Retrieving Biomedical Literature: An Open Source Search Engine Based on Open Access Resources

Hayda Almeida, Ludovic Jean-Louis and Marie-Jean Meurs

The retrieval of biomedical literature is a critical task for scientific researchers and health care practitioners. Open scientific literature databases contain a massive amount of data, which is extensively used to support various research activities in life sciences. Lots of research efforts have been made towards improving the retrieval of bioliterature, but the task is still challenging.

PubMed and PubMed Central (PMC) are scientific literature databases maintained by the U.S. National Library of Medicine. As of February 2016, PubMed holds over 25 million records, allowing users to search the content of article abstracts, while PMC holds over 3.7 million of free full-text articles. When utilizing databases such as PubMed and PMC to retrieve relevant information, researchers generally need to express their search needs using a specific query language. This makes the task difficult for users not experienced with query languages, and can compromise the knowledge discovery process.

In this work, we present an open source search engine that aims to address two different aspects related to the retrieval of biomedical literature: improve the content access offered by PubMed or PMC, and facilitate the query formulation for users by processing queries in natural language. The system is composed of two modules: the indexation module and the complex query module. Based on the search platform Solr/Lucene, the indexation module generates the inverted index of the dataset, representing all documents using relevant content found in the article content (titles, abstract, body, keywords, references, etc.).
The complex query module handles complex user queries, which are processed according to different query types. For each type, a specific search strategy is applied to better meet the user needs. In addition, query terms can be expanded using UMLS concepts.

Our search engine was created based on the open-access scientific literature made available by the PubMed Baseline Database (BD), and the PMC Open Access (OA) Subset repository. A total of 25,403,053 articles from these sources was indexed as of February 2016. Information retrieval systems are often evaluated using reference judgments or pseudo-judgments. Here we proposed an evaluation method based on pseudo-judgments, and sets of annotated queries. Our evaluation dataset is composed of query-document sets manually annotated by curators working on the mycoCLAP database.
The dataset utilized for preliminary evaluation has 19 query-document relations. From the total, 9 queries have a correct response document mapped to a PMC OA entry (full text article). The other 10 have a correct response document mapped to a PubMed BD entry (article abstract). For each query, we analyzed the first 20 ranked documents, and computed a Mean Reciprocal Rank (MRR) score for the correct response document, considering the position where it was found in the search result list. The MRR score over 0.5 indicates that the system retrieved the correct response document in first or second positions for more than half of the requests. Our work currently focuses on improving the retrieval of full-text documents.