



FULLTEXT SEARCH IN MEDICAL REPORTS

Volšička Vít

Institute of Biostatistics and Analyses, Masaryk University

D2.2 MEDICAL INFORMATICS AND INFORMATION SCIENCE

Keywords: *fulltext, search, Apache, Solr*

Search in already existing documents is an integral part of the educational process and acquiring new information. Unfortunately, finding relevant information is not always easy, because many documents are saved only in the form of free text and their search is therefore difficult. We face to the same problem when we are trying to scan medical documents such as medical reports.

And those are stored with exceptions in hospital information systems as free text. However, for example, a task to find relevant information about patients who had been in hospital with specific diagnosis and underwent specific medical procedure, makes a great sense. Even in the case of medical reports we are facing to situation, when we need to search and identify relevant records in large volumes of free text.

The solution of selection of relevant records and finding the necessary information in them is called fulltext search. It makes possible efficiently search in large amount of documents and return those that contain the search phrase in a short time. The advantage of this approach over other option is a possibility to enter long queries, including advanced operators, while maintaining a short waiting time for the result. It allows easily and quickly find relevant documents for study.

All leading database systems currently offer support for fulltext search, but some do not support the complex morphology of the Czech language, which is solved by converting words to their basic form. It is called standardization. Full support provides some fulltext libraries, in the forefront with Apache Solr.

Full-text searching is currently the best solution for searching in unstructured text. Fulltext search is, unlike common search method finding of exact matches, faster and it is in the case of normalization of words many times more precise. The proposed library Apache Solr for implementation of full-text search in medical reports is advantageous due to the good support of the Czech language in the basic installation, a wide range of settings and options for deployment over any platform; does not depend on the used database, programming language or information system. The library has been satisfactorily tested over real data from the hospital. However, there is still a need to make adjustments for effective search in this issue, especially correction of typos made not only in writing but also in entering of words to search box and create a list of used abbreviations and synonyms for more accurate results.

Search in medical text information has its own specifics that we can solve by some tools designed for text searching. An appropriate tool is Apache Solr — fulltext library that can be easily modified to the needs of the issue and provides enough power to work effectively.