Query term suggestion in academic search

Suzan Verberne[1], Maya Sappelli[1,2], and Wessel Kraaij[2,1]

1. Institute for Computing and Information Sciences, Radboud University Nijmegen
2. TNO, Delft

Abstract. In this paper, we evaluate query term suggestion in the context of academic professional search. Our overall goal is to support scientists in their information seeking tasks. We set up an interactive search system in which terms are extracted from clicked documents and suggested to the user before every query specification step. We evaluated our method with the iSearch collection of academic information seeking behaviour and crowdsourced term relevance judgements. We found that query term suggestion can significantly improve recall in academic search.

Keywords: Professional search, user interaction, query term suggestion

1 Introduction

Academic search is a form of professional search that is carried out by scientists. It has the following characteristics [6]: (1) it is interactive: multiple queries are needed to satisfy one information need; (2) it takes place in a specific domain and (3) it tends to be recall-oriented. In our project, we aim to support academic searchers in their information seeking tasks. In the current paper, we focus on support in the form of query term suggestion. Query formulation is an interactive process in which the user adapts the initial query after inspection of the result list. We focus on query *specification* by adding terms. We propose a set-up for academic information seeking in which the user gets suggestions for additional query terms (both single-word and multiword) after inspection of the result list for his previous query. For our experiments, we use the iSearch data collection [8]. This collection contains elaborate descriptions of real search tasks by academic researchers together with a domain-specific corpus and relevance judgements. Information on actual user interaction is not available. To compensate for this, we use a click model for simulating user interaction with the system. The main contribution of the current paper is that we show the potential value of query term suggestion in academic search.

2 Related work

In previous work on query term suggestion, three different sources for query terms are used: search engine query logs, documents in the retrieval corpus or external knowledge sources. Query logs are especially useful when the queries of other users can be reused by the current user, for example because they occur in similar sessions [5]. For personalization purposes, the user's own previous queries

are sometimes used as a source for query term recommendation, but this data is sparse and topic-dependent [2].

Documents in the corpus are an often-used source for query term suggestion if there are no relevant query logs available. The idea is to extract terms from the documents in the corpus that are most relevant to the user's current query. Relevance can either be defined by the search engine itself, using the top-n highest ranked documents ('pseudo-relevance feedback'), or by the user's clicks, extracting terms from the documents that are clicked by the user ('relevance feedback') [1]. For a topic domain with a controlled vocabulary, terms from a domain-specific thesaurus prove to be a good additional source for query term suggestion [3]. In all cases, the extracted terms are recommended to the user in a short suggestion list.

The current work is — to our knowledge — the first to evaluate query term suggestion for academic search. In large-scale academic search (such as Google Scholar or Microsoft Academic Search), query logs may be helpful for frequent queries, but for the long tail of low-frequency queries we need other sources of query terms. We experiment with clicked documents as source for query terms in a search session. We combine query terms into free text queries in a Language Modelling (LM) framework. The previous work that is most similar to the current work is that of Kim et al. (2011) [6], who evaluate boolean query term suggestion for patent retrieval and medical information search, tasks in which boolean operators are considered very important for reasons of reproducibility and full control. Like Kim et al., we apply user simulation to the evaluation of our approach, but instead of assuming that the user always selects the highest-ranked selected term, we use human judgments for term selection.

3 Methodology

Data collection The iSearch collection of academic information seeking behaviour [8] consists of 65 natural search tasks (topics) from 23 researchers and students from university physics departments. The topic owners were given a search task description form with five fields:

(a) What are you looking for? (information need)
(b) Why are you looking for this? (work task context)
(c) What is your background knowledge of this topic? (knowledge state)
(d) What should an ideal answer contain to solve your problem or task? (ideal answer)
(e) Which central search terms would you use to express your situation and information need? (search terms)

A collection of 18K book records, 144K full text articles and 291K metadata records from the physics field is distributed together with the topics. For each topic, 200 documents were manually assessed on their relevance for the information need using a 4-point scale.

For the purpose of evaluation and the simulation of term selection by independent human users, we set up a Human Intelligence Tasks (HIT) on Amazon Mechanical Turk to judge the relevance of automatically generated terms. The subjects were presented with the information need, work task context, background knowledge and ideal answer of a topic from the iSearch data, and a

list of terms that were automatically extracted from clicked documents in inital system runs. The top-15 highest scoring terms (see the next section for a description of the term extraction and scoring method) were presented in alphabetical order to the participants and we asked them to select the relevant terms. The workers were told that their queries were reviewed by an expert before their task was approved and they would receive their payment in order to prevent spam submissions. Each topic was assessed by two workers in order to be able to calculate inter-rater agreement. 17% of the suggested terms were judged as relevant. The two workers agreed in their judgment for 78% of the terms. Cohen's κ for inter-rater agreement is 0.50, which implicates a moderate agreement. The workers were also asked to indicate their level of familiarity with the topic on a three-level scale. All workers indicated that they were unfamiliar with the topic. This implicates that we cannot claim that we have collected expert judgements, but the fact that they come from human workers with knowledge of the original user's information need and background knowledge, makes them more reliable than automatically estimated judgements.

Retrieval process and query term suggestion We indexed the iSearch collection with the Indri search engine[1]. We used the Indri API to set up a query interface to the combined index of Metadata, Book and Article records. All characters that are not alphanumeric, no hyphen or whitespace[2] are removed from the query terms. Multiple query terms are concatenated and combined using the `combine` function in the Indri query language. For example, the two terms 'ZNO' and 'Transparent conductive oxides' together form the Indri query `#combine(zno transparent conductive oxides)`. As ranking model, we use the Indri LM with Dirichlet smoothing. Per query, we retrieve 100 results from the combined index.

The first user query is the first term from field e 'search terms'. A result list of 100 documents from Indri is presented to the user. In the simple baseline setting without query term suggestion, each follow-up query is the previous query expanded with the next term from the search terms field, until all the terms have been added. For example, the initial query 'zno' is first expanded to 'zno transparent conductive oxides' and then to 'zno transparent conductive oxides magnetron sputtering'[3]

In the experimental setting, the simulated user gets 10 suggestions for query terms to be added to the next query. These terms have automatically been extracted from the documents that the user clicked on in the current search session.[4] All n-grams with $n = 1, 2, 3$ in these documents were considered candidate terms. We scored them with Kullback-Leibler divergence for informativeness and

[1] http://www.lemurproject.org/indri/

[2] Note that a term can consist of multiple words

[3] Although the queries get longer and longer, no empty result sets will occur because in Indri not all search terms necessarily have to be present in the result.

[4] In the case of metadata and book records we took the fields title and description to extract the terms from; in the case of articles in PDF, for which no metadata is available, we extracted the terms from the first 200 words of the document.

Table 1. An example of the queries that the simulated user builds with and without term suggestion. In the first row, all consecutive terms that are added come from field e in the iSearch data; after 4 queries the user is out of terms. In the second row, the first three terms come from the list of query suggestions provided by the system. After these three, the user considers the suggested queries to be not relevant and he adds the remaining terms from field e.

setting	initial query	terms added in consecutive queries, comma separated (Recall for query after the addition of this term)
without term suggestion	zno (0.106)	transparent conductive oxides (0.126), magnetron sputtering (0.136), doping (0.136)
with term suggestion	zno (0.106)	ferromagnetic (0.131), doped (0.162), ferromagnetism (0.162), transparent conductive oxides (0.172), magnetron sputtering (0.177), doping (0.177)

phraseness (KLIP) [9] in order to rank the terms. For the informativeness component of KLIP, KL divergence is calculated between the probability distributions for a term in a foreground collection and in a background collection. In our case, the foreground collection is the collection of documents clicked by the user and as background collection we use the Corpus of Contemporary American English.[5] In a comparison with two other term scoring methods [10], KLIP was found to generate the most descriptive terms for a given document set according to the owner of the document set.

We use a click model to decide whether a user clicks on a document. The click model that we use is the perfect click model from [4][6]: it assumes that a user never clicks on an irrelevant document and the user does not stop inspection of the result list before he has seen all 100 results. When the user has finished inspection of the result list, the system presents a ranked list of terms extracted from the clicked documents. The user selects the highest ranked term that was judged as relevant by at least one human judge in the crowdsourcing task and that has not been added to the query before. If none of the terms is relevant, the user adds the next term from the search terms field to the query. With the additional term, the query is issued again. This process is repeated until the user is out of query terms. An example is shown in Table 1.

4 Results

We evaluate our term suggestion in two ways: First, we evaluate the relevance of the suggested terms using the crowdsourced human judgments. The measures that we use are precision (what proportion of the terms that are suggested during the interactive retrieval process have been judged as relevant) and success rate (the proportion of term suggestion lists from which the user chose a suggested term instead of a term from the search terms field). Second, we compare the effectiveness of the retrieval process when query terms are suggested by the

[5] We also experimented with the iSearch corpus as background collection but this gave significantly poorer results.

[6] We adapted the model from a 3-level relevance scale to a 4-level relevance scale: the probability that the user clicks on a document with relevance 0 is 0, with relevance 1 it is 0.33, with relevance 2 it is 0.67 and with relevance 3 it is 1.0.

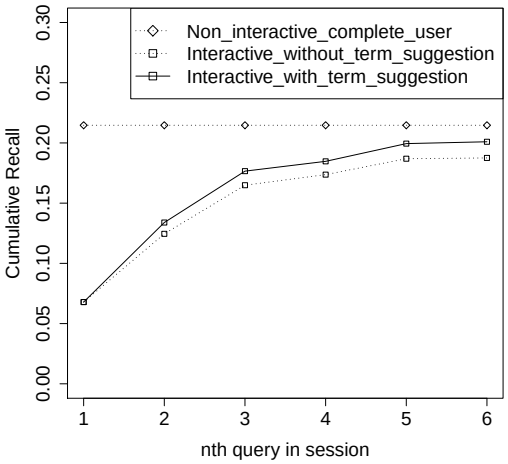

Fig. 1. Cumulative recall over session, averaged over all topics

system to the effectiveness of the retrieval process when the search terms from the iSearch data are used for all query specifications. Since the search terms have been formulated by the owner of the information need, they can be considered a reference for the automatically extracted terms. We measure effectiveness in terms of cumulative recall (over the session).

We obtained a success rate for query term suggestion of 12.8%. The precision of the suggested query terms was 17.0% according to the MTurk workers. These results seem poor at first sight. The only way to judge their real value is to measure the effect of query term suggestion on the retrieval process. This is shown in Figure 1. We show the average results over the course of the session up until the sixth query, because very few topics have more than six queries. As a reference, we compare our results to the results for the *non-interactive complete user*: This user issues one 'complete' query per topic, consisting of all terms from the fields 'information need', 'work task context' and 'search terms' from the iSearch data (average query length: 106 words) [7].

According to a two-sided paired t-test ($n = 66$ topics), the difference between the best-scoring interactive setting and the non-interactive complete user is highly significant with $P < 0.00001$, meaning that the non-interactive complete user achieves significantly better recall than both the interactive users. The difference in the final recall reached with and without term suggestion is significant on the 0.05 level ($P = 0.019$). The results show that the use of query term suggestion can lead to a higher recall compared to using the search terms that were formulated by the topic owner without the help from query term suggestion. This is surprising given the finding that only in 12.8% of the query specifications, the user picked a term from the suggestion list.

5 Conclusion and future work

We showed that query term suggestion with terms extracted from clicked documents can have a positive effect on the effectiveness of interactive academic

search. Our next steps include improvement of the term extraction algorithm, the application of other (non-perfect) click models, comparing our term suggestion method to pseudo-relevance feedback (PRF), and the evaluation of our query term suggestion method on real users.

Acknowledgements

This publication was supported by the Dutch national program COMMIT (project P7 SWELL).

References

1. Belkin, N.J., Cool, C., Kelly, D., Lin, S.J., Park, S., Perez-Carballo, J., Sikora, C.: Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. Information Processing & Management **37**(3) (2001) 403–434
2. Feild, H., Allan, J.: Task-aware query recommendation. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. SIGIR '13, New York, NY, USA, ACM (2013) 83–92
3. Hienert, D., Schaer, P., Schaible, J., Mayr, P.: A novel combined term suggestion service for domain-specific digital libraries. In: Research and Advanced Technology for Digital Libraries. Springer (2011) 192–203
4. Hofmann, K., Schuth, A., Whiteson, S., de Rijke, M.: Reusing historical interaction data for faster online learning to rank for ir. In: Proceedings of the sixth ACM international conference on Web search and data mining. WSDM '13, New York, NY, USA, ACM (2013) 183–192
5. Huang, C.K., Chien, L.F., Oyang, Y.J.: Relevant term suggestion in interactive web search based on contextual information in query session logs. Journal of the American Society for Information Science and Technology **54**(7) (2003) 638–649
6. Kim, Y., Seo, J., Croft, W.B.: Automatic boolean query suggestion for professional search. In: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. SIGIR '11, New York, NY, USA, ACM (2011) 825–834
7. Lioma, C., Larsen, B., Ingwersen, P.: Preliminary experiments using subjective logic for the polyrepresentation of information needs. In: Proceedings of the 4th Information Interaction in Context Symposium, ACM (2012) 174–183
8. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K., eds.: Advances in Information Retrieval. Volume 5993 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2010) 627–630
9. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. In: Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18, Association for Computational Linguistics (2003) 33–40
10. Verberne, S., Sappelli, M., Kraaij, W.: Term extraction for user profiling: evaluation by the user. In: Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization, CEUR (2013) 51–57