ALGORITHM OF CLASSIFICATION OF SCIENTIFIC TEXTS BY THE METHOD OF LATENT SEMANTIC ANALYSIS

Ismukanova A.N.[1], Lavrov D.N.[2] (Russian Federation), Keldybekova L.M.[3], Mukunova M.Zh.[4] (Republic of Kazakhstan) Email: Ismukanova433@scientifictext.ru

[1]Ismukanova Aygerim Nauryzbayevna - Postgraduate Student,
DEPARTMENTS OF COMPUTER TECHNOLOGIES AND NETWORKS;
[2]Lavrov Dmitry Nikolaevich - PhD in Engineering, Associate Professor, Head of the Department,
DEPARTMENTS OF COMPUTER TECHNOLOGIES AND NETWORKS, FACULTY OF COMPUTER SCIENCES,
OMSK STATE UNIVERSITY NAMED AFTER F.M. DOSTOYEVSKY,
OMSK;
[3]Keldybekova Liliya Muratbekovna - Teacher;
[4]Mukunova Marzhan Zhumanovna – Teacher,
DEPARTMENT OF FOREIGN LANGUAGES, FACULTY OF "PHILOLOGY",
KOKSHETAU STATE UNIVERSITY NAMED AFTER SH. UALIKHANOV,
KOKSHETAU, REPUBLIC OF KAZAKHSTAN

Abstract: *during the past two decades the study of scientific texts focused on the factors affecting the understanding of the language. However, at presents there are no studies in the field of computer technology, capable of accurate assessment of the classification of the scientific text. New technologies for the LSA model could represent a important advance of the assessment of scientific texts.*

LSA model despite the complexity of the opacity and can be used for a number of different tasks with a generalization or extension of the meaning of the search query.

The Latent semantic analysis (Latent Semantic Analysis (LSA)) - the theory and a method for extraction, and submission of the contents of contextual use of words statistical calculations was applied to a large number of texts. The latent semantic analysis (LSA) – is the semantic domain defining mathematical representation of computing linguistic model. Works on improvement and adaptation to various tasks of the latent semantic analysis (LSA) are conducted long ago languages.

Keywords: *latent semantic analysis (LSA), artificial intelligence (AI), artificial neural network (ANNs), machine learning (ML), classification.*

АЛГОРИТМ КЛАССИФИКАЦИИ НАУЧНЫХ ТЕКСТОВ ПРИ ПОМОЩИ ЛАТЕНТНО-СЕМАНТИЧЕСКОГО АНАЛИЗА

Исмуканова А.Н.[1], Лавров Д.Н.[2] (Российская Федерация), Кельдыбекова Л.М.[3], Мукумова М.Ж.[4] (Республика Казахстан)

[1]Исмуканова Айгерим Наурызбаевна - аспирант,
кафедра компьютерных технологий и сетей;
[2]Лавров Дмитрий Николаевич - кандидат технических наук, доцент, заведующий кафедрой,
кафедра компьютерных технологий и сетей, факультет компьютерных наук,
Омский государственный университет им. М.Ф. Достоевского,
г. Омск;
[3]Кельдыбекова Лилия Муратбековна - преподаватель;
[4]Мукумова Маржан Жумановна – преподаватель,
кафедра иностранных языков, факультет филологии,
Кокшетауский государственный университет им. Ш. Уалиханова,
г. Кокшетау, Республика Казахстан

Аннотация: *в течение прошлых двух десятилетий исследование научных текстов фокусировалось на факторах, влияющих на понимание языка. Однако в настоящее время отсутствуют исследования в области компьютерных технологий, способных к точной оценке классификации научного текста. Новые технологии для модели LSA (латентно-семантического анализа) могли представлять важное усовершенствование в исследовании оценки научных текстов.*

Модель LSA, несмотря на трудоемкость и непрозрачность, может использоваться для разного ряда задач при обобщении или расширении смысла поискового запроса.

Латентно-семантический анализ (Latent Semantic Analysis (LSA)) - теория и метод для извлечения и представления содержания контекстного использования слов статистическими вычислениями применялось к большому количеству текстов. Латентно-семантический анализ (LSA) является семантическим доменом, определяющим математическое представление вычислительной *лингвистической модели. Работы по усовершенствованию и адаптации к различным задачам латентного семантического анализа (LSA) ведутся давно.*

Ключевые слова: *латентный семантический анализ (ЛСА), искусственные нейронные сети (ANNs), машинное обучение (МА), классификация.*

Works on improvement and adaptation to various tasks of the latent analysis are conducted long ago. The essence of the method is rather simple. In the beginning on an entrance to an algorithm there is a set of texts which will be transformed to a matrix of frequency of occurrence of words in these texts. The line number corresponds to a word, and number of a column corresponds to the text. By means of an algorithm of singular decomposition (SVD) at the received matrix the rank goes down. It allows to reject dependence of words and to allocate a so-called semantic kernel. Then on the basis of the received matrix with the lowered rank correlation coefficients between texts are calculated. In one of the first works the line in which it is possible to group texts on similar subjects is empirically defined. If a part from these texts already has the universal decimal code (UDC) which of course is correctly exposed by the author or the editor, all group will have this code or close to it. It will allow to calculate UDC in the automatic mode.

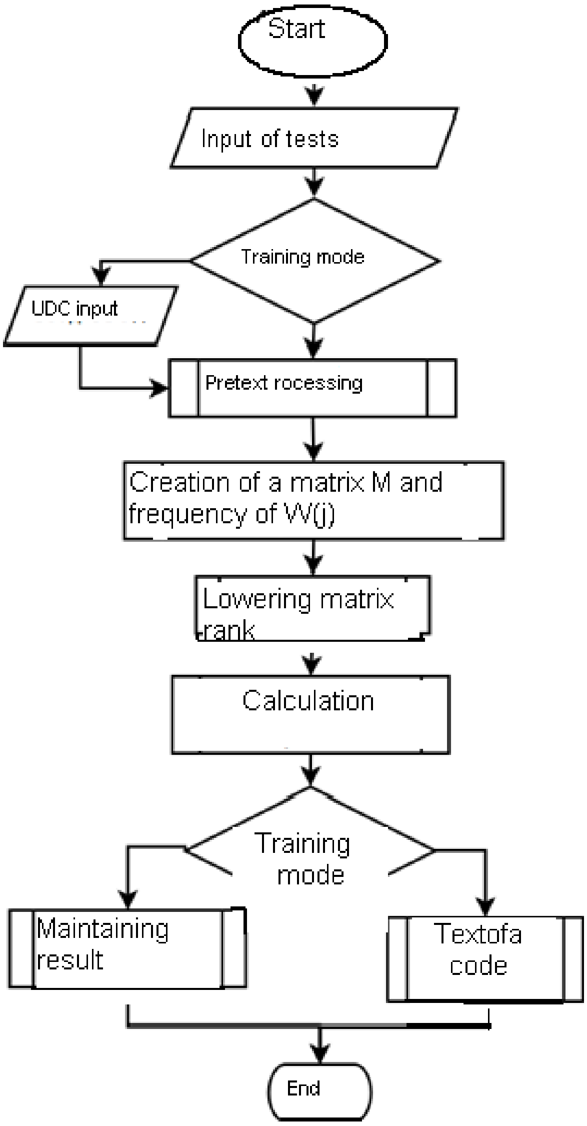

Fig. 1. The block diagram

The classification of scientific texts in the Russian and Kazakh languages, by means of assignment of a universal decimal code (UDC) is an actual problem nowadays. The problem of the classification of scientific texts is easily solved for the English language due to its simplicity of morphology and syntax. On this way there is a number of unresolved tasks for the Russian language and practically the usage of analogical reception for the Kazakh language is not investigated. For the Russian language several researches of applicability of different approaches were conducted.

One of the most common of clustering of texts is the method of the latent semantic analysis (LSA) [1].

The language of Python was chosen for program realization of the tool of the analysis of languages. The demonstration of the system is presented by realization of “machine learning “ on the basis of the latent semantic analysis.

The library of work with natural languages Natural language Toolkit (NLTK) was actively used by application programming in Python.

The main method of application of NLTK is inclusion in a set of parameters of widespread phrases of two-three words. In NLTK there is a support of these options such as: ntkl.bigrams(…) and ntlk.trigrams(…).

NTLK has the following options:

1) Tokenization of the text;
2) The choice from all data set of the most frequently-used words;
3) The identification of the most frequently-used two- or three-word phrases.

Tokenization is a division of the text into small portions, tokens. Words, prepositions and signs of a punctuation belong to tokens. Rather often there is a task to present the text in the form of the massif of significant words. Aftertokenization, it is necessary to make cleaning regarding signs of a punctuation and insignificant words (for instance: prepositions). It becomes by means of transfer of the list of stop-words to library, which are automatically excluded from consideration. And that significantly increases productivity of the method of LSA. After that the gathered statistics of words is transferred to LSA. Thus, all preliminary processing and preparation of data is carried out by means of NLTK.

The analysis of texts can be carried out both without a dictionary (taking into account any words, met in a text), and with a dictionary (taking into account only words, presented in the configured dictionary).

Besides usage of dictionaries, while analyzing texts, grammar peculiarities of different human languages are not used. The exception of the grammatical analysis allows to provide the high speed of classification at the large volume of entrance data. However, lack of grammatical analysis doesn’t allow to use particular factors, related to syntax and morphology of texts in the Kazakh and other languages. In the current version of system words act as features. In the experimental version of the system which is in development, the sequence of words, their phrases and also sets of alternative words (synonyms). At the initial stage of work training of system of automatic classification, requires preparation of a training data set. Training data set is to include a set of texts with the related categories by results of preliminary manual classification [2].

During work, training set is updated, taking into account the specification and correction made in manual by results of automatic classification.

References in English / Список литературы на английском языке

1. *Denhière G., Lemaire B., Bellissens C. & Jhean-Larose S.*, 2007. A semantic space modeling children’s semantic memory. In T.K. Landauer, D.McNamara, S. Dennis & W. Kintsch (Eds.). The handbook of latent semantic analysis. Mahwah, NJ: Erlbaum. PP. 143-167
2. *Landauer T.K., Foltz P., Laham D.* An Introduction to Latent Semantic Analysis. Discours Processes. 25, 1998. PP. 259-284.

Список литературы / References

1. *Денхиер Г., Лемер Б., Беллиссенс К. & Джхин-Лэроз С.*, 2007. Семантическое пространство, моделируя детскую семантическую память. В Т.К. Лэндоере, Д. Макнамаре, S. Dennis, & W. Kintsch (Редакторы). Руководство скрытого семантического анализа. Мово. Нью-Джерси: Erlbaum. С. 143-167.
2. *Ландауер Т.К., Фолц П., Лэхэм Д.* Введение в скрытый семантический анализ. Процессы Discours, 25, 1998. С. 259-284.