Document Recommendations in Slovenian Academic Digital Libraries

Mladen Borovič(✉) and Milan Ojsteršek

University of Maribor, Maribor, Slovenia
{mladen.borovic,milan.ojstersek}@um.si

Abstract. With the aim to improve the overall visibility of Slovenian research, one of the main features in the Slovenian open-access infrastructure is providing recommendations of similar documents to researchers, students and other interested parties across all included digital libraries, and other digital archives and journal repositories. In this work we describe the architecture of our hybrid recommender system along with observations of its use in the years after inclusion on real-world data between 2015 and 2018 by investigating which types of recommended documents get recommended the most.

Keywords: Hybrid recommender systems · Real-world recommender systems · Open-access · Digital libraries

1 Introduction

The use of recommender systems in academia has recently been on the rise. Students and researchers use them in digital libraries to find relevant theses, articles, studies, datasets and other documents. As an important feature of digital libraries, quite a few recommender systems have been developed for use in academia. Recommender systems such as presented in [1,3,5] were developed specifically for use in academic digital libraries and repositories to aid researchers in finding relevant publications. Moreover, such recommender systems can also be found in academic social networks like Mendeley [6]. In Slovenia, research regarding recommending documents in the Slovenian language for academic purposes is very scarce. The reason for this was the lack of a structured dataset of documents. This has improved since the introduction of the Slovenian Open-Access Infrastructure [4] which provided a large structured dataset with approximately 200,000 documents[1]. As a part of the infrastructure, a hybrid recommender system has been developed with the aim to improve the visibility of research in Slovenia and encourage researchers from all Slovenian universities to collaborate. This work presents the architecture of our hybrid recommender system included in the Slovenian Open-Access Infrastructure and some observations we made on the digital libraries that are using our recommender system.

[1] Open Science Slovenia Dataset, https://www.openscience.si/OpenData.aspx.

A. Doucet et al. (Eds.): TPDL 2019, LNCS 11799, pp. 361–364, 2019.
https://doi.org/10.1007/978-3-030-30760-8_34

2 Slovenian Open-Access Infrastructure

In 2013, the Slovenian Open-Access Infrastructure was established and has provided researchers, students and the public with access to the publications of Slovenian educational and research institutions. The infrastructure consists of a national web portal, institutional repositories for each of the four Slovenian universities, a repository for research institutions and a repository for colleges and higher education institutions. Metadata from other digital archives are also aggregated within the infrastructure. By type, the infrastructure contains diploma, master's and doctoral theses, journal and conference articles, proceedings, datasets, scientific and technical reports, books, lecture materials and videos of lectures. Because a great majority of publications are in Slovenian, an extensive full-text corpus of Slovenian language in different research domains was created. Currently it represents the largest corpus of texts in Slovenian language [2].

3 Method

We use a cascade approach in our hybrid recommender system (Fig. 1) with content-based filtering acting as a primary recommendation technique and collaborative filtering as its cascading re-ranking method. Documents are represented with titles, keywords, abstracts, typologies and year of publication. We use *tf-idf* weights that are the basis for the calculation of BM25 similarity values for each document pair, forming a document similarity index. New documents are periodically processed as they are included to the system daily. Finally, the user activity data and the calculated similarities between documents are also considered before ranking the documents into a list that is presented to the end-user. The ranking process is where the hybridization occurs, applying content-based filtering and collaborative filtering in cascade.

In our content-based filtering method, we use a collection of metadata, which describes the documents with titles, keywords and abstracts, document typology [8], issue year, authors, repository and the language of the document. Our content-based filtering method uses two scores to return an initial ranking of the documents. A BM25 score is used as a relevance measure between the documents multiplied by a Jaro-Winkler [7] distance score (Eq. 1) acting as a document typology similarity.

$$Score_{CBF} = BM25(d_A, d_B) \cdot d_{jw}(t_{d_A}, t_{d_B}) \tag{1}$$

In our collaborative filtering method, we use the user activity for a document a_d. As actions include views and downloads, the counts of these actions are stored for each document and regularly updated as users use the digital libraries. A feedback value $f(a_d)$ is calculated with the sum of all values of actions on each document. A similar feedback value $f(a_r)$ is calculated with the sum of all values of action on each clicked recommended document. The final score for this method

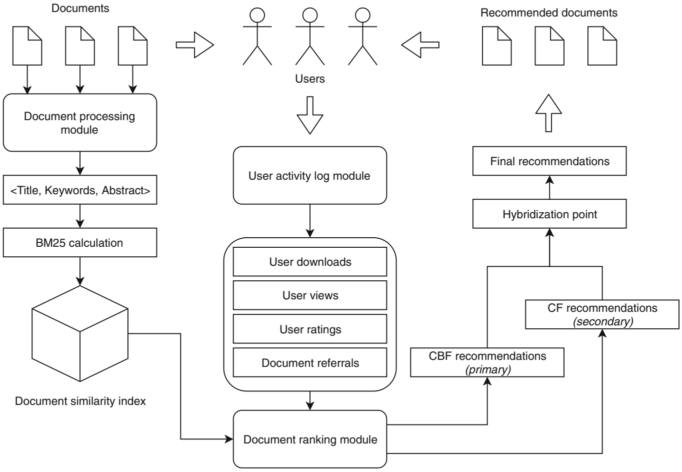

Fig. 1. Architectural diagram of the hybrid recommender system.

is calculated with the sum of feedback values $f(a_d)$ and $f(a_r)$ multiplied by the respective download to view ratios h_d and h_r as shown in Eq. 2.

$$Score_{CF} = f(a_d) \cdot \frac{downloads(d)}{views(d)} + f(a_r) \cdot \frac{downloads(d_r)}{views(d_r)} \tag{2}$$

The hybrid recommender system is implemented in two phases. The content-based method is first used to obtain an initial relevant set of documents which can be recommended. At this stage, an additional exponential temporal decay is applied to increase the ranks of recently published documents. The resulting set of ranked documents is then re-ranked using the feedback values of user actions obtained with our collaborative filtering method.

4 Observations and Conclusions

The goal of the recommender system was to provide recommendations in repositories across the national open-access infrastructure and encourage collaboration of researchers from different Slovenian universities. We investigated the types of documents which get recommended the most. This ties into the logic of the recommender system, which is configured to recommend similar types of documents and it reflects what types of documents are the most popular among our users.

We found that two groups of documents emerged as the most recommended. The first group consists of undergraduate theses, followed by master's theses and doctoral dissertations. The second group consists of scientific articles, review articles, professional articles and other reviews. An increase of recommendations through the years for these two groups can also be observed from Fig. 2. This is due to natural accumulation of new documents in our digital repositories which

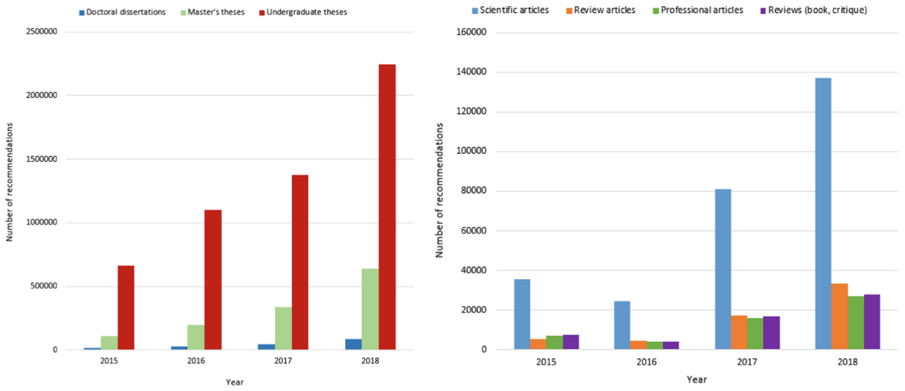

Fig. 2. Recommendations per year (left: group 1; right: group 2).

is on average approximately 13000 per year. We conclude that recommendations in digital libraries have a positive effect on students and researchers looking to broaden their research or acquire different views on the same topic. A unified framework is to be developed in the future in order to perform a more extensive evaluation of our recommender system's contribution to knowledge exchange.

References

1. Beel, J., Aizawa, A., Breitinger, C., Gipp, B.: Mr. DLib: recommendations-as-a-Service (RaaS) for academia. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 1–2, June 2017. https://doi.org/10.1109/JCDL.2017.7991606
2. Erjavec, T., Fišer, D., Ljubešić, N., Logar, N., Ojsteršek, M.: Slovenska znanstvena besedila: prototipni korpus in načrt analiz. In: Proceedings of the Conference on Language Technologies & Digital Humanities, 29th September–October 1st 2016, pp. 58–64. Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia (2016)
3. Knoth, P., et al.: Towards effective research recommender systems for repositories. In: Proceedings of the 12th International Conference on Open Repositories. Brisbane, Australia, June 2017. https://arxiv.org/abs/1705.00578
4. Ojsteršek, M., et al.: Establishing of a Slovenian open access infrastructure: a technical point of view. Program **48**, 394–412 (2014)
5. Porcel, C., Moreno, J., Herrera-Viedma, E.: A multi-disciplinar recommender system to advice research resources in University Digital Libraries. Expert. Syst. Appl. **36**(10), 12520–12528 (2009). https://doi.org/10.1016/j.eswa.2009.04.038. http://www.sciencedirect.com/science/article/pii/S0957417409003698
6. Vargas, S., Hristakeva, M., Jack, K.: Mendeley: recommendations for Researchers. In: RecSys 2016 Proceedings of the 10th ACM Conference on Recommender Systems, Boston, MA, USA, p. 365 (2016)
7. Winkler, W.: String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. In: Proceedings of the Section on Survey Research Methods, pp. 354–359 (1990)
8. Typology of documents/works for bibliography management in COBISS. https://home.izum.si/COBISS/bibliografije/Tipologija_eng.pdf. Accessed 10 Apr 2019