Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210174

A Recommender System to Help Refining Clinical Research Studies

João Rafael ALMEIDA[a,b,1] , João Figueira SILVA[a,1], Sérgio MATOS[a], Alejandro PAZOS[b] and José Luís OLIVEIRA[a]
[a] *DETI/IEETA, University of Aveiro, Portugal*
[b] *Department of Computation, University of A Coruña, Spain*

Abstract. The process of refining the research question in a medical study depends greatly on the current background of the investigated subject. The information found in prior works can directly impact several stages of the study, namely the cohort definition stage. Besides previous published methods, researchers could also leverage on other materials, such as the output of cohort selection tools, to enrich and to accelerate their own work. However, this kind of information is not always captured by search engines. In this paper, we present a methodology, based on a combination of content-based retrieval and text annotation techniques, to identify relevant scientific publications related to a research question and to the selected data sources.

Keywords. Research Question Refinement, Medical Studies, NLP, Medical Studies

1. Introduction

Medical research studies can be divided in categories, such as observational studies and clinical trials. Observational studies consist in documenting the relationship between the exposure and outcome in the study [1]. With the objective of reducing the time spent in the execution of these studies, some organisations and projects started harmonising Electronic Health Records (EHR) into standardised data schemes. This process allows researchers to reuse stored information to perform observational studies without having to collect patient data [2].

The European Medical Information Framework (EMIF)[2], was one of the first initiatives that aim to improve the access of researchers to patient-level data from distinct health databases across Europe. In this project, we developed the EMIF Platform to provide metadata information about each database, including scientific literature related to each database. In addition, this platform also integrates computational tools that researchers can use to conduct a study from the design stage up to the result aggregation phase [3].

Despite removing the need to collect the data, correctly selecting the study design is essential to increase study feasibility. This selection is made according to the target exposure in the study and the current advances in the field. Therefore, this initial task

[1] Corresponding author, João Rafael Almeida – Universidade de Aveiro, Campus Universitário do Santiago, 3810-193 Aveiro, Portugal, Both authors contributed equally to this work.; E-mail: joao.rafael.almeida@ua.pt.

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

requires a deep background analysis that is currently optimised with search engines such as PubMed [3]. However, we identified an opportunity to improve this stage by automatically recommending literature based on the target outcome and databases chosen for the study. In the present paper, we propose a methodology to be integrated into research platforms similar to EMIF Platform, aiming to recommend literature by combining content-based retrieval techniques with annotated abstracts from scientific publications. The system was validated in a controlled environment using the history of studies performed in one community of the EMIF Platform, which includes metadata about 19 EHR databases.

2. Methods and Materials

The proposed methodology can be divided into 3 main components: 1) Cohort definition, in which the researcher defines the target outcome and the data sources; 2) Literature annotation where, based on the study design, abstracts are extracted from PubMed and annotated; and 3) Content-based retrieval, that aims to rank the most relevant publications based on the annotations.

2.1 Cohort Definition

A cohort is defined as a group of subjects that share similar characteristics. The design of these studies contains several features that are identified and observed over time in the group of selected subjects [4]. Subject data can be collected in follow-up visits, or extracted from observational databases.

OHDSI (Observational Health Data Sciences and Informatics) [4] is one of the initiatives that aim to develop methodologies and solutions for supporting large-scale observational studies in health care data [2]. From their large ecosystem of solutions, OHDSI released a tool named ATLAS[5] which consists of a web-based platform to design cohorts and to allow population-level analysis of observational data.

In ATLAS, researchers can define cohorts by identifying groups of people based on particular health conditions, in which the inclusion and exclusion criteria, the target population and the expected outcome are defined. The output of ATLAS is a query that can be shared with other researchers in order to replicate the study schema in their own databases, and can be provided via a JSON formatted file. However, this output contains metadata information about the study, namely the concepts used during query definition which are normalised as they belong to established standard vocabularies. This metadata can be explored for other additional purposes.

2.2 Literature Annotation

ATLAS research queries are stored in JSON format and contain relevant information about the study. These files were processed to extract keywords of interest, and the resulting words were used in Entrez to retrieve associated literature from PubMed. Additionally, the EMIF Platform contains a curated list of the publications associated

[3] https://pubmed.ncbi.nlm.nih.gov/

[4] http://www.ohdsi.org/

[5] http://www.ohdsi.org/web/atlas/

with each database, *i.e.* a manually inserted list of scientific papers that made use of a given database. Therefore, using the information contained in the cohort selection script and the databases to be used in the study, the corresponding lists of related publications were also retrieved. A final list of publications of interest was compiled by aggregating the output of both aforementioned sources.

With the constant growth in terms of production volume of biomedical literature, it becomes of paramount importance to develop solutions that are capable of selecting and summarising existing scientific publications into more condensed representations. One possible methodology to summarise information in biomedical text is by annotating relevant content using Natural Language Processing (NLP) solutions. While most of these solutions are developed considering general text, these typically incur in decreased performance when applied to domain specific text such as biomedical literature. This problem was tackled with the development of biomedical NLP solutions, that are designed to efficiently annotate biomedical text, with an example of such solution being Neji [5].

The list of publications resulting from the aggregation procedure was annotated using an approach similar to that explored in [6], where Neji was used to annotate relevant concepts in clinical notes. Here, we firstly created custom dictionaries from the standard concepts used in the research studies during the previous stage (these codes are obtained from Athena[6]). The resulting dictionaries were configured to be used by Neji, and then the Neji annotator was used to annotate all abstracts from the previously selected list of scientific publications. Next, a visualisation component from Becas [7] was also integrated so that a visual overview of each annotated abstract could be provided to the user.

2.3 Content-Based Retrieval

A content-based recommendation system provides suggestions based on the content of a item and the user's rating. The similarity between the items is calculated based on the analysis of selected features [8]. Suggestions are defined considering the interests of the user, and this can be solved as a classification problem if the items the user likes and dislikes are considered [9].

In our proposal, the items are the annotated publications, while Neji annotations are the features used to calculate the similarity. User interests are the concepts defined in the research question, which were established in the process of cohort definition. With a Bayesian classifier, a probabilistic model was defined to estimate a *posterior* probability $P(c|s)$, of publication s belonging to class c. The calculation was based on the following probabilities: 1) observing an item with the label c, $P(c)$; 2) observing item s given class c, $P(s|c)$; and 3) observing item s, $P(s)$. The Bayes theorem was then used to calculate $P(c|s)$ as:

$$P(c|s) = \frac{P(c)P(s|c)}{P(s)} \quad (1)$$

[6] https://athena.ohdsi.org/

This approach can classify the publications based on the Neji annotations, and suggest the most similar. In this case, it was not necessary to have the user ranking since it is assumed that the interest in each of the cohort's concepts is equally distributed.

3. Discussion

The proposed methodology is represented in Figure 1, which illustrates an overview of the workflow. The research starts with the definition of the cohort and with the selection of the databases of interest. The research query is exported from ATLAS in a JSON file, from which the concepts can be extracted and then inserted in a PubMed query. Even though these queries usually return a huge number of publications, cache mechanisms in the Neji service process these publications quickly.

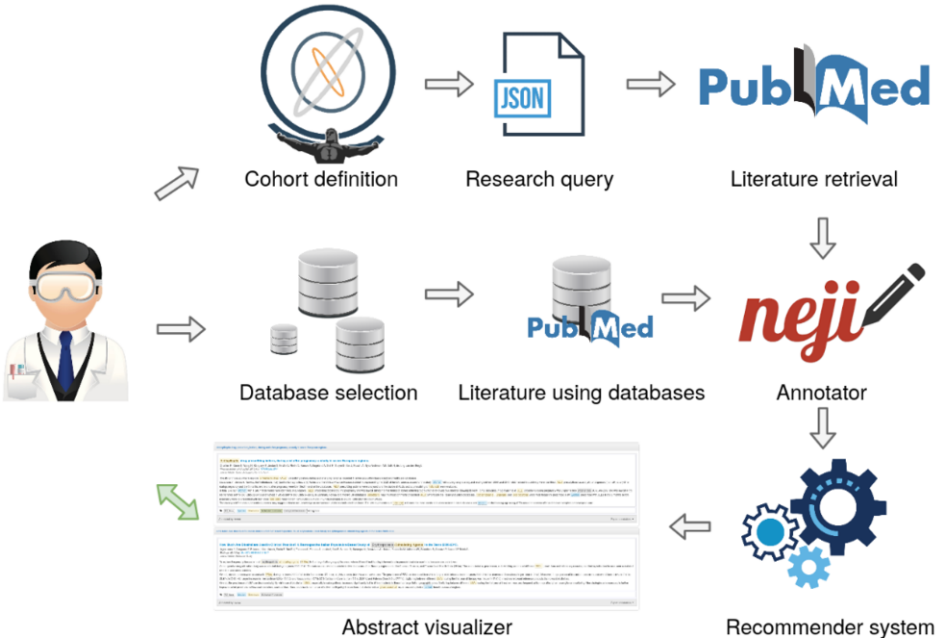

Figure 1. System architecture overview. The green arrow represents the last phase in which the researcher can interact with the recommended literature and visualise the provided annotations.

Simultaneously to this flow, the literature associated with the chosen databases is also annotated using Neji together with the output of the PubMed query. Once the annotation stage is complete, the recommender system calculates the similarity between each publication with the concepts in the research query, and provides a recommendation of the most similar. We did not define a threshold for the minimum level of similarity, because the most important aspect in this process is not to eliminate undesired publications, but instead to provide the most relevant first. Finally, the abstracts are presented in the EMIF Platform with their corresponding Neji annotations.

This methodology was validated on the EHR community in the EMIF Platform since it contains the metadata of 19 databases, in which almost all of them have a list of curated literature. The research questions used to validate this methodology were extracted from

previous studies. Since there is no gold standard available, we were not able to use standard metrics to objectively evaluate system performance. However, our main motivation was to develop and validate a methodology to support future studies and help researchers refining their research question in order to obtain more impactful findings, which we believe that was accomplished by performing a manual analysis of the first 10 suggested articles for each research question.

4. Conclusions and Future Work

The proposed recommender system can be integrated into distinct platforms designed to conduct research studies. The possibility of gathering and recommending literature regarding a study accelerates background analysis, which fosters the refinement of the research question at earlier stages. Despite PubMed and other search engines already providing good filtering features, having a literature recommender system integrated into the ecosystem used to conduct studies may potentially increase the impact of the findings.

We believe that automatically suggesting and annotating the most relevant literature about the study at an initial stage will save time and help researchers in the important process of refining the study protocol.

Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. JFS and JRA are funded by the FCT - Foundation for Science and Technology (national funds) under the grants PD/BD/142878/2018 and SFRH/BD/147837/2019 respectively.

References

[1] Ranganathan P, Aggarwal R. Study designs: Part 1 - An overview and classification. Perspect Clin Res. 2018 Oct-Dec;9(4):184-186. doi: 10.4103/picr.PICR_124_18

[2] Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li YC, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8

[3] Almeida JR, Fajarda O, Pereira A, Oliveira JL. Strategies to Access Patient Clinical Data from Distributed Databases. Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF; 2019 Feb 22-24; Czech Republic; p. 466-473. doi: 10.5220/0007576104660473

[4] Ranganathan P, Aggarwal R. Study designs: Part 3 - Analytical observational studies. Perspect Clin Res. 2019 Apr-Jun;10(2):91-94. doi: 10.4103/picr.PICR_35_19

[5] Campos D, Matos S, Oliveira JL. A modular framework for biomedical concept recognition. BMC Bioinformatics. 2013 Sep 24;14:281. doi: 10.1186/1471-2105-14-281

[6] Silva JF, Almeida JR, Matos S. Extraction of Family History Information From Clinical Notes: Deep Learning and Heuristics Approach. JMIR Med Inform. 2020 Dec 29;8(12):e22898. doi: 10.2196/22898.

[7] Nunes T, Campos D, Matos S, Oliveira JL. BeCAS: biomedical concept recognition services and visualization. Bioinformatics. 2013 Aug 1;29(15):1915-6. doi: 10.1093/bioinformatics/btt317

[8] Pazzani MJ, Billsus D. Content-based recommendation systems. In The adaptive web. Springer, Berlin, Heidelberg. 2007; p. 325-341. doi: 10.1007/978-3-540-72079-9_10

[9] Champiri ZD, Shahamiri SR, Salim SB. A systematic review of scholar context-aware recommender systems. Expert Systems with Applications. 2015; 42(3): 1743-1758. doi: 10.1016/j.eswa.2014.09.017