Using Data Marts to Select Related Research Articles: A Case Study for the Prioritization of Drug Targets

Marlon Amaro Coelho Teixeira, Kele Teixeira Belloze, Maria Cláudia Cavalcanti, Floriano Silva-Junior

Acre Federal Institute, CEFET/RJ, Military Institute of Engineering, Oswaldo Cruz Institute,

Protozoan trypanosomiasis are among the ethiological agents of major tropical diseases such as leishmaniasis, Chagas disease, malaria, sleeping sickness and amebiasis. These parasite infections affect the poorest populations of the third world countries with limited access to effective treatments and, therefore, to find novel drugs is of vital importance for them. The research efforts to combat these protozoa grow every day and consequently a large amount of unstructured data has been made available through scientific articles. These articles are accessed in the vast majority of cases by tools that are keyword-based queries, but they are limited and can not meet the needs of researchers. Simple searches performed through these interfaces can return more than a thousand hits. Tools that combine large amounts of data with high performance, enabling users to manipulate and analyse information from different perspectives are more appropriate to deal with this information. However, in the context of a scientific research, these approaches are not quite exploited. The main innovation of this work is to demonstrate that a widely used approach in the analysis of trade data can be applied in analysis of scientific data supporting decision making researcher. Initial experiments were run on a scientific scenario where a corpus of selected papers was annotated using three distinct ontologies with focus on the research of five protozoan organisms: Entamoeba histolytica, Leishmania major, Plasmodium falciparum, Trypanosoma brucei and Trypanosoma cruzi. Then, the annotation data was extracted, organized and aggregated into a dimensional schema of a demo Data Mart. Finally, based on some simple queries over these data, it was possible to verify that this approach helps the scientist on his/her research, correlating terms and preventing that articles are not accessed. In contrast, using a key-based tool, such query misses many articles and also return many false positives for example, consulting the Gene knockout and knock-out synonymous terms in PubMed, 64017 and 10027 articles are obtained respectively, then if the researcher to use the term knock-out in his query, 53990 articles will no longer be accessed.