A NOVEL METADATA BASED META-SEARCH ENGINE

Jianhan Zhu, Dawei Song, Marc Eisenstadt and Cristi Barladeanu
Knowledge Media Institute, The Open University, U.K.

Keywords: Meta search engines, collection fusion, metadata.

Abstract: We present a novel meta-search engine called DYNIQX for metadata based cross search in order to study the effect of metadata in collection fusion. DYNIQX exploits the availability of metadata in academic search services such as PubMed and Google Scholar etc for fusing search results from heterogeneous search engines. Furthermore, metadata from these search engines are used for generating dynamic query controls such as sliders and tick boxes etc for users to filter search results.

1 INTRODUCTION

In the light of large scale powerful search engines such as Google, which have achieved tremendous success in recent years, thanks to their effective use of the PageRank algorithm (Brin and Page 1998), smart indexing, and efficiency in searching terabytes of data (Ghemawat et al. 2003), how can traditional professional, academic and library repositories survive and keep their successes within their specific domain? Even given the success of these big search engines, in fact it is still very difficult for them to work effectively with repositories that belong to specific professional or proprietary domains. We think there are two main reasons for this.

Firstly, due to legal or proprietary constraints, sometimes search engines cannot get hold of full content of information and may provide only the link to the place where the information can ultimately be found.

Secondly, big search engines, which tap into heterogeneous resources, sometimes cannot perform as well as some domain or context specific search services (for example, in the context of arranging air travel between London and New York, the British Airways website will provide much better search services than Google). We think that the key for successful domain specific specialized search services is to fully utilize the domain context and metadata which describes the domain context.

A limitation of current niche search services is the wide existence of information islands, which results in a contextual "jump" for users when they are searching different repositories (Awre et al. 2005). It is important to give users a unified search interface, which has been successfully used by big search engines.

We treat the problem of building a meta search engine on top of a number of search engines as a collection fusion problem as defined by Voorhees et al. (1994; 1995). The research questions we would like to answer are: How to generate a single ranked search result list based on a number of ranked lists from search engines? How to take into account relevance of each result to the query and the original rankings of the search results in the integrated ranked list? How to integrate metadata in ranking?

After reviewing existing work, we found the necessity for a meta-search system that can seamlessly integrate multiple search engines of different natures. To tackle this, we propose a novel dynamic query meta-search system called DYNIQX that integrates search results from multiple sources by taking into account search results' relevance to the query, original rankings, and metadata, in collection fusion, and provides a unified search interface. DYNIQX also provides plug-in interfaces for new search engines. DYNIQX can help facilitate our investigation of current cross-search and metadata-based search services, identification of resources suitable for cross-search or metadata-based search, and comparison of single source search, cross-search, and metadata-based search.

2 DYNIQX

Currently many niche search engines have adopted what we call a linear/top-down/hierarchical approach. For example, in the Intute search

Zhu J., Song D., Eisenstadt M. and Barladeanu C. (2008).
A NOVEL METADATA BASED META-SEARCH ENGINE.
In *Proceedings of the Third International Conference on Software and Data Technologies - ISDM/ABF*, pages 312-315
DOI: 10.5220/0001896403120315

(http://www.intute.ac.uk), a popular search engine among students for finding high quality educational websites, a searcher may select from a list of subject areas and/or resource types for his/her search, and he/she is then taken to the result page. We think the rigidity of this approach may limit the user to search within the classification of the resources. Additionally, there are many forms of metadata which have not been fully exploited during the search process.

To overcome the above limitations, we propose to experiment with the dynamic query approach based on Shneiderman's philosophy (Shneiderman 1994) of letting users experiment in real time to tune search results. Dynamic queries help users search and explore large amounts of information by presenting them with an overview of the datasets, and then allow them quickly to filter out unwanted information. "Users fly through information spaces by incrementally adjusting a query (with sliders, buttons, and other filters) while continuously viewing the changing results." A popular example of this approach is that of Kayak.co.uk, a meta-search engine which searches over 100 travel sites.

In DYNIQX, search results from a number of search engines are fused into a single list by both the relevance of each result to the search query based on our indexing of top results returned from these search engine, and the rankings of the result provided by one or more search engines as below:

$$p_{fuse}(q \mid d) \propto (1-\lambda) p(q \mid d) + \lambda / (\log(Rank_{average}(d) + 1)$$

where q is the query, $p_{fuse}(q|d)$ is the fused conditional probability of document d used to rank it in the final list, $p(q|d)$ is the conditional probability of d based on our index, λ is a parameter adjusting the effect of the two components in the final probability, and $Rank_{average}(d)$ is the average ranking of document d given by search engines. In the equation we take the log of the average ranking in order to transform the linear distribution of the rankings of d for integrating with the document conditional probability.

DYNIQX provides a novel way of meta-searching a number of search engines in terms that high quality search results from a number of search engines are integrated, metadata from heterogeneous sources are unified for filtering and searching these high quality search results, high quality results based on a number of queries covering a topic are all integrated in DYNIQX, and features such as metadata-driven controls and term clouds are used for facilitating search.

The architecture of our DYNIQX system is shown in Figure 1. In Figure 1, first, a user sends a query to the DYNIQX system. The query is processed and translated into the appropriate form for each search service, e.g., PubMed. For each query, each search engine, e.g., Intute, PubMed, or Google Scholar, returns a ranked list of search results. Results from all these ranked lists are indexed and searched by Lucene (Hatcher and Gospodnetic 2004).

Unlike typical search engines where the user can only specify one query at a time, in DYNIQX, the user can specify a number of queries describing different aspects of a search topic, e.g., "bird flu", "avian influenza", and "H5N1" etc for finding documents on "bird flu". Each query is translated into the appropriate form for each search service, e.g., PubMed. For each query, each search engine returns a ranked list of results. Results are ranked by their overall relevance scores to the topic in a single ranked list. For each result, its overall relevance score integrates the relevance of the result to the queries based on the Lucene index, rankings and relevance scores of the result by each search engine, and metadata associated with the result. Metadata from these heterogeneous sources are unified for filtering results. This is illustrated in the DYNIQX search interface shown in Figure 2.

In Figure 2, in Section A, a user adds a number of search queries shown in Section B. Statistics of search results from different search engines are shown in a table in Section B. The user can select/deselect search engines in Section E for meta-search. Once search results are retrieved from search engines, the user can view a single ranked list in Section G. When more results arrive, the user clicks a refresh button in Section A to refresh the single ranked list. Based on the significance of terms measured by their document frequencies, a term cloud is displayed in Section F for filtering result. In Section D, the user can exclude/include queries in the meta-search. Metadata associated with search results are used for re-ranking search results in Section C.

3 CONCLUSIONS

In this paper, we propose a novel metadata based search engine called DYNIQX which fuses information from data collections of heterogeneous nature. Metadata from multiple sources are integrated for generating dynamic controls in the forms of sliders and tick boxes etc for the users to further filter and rank search results. Since the effect of metadata in IR has not been sufficiently studied

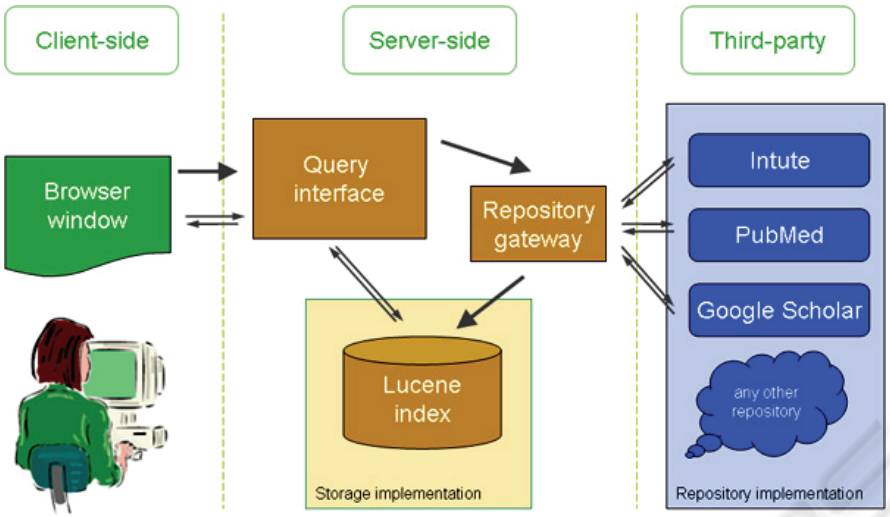

Figure 1: Architecture of DYNIQX.

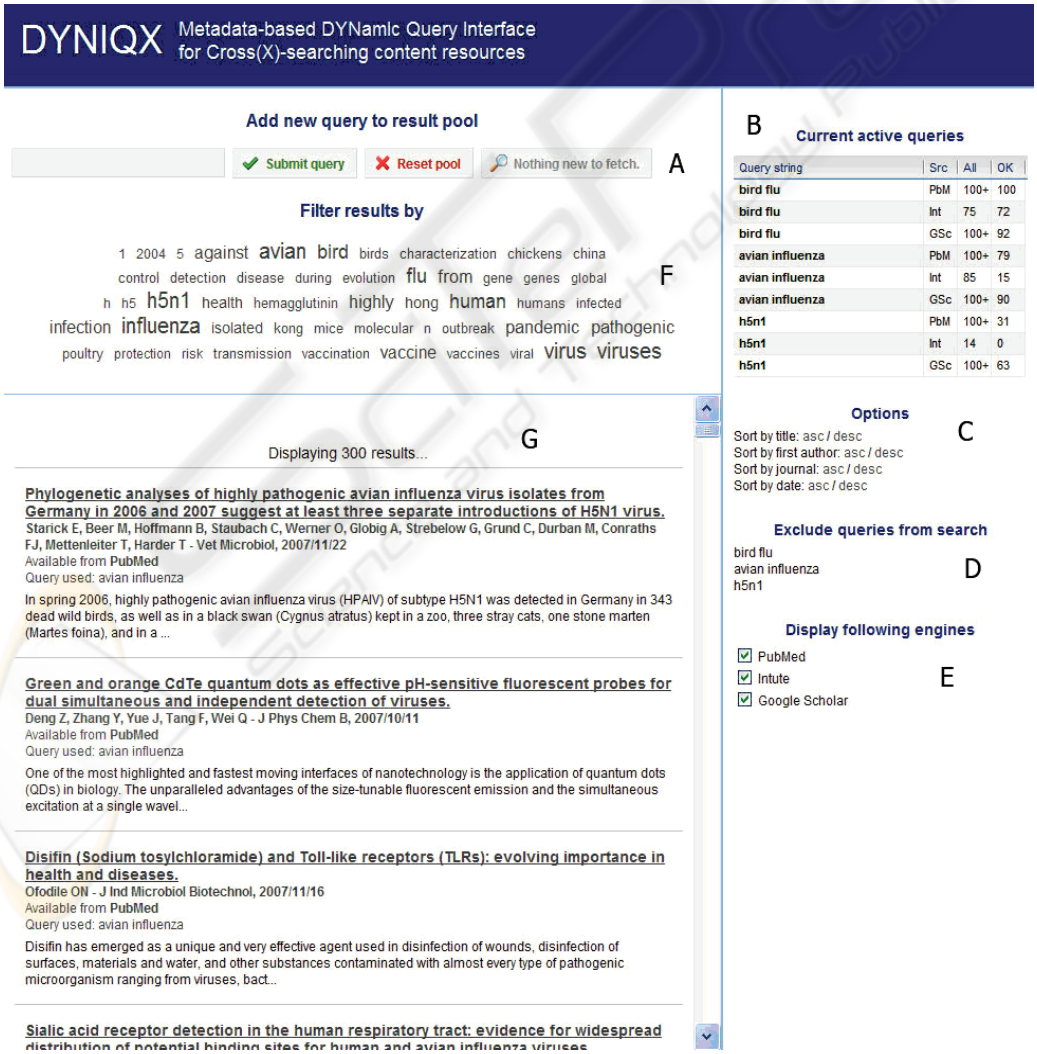

Figure 2: DYNIQX dynamic query interface.

previously, our work provides insights into how to integrate metadata with mostly content based information retrieval systems. Our preliminary user evaluation reported in (Zhu et al. 2008) shows that DYNIQX can help users to complete real world information search tasks more effectively and efficiently than individual search engines. In the future, we will carry out more formal user evaluation of DYNIQX and study the effect of different ranking algorithms in collection fusion in DYNIQX.

ACKNOWLEDGEMENTS

The work reported in this paper is funded in part by the JISC (Joint Information Systems Committee) funded DYNIQX (Metadata-based DYNamIc Query Interface for Cross(X)-searching content resources) project.

REFERENCES

Awre, C. et al. (2005) The CREE Project: investigating user requirements for searching within institutional environments. *D-Lib Magazine*, October 2005, 11(10).

Brin, S., and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks* 30(1-7): 107-117

Ghemawat, S. et al. (2003) The Google File System. In *ACM Symp. on Operating Systems Principles*.

Hatcher, E., and Gospodnetic, O. (2004) *Lucene in Action*. Manning Publications Co, ISBN: 1932394281.

Shneiderman, B. (1994) Dynamic Queries for Visual Information Seeking. *IEEE Software* 11(6): 70-77.

Voorhees, E.M. et al. (1994) The Collection Fusion Problem. In *Text REtrieval Conference (TREC)*.

Voorhees, E.M. et al. (1995) Learning Collection Fusion Strategies. In *SIGIR*: 172-179.

Zhu, J., Song, D., Eisenstadt, M., Barladeanu, C., and Rüger, S. (2008) DYNIQX: A novel meta-search engine for metadata based cross search. In *First IEEE International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008)*, VSB- Technical University of Ostrava, Czech Republic.