PyTerrier-based Research Data Recommendations for Scientific Articles in the Social Sciences

Narges Tavakolpoursaleh[1], Johann Schaible[1]

[1]GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

Abstract

Research data is of high importance in scientific research, especially when making progress in experimental investigations. However, finding appropriate research data is difficult. One possible way to alleviate the situation is to recommend research data to scholarly search system users based on the research articles they are searching. With LiLAS, the lab organizers provide the opportunity i) to present such recommendations to users of the live system *GESIS Search* and ii) to evaluate the experimental recommender system in this live system with its actual users. As part of our participation in LiLAS, we computed a simple method for recommending research data and evaluated our approach in two rounds each lasting approximately one month. For our approach, we applied the classical TF-IDF method to rank the research data by their relevance to existing publications. We measure our method's usefulness using user feedback, i.e., simple clicks on the recommendations. In both rounds, our experimental system obtained almost the same outcomes as the baseline.

Keywords

recommender systems, information retrieval, online evaluation, research data, living lab

1. Introduction

Evaluating recommender systems and its underlying approaches is a crucial step in assessing the overall quality of recommendations. Typically, such approaches are evaluated *offline* using test collections. Using such collections including relevance assessment, we can specify how well a recommendation fits the users' information need. The CLEF-lab *LiLAS* makes use of the STELLA framework [1, 2, 3], which allows the organizers to provide an environment to evaluate recommendations *online*. This means that recommendations are presented to real users in a live system and the recommendation quality is assessed by how well the users perceive the recommendations. Thus, no test collection or manual relevance assessment by domain experts is needed. Instead, solely the actions by the actual users of the search system "indicate" whether a recommendation fits the information need or not. This pseudo-relevance is estimated by implicit user feedback, such as clicking on well-fitting recommendations.

Recommending research data based on a currently viewed publication aims at alleviating the situation of finding appropriate research data for a specific use. However, to do so, a scholarly search system must contain information on both scientific publications and research data in a given domain. Luckily, the live search system for the broad domain of social sciences *GESIS Search* [4] is integrated in the STELLA framework provided in the LiLAS lab.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

In this paper, we present our recommendation approach for GESIS Search that suggests research data based on a currently viewed scientific publication. We considered the basic content representing metadata of publication and research data as the features. After extracting the features, we compute term frequency and document frequency (TF-IDF) for each term in each dataset for text matching and making recommendations. The implemented approach is provided as a Docker image in the format required by LiLAS for reproducibility. This way, our recommendation approach is able to compute recommendations on-the-fly instead of using pre-computed result lists for only a small portion of scientific publications. The results illustrate that our approach is not outperforming the provided baseline in a statistical significant manner. However, it shows its general usefulness, especially being a simple approach that can be improved further in many ways, e.g., using multi-lingual features to detect more precise similarities.

In the following, we briefly present the scientific search system GESIS Search as well as the data and task provided by LiLAS (cf. Section 2). We depict and describe our approach in detail in Section 3. The results of the evaluation and the discussion of the results are described in Section 4 and Section 5, respectively, before we conclude the paper in Section 6.

2. System, Data and Task

2.1. System and the Data

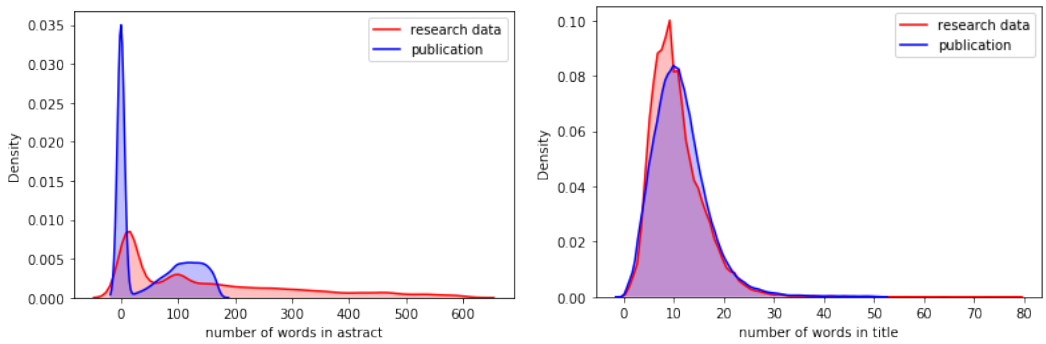
GESIS Search [1] [4] offers an integrated search system for information on the broad topic of social sciences and facilitates finding research data and publication in one portal.

To train a recommendation approach, GESIS provides a corpus of social science publications and research data for the LiLAS participants. The research data set consists of about $78k$ records in the first round and $99k$ records in the second round. The number of publications provided increased from $93k$ records in the first round to 110k in the second round. The records are composed of metadata of the documents in different languages. This metadata consists of a title, an abstract, and topics for both research data and publications and some specific metadata like authors and DOI for publication and temporal and geographical coverages for research data. Examples of such research data and publications can be seen in Figure 2.

2.2. LiLAS Task

The recommendation task is defined as ranking the most relevant research data to the top for a given source publication. In submission type B, the participants should provide a REST-API for sending requests and getting the ranking. The system should be prepared as a Docker container service which LiLAS integrates into the evaluation environment in GESIS Search. LiLAS also provides sample templates that implement minimal REST-based web services and can be developed by the participant. Finally, the participants register their public accessible GitHub repository at the central dashboard service of the LiLAS for the Living lab evaluation. To perform the task, the participants are provided with a list of seed publications (i.e., the

[1] http://search.gesis.org



(a) Abstracts (b) Titles

Figure 1: Density of number of words in abstract and titles

publication IDs), a list of research data IDs, the metadata for both obtained from GESIS Search, as well as a list of research data candidates for $1k$ seed publications.

3. Approach

Krämer et al. [5] classified the relevancy of research data to the research question in the social science domain into three main aspects: relevance of the content, relevance criteria related to data characteristics or factors, and documentation needed to assess relevance. In their study, the participants assess the topical relevance of research data based on how well the research data content fits the research questions. Besides the topical relevance, participants assess the relevance based on other metadata as the type of publication of the primary research connected to a research data, temporal and spatial extent of the data, and characteristics of the sample.

The context-based recommendation of contextual data is rarely covered due to the lack of standard evaluation data and the cold start problem [6]. Given the opportunity to participate in a living lab evaluation infrastructure, we aim to assess our experimental content-based recommendation methods for recommending research data based on publications with actual users of the GESIS portal.

The records (research data and publications) are rarely described with the whole metadata set. But most of the publications have at least a title with an average of 11.5 words. More than 40% of publications have no abstract. On average, the number of words in the abstract amounts to 86.7 terms. We selected a few descriptive metadata from the publications and research data as the entities' features. These metadata of research data and publication include information that can characterize the entities and semantically connect the two types of entities. These include the title, abstract, and the topics for research data and publications, which resemble the set of features used by our approach. The title as a noteworthy minimal description of data expresses the very brief data content and is scanned when looking for data [5]. Unlike the abstract and topics, titles are always available for both data types.

3.1. Experimental System: gesis_rec_pyterrier

We collected three fundamental descriptive metadata elements to identify the resources of both types: title, abstract, and topics. The publication's titles are, in most cases, highly representative and informative and represent the content of the paper. However, the titles of research data are not always informative; for example, "German general Social Survey - ALLBUS 2012". Nevertheless, the abstract information of both resources is limited but appropriate to identify them. Also, the topics or keywords hold compact essential descriptive information about the content of the resources. Figure 3 and 1 depicts the distribution of topics, titles, and abstract lengths (number of words) in both types of documents.

As the first experimental system, we decided to utilize *Pyterrier*, a Python wrapper on top of Terrier for performing information retrieval experiments [7] and compare it with the baseline. We chose this system to establish a comparison between the baseline and this simplistic out-of-the-box approach. Pyterrier provides easy to conduct IR experiments with different weighting models, such as TF-IDF and BM25. Terrier also supports non-English language texts since it represents terms as UTF. It has additional plugins for Bert, EPIC, ColBert, and other methods. However, we did not apply them for the first two experimental rounds. We considered the simple term weighting model of TF-IDF, which scores a document regardless of term position in the text, in order to compare it directly to the baseline using BM25. We collected the title, abstract and the topics of the publications for issuing the queries, and the research data for the indexing.

The research data recommendations are based on the terms in the title, topic, and abstract of the research data as well as of the publications. When providing a publication identifier (seed item of the recommendation), it will be translated into the corresponding publication title, abstract, and topics, which, in turn, are used to query the index of research data with a basic TF-IDF-based algorithm without extra features. This means, during the indexing as well as retrieval process, the text from the title, abstract, and topics of research data and publications is analyzed using the standard tokenizer, stemmer, as well as stop-word-removal provided by Pyterrier. The Terrier weighting model employs Robertson's TF (the term frequency of the term in the document) and standard Sparck Jones' IDF [7]. The experimental system implements an API for indexing and searching, and it is provided as a Docker image with the LiLAS required format for reproducibility[2].

3.2. Experimental Setup

The STELLA infrastructure used for the LiLAS lab contributes the participants' experimental recommender systems in two forms: A) the pre-computed runs and B) the Docker container (Dockerfiles and their source code) [8]. The participants can decide whether they submit type A) or type B). We chose to submit type B), i.e., a Docker container comprising our recommendation approach.

Our experimental ranking is merged with the baseline through the STELLA interleaving mechanism to generate the final result list and to present it to the users (see Figure 2). User feedback in the form of clicks is collected and sent to the central STELLA server. There the

[2]https://github.com/stella-project/gesis_rec_pyterrier

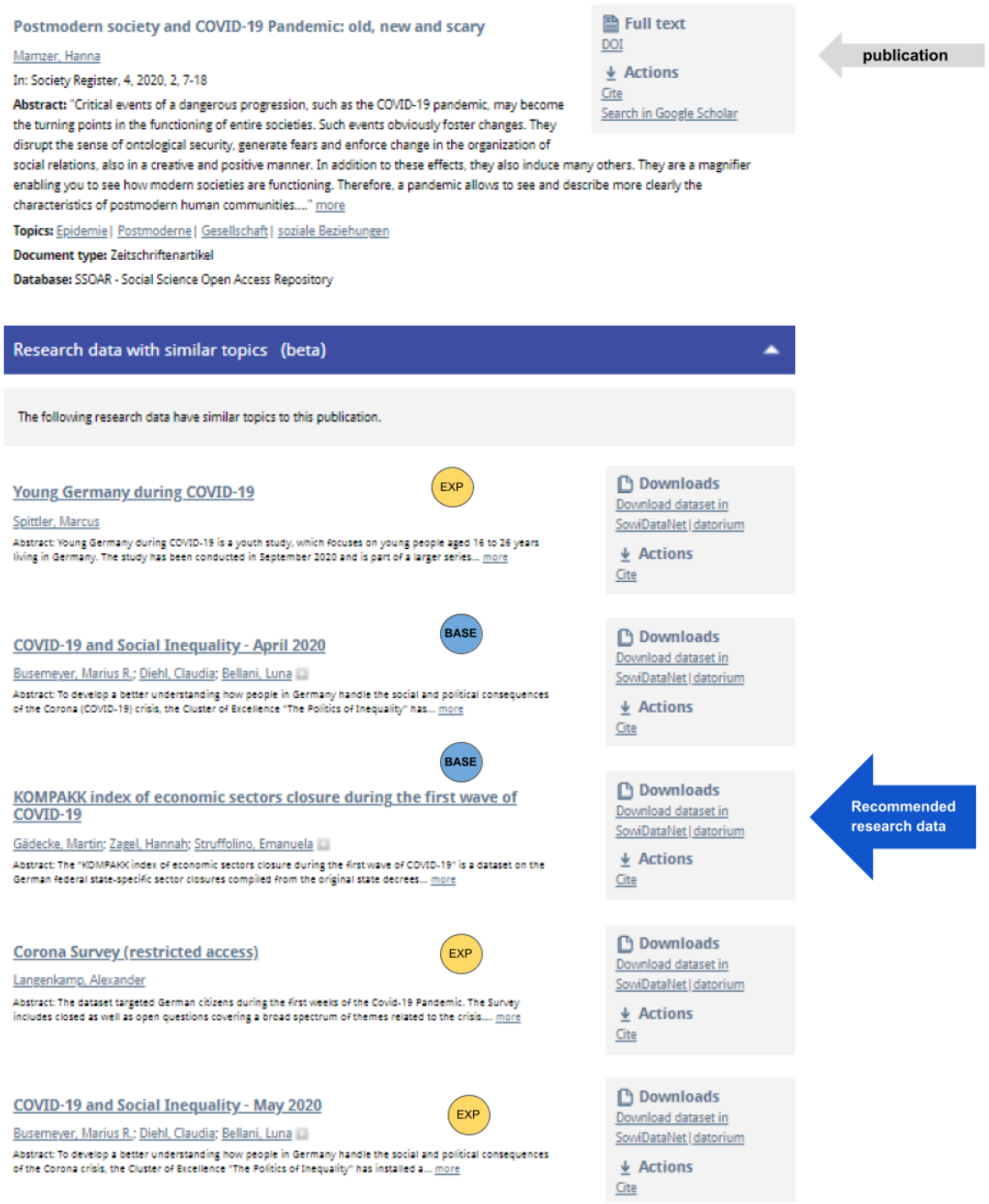
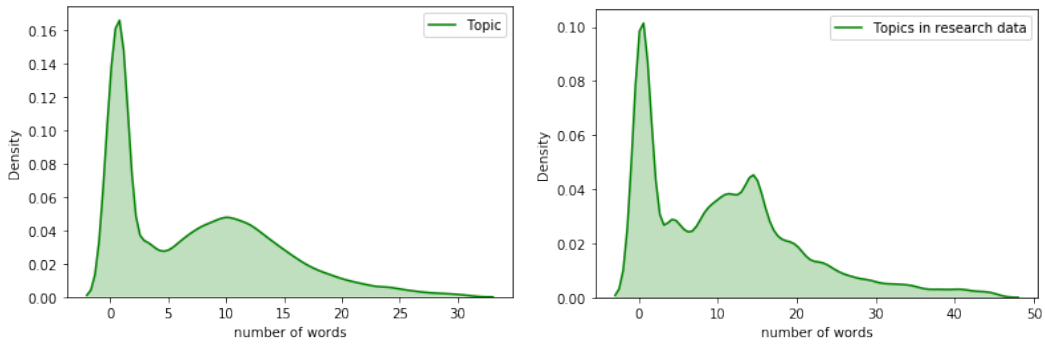


Figure 2: Screenshot of GESIS Search: an example of experimental recommendation ranking, gesis_rec_pyterrier, interleaved with Baseline ranking



(a) publication (b) research data

Figure 3: Density of number of words in topics of documents

evaluation metrics are calculated with some statistics and are displayed as well as reported to the participants (Figure 4).

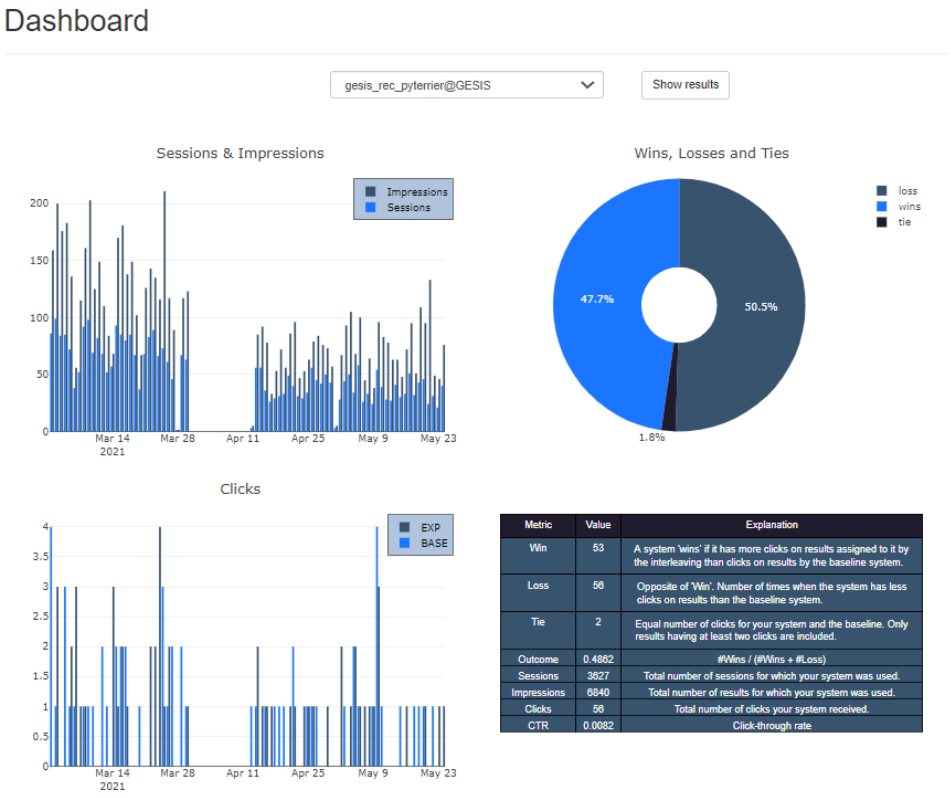
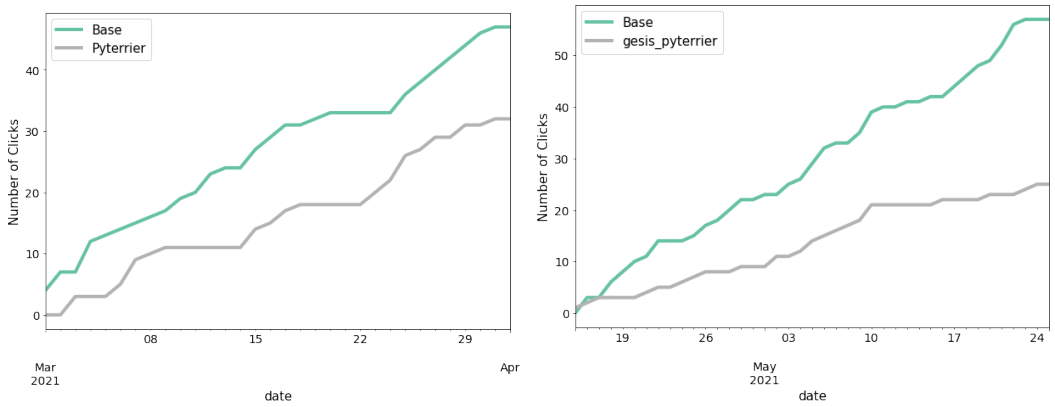


Figure 4: Screenshot of LiLAS dashboard contains the evaluation metrics for the experimental system

4. Result



(a) first Round (b) Second Round

Figure 5: The cumulative number of user clicks on the recommended items during the first round

In the first round in March 2021, the first 100 most frequently viewed publications in the GESIS search were clicked between 4 to 29 times per session. In some sessions, items have been viewed several times. The recommended research data of all systems got about 0.013 CTR with 91 clicks (which includes 82 unique research data) in 6,765 impressions. The CTR for gesis_rec_pyterrier and baseline are respectively 0.0055 and 0.0069. (Figure 5).

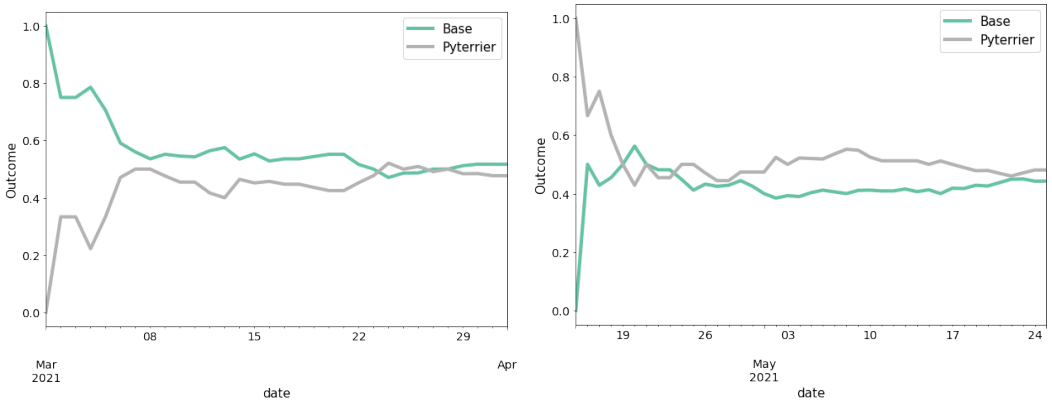
In the first round, the two systems performed almost the same (Figure 6). It is observed that the baseline is showing an not substantially better performance than the experimental system with the outcome of 0.5168.

In the second round from 15 April to 25 May, we proceeded with our single experimental system *gesis_rec_pyterrier*. The experimental setting remained unchanged, and only new records of publication and research data are included in the corpus. In the second round GESIS received 131 user clicks on recommended items of all three systems. The CTR is 0.016, with the whole impressions of 7,753. Our pyterrier system received 25 clicks and got a CTR of 0.0068. In this round, the CTRs for the baseline and new experimental system are 0.007 and 0.011.

Our pyterrier system and the baseline, as for the first round, perform almost the same (see Figure 6).

5. Discussion

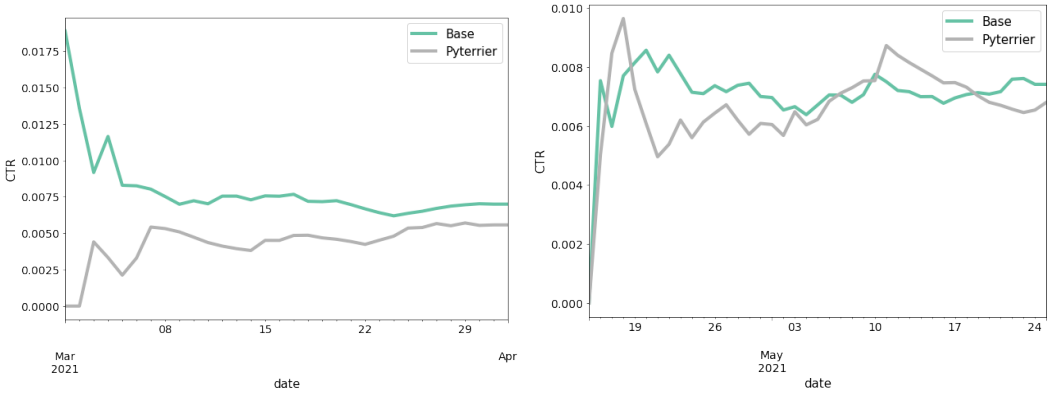
Although the pair of experimental ranking systems are interleaved starting randomly from one system ranking, as shown in Figure 8, the number of clicks on the top-1 ranked item (highest rank) is greatly different from the items in the other places. 52.8% of clicked items have the ranking position one. It shows that the first item in the ranking list has been clicked the most, regardless of the ranker system. Recommendations on the second (13.5%) and third positions (15.7%) have been clicked almost the same. Different studies ([9, 10]) have also specified that



(a) first Round (b) Second Round

Figure 6: Outcomes for Baseline and Pyterrier during the first and the second round

	WIN	LOSS	TIE	OUTCOME
Round#1				
BASE	46	43	1	0.5168
gesis_rec_pyterrier	31	34	0	0.4769
Round#2				
BASE	54	68	4	0.442
gesis_rec_pyterrier	25	27	1	0.48

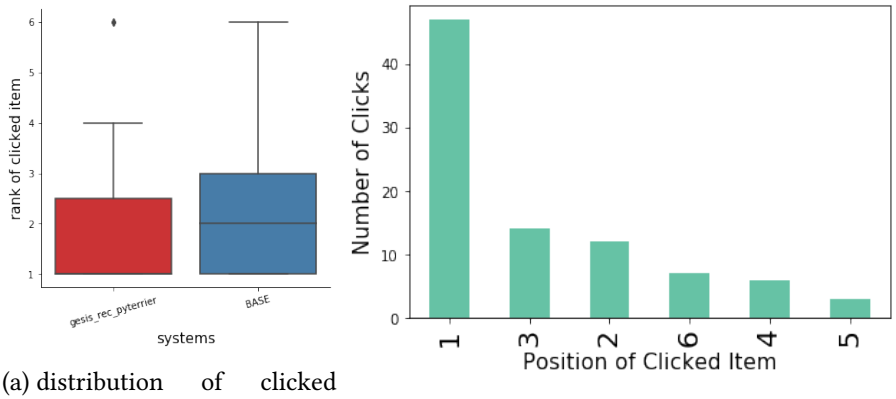


(a) first round (b) second round

Figure 7: Click-Through Rate in the first and second rounds

ranking higher yields in higher CTR. However, in the LiLAS setting, the systems have an equal chance of being represented first in the highest rank regarding the interleaving algorithm.

The recommendation service of GESIS displays just the first three rankings of two systems,



(a) distribution of clicked item's rank in two systems

(b) Rank of clicked item

Figure 8: Position of clicked item in the recommendation page

the baseline constantly and an experimental system interleaved by LiLAS. Due to the low traffic of active users viewing the publication (the number of impressions per day) and the limited number of recommended items, the number of user clicks is deficient. Therefore, the collected clicks might not be entirely satisfactory for two months of evaluation for several test systems (Figure 9).

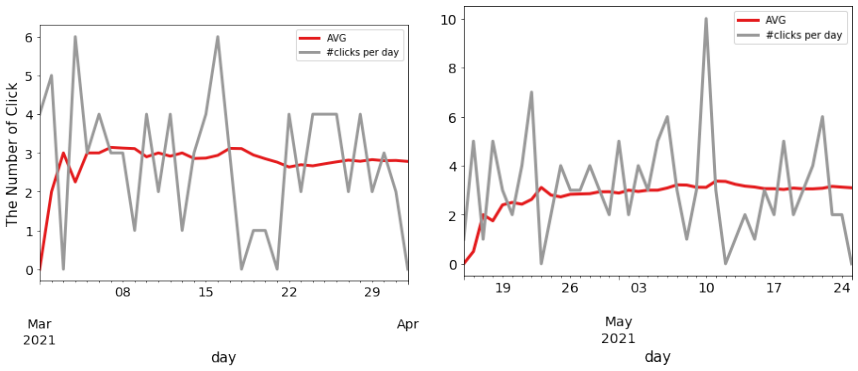


Figure 9: Daily number of clicks on the all recommended item during the first and second Round

6. Conclusion

We succeeded in the first-hand experiment on online evaluation of cross-domain recommendations between publication and research data. As the experimental system, we have implemented a naive content-based recommendation using the metadata available in most documents, i.e., title, abstract, and topics, primarily to compare this out-of-the-box approach to the baseline. We submitted a dockerized system capable of reproducing the ranking with new data. Our recommender system is implemented using the pyterrier library, and we applied the weighting

model based on TF-IDF, which resembles a direct comparison to the baseline's BM25 ranking.

This simplistic approach was not able outperform the baseline. However, we used only the out-of-the-box Pyterrier system without any further configuration or inclusion of other features, such as the multi-lingual support. Future work comprises to extend this simplistic approach step by step. For example, we will focus on utilizing the Bert plugin to integrate our embedding-based recommendation approach [11] into Pyterrier. We will also focus on user needs and decision factors for considering the research data and apply information extraction, translation and semantic representations, and contextual text representation methods for the research data recommendation.

On a general note, the low number of clicks during the first two evaluation rounds (Figure 9) indicate that more "traffic" is needed to better evaluate recommendation approaches in a live setting, as a recommended item is clicked only after the users have searched for a publication, clicked on a publication of interest, and then only clicked on the recommended research data. However, we still believe that LiLAS provides an excellent opportunity for researchers to evaluate their approaches with real users in a live system. It supports both researchers and the portal to develop and evaluate their experimental system to recommend cross-domain data.

References

[1] T. Breuer, P. Schaer, N. Tavakolpoursaleh, J. Schaible, B. Wolff, B. Müller, Stella: Towards a framework for the reproducibility of online search experiments., in: OSIRRC@ SIGIR, 2019, pp. 8–11.

[2] J. Schaible, T. Breuer, N. Tavakolpoursaleh, B. Müller, B. Wolff, P. Schaer, Evaluation infrastructures for academic shared tasks, Datenbank-Spektrum (2020) 1–8.

[3] P. Schaer, T. Breuer, L. J. Castro, B. Wolff, J. Schaible, N. Tavakolpoursaleh, Overview of lilas 2021 - living labs for academic search, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Twelfth International Conference of the CLEF Association (CLEF 2021), volume 12880 of *Lecture Notes in Computer Science*, 2021.

[4] D. Hienert, D. Kern, K. Boland, B. Zapilko, P. Mutschke, A digital library for research data and related information in the social sciences, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), IEEE, 2019, pp. 148–157.

[5] T. Krämer, A. Papenmeier, Z. Carevic, D. Kern, B. Mathiak, Data-seeking behaviour in the social sciences, International Journal on Digital Libraries (2021) 1–21.

[6] Y. Li, J. Nie, Y. Zhang, B. Wang, B. Yan, F. Weng, Contextual recommendation based on text mining, in: Coling 2010: Posters, 2010, pp. 692–700.

[7] C. Macdonald, N. Tonellotto, Declarative experimentation ininformation retrieval using pyterrier, in: Proceedings of ICTIR 2020, 2020.

[8] P. Schaer, J. Schaible, L. J. G. Castro, Overview of lilas 2020–living labs for academic search, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 364–371.

[9] Y. Chen, T. W. Yan, Position-normalized click prediction in search advertising, in: Pro-

ceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2012. URL: https://doi.org/10.1145/2339530.2339654. doi:10.1145/2339530.2339654.

[10] D. Yankov, P. Berkhin, L. Li, Evaluation of explore-exploit policies in multi-result ranking systems, arXiv preprint arXiv:1504.07662 (2015).

[11] N. Tavakolpoursaleh, J. Schaible, S. Dietze, Using word embeddings for recommending datasets based on scientific publications, in: LWDA, 2019.