*Digital Personalized Health and Medicine*
*L.B. Pape-Haugaard et al. (Eds.)*


*doi:10.3233/SHTI200353*

# A Recommender System Based on Cohorts' Similarity

João Rafael ALMEIDA[a,b], Eriksson MONTEIRO[c], Luís Bastião SILVA[c], Alejandro PAZOS[b], and José Luís OLIVEIRA[a]

[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*
[c] *BMD Software, Portugal*

**Abstract.** Aiming to better understand the genetic and environmental associations of Alzheimer's disease, many clinical trials and scientific studies have been conducted. However, these studies are often based on a small number of participants. To address this limitation, there is an increasing demand of multi-cohorts studies, which can provide higher statistical power and clinical evidence. However, this data integration implies dealing with the diversity of cohorts structures and the wide variability of concepts. Moreover, discovering similar cohorts to extend a running study is typically a demanding task. In this paper, we present a recommendation system to allow finding similar cohorts based on profile interests. The method uses collaborative filtering mixed with context-based retrieval techniques to find relevant cohorts on scientific literature about Alzheimer's diseases. The method was validated in a set of 62 cohorts.

**Keywords.** Alzheimer cohorts, Cohort catalogue, Recommendation systems

## 1. Introduction

There is a need for discovering new biomarkers to diagnose Alzheimer's disease in the predementia stage to predict the rate of decline [1]. Therefore, several institutions increased the efforts in collecting Alzheimer's patients' data, as well as to build a harmonised structure to store these cohorts.

In the EMIF project[1], we have developed a platform to collect and publish metadata about EHR databases and disease specific cohorts [2]. The EMIF Catalogue includes several communities, one of them dedicated to the study of Alzheimer's disease. In this centralised platform, data sets are classified through a set of general characteristics, which allow the identification of similar patterns across the collection. However, the manual analysis of these similarities is time consuming and hard to perform.

In this paper, we propose a methodology that, based on previous knowledge gathered from user's preferences and user's data sets, is able to suggest similar cohorts and related scientific publications. The method uses collaborative filtering and context-based retrieval techniques, taking as inputs the cohorts' metadata in an predefined ontology, its concepts and relationships.

[1] http://www.emif.eu

## 2. Methods and Materials

The EMIF Catalogue is a web platform for biomedical data sharing, where data custodians can publish information about their data sets [2]. This information is mainly statistical information and aggregated databases' metadata. In the platform, researchers can search for databases of interest and request access to the raw data, data subsets, or ask to perform certain research questions.

Most of the research studies do not follow a standard structure or format, because they are built with a specific purpose, with inclusion/exclusion criteria. To allow the reproducibility of research questions in different cohorts, our strategy was the adoption of a taxonomy implemented using an ontology with the recorded variables and also the values (columns and rows). This method allows comparing multiple cohorts, i.e., all the variables will be mapped into ontology classes, and this ontology supports relationships between concepts, following a hierarchical tree with root entities that represent core categories that are being followed (i.e. Patient Demographic Data, Neuropsychiatry, Laboratory Results, etc.). Moreover, it is also possible to create constraints in the cohort scope, and it is flexible enough to extend or create new variables. The ontology management is maintained by community using a common RDF (Resource Description Framework) specification.

## 3. Results and Conclusions

We proposed a recommendation system that combines two techniques. The collaborative filtering can detect similar users and provide the recommendation based on the analysis when the cohorts structure of the users interest are too disperse. On the other hand, the context-based retrieval can predict suggestions when the user is more singular, by relying only in the cohorts' concepts. We applied evaluation metrics to each individual approach and then combining both strategies. The methodology was integrated in the EMIF Catalogue platform, using the statistical data collected from the Alzheimer's community. This group gathered metadata information from 62 cohorts, representing in total 661 concepts, which have also been indexed through our system. This community has more that 500 registered users, allowing a first test-bed to apply our methodology.

## Acknowledgments

This work was also funded by the project POCI-01-0145-FEDER-016385 (NETDIAMOND) co-funded by Centro 2020 program, Portugal 2020, European Union, through the National Science Foundation. JRA is funded by the National Science Foundation (FCT), under the grant SFRH/BD/147837/2019.

## References

[1] Isabelle Bos, Stephanie Vos, *et al.* The emif-ad multimodal biomarker discovery study: design, methods and cohort characteristics. *Alzheimer's research & therapy*, 10(1):64, 2018.

[2] José Luís Oliveira, Alina Trifan, and Luís A Bastião Silva. Emif catalogue: A collaborative platform for sharing and reusing biomedical data. *International journal of medical informatics*, 126:35–45, 2019.