Interactive Query Expansion in Meta Search Engines

Wolfgang Köhler and **Daniel Backhausen** and **Claus-Peter Klas** and **Matthias Hemmje**
Distance University in Hagen, Germany
DE-58084, Hagen, Germany
Wolfgang.Koehler1, Daniel.Backhausen, Claus-Peter.Klas, Matthias.Hemmje@FernUni-Hagen.de

Abstract

For meta search systems like digital library solutions, techniques like recommendation and especially query expansion are complex to realize because often the content of the information objects is not present or directly accessible. This approach takes new roads by integrating suggestion terms from two distinct sources in an interactive hybrid recommendation system. The terms are acquired through lexical-syntactical analysis using WordNet, as well as through association rule mining among the query logs.

1 Introduction

Recent research in query expansion techniques for information retrieval systems has seldom taken the particular situation of meta search engines into account. There are several reasons for this: Meta search engines usually do not provide the content of information objects that is often necessary for expanding user queries. The well-known recommender systems based on similarity measures do not work here. Because of the lack of data, the cold start problem is even more aggravated. An additional challenge is the long time of processing a query, since the engine needs to wait on other search engines. Therefore, it is advisable to involve the user more in the process of query formulation rather than relying on blind feedback techniques. Our related works section illustrates the problems of recommendation in meta search engines and shows the approaches that we developed further to tackle those issues. The implementation section explains the concrete steps we try to take in implementing a hybrid interactive recommendation system based on lexical-syntactical analysis and association rule mining. We tested our system in EzDL, a digital library meta search engine. After discussing the user evaluation of the system, we conclude by lining out the possibilities for future research.

2 Related Work

There are several techniques to find suitable terms for query expansion. One of them is the so called automatic query expansion. Well-known approaches have been developed by Mita et. al in [Mitra *et al.*, 1998], Xu et. al. [Xu and Croft, 1996], and Qiu and Frei in [Qiu and Frei, 1993]. They use either a refined form of blind feedback, local context analysis based on a concept database, or a similarity thesaurus to increase the effectiveness of this procedure significantly.

However, none of the presented approaches is useful for query expansion in meta search engines. They all require some kind of information object content, which is typically not present in meta search engines. Instead, other additional content needs to be provided, for example a lexicon or a word net. Yet another way would be the use of user interaction data, which is stored in some kind of activity log.

In addition to the first mentioned way to expand queries, different attempts have been made to automatically expand queries using WordNet [Fellbaum, 1998] by exploiting lexical-semantic relations [Voorhees, 1994]. Even though these experiments did not show a significant improvement in query performance by just linking WordNet to an Information Retrieval system, this effect can indeed by reached with a more refined approach. [Kim *et al.*, 2004] demonstrated that performance can be enhanced by disambiguating the query terms first before expanding them. The authors suggest to disambiguate query terms by determining their root sense according to their context. Obviously, in a query for a meta search engine, there is not much context to draw from, so the usefulness of this approach would be limited in our scenario.

A more promising approach has been made by Liu et al. All synonyms that have a similar meaning are saved in a synset in WordNet. The correct meaning of a given term can be found by determining the most appropriate synset. According to [Liu *et al.*, 2004], this can be done effectively by using the information in the synset definition. For two words that are part of a nominal phrase, a check is made whether their synsets contain any information that helps to determine the correct meaning in this context. The synset definition might provide terms useful for query expansion. The approach has lead to a precision improvement of 15.6 to 21.5 % on the TREC 9, 10 and 12 datasets.

Recently, some more effort has been made to analyze query logs in order to identify good query expansion terms. As proposed by [Cui *et al.*, 2002], this can be even more effective than local context analysis. Here the authors use query logs as a basis for query expansion. However, the disadvantage for meta search systems regarding the need of content remains. One idea to address this issue is the approach proposed by Fonseca et al. in [Fonseca *et al.*, 2005]. Here the query logs are mined for association rules by inspecting which queries frequently co-occur in a user session. The brilliant idea coming from this approach is to equate itemsets and transaction sets known from association rule mining with sets of queries and sets of user sessions. For a given query, a relation graph is built, starting from the user query and showing the transitive associations between the queries. Circuits in the graph are called concepts. The concepts are candidates for query expansion. Compared to other approaches presented in this paper, this

one involves the interaction of the user. That means that a suitable concept needs to be explicity chosen by the user among the given options. Furthermore, the user can also specify the kind of relation between query and chosen concept, leading to a different Boolean connection between query and concept. Synonyms and specializations are connected to the query via the `OR` operator. Generalizations and associations are connected via the `AND` operator. According to Fonseca et al. this approach leads to an increase in precision of 53 % on average when tested with a web search engine.

Within our research work we implemented the approach presented by Fonseca et al. in [Fonseca *et al.*, 2005] and evaluated its usefulness for meta search engines. To use this approach in the context of meta search systems, different modifications have to be made, which we will present in this paper.

Before elaborating on this, however, we need to discuss how an effective interactive recommender system for query expansion should look like. Harman shows through experiments in the Cranfield 1400 test collection in [Harman, 1988] that the effectiveness of the system significantly increases, if it draws on two distinct sources for the expansion terms. This is what we want to call the two-window approach. A third source did not bring as much improvement, possibly because the terms presented are already included in the first two sources.

3 Implementation

This leads us to our concept of building a hybrid recommendation system that is based on 1) a lexical-semantic analysis and 2) query log analysis using a mixture of the approaches proposed by [Liu *et al.*, 2004] and [Fonseca *et al.*, 2005]. In presenting the suggested query terms, we want to follow the two-window approach by [Harman, 1988].

For our research prototype we used EzDL[1], a meta search system for digital libraries. EzDL has already proven its usefulness in different research activities such as the implementation and evaluations of interactive information retrieval scenarios [Klas *et al.*, 2004; Klas and Hemmje, 2009] at the University of Duisburg-Essen and the DistanceUniversity of Hagen. As described by Beckers et al. in [Beckers *et al.*, 2012], EzDL is a service-oriented system that can be used as a meta-search system for heterogeneous sources or digital libraries. In addition it provides an evaluation framework with already existing tools and rich user logs. EzDL consists of a backend agent based system and a rich user client, giving access to the services.

To address the cold start problem present in collaborative recommender systems, we initially use WordNet. It is used here as a database for recommendations based on lexical-semantical relationships. First, after performing stemming on the query term, the term is looked up in WordNet for all parts of speeches. Potential neighboring words in the query are used to disambiguate the synset following the approach of [Liu *et al.*, 2004]. The most appropriate synset is given the highest rank in the list. Then, all synsets related to the best synset are added to the list of suggestions as related synsets. The suggestions are displayed in a drop down box which opens when the user clicks on the term within the query (see figure 1). This disambiguation helps to avoid information overload.

[1] http://www.ezdl.de

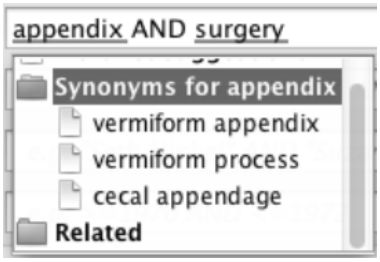

Figure 1: Popup box with search suggestions for the word *appendix*.

Since a meta search system often integrates different heterogeneous data sources, information search takes longer than in local retrieval solutions. In this case, users are encouraged to carefully formulate their query. For this purpose, EzDL offers the possibility to specify query terms more precisely by using various more meaningful input fields like title, author, and year of publication. We argue that when using meta search systems, queries are more sophisticated than in other systems like web search solutions where users usually submit very short and ambiguous queries. This means that the probability that two distinct users will enter exactly the same query is not as high as in many web retrieval systems.

As a consequence to this, expanding the query as it is suggested by [Fonseca *et al.*, 2005] is not sufficient. Rather, the particular query terms should be used as the basis for an association rule mining among the query logs. Association rule mining is a computationally expensive process. Therefore, the computation is performed during the start of the client application of EzDL. The results are stored in a database so that they are immediately available upon the next start of the client. In this way, the client is instantly supplied with working data.

The ChARM algorithm introduced by [Zaki and Hsiao, 2005] provides a very efficient method for computing association rules. The big advantage of the algorithm is that it only acquires the closed frequent itemsets, which avoids a lot of redundancy found in other association rule mining algorithms. The algorithm does so by taking a "round trip" over the sets of items and transactions through a Galois connection. Considering also the set of transactions avoids having to solve an NP-complete problem, namely, finding all frequent itemsets. The ChARM algorithm, instead, only finds the closed frequent itemsets. In the next step, ChARM mines non-redundant association rules by utilizing the concept of minimal generators which is applied to the closed frequent itemsets, as explained in [Zaki, 2004]. Reducing redundancy is key in making this algorithm so efficient. In comparison to Apriori, ChARM reduces the generation of redundant rules up to a factor of 66.

The ChARM algorithm can be configured by setting the values of minimum support and minimum confidence. For testing purposes, we used a minimum support of 2 and a minimum confidence of 30%.

Using the association rules stored in the database, a query relation graph is built considering each term in the query. For each term, binary association rules containing the term are considered for expanding the tree. Binary association rules are rules that identify a mapping between exactly two terms. Each term is represented by a node, but no term is represented by more than one node. This way of building the graph reflects transitive relationships between

the query terms. The transitive relationships can be identified by finding all elementary circuits in the graph. After all of those are found, it could be decided whether the graph is Hamiltonian, i.e. if the graph consists of one circuit containing all the nodes of the graph. This question in itself is not of our concern, but it is interesting to note that the question whether a graph is Hamiltonian is an NP-complete problem.

The procedure of building a query graph is induced every time the query is changed, and thus it needs to be executed very fast. Clearly, messing with NP-complete problems would not be something we would like to deal with on a regular basis. The algorithm by Tarjan presented in [Tarjan, 1973] finds all elementary circuits with a complexity of $O((|V| \cdot |E|)(|C| + 1))$ for $|V|$ nodes, $|E|$ edges, and $|C|$ circuits. Thus, by using this algorithm we can reach polynomial complexity as long as we do not have to deal with a huge number of circuits.

If all the nodes of a given circuit are already contained in another circuit, the first circuit is redundant, and it will be removed from the set of circuits. The remaining circuits are presented to the user as concepts for query expansion. In the presentation, words of the concept that are also part of the current user query are omitted. The user can choose the kind of relationship between concept and current query, which determines how the concept is linked to the current Boolean query (see figure 2).

4 Evaluation

The modified client was evaluated with six undergradute and graduate students, but none were experts in computer science or literature. The group included representatives of both genders. The students were asked to search for information objects about deadlocks. They were instructed to find relevant sources for writing a research paper about this topic. Before they started, they were briefly introduced about the concept of deadlocks in computer science. While searching, they verbalized their thoughts. In addition, the screen was recorded. In the beginning, the students had trouble identifying the recommendation tool at all, due to misplacement on the screen or because they did not realize the functionality of the tool. The first three had to be encouraged to take a look. As a response to this, the user interface was changed to highlight the query expansion tool.

The result regarding the suggestions given by WordNet were only noticed by some users, and they were quickly dismissed as not relevant to the query. The suggestions created from the query logs were treated differently. While the query expansion tool integrated in EzDL (figure 2) was often treated with initial scepticism, the suggestions proposed lead at the end to relevant search results in most of the cases. Since the users were not familiar with the concepts of deadlocks before the evaluation, the suggestions helped them to see which other concepts might be related to the topic, and which are the key authors on this topic. Even if the suggestions were not used via the tool, the students read and reused the suggested terms in new queries.

As a side finding, the Boolean expressions of the suggestions were mostly not understood. In fact, the users expected queries to be linked exactly the other way around than how Fonseca et al. did it in their recommendation system, i.e., they expected generalizations to be linked by the `OR` operator. Another evaluation on a larger scale needs to show if the sample in this case was too small, if the evaluation by Fonseca et al. was somehow faulty, or if the situation of this case affects the evaluation to turn out differently.

5 Discussion and Next Steps

In this paper we proposed, implemented and evaluated a two step recommendation system for query expansion in meta search engines. The system adapts to all users, as the query base increases.

The next steps will be manyfold. First, further evaluative research will show whether the linking of Boolean expression needs to be done in a different way, and whether the values that we have used for minimum support and minimum confidence are appropriate. Secondly, the integration of other services like Wikipedia for a better disambiguation and suggestion of query terms will be tried. Thirdly, we will investigate, from the human-computer interaction point of view, how to better highlight the suggestions components and how to make them more recognizable without disturbing the work flow of the user. And fourthly, we will make the system task aware, as described in [Backhausen, 2012], in order to learn task based and not with respect to all user logs. This way, the suggestions should be more focused. Finally, we will investigate query formulation which can be assisted by building suggestions using the meta information of objects that are marked as relevant by other users or stored in their personal library, as described in [Landwich *et al.*, 2009].

References

[Backhausen, 2012] Daniel T. J. Backhausen. Adaptive ir for exploratory search support. In *SIGIR: Doctoral Consortium*, page 992, 2012.

[Beckers *et al.*, 2012] Thomas Beckers, Sebastian Dungs, Norbert Fuhr, Matthias Jordan, and Sascha Kriewel. ezdl: An interactive search and evaluation system. In *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*, pages 9–16, Dunedin, New Zealand, August 2012. Department of Computer Science, University of Otago.

[Cui *et al.*, 2002] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, WWW '02, pages 325–332. ACM, 2002.

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet*. Language, speech and communication. MIT Press, Cambridge, MA, 2 edition, 1998.

[Fonseca *et al.*, 2005] Bruno M. Fonseca, Paulo Golgher, Bruno Pôssas, Berthier Ribeiro-Neto, and Nivio Ziviani. Concept-based interactive query expansion. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 696–703. ACM, 2005.

[Harman, 1988] D. Harman. Towards interactive query expansion. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '88, pages 321–331. ACM, 1988.

[Kim *et al.*, 2004] Sang-Bum Kim, Hee-Cheol Seo, and Hae-Chang Rim. Information retrieval using word senses: root sense tagging approach. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 258–265. ACM, 2004.

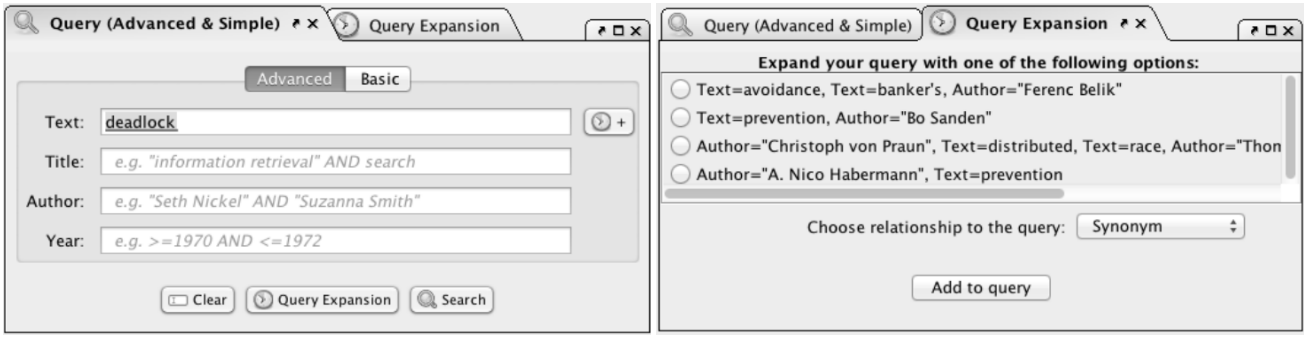

Figure 2: Concepts for query expansion suggested for the query "deadlock".

[Klas and Hemmje, 2009] Claus-Peter Klas and Matthias Hemmje. Catching the user - user context through live logging in daffodil. In *SIGIR 2009 Workshop: Understanding the User, Boston, USA*, 2009.

[Klas *et al.*, 2004] Claus-Peter Klas, Norbert Fuhr, and André Schaefer. Evaluating strategic support for information access in the DAFFODIL system. In *Research and Advanced Technology for Digital Libraries. Proc. European Conference on Digital Libraries (ECDL 2004)*, 2004.

[Landwich *et al.*, 2009] Paul Landwich, Tobias Vogel, Claus-Peter Klas, and Matthias Hemmje. Model to support patent retrieval in the context of innovation-processes by means of dialogue and information visualisation. *Electronic Journal of Knowledge Management*, 7:87–98, 1 2009.

[Liu *et al.*, 2004] Shuang Liu, Fang Liu, Clement Yu, and Weiyi Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 266–272. ACM, 2004.

[Mitra *et al.*, 1998] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 206–214. ACM, 1998.

[Qiu and Frei, 1993] Yonggang Qiu and Hans-Peter Frei. Concept based query expansion. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, pages 160–169. ACM, 1993.

[Tarjan, 1973] Robert Tarjan. Enumeration of the elementary circuits of a directed graph. *SIAM Journal on Computing*, 2(3):211–216, 9 1973.

[Voorhees, 1994] Ellen M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 61–69. Springer-Verlag New York, Inc., 1994.

[Xu and Croft, 1996] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11. ACM, 1996.

[Zaki and Hsiao, 2005] M.J. Zaki and C.-J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *Knowledge and Data Engineering, IEEE Transactions on*, 17(4):462–478, April 2005.

[Zaki, 2004] Mohammed Zaki. Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248, 2004.