possible, so the parameters such as speed and volume of the dictionaries are not decisive.

2. Task Formulation

The object of the research is a text in Russian, no longer than 20 sentences, with the topic which is unambiguously understandable for human. The purposes of the research are to evaluate the execution of the four selected methods of semantic analysis the developed system is based upon and to compare efficiency and speed of analysis for the methods.

3. Methods of Text Semantic Analysis

All variety of text analysis methods can be divided into two groups:

- linguistic analysis – is based on extracting the meaning of the text from its semantic structure;
- statistical analysis – is based on extracting the meaning of the text and core words by the frequency distribution of words in the text.

The division into two groups is conditional, since in real problems a combination of methods is always used to achieve the result [1, 2].

In this paper, algorithms of semantic analysis from both groups, which are most often used for tackling practical problems, are realized.

3.1. Frequency Semantic Analysis

Method of Frequency Semantic Analysis (FSA) is based on calculating of frequency of word occurrence in the text. There are several refinements for the correct operation of the algorithm [3]:

- since not every word in the text can be the core word, only nouns are counted;
- to distinguish nouns in the text, a dictionary should be used.

The steps of algorithm work: firstly, all words of the text are compared with the dictionary; secondly, the matches are entered into the array, and then they are compared by the number of occurrences. Finally, the words with the largest number of occurrences are the core words of the text.

3.2. Porter Stemming

"Stemming" is a clipping the ending and suffix of the word so that the rest part becomes the basis for all grammatical forms of the word. "Porter Stemmer" or "Porter Stemming" is the algorithm of stemming that determines the basic part of a word. The stemmer can only work with languages that implement word modification through affixes, for example, Russian or English. The main advantage of this algorithm is that it does not need any dictionary or library.

First of all, there are several notations introduced for the parts of the word for stemming process:

- RV – the part of the word after the first vowel. It can be empty if there are no vowels in the word;
- R1 – the part of the word after the first "vowel-consonant" combination;
- R2 – the subpart of R1 after the first "vowel-consonant" combination.

In the article [4], the Porter's algorithm for a word's basic part (stem) determining is described. The algorithm includes the deleting of prefixes, endings and suffixes:

- if there is a gerund ending in the word, it must be removed. Otherwise, if the endings "sia" or "sj" are in the word they must be removed. Next, an adjective/verb/noun ending are looked for. If one of them is found – it must be removed;
- the ending "i" should be found and removed if it is there;
- the endings "ost" or "ostj" must be found and removed if one of them is there;
- if the word ends with "nn" – the last letter "n" must be removed;
- if the word ends with "eyesh" or "eishe" – this part must be removed and then, the last letter "n" must be removed if the word ends with "nn" again;
- if the word ends with "ь" – it must be deleted;

To determine the theme of the text using an algorithm based on Porter Stemming, it is necessary to carry out the stemming process for all words of the text being analyzed. As a result, an array of stems is obtained. The words in the text that are derived from the stem with the most frequent number of occurrences are marked as the theme of the text [5].

3.3. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a method of processing data in a natural language. The method analyzes the relationship between the set of documents and the terms in them and juxtaposes some factors (themes) to all documents and terms. The LSA method is based on the principles of factor analysis. As an input, the LSA uses a term-to-document matrix (terms – words or phrases) [6].

Elements of this matrix contain coefficients (weights) taking into account the frequency of occurrences of each term in each document. The most common version of LSA is based on the using of the singular decomposition of the diagonal matrix (SVD - Singular Value Decomposition). Using the SVD-decomposition, any matrix decomposes into a set of orthogonal matrices, the linear combination of which is a quite accurate approximation to the original matrix.

More formally, according to the singular decomposition theorem, any real rectangular matrix can be decomposed into a product of three matrices:

$$A=USV^T,$$

where matrixes U and V are orthogonal, and the matrix S is diagonal, values in diagonal of the matrix S are called "singular values" of matrix A. Letter "T" means transpose for matrix V.

Such decomposition has a significant feature: if in the matrix S retain only "k" largest singular values, and in the matrices U and V retain only columns corresponding to these values, then the product of the resulting matrices S, V and U is the best approximation of the original matrix A to the matrix Â of "k" rank:

$$\hat{A} \approx A=USV^T.$$

The main idea of the LSA is that if the matrix A is the term-to-document matrix, then the matrix Â containing only the first "k" linearly independent components of the matrix A reflects the basic structure of the dependencies presenting in the original matrix. Proceeding from this decomposition, the dependence between terms and documents is analyzed and the theme of the text is determined [7].

3.4. Syntactic Semantic Analysis

Syntactic Semantic Analysis is a method of processing textual information, which creates templates for comparison with words of text. As a result of the method a list of pairs is created for each sentence [8]. Pair includes:

- the type of word in the sentence;
- the position of the main word for this dependent word.

It is assumed that the basic templates are formed from the most important and often used semantic relations in the text. The basic semantic template is the rule by which the semantic relation is determined in the text being analyzed. The basic semantic template consists of 4 main parts:

- a sequence of words or indivisible semantic units for which their morphological features are indicated;
- the name of the semantic relation that should be formed if the sequence described in the previous item is found in the text;
- a sequence of numbers that determines the positions in a sequence whose elements should be added to the queue with priority. According to the queue the words from the sentence being analyzed are deleted;
- a number indicating the value of the priority, the group of semantic dependencies to which this semantic relation relates.

Using the basic semantic templates, the priority queue is composed. This queue is used to store words that are the argument on the right side of a semantic collocation found in the sentence being analyzed.

To determine the theme of the text from each sentence, according to the priority queue, the word with the biggest number of dependencies is selected and the number of its occurrences in the text is calculated. The word with the maximum number of occurrences is the theme of the text.

4. System Operation

To conduct the research on the methods of text analysis, an automated system was developed. The system operation includes several steps.

At the first step the system splits the text into elements (words or sentences, it is depends on the algorithm chosen by the user), and sends them for processing. The second step is handling the elements and selection the core words.

If the FSA has been chosen by the user, the system compares words from the text with words from the dictionary and finds among them words with the maximum number of occurrences in the text. Next, it displays the finding result – core words of the text. Additionally, system displays the list of words has not been found in the dictionary. It is possible to add that words to the dictionary and run the algorithm anew.

If the algorithm based on Porter Stemming has been chosen, the system defines the basics of original words in the text and looks for the most frequent among them. In this way the core words of the text are found by this algorithm.

In the case of LSA the system constructs a word-on-sentence matrix using the sentences of the text and carries out the SVD decomposition. Then only the first two columns of the resulting matrices are used. From the first two columns of the matrix V^T corresponding to the sentences, a maximum and a minimum are selected. These values correspond to the maximum and minimum coordinates x and y on the coordinate plane. In this way, the area indicated, the entry into which for points from the first two columns of the matrix U corresponding to words means the inclusion into core words.

If the SSA has been chosen by the user, in each sentence words are checked for matching patterns. After that the weight value sets for every word according to the pattern. The more word dependent words, the weight is less and priority is higher. Next, the word with a minimum weight is determined in each sentence, the words with the most number of occurrences form the core of the text.

5. Results

As objects for research, texts for the essays of the Unified State Examination in the Russian language were chosen. These texts were chosen because of their simplicity and small size, and also because their themes are clearly defined.

In tables 1-5 core words for text examples are presented. In addition, time of processing for each method of analysis is given.

Table 1. «Send your head on vacation!» (P. Izmaylov).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	vacation
Porter Stemming	1	feelings each other another series head heads vacation
Latent Semantic	210	head feelings love series passion time rhythm rubles vacation
Syntactic Semantic	720	series publishing vacation

The main idea of the first text is “the influence of mass literature on the human intellectual development”. No methods produced similar theme, but the most suitable core words were given by the latent-semantic method and Porter stemming method.

Table 2. «Things and books, books and things...» (L. Lickhodeev).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	locomotive light book thing time interlocutor
Porter Stemming	2	book
Latent Semantic	240	think idea interlocutor light book thing things time another each other
Syntactic Semantic	840	think things time interlocutor

The main idea of the text could be defined as “relationship between book and time”. The most appropriate core words got the latent semantic method of analysis.

Table 3. « Earth is a cosmic body, and we are cosmonauts...» (V. Solouckhin).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	life support system spaceship cosmonaut Earth communication possibility sides human disease
Porter Stemming	1	life support cosmic cosmonaut cosmonauts spaceship human
Latent Semantic	120	Solar life support cosmic cosmonaut cosmonauts small spaceship Earth river nature disease outside human inner life
Syntactic Semantic	540	cosmonauts cosmonaut spaceship human

The text's main idea is "relations between human and nature". As in previous example, the latent semantic analysis gave the most similar core words.

Table 4. «Books...» (A. Yetoyev).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	life people human childhood book friend
Porter Stemming	1	book people
Latent Semantic	120	measure meet human people similar similarly book books enable
Syntactic Semantic	540	human space life population people book

Core words given by syntactic semantic analysis are the most similar to theme of the fourth text "the role of book in human life"

Table 5. «About the soul…» (M. Prishvin).

Method of analysis	Approximate time of processing (sec)	Core words
Frequency Semantic	5	soul raincoat
Porter Stemming	1	soul
Latent Semantic	90	soul raincoat
Syntactic Semantic	600	soul year raincoat

The topic of the fifth text is "soul of human". All algorithms gave the satisfactory results, the most accurate of which gave the Porter stemming.

6. Conclusion

In the article methods of classification of texts, such as Porter stemming, syntactic semantic, frequency semantic and latent semantic analysis, are considered. The results of the analysis of little complexity texts are given. Based on these results it can be concluded that the usage of methods for determining the topic of a text depends on the complexity of the text – the more accurate analysis for the more complex text should be.

The same applies to trivial texts. The using of complex methods for simple texts leads to unnecessary waste of time and resources, and the result is superfluous in comparison with simple algorithms. Thus, the research shows that for short texts the most effective method is the latent semantic analysis, the fastest method is the Porter stemming. Finally, it should be mentioned that the combination of text analysis methods, for example, combining the Porter method of stamping and frequency-semantic analysis, can be appropriate for effective and accurate core words determination.

References

[1] Velichkevich AG, Cherepackhina AA. Latent semantic analysis of text using Porter algorithm. Youth scientific and technical herald 2015; 10: 38 p. (in Russian)

[2] Mikhaylov DV. Kozlov AP, Emelyanov GM. An approach based on tf-idf metrics to extract the knowledge and relevant linguistic means on subject-oriented text sets. Computer Optics 2015; 39(3): 429–435. DOI: 10.18287/0134-2452-2015-39-3-429-438.

[3] Understanding and synthesizing text by computer. URL: http://compuling.narod.ru/index2.html (11.12.16). (in Russian)

[4] Russian stemming algorithm. URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (11.12.16). (in Russian)

[5] Silva G, Oliveira C. A lexicon-based stemming procedure. Lecture Notes in Computer Science 2003; 2721: 159–166.

[6] Zaboleeva-Zotova AV. Latent semantic analysis and new solutions in Internet. Moscow: Information Technologies, 2001; 22 p. (in Russian)

[7] Kuralenok I, Nekrest'yanov I. Automatic document classification based on latent semantic analysis. Programming and Computer Software 2000; 26(4): 199–206.

[8] Rabchevsky EA. Automatic construction of ontologies based on lexical-syntactic templates for information search. Petrozavodsk, 2009; 107 p. (in Russian)