FULL-TEXT SEARCH OF DOCUMENTS IN THE CONTENT MANAGEMENT SYSTEM ELAB-SCIENCE

Dunets A. P., Sytova S. N., Kavalenka A. N., Mazanik A. L., Sidorovich T. P., Charapitsa S. V.

Institute for Nuclear Problems, BSU, Minsk, Belarus, e-mail: dunets@gmail.com

The function of searching documents and information in databases of various types arises quite often when implementing information systems [1]. In many cases, it is necessary to search for information on query in natural language. It should be noted that existing search engines such as Google and Yandex have very sophisticated search algorithms. However, there are specific practical situations when the use of popular search engines is impossible. For example, when the developed information system does not have access to the Internet for some reason (in this case it does not matter which one). Another example is if the search needs to be run in some subset of the data, when the subset is given by a set of hard constraints. In this situation, it is necessary to integrate search tools into the software that will be developed, which will make it possible to circumvent these shortcomings. Similar solutions exist. But at the same time they have a common disadvantage - the limited possibilities of searching for queries in natural language.

Files and data chunks for extracting text can come from different sources. At the moment implemented the processing of documents in PDF, DOC, DOCX, RTF. The extraction of text from them is performed by means of Apache Tika. Extraction of word stems (base or root form) is performed using Snowball algorithms [2]. The collected data is stored in the database for use in the search process.

The Okapi BM25 [3, 4] algorithm is used to rank the results by relevance. This algorithm is considered more relevant documents that contain rare words from the query. The less common the word in the set of texts, the more important it is. Despite the presence of some shortcomings, the algorithm is effectively used in real-world tasks. To improve the efficiency of the search, the user's original search query expands to synonyms for words that participate in the query already. In this case, the ranking function covers more documents in the analysis. As a search algorithm for synonyms, Word2Vec [5] is used. The key advantage of this algorithm is that its result is a list of synonyms with an estimate of the degree of proximity to the original word. This estimation is used in the ranking of documents.

References

1. Information system eLab for accredited testing laboratories / S.N. Sytova [et al.] // Informatics. – Minsk: UIIP NAS Belarus, 2017. – № 55. – pp. 49-61.

2. Snowball Stemmers [Electronic resource] / ed. Dr. Martin Porter. – Mode of access: http://snowballstem.org/ – Date of access: 15.05.2017.

3. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, An Introduction to Information Retrieval. – Cambridge, UK: Cambridge University Press, 2008. – p. 213.

4. The BM25 Weighting Scheme [Electronic resource] / Oligarchy Limited – Cambridge, UK. – Mode of access: https://xapian.org/docs/bm25.html – Date of access: 15.05.2017.

5. Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient Estimation of Word Representations in Vector Space [Electronic resource] / Cornell University Library – Ithaca, NY, USA. – Mode of access: https://arxiv.org/abs/1301.3781 – Date of access: 15.05.2017.