Text Mining for Biological Data: An Update from the Last Decade

Camilla Reginatto De Pierri, Diogo de Jesus Soares Machado, Bruno Thiago de Lima Nichio, Antonio Camilo da Silva Filho, Fabio de Oliveira Pedrosa, Roberto Tadeu Raittz

Federal University of Paraná

Abstract

Scientific literature is the basis of research in any field of study. Analysis of information from scientific texts is an important strategy to define the starting point and evaluate the state of the art in a given field of research, as well as assisting in the construction of hypotheses and interpretation of results. In the biological area, with the growing number of scientific texts deposited in public databases, the task of identifying relevant studies becomes complex and time consuming. To deal with this large amount of information, Text Mining (TM) approaches efficiently handle knowledge seeking. TM is a process that refers to the extraction of information found in texts. The advantage of TM techniques is the ability to improve bibliographic search, facilitate analysis and data storage, making the search process refined and accurate. Currently, there are series of TM tools that contemplate different methodologies with potential for study in the most varied biological scenarios. However, we noticed that most studies using TM as a research tool do not relate findings to a set of strategies, which in our view, may limit the discovery of knowledge. To assist researchers in choosing the best TM strategy, we conducted a literature review on the topic TM and the main tools available. The criteria for study selection were: 1) TM tools developed and / or implemented from 2009 to 2019; 2) Only tools available through scientific publication; 3) Only tools that the research was published in PubMed database. We have identified 41 TM tools with the most varied applications. These include MedlineRanker, DataShield, FACTA +, SAPIENTA, BioC and DISEASES with the highest number of citations, according to Google Scholar, Scopus and Web of Science. The methodology of the tools selected in this research involves processes of information retrieval, machine learning, natural language processing and computational and statistical language, focusing mainly on the study and identification of events related to genes and proteins. We found that text mining is not a simple keyword search in databases. Several automated processes and methods are required for the extraction of knowledge from texts. The use of one or more TM approaches is valuable for identifying relevant concepts and uncovering hidden knowledge in light of unexplored subjects.

Funding: CAPES, Fundação Araucária