DBLP-Filter: Effectively Search on the DBLP Bibliography

Jiang Du, Peiquan Jin*, Lizhou Zheng, Shouhong Wan, Lihua Yue
University of Science and Technology of China, Hefei, China, 230027
jpq@ustc.edu.cn

ABSTRACT

DBLP is a well-known online computer science bibliography. As nearly all important journals and conferences on computer science are tracked in DBLP, how to effectively search DBLP records has become a valuable topic for the computer science community. In this paper we present DBLP-Filter, a new DBLP search tool. The major features of DBLP-Filter are: (1) it provides new search options on concepts and literature importance; (2) it can maintain user profiles and can support user-area-aware search; (3) it provides the service of new literatures alert. Compared with the existing DBLP search tools, DBLP-Filter is more functional and also shows better effectiveness in terms of MAP and F-measure when tested under a set of randomly-selected queries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval.

Keywords: DBLP; Literature search; Concept search; Literature importance; User-aware search

1. INTRODUCTION

DBLP has been one of the most important sources for researchers in computer science to find interested literatures, as almost all the major journals and conferences in computer science are tracked in DBLP.

In order to provide effective search service on DBLP bibliography, some researchers have developed several tools for searching DBLP records, which are also lined on the DBLP website (http://www.informatik.uni-trier.de/~ley/db/). Those tools include the Faceted search tool, the Free search tool, and the CompleteSearch search tool [1]. Although those tools provide a lot of search options on DBLP records, there are still some problems existed. For example, they cannot support search on synonyms, e.g., when searching on "*Microblog*", papers about "*Twitter*" will not be returned. In general, the major problems in current DBLP search can be summarized as follows:

(1) Concept-based search is not provided, i.e., when you search "*Micorblog*", literatures about "*Twitter*" but without an explicit word "*Microblog*" will not be returned. This will lower the recall of the search results.

(2) Literature importance is not considered. As there are thousands of journals and conferences in the computer science area, researchers and especially fresh students need to find those much important literatures in their research field. The ranking of journals and conferences is also concerned by many organizations and programs, such as the China Computer Federation (CCF) and the Excellent in Research for Australia (ERA).

Copyright is held by the author/owner(s).
WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577329

(3) User preference as well as new literatures alert is not supported. Researchers have to track the state-of-the-art literatures, which is generally performed manually so far. However, this can be improved if we can develop new literatures alert service based on the DBLP records.

Some keyword-based search systems on relational databases may also be used for DBLP search, such as SPARK [2] and PerK [3]. However, they were focused on introducing textual relevance into relational database, and did not consider the special needs in literatures search.

In this paper, we are aiming at developing a new search tool for DBLP records called DBLP-Filter. The new tool has the following new features:

(1) It provides new search options on concepts and literature importance.

(2) It can maintain user profiles and can support user-area-aware search.

(3) It provides the service of new literatures alert.

The DBLP-Filtering search tool has been implemented, and we compared DBLP-Filter with other three existing search tools. The evaluation results including F-measure and MAP (Mean Average Precision) on a set of randomly-selected queries show that DBLP-Filter outperforms its competitors, and provides more functions than previous tools.

2. DESIGN OF DBLP-FILTER

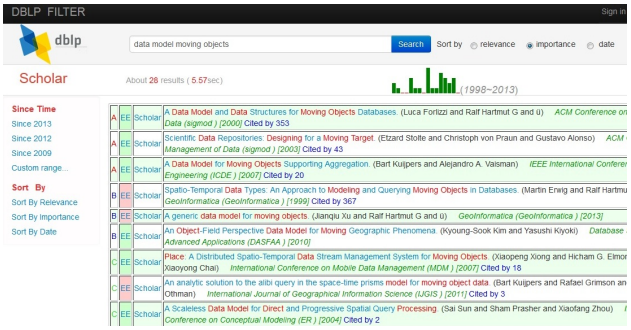

Figure 1. The snapshot of DBLP-Filter

Figure 1 shows the snapshot of DBLP-Filter. The implementation of DBLP-Filter is similar with other search tools, namely first parsing the DBLP records into a relational database, and then processing user queries in a Web-based interface as well as returning the ranked results. The key features of DBLP-Filter lie in the information enhancement, ranking, and the user-aware services.

Concept and Citation Enhancement. The basic information in a DBLP search tool comes from the DBLP raw data. However, those raw DBLP records are not sufficient for literature search. For example, there is no citation information for each literature in the original DBLP data. Therefore, in DBLP-Filter, we first add the citation information for each crawled literature. The citation information is crawled from the ACM Digital Library and is stored in a table in the database. We

also construct an ontology from WordNet [4] to reflect the association among different words. For instance, "Twitter" is a type of "Microblog". Therefore, when we process a query including "Microblog", the system will automatically match the literatures about "Twitter".

Ranking on Literature Importance. Literatures with high citation or published in prestigious journals or conferences are usually more influential in one's research filed. Therefore, in DBLP-Filter we introduce the importance-based search option (the "*importance*" search option shown in Fig.1), which aims to return important literatures and thus enable researchers focusing on those most important papers in their research topics. Basically, we measure the literature importance according to three factors: the ranking of published journal/conference, citation number. Currently, the ranking of journal/conference is manually maintained in the database. The current version of journal/conference ranking comes from the China Computer Federation (CCF), which can be found in http://www.ccf.org.cn/sites/ccf/paiming.jsp (in Chinese). Other rankings can also be easily integrated into our system. The citation number is crawled from the ACM Digital Library.

When using the default relevance-based ranking method (as shown in Fig.1), we consider a lot of factors including textual relevance, concept relevance, published date, literature importance, and user relevance (will be discussed in the next paragraph).

User-Aware Search Services. Users concerning the DBLP records are with different research fields. However, one keyword will frequently appear in the literatures within different fields. For example, the keyword "*moving objects*" not only frequently appears in database area (particularly in moving objects databases), but also very common in the research on video-based object tracking. In order to solve this problem, we allow user registration in DBLP-Filter. A new user will be asked to select his/her research area or input research topics if necessary. The user's profile will be automatically matched with the DBLP records during the query processing and ranking process.

Another type of user-aware search services in DBLP-Filter is new literatures alert. When a new user is to register in DBLP-Filter, he/she can choose whether to receive new literatures alerting. If yes, our system will send new literatures bibliography to the user via Email in case that new related literatures are indexed by DBLP. In our system, we use an incremental method to append new DBLP records into the database, i.e., only the new or updated records will be inserted into the database.

3. EVALUATION

For presenting a concrete comparison between DBLP-Filter and the three existing DBLP search tools including the Faceted search tool (*Faceted*), the Free search tool (*FreeSearch*), and the CompleteSearch search tool (*CompleteSearch*), we summarize the key features of them in Table 1. The *Results Grouping* column in Table 1 means that the tool can group the results in terms of published year, type, author, venue, and so on. Although we have not added this feature into DBLP-Filter, this can be done in the near future, because all the grouping fields are included in the original DBLP records.

Table 1. Key features of DBLP-Filter and other tools

Feature / Tool	Concept Search	User Awareness Search	New Literature Alerting	Ranking Metrics	Results Grouping	Query Expansion
DBLP-Filter	Yes	Yes	Yes	1.Text Relevance 2.Time 3.Importance 4.User Relevance	No	No
Faceted	No	No	No	1.Text Relevance 2.Time	Yes	Yes
FreeSearch	No	No	No	1.Text Relevance 2.Time	Yes	No
CompleteSearch	No	No	No	1.Text Relevance 2.Time	Yes	No

To further evaluate the effectiveness of the proposed search tool, we randomly selected 16 queries and tested them on DBLP-Filter and the three existing DBLP search tools. Figure 2 shows the average F-measure and MAP (Mean Average Precision) of DBLP-Filter and other three tools. When calculating the F-measure, the set of relevant results are constructed using the buffering technique, which unions the first 50 results of all the four search tools. And the MAP score is the average value of all the P@50 scores for the 16 queries.

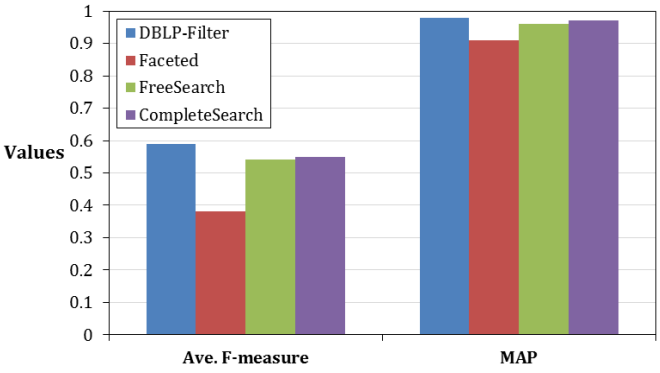

Figure 2. Ave. F-measure and MAP under 16 queries

4. CONCLUSIONS

In this paper, we proposed DBLP-Filter for effectively search DBLP records. Compared with existing search tools, our system is more functional in supporting concept-based search, literature importance based search, user-aware search, and new literature alert. Furthermore, our preliminary experimental results show that it has higher F-measure and MAP under 16 randomly-selected queries. The future work will be focused on enhancing the system with results grouping function.

5. ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China under the grant no. 71273010 and no. 61073039.

6. REFERENCES

[1] The DBLP Bibliography, in http://www.informatik.uni-trier.de/~ley/db/.

[2] Luo Y, Wang W, Lin X, SPARK: A Keyword Search Engine on Relational Databases. In *Proc. Of ICDE*, 2008, 1552-1555

[3] Stefanidis K, Drosou M, Pitoura E, PerK: personalized keyword search in relational databases through preferences. In *Proc. Of EDBT*, 2010, 585-596

[4] Fellbaum C, WordNet: An electronic lexical database. in http://www.cogsci.princeton.edu/wn