Recommender Systems for Scientific Fields

Márcia Barros (PhD Student)[1, 2], André Moitinho (Co-supervisor)[2], Francisco M. Couto (Supervisor)[1]

Email: marciabarros@edu.ulisboa.pt

1 - LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal

2 - CENTRA, Faculdade de Ciências, Universidade de Lisboa, Portugal

Abstract: Databases for scientific entities, such as chemical compounds, diseases and astronomical objects, are growing in size and complexity, reaching the billions of items per database. The researchers need new and innovative tools for assisting the choice of these items. In this work, we propose the use of Recommender Systems (RS) approaches for helping the researchers to find new items of interest.

RS have been successfully explored in a vast number of domains, e.g. movies and tv shows, music, or e-commerce. In these domains we have a large number of datasets freely available for testing and evaluating new recommender algorithms. For example, Movielens and Netflix datasets for movies, Spotify for music, and Amazon for e-commerce, which translates into a large number of successful algorithms applied to these fields.

However, RS are not being used so often in scientific fields, such as Health, Chemistry and Astronomy.

We identified as one of the major challenges for applying RS in scientific fields the lack of standard and open access datasets with the information about the preferences of the users.

To overcome this challenge, we developed a methodology called LIBRETTI - LIterature Based RecommEndaTion of scienTific Items, whose goal is the creation of <user, item, rating> datasets, related with scientific fields. These datasets are created based on the major resource of knowledge that Science has: scientific literature. We consider the users as the authors of the publications, the items as the scientific entities (for example chemical compounds or diseases), and the ratings as the number of publications an author wrote about an entity. The first case studies conducted with LIBRETTI were in the fields of Astronomy and Chemistry, having as items open clusters of stars and chemical compounds, respectively. More recently, LIBRETTI methodology was applied to phenotypes, diseases, and gene terms, particularly related to the COVID-19 disease.

With these datasets available, it is now possible to start testing and developing new recommender algorithms. In the field of Chemistry, we developed a hybrid recommender model suitable for implicit feedback datasets and focused on retrieving a ranked list according to the relevance of the items. The model integrates collaborative-filtering algorithms for implicit feedback (Alternating Least Squares and Bayesian Personalized Ranking) and a new content-based algorithm, using the semantic similarity between the chemical compounds in the ChEBI ontology. The algorithms were assessed on an implicit dataset of chemical compounds, CheRM-20, with more than 16 000 items.

However, we know that science is mutable along the time, and relevant items in the past may not be relevant for a user anymore. Thus, we are now working on the recommendation of scientific items taking in account the time when each preference was published. Instead of <user,item,rating>, we are considering the sequence of scientific entities a user had interest along a time period and trying to predict the best next entity for this user. This is an ongoing work, but we already have promising results.

Presentation mode: Video

Acknowledgements: This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017), LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020), CENTRA Research Unit (ref. UIDB/00099/2020), and PhD Scholarship ref. SFRH/BD/128840/2017