

TEKMA at CLEF-2021: BM25 based rankings for scientific publication retrieval and data set recommendation

Jüri Keller¹, Leon P. M. Munz¹

¹*Technische Hochschule Köln, Ubierring 48, 50678 Cologne, Germany*

Abstract

In this paper we report the results of our participation in the Living Labs for Academic Search (LiLAS) CLEF Challenge, which is aimed at strengthening the concept of user-centered living labs for the academic search domain. We made one submission for each of the two tasks. For both submissions we focused on data enrichment and Solr's implementation of the probabilistic BM25 ranking function. The proposed systems were evaluated live using the STELLA infrastructure. These live results show that the submitted pre-computed ranking for ad-hoc search (tekma_s) cannot compete with the live baseline system. However, our approach of a pre-computed hybrid recommendation system for research data sets (tekma_n) produced better results than the baseline system.

Keywords

Living Labs, Social Science, Life Science, (Online) Evaluation in IR

1. Introduction

Due to the continuing flood of information and the steadily growing number of scientific publications and research data sets, the ability to find them is an ongoing challenge. In order to find suitable publications in a multilingual scientific database, sophisticated search systems are required that can rank the most relevant results for a search query to the top. In addition, recommendations of suitable research data sets can be equally relevant to completely cover the information need. Since the search for data sets, even using designated search engines, can be tedious, a possible solution may be to recommend relevant research data sets directly to corresponding publications. For this reason, as participants in the Living Labs [1] for Academic Search (LiLAS) CLEF Challenge, we decided on submitting pre-computed rankings for both tasks presented below. An introduction to the LiLAS lab at CLEF can be found in the corresponding overview paper [2].


We participated in both tasks of LiLAS 2021:

- **Task 1 - Ad-hoc retrieval of multilingual scientific documents.** The goal of Task 1 is to support researchers to find the most relevant documents regarding a head query. Participants are asked to create an experimental ranking system for the multi-lingual life science search portal LIVIVO ¹. A good ranking system should present users the most

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



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.livivo.de/>

relevant documents regarding a query on top of the result set. Multiple languages can be used for querying (e.g. English, German, French, etc.); regardless of the language used on the query, the retrieved results can include candidate documents in other languages.

- **Task 2 - Research Data Set Recommendations.** The main task here is to provide a recommendation system for the social science portal GESIS Search². Regarding a seed publication, relevant research data sets should be recommended. For example, the user is interested in the impact of religion on political elections and found a publication regarding that topic, she will be presented with a list of research data sets regarding the same topic.³

Both proposed systems are based on an approach that uses the probabilistic BM25 ranking function [3] to determine the similarity between index and query. Results from the TREC-COVID Challenge⁴ described by Roberts et al. [4] show that almost all top-performing systems used BM25 as first stage ranker to produce already good baselines. As part of a semester project, we have successfully implemented a similar approach on the TREC-COVID data set. While these evaluations have been offline, it is especially interesting to see online performances using live data from real users in the Living Labs for Academic Search (LiLAS) Challenge at CLEF using the STELLA [5] infrastructure.

Furthermore, we decided to modify the approach to function as a recommender system as well. In general, there are three approaches to recommender systems: Content-based recommendations, collaborative recommendations and hybrid approaches [6]. Since no user or profile data was initially available to accomplish the task, we used a type of content-based recommendation. After completing the first round, we could use the obtained click data to rerank the results.

The remainder of this paper is structured as follows. In Section 2 and 3 we outline the submitted systems `tekma_s` for ad-hoc retrieval and `tekma_n` for recommendations. In these Sections, the corresponding corpora, enrichment approaches and experiments are described for each system. The results achieved are summarized in Section 4. In Section 5 this paper ends with a conclusion.

2. Task 1: Ad-hoc retrieval of scientific documents

For the first evaluation round in Task: 1 Ad-hoc Search Ranking, a pre-computed ranking approach was proposed. The system was implemented using Apache Solr⁵ and Pseudo-Relevance Feedback. To evaluate its ranking ability, the head queries and corresponding candidates are used to replicate the baseline system. Based on the given head queries and the full document corpus, multiple runs are pre-computed and evaluated.

2.1. The LIVIVO corpus

Through the lab organizers, two data sets are provided by the cooperating research infrastructure platform LIVIVO for task 1, documents and candidates.

²<https://search.gesis.org>

³<https://clef-lilas.github.io/tasks/>

⁴<https://ir.nist.gov/covidSubmit/index.html>

⁵<https://solr.apache.org/>

With the documents data set, metadata for over 22 million documents from the bio-medical field, from the LIVIVO search portal were provided. The metadata includes, among several others, titles, abstracts, tags from controlled vocabularies like the Medical Subject Headings (MESH)⁶ and Chemical Thesaurus (CHEM)⁷ as well as the language of the document. Even though over three-quarters of the documents are labelled as English, documents from over 30 other languages are provided as well. The metadata is not distributed consistently, leaving some documents even without a title.

The candidate data set contains the head queries from the LIVIVO search portal and the ranked document identifiers. Every head query includes the query string and its frequency. The query strings are multilingual and sometimes include boolean operators. Since the candidates are ordered based on the current LIVIVO ranking system, they can serve as a baseline ranking.

2.2. `tekma_s`

To pre-compute rankings, the full document corpus is indexed as provided using Apache Solr. The documents are processed by a Solr analyzer stack and then queried using Pseudo-Relevance Feedback. To fine-tune the queries, several fields are boosted. In general, only English documents are considered for the ranking.

The same analyzer is used for indexing and querying. This includes the Solr standard tokenizer, the Solr classic filter, a stopwords filter with a corresponding English stopwords list, the Porter stem filter and an English possessive filter.

Queries are generated by searching multiple document fields with various boosting. Besides the TITLE and ABSTRACT fields, the AUTHOR, MESH, CHEM and LANGUAGE fields are considered for querying. Since the query and document analyzers are designed for English documents only and the vast majority of documents are English anyway, all other documents are ignored while querying.

In order to improve the baseline ranking, Pseudo-Relevance Feedback is used to extend the query. Based on the assumption that the best-ranked documents are somehow relevant, information on them is used to rewrite and extend the query [7]. Using the base query, a ranking is generated. The MESH terms are extracted from the ten best-ranked result documents. These MESH terms are ranked by frequency and the most frequent five terms are added to almost all fields in the final query, except the "author" and "language" fields. These fields do not contain standard information and therefore should not be expanded with MESH terms. Thus, to retrieve the final ranking actually two queries are sent to the system. The first one gathers information from the first search results and the second query uses this information in addition to produce the final ranking.

By using the provided head query candidates as a baseline, several query configurations and field boosting are tested. The submitted run `tekma_s` queries the fields TITLE, ABSTRACT, AUTHOR, MESH, CHEM and LANGUAGE. As described in Section 2.2, only English documents are utilized. Therefore, all fields except the LANGUAGE field are optional and are boosted. The fields MESH and CHEM are boosted by 1.5. If they exist, they are considered highly relevant, since they precisely classify the content of the document. The fields TITLE and AUTHOR are boosted

⁶<https://www.nlm.nih.gov/mesh/meshhome.html>

⁷https://images.webofknowledge.com/WOKRS534DR1/help/MEDLINE/hp_chemical_thesaurus.html

Table 1
Research data data sets statistics

	Title	Title_en	Abstract	Abstract_en	Topic	Topic_en
Before preprocessing	99541	6320	94479	4725	83384	5067
After preprocessing	99541	6320	83957	4725	83957	7426

by 1.0 and because some head queries include author names, the `AUTHOR` field is included. The `ABSTRACT` field is included as well, but is boosted down by 0.3 because much more words are in the abstracts and chances are higher that they are irrelevant for the document. The corresponding source code can be found in a public repository.⁸

3. Task 2: Research data set recommendations

Building on the pre-computed ranking approach from round one, a variation of the system was proposed for this task. Variations are made to adapt the system to the recommendation task. Furthermore, different re-ranker are added and the data sets are enriched. These changes are evaluated following the same strategy as described in 2. The smaller document corpus for Task 2: allowed for pre-compute rankings for every single seed document, making this task more suitable for the pre-computed system type. Before indexing, the baseline data sets are enriched by translations and additional topics from the Consortium of European Social Science Data Archives (CESSDA)⁹. By that multiple languages can be used to query the corpus and the topic distribution is more complete. To pre-compute the recommendations, the `tekma_s` system utilized the whole data set and not just the provided candidate lists. Instead of user-generated queries, for this task the queries are generated by the system itself from the seed documents. Retrieved results are re-ranked and then serve as recommendations for the seed document. The corresponding source code can be found in a public repository.¹⁰

3.1. The GESIS corpus

Three data sets are provided by the lab organizers originating GESIS Search for task 2, publications, data sets and candidates.

The publication data set contains metadata for 110420 documents from GESIS-Search, a social science database. The metadata includes 11 attribute fields e.g. title, abstract and authors of the documents. Again, the metadata field were inconsistent. 56% of the publication data contains an abstract and topics are assigned in 67% of the cases. The metadata of the publications are the seed documents given data set recommendations should be made for.

In addition, metadata for 99541 research data sets is provided. This metadata contains 16 fields containing `TITLE`, `ABSTRACT` and `TOPIC` and other fields for English data sets. The distribution of the content is shown in Table 1.

⁸https://github.com/stella-project/tekmas_precom

⁹<https://www.cessda.eu/>

¹⁰https://github.com/stella-project/tekma_n_precom

Like the LIVIVO corpus described in Section 2.1, the GESIS corpus also contains collections of candidates. The top 100 most used seed documents and its data set recommendation are listed here. By that, they can as well be used as a baseline ranking.

3.2. Data enrichment

3.2.1. Field translation

The publication metadata fields for title and abstract are language inconsistent. Using the Python library `langdetect`¹¹, we found that 53% of the titles and 46% of the abstracts are in German, while 40% of the titles and 35% of the abstracts are in English. In the metadata of the research records, in addition to the fields for titles and abstracts, there are also fields for English titles and abstracts. However, different languages are mixed here as well. To solve this problem of multilingualism and to homogenize and extend the publications' metadata, all titles and abstracts of the publications are translated into both languages using Python library `Deep_translator`¹². For this purpose, two additional fields were created: `TITLE_EN` and `ABSTRACT_EN` and filled with the respective translated content. Since 93% of the titles and 81% of the abstracts are in German or English, we narrowed down the translation to these and ignored all other languages. Using this method, we were able to sort the publication metadata linguistically and populate the expanded fields `TITLE_EN` with 110420 and `ABSTRACT_EN` with 62013 entries.

3.2.2. Assigning missing topics

Not all metadata records have topics assigned. The assigned topics are from a controlled vocabulary managed by CESSDA. To assign appropriate topics in a simple way automatically, just existing topics are used for assigning. Therefore, a collection from all topics in the corpus is created and then translated into German or English depending on their source language. Since it should be avoided to overwrite existing information or attributes, two additional attribute fields were added: `TOPIC_EXT_GER` and `TOPIC_EXT_EN`. For these collections, a matching procedure was performed on the title. If one of the topics appeared in the title of a metadata record, it was added to the corresponding attribute field. These approaches should result in more matches being generated between the topics of the data sets. By that method, the German topics are expanded by 556 and the English topics by 2359.

3.3. Indexing

Through separating fields with multiple languages in dedicated fields for each language, language depending analyzers could be used on one index. The same index and query analyzers were applied to the respective field types to achieve as many matches as possible in the search query. The filters and tokenizers correspond to the standard repertoire of Solr. For the German fields, the tokens were separated at the blanks by a whitespace tokenizer. For the English fields, the

¹¹<https://pypi.org/project/langdetect/>

¹²<https://pypi.org/project/deep-translator/>

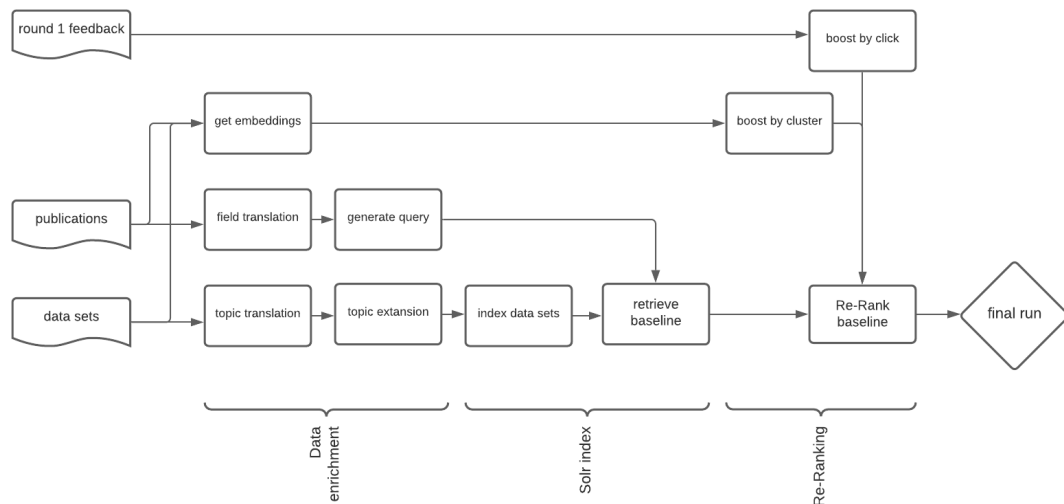


Figure 1: Visualization of the full system used to pre-computing the tekma_n run, from data input on the left to the final output on the right. Curvy boxes represent data inputs, rectangular boxes processing steps.

standard tokenizer of Solr¹³ is used. For the German, as well as for the English field type a lower case filter was used. For the German and English field type, a stopwords filter, with a corresponding stopwords list is applied. The English stop word list is based on Wordnet¹⁴. The Snowball Porter Stemmer algorithm¹⁵ is used to shorten the tokens uniformly to their root words for the German field type under specification of the language "German2". For the English field type, the Porter Stem Filter¹⁶ and the English possessive filter¹⁷ is used.

3.4. tekma_n

To generate recommendations for a publication, the publication is used as query to search the created Solr data set index. As baseline search, Solr default BM25 ranking function and a variety of field combinations and boosting factors are used.

3.4.1. Querying

Since not all fields are given for all seed publications, the queries are generated dynamically considering all available field data and therefore differ in length and complexity. If available, the fields TITLE, ABSTRACT and TOPIC as well as their language variations TITLE_EN, TITLE_DE, ABSTRACT_EN, ABSTRACT_DE, TOPIC_EN and TOPIC_DE and the extended topic fields EXT_TOPIC_DE and EXT_TOPIC_EN are used for the search. Each searched field is boosted individually for a

¹³https://solr.apache.org/guide/8_8/tokenizers.html#standard-tokenizer

¹⁴<https://wordnet.princeton.edu/>

¹⁵<https://snowballstem.org/>

¹⁶https://solr.apache.org/guide/6_6/filter-descriptions.html#FilterDescriptions-PorterStemFilter

¹⁷https://solr.apache.org/guide/6_6/filter-descriptions.html#FilterDescriptions-EnglishPossessiveFilter

run, considering its ability to describe the searched data set. In general, title fields are boosted higher than abstract fields for example.

3.4.2. Re-Ranking

To improve recommendation quality the baseline results are re-ranked in two ways. First, a re-ranker based on the results from round one is applied. On top of these re-ranked results, a second re-ranker is applied considering similarity based on document embeddings.

3.4.3. Re-Ranking by User Feedback

As direct proof of relevance, the click feedback from round one is used to boost certain data sets. Given a ranking from the baseline, system data sets are boosted that were clicked in round one, considering the same query document. Due to click sparsity and importance, a strong, static boost is added.

3.4.4. Embedding based Similarities

Since documents and data sets have broad similarities in structure and nature, the overall document similarity is considered as another factor of relevance. To calculate similarity across documents and data sets, document embeddings and a k-nearest neighbors (k-NN) [8] algorithm is used. The document embeddings are calculated using SPECTER [9] a transformer-based SciBERT language model through its available web API ¹⁸. From the title and abstract of a document, this language model calculates a vector that represents the document. With vectors for all documents, the documents can be mapped in a multidimensional space and the distances between them can be measured. The closer the documents are, the bigger the similarity between them. By means of the k-NN algorithm, using the euclidean distance to measure the distance between the documents, the closest documents to a seed document are calculated. Given a baseline ranking, the most similar data sets are calculated for that query document and all matches gain a strong static boost.

3.5. Experiments

Multiple experiments were made to test different system configurations. By that, optimal field combinations and parameter settings for boostings and re-ranker should be determined. Therefore, two test collections are created from the given head queries and candidates, resembling the baseline system. This data holds no ground truth, but can help to put the results in context. The overall goal is to determine system settings, returning results not too far off from the baseline system, but still providing enough variation for different results. All runs are evaluated using `pytrec_eval`¹⁹.

The supplied head queries and the candidates ranked for that query were used to create two baselines representing the production system. For the first baseline, all candidates for a given

¹⁸<https://github.com/allenai/paper-embedding-public-apis>

¹⁹https://github.com/cvangysel/pytrec_eval

Table 2
Evaluation results for different system settings

Run	re-ranked	map	ndcg	recip_rank	P_5	P_10	R_5	R_10	num_rel_ret
1	False	0.077	0.281	0.441	0.273	0.241	0.014	0.024	43380
1.1	True	0.077	0.280	0.425	0.269	0.239	0.014	0.024	43380
2	False	0.070	0.266	0.432	0.256	0.224	0.013	0.023	41141
2.1	True	0.070	0.266	0.423	0.255	*0.225	0.013	*0.023	41141
3	False	0.082	0.292	0.446	0.278	0.249	0.014	0.025	45156
3.1	True	0.082	0.291	0.427	0.272	0.246	0.014	0.025	45156
4	False	0.074	0.274	0.429	0.263	0.230	0.013	0.023	42482
4.1	True	0.073	0.273	0.415	0.253	0.229	0.013	0.023	42482
5	False	0.083	0.293	0.447	0.277	0.246	0.014	0.025	45330
5.1	True	0.082	0.292	0.426	0.266	0.243	0.014	0.025	45330

head query are marked as relevant. For the second baseline, the relevance scores, provided for every candidate, are used to rank the candidates.

In early experiments the first baseline was used to construct the field composition for the query. Fields were added gradually to improve overall results retrieved and relevant results retrieved. Furthermore the fields TITLE, TITLE_EN and TOPIC from seed publications were used to query the fields TITLE, ABSTRACT, TOPIC and the created field EXT_TOPIC_DE as well as the English fields TITLE_EN, ABSTRACT_EN and the created field EXT_TOPIC_EN.

Evaluating re-ranking and boosting was done using the second baseline. Results are shown in Table 2. Each line contains the results for a single run. Runs marked as .1 and re-ranked "True" contain results for the same run above, but are re-ranked. The first three runs compare the boostings for the topic fields. The boosts 0.5, 0.7 and 0.3 are tested. Surprisingly, boosting the topics down to 0.3, tested in run 3, showed the best results. In the remaining runs 4 and 5 negative boosts are applied to the abstract field. In run 4 abstract fields are boost down to 0.3 and in run 5 slightly less harsh, down to 0.5. Results in general, are close to each other; run 3 and 5 are all almost the same. Even though run 5 performed slightly better in overall metrics like *ndcg*, the *P@5* and *R@5* for run 3 were slightly better. Since just little recommendations can be provided, these metrics were privileged and the configuration of the highlighted run 3.1 was used for the final system `tekma_n`. For the experiments in Table 2 almost all runs without re-ranking performed slightly better. *P@10* and *R@10* from run 2.1 being an exception are marked with an asterisk. For the final run, re-ranking was included in any way to test its performance in a live system and on the full data.

4. Results

tekma_s As the results in Table 3 show, the ranking ability of the proposed pre-computed run is limited by multiple factors. Overall, the system received 124 impressions in 80 days. This is mostly because it could only be utilized for former head queries which therefore were pre-computed. Furthermore, 61 pre-computed rankings from the submitted run have ten or fewer results, so the chance of being clicked is even smaller. Another limiting factor is the

Table 3Final results from system `tekma_s` after 80 days

Metric	Win	Loss	Tie	Session	Impression	Clicks	CTR
Value	12	17	2	104	124	15	0.121

Table 4Final results from system `tekma_n` after two rounds

Metric	Win	Loss	Tie	Session	Impression	Clicks	CTR
Value	26	17	1	1144	2026	28	0.0138

Table 5

Ranking position distribution of data set recommendations clicked from task 2.

Ranking Position	1	2	3	4	5	6
Amount documents clicked	21	8	6	5	2	5

language. By only including English documents, even for German queries, highly relevant documents were ignored. Because of the few resulting data, further investigations do not promise any information gain and therefore haven't been done.

tekma_n Over a course of 28 days, from 12. April to 9. May 2021 the system `tekma_n` received 1980 Impressions. Recommendation rankings for 57 seed documents received clicks, by that an overall click-through rate of 0,0227 was achieved. Two recommendation rankings received more than one click, resulting in a total of 45 data sets clicked. With 28 clicks total, the experimental system `tekma_n` wins 24 times, while the baseline system wins 16 times. The results are summarized as well in Table 4. One recommendation ranking revived equally one click for the experimental and one for the baseline system.

The clicks are distributed unevenly favoring the first ranking positions. While data sets in the first position were clicked 21 times, data sets ranked lower were clicked less often. The full distribution of ranking positions documents were clicked is shown in Table 5. Considering just clicked recommendation lists both systems, the baseline and the experimental were utilized almost equally for the first ranking. The baseline system could rank 11 times first and the experimental system 10 times. Comparing all recommendation ranking, this finding amplifies, resulting in 1021 by 958 in favor of the baseline system.

To analyze the recommendations, the experimental system `tekma_n` performed worse than the comparative system. The originally submitted recommendations are compared with the actual clicked recommendations. The experimental system does not rank 9 clicked data sets at all, but ranked four data sets at the exact same position they were ranked by the baseline system and were clicked.

Investigating clicked data sets and seed publications help to understand the experimental system better, recommending data sets based on similarity. Given the publication with the

German title "Kriminalität im Deutschen Kaiserreich, 1883-1902: eine sozialökologische Analyse", tekma_n recommends the data set "Sozialökologische Analyse der Kriminalität in Deutschland am Ende des 19. Jahrhunderts unter besonderer Berücksichtigung der Jugendkriminalität" and got clicked.

Seed publication:

TITLE: **Kriminalität im Deutschen Kaiserreich, 1883-1902: eine sozialökologische Analyse** [10]

TOPIC: GESIS-Studie

Clicked data set recommendation from tekma_n:

TITLE: **Sozialökologische Analyse der Kriminalität** in Deutschland am Ende des 19. Jahrhunderts unter besonderer Berücksichtigung der Jugendkriminalität [11]

ABSTRACT: "Daten zur **Kriminalität im Kaiserreich**. Die Untersuchungseinheiten sind die Stadt- und Landkreise des **Deutschen** Reiches unter Berücksichtigung von Gebietsänderungen. Für alle erfassten Kreise wurden Kriminalitätsraten in den Kategorien Gesamtkriminalität, gefährliche Körperverletzung, sowie einfacher und schwerer Diebstahl erhoben.

Themen: Entwicklung der **Kriminalität** in den Untersuchungsperioden 1893. (...)"

EXTENDED TOPIC: Kriminalität

This match came together by multiple matching tokens, highlighted in the text above. The matching title tokens already describe the broader topic of both documents. This gets enhanced by the extended topic. The temporal dimension is added through matching abstract tokens.

Creating the tekma_n system, one focus was data enrichment. Described in Section 3.2, titles and abstracts from the publications were translated and the topics were extended. To measure any effect of these approaches resulting in data sets clicked the interleaved recommendations returned from the STELLA are compared with recommendations with data enrichment like the submitted one and without data enrichment. If a data set is ranked lower without data enrichment, this directly impacts being clicked for that query. Surprisingly no applied data enrichment method, neither the translations nor the new assigned, formerly missing, topics resulted in changed positions for the clicked documents. Remembering the small basis of data, data enrichment did not affect the results.

The same methods are used to measure any effects of the applied re-rankers. Results are the same. The clicked documents were not re-ranked.

5. Conclusion

The goal of our participation in the Living Labs for Academic Search (LiLAS) CLEF Challenge was to extend our existing approach from the TREC-COVID Challenge and evaluate how well it performs on various tasks and in a live environment.

Building on the baseline system for the first task, we developed a recommender system for the second task with the same underlying calculation of token similarity and implemented by Solr. We extracted terms from the seed publications to generate queries. We paid special

attention to the translation and expansion of data sets as well as re-rankings based on click data and embedding similarities.

It showed that our data enrichment methods and re-ranking did not affect the position of the clicked documents. Nevertheless, our experimental system with 28 clicks performed better against the baseline system with 16 clicks. The better performance of our system can therefore be attributed to the BM25 function and the set analyzers.

The findings can be used as a guide for future experiments. Thus, the data expansion procedure could be extended, so that significantly more topics are expanded. Multilingual processing and translation could also be extended to other languages. It would also be interesting to see how well the system would perform over a longer runtime.

References

- [1] P. Schaer, Living labs - an ethical challenge for researchers and platform providers, in: M. Zimmer, K. Kinder-Kurlanda (Eds.), *Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts, Digital Formations*, Peter Lang, 2017.
- [2] 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. Goeriot, 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.
- [3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al., Okapi at trec-3, *Nist Special Publication Sp 109* (1995) 109.
- [4] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. M. Voorhees, L. L. Wang, W. R. Hersh, Searching for scientific evidence in a pandemic: An overview of TREC-COVID, *CoRR abs/2104.09632* (2021). URL: <https://arxiv.org/abs/2104.09632>. arXiv:2104.09632.
- [5] 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.
- [6] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions, *IEEE Transactions on Knowledge and Data Engineering* 17 (2005) 734–749. doi:10.1109/TKDE.2005.99.
- [7] I. Ruthven, M. Lalmas, A survey on the use of relevance feedback for information access systems, *Knowl. Eng. Rev.* 18 (2003) 95–145. URL: <https://doi.org/10.1017/S0269888903000638>. doi:10.1017/S0269888903000638.
- [8] E. Fix, J. L. Hodges Jr, *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1, USAF school of Aviation Medicine, 1951.
- [9] A. Cohan, S. Feldman, I. Beltagy, D. Downey, D. Weld, SPECTER: Document-level Representation Learning using Citation-informed Transformers, in: *ACL*, 2020.
- [10] H. Thome, *Kriminalität im deutschen kaiserreich, 1883-1902 : eine sozialökologische analyse*, *Geschichte und Gesellschaft* 28 (2002).

- [11] H. Thome, Sozialökologische analyse der kriminalität in deutschland am ende des 19. jahrhunderts unter besonderer berücksichtigung der jugendkriminalität, GESIS Datenarchiv, Köln. ZA8100 Datenfile Version 1.0.0, <https://doi.org/10.4232/1.8100>, 2006. doi:10.4232/1.8100.