Specialty Meta-Search Engine for the Academic Community/ Yaffa Aharoni

Abstract

The Web is growing rapidly and has become a unique retrieval base for a large portion of the world's population. This includes members of the academic community who use it to locate scientific and research materials to serve their information needs. Search engine technologies are developing fast as well. There are three main types of search engines: Index search engines that use programs which crawl the Web and create search engines indices; subject directories that depend on humans for their listings, and meta-search engines which submit queries, in parallel, to several other search engines and display the search results to the user, usually after merging and sorting them into a single list. Search engines – both indices and directories – can be further classified into two categories according to subject coverage: general search engines that cover a wide range of topics, and specialty search engines whose repositories focus on a particular field or topic, a target audience or a specific media type. Because of this specialization, coverage is usually more in-depth and controlled.

Despite the development of these search technologies, some problems still remain, principally, the inability to find relevant, reliable information. This problem is exacerbated in the case of the academic community, which requires reliable scientific materials in various specialized research areas. General search engines, intended to meet the needs of the general population, are more likely to index commercial sites than academic/educational ones. Specialty search engines can better meet the needs of the academic, yet these search engines remain unknown to the majority of the population. Even if users are acquainted with them, they encounter many problems, particularly low recall since the specialty search engines repositories are small, requiring users to move from one search engine to another. They therefore need to know which ones to use and how to use different interfaces, whereas they would prefer a single source.

In light of this situation, this research proposes that the solution for the academic community could be a meta-search engine that would allow searches to be

Library of Information Science הספריה ללימודי מידע
Bar-Ilan University, Ramat-Gan, Israel 52900 אוניברסיטת בר-אילן, רמת גן, ישראל
Websitewww.is.biu.ac.il/library Email: scheinr@mail.biu.ac.il Tel. 972-3-5318163 טל.

sent to several specialty search engines most relevant for the information needs of the academic community. The basic premise is that since the material in the specialty search engines repositories is usually controlled, it is more reliable and of better quality. A search engine of this type for the academic community would be frendlier and more efficient in terms of precision and recall.

For the purposes of this research, an appropriate algorithm for specialty meta-search engine, focusing on the server selection problem, was developed, based on findings in the literature, search patterns of the academic population and features of the specialty search engines. This algorithm was implemented in a prototype of a specialty meta-search engine for the medical community called AcadeME. AcadeME's performance was compared to that of Google and Queryserver. Hence, there were four stages in the research: 1. locating the search engine use patterns of the academic community. 2. identifying the features of the specialty search engines. This stage included a sub-stage of identifying the relevant specialty search engines and storing their details in a database. 3. developing an algorithm for a specialty search engine. 4. evaluating the specialty search engine.

Analysis of search engine use patterns covered four fields of research: medicine, life sciences, nursing and computer sciences. The focus of the hypotheses in this part of the research was on search patterns that could highlight the need for a specialty search engine, and its desired features in relation to materials, search language and interfaces. The variables related to use patterns were investigated at the univariate, bivariate and multivariate level. Main hypotheses at the univariate level were: "Much search engine use is for research purposes", "general search engines are the most used and specialty search engines are unknown", "most search options are not used", "number of checked search engine results is small" and "success in search engine use is reasonable but there is still much to be done".

At the bivariate and multivariate level, the correlation between user characteristic variables and search patterns were examined. The independent variables examined can be classified into two main categories: cognitive personal variables and discipline variable. The goal was to discover whether a specialty meta-search engine required unique features for each discipline or whether its features could be common to all disciplines. Ten models dealing with various search patterns were investigated – "search engines use for general purposes and for research and teaching purposes", "use of the various materials", "the desired material lan-

guage", "the search language", "the preferred search interface", "use of the same queries for updating", "the desired material age" and "the number of checked results". It was hypothesized that two groups of variables – personal cognitive variables and the discipline variable would be correlated at different levels with the various use patterns. User search patterns were examined by a survey and by TLA. The survey sample included 600 users in the following fields of research: medicine, life sciences, nursing and computer sciences. Use of the TLA research method contributed to the validity and reliability of the research findings.

The above research hypotheses were confirmed. They supported the need for a specialty meta-search engine and led to certain conclusions concerning its desired features. The main conclusions stressed the importance of using a simple search strategy and considering it by default as a Boolean search; using simple interfaces; and displaying only the most relevant results to the user. These findings were common to all the research disciplines investigated since no significant correlation was found between them and the discipline variable that would indicate different treatment for each discipline.

In the second stage, a comparative study was conducted. Its goal was to compare in terms of precision, recall and F measure, the performance of the seven specialty search engines selected for input to the specialty meta-search engine. The main features of the specialty search engines were investigated by conducting 106 queries in each search engine and checking the results using the recall precision and F measures for evaluation. The overlap among the results of the seven search engines and the causes for retrieval failure were examined as well. Low recall scores and precision problems were detected in most of the search engines as well as low overlap. These findings confirmed the important contribution of a specialty meta-search engine, especially in light of the low recall scores found in most of the search engines. As for the features of a meta search engine, the findings supported an independent mechanism for ranking the results according to relevance; a small number of specialty search engines for input to the specialty meta-search engine, provided that one or two specialty search engines with big repositories are included; and two separate indices – "a domain-search engine index" and "a query term – search engine index". These indices would store the search engine results for use in selecting the appropriate specialty search engines. Two indices are required since it was found that the best search engine for a par-

ticular domain is not always the most suited to all queries. The findings for retrieval failure indicated the need to consider the user simple search as a request for a Boolean search and to give high priority to proximity in retrieving the items; i.e., those that contain query terms in proximity should be ranked higher than others. The main conclusions reached from the literature findings concerned the need for a dynamic meta-search engine, like the dynamic web, and for using a learning method for selecting the most appropriate search engines.

All these conclusions were implemented in an algorithm for a specialty meta-search engine. The main features of this search engine were: 1. selection of the best search engines for a query, based on cumulative scores that are updated continuously. This feature is compatible with the dynamic nature of the Web and enables meta-search engine autonomy in relation to the input search engines. 2. two indices for storing the search engine scores – scores for query terms and general scores for each search engine. 3. an independent mechanism for ranking the results according to relevance and automatic classification of the relevance of search engine results, based on the number of query terms included and the proximity factor, and ignoring the ranking of the input search engines. 4. displaying the most relevant search results – but not too many.

Evaluation the performance of AcadeME confirmed the thesis regarding the importance of a specialty search engine for the academic community. This evaluation was conducted by comparing AcadeME to Google and Queryserver. The comparison between AcadeME and Google was conducted by submitting the same 50 queries to each search engine, following the links of the results and evaluating each site. This comparison emphasized the contribution of AcadeME to the quality of the results. A significant correlation was found between search engine and type of material retrieved. A high percentage of reliable sites containing high-quality material, including scientific articles and material from the invisible Web, were found by the specialty meta-search engine. In addition, a higher percentage of information pages was found in the specialty meta-search engine compared to Google. However, Google includes a higher percentage of navigation pages in the results. Taking into account the relevance of the results using an ordinal scale for relevance judgment and the academic user's point of view, it was found that the meta search engine score was 20 percent higher than Google's. Yet, taking into consideration the low overlap of results and the high percentage of

relevant results in Google, it may be said that the search engines complemented each other.

The comparison between AcadeME and one of its input specialty search engines – Queryserver was conducted by checking the results of the same 50 queries in each search engine and by checking the results retrieved from Queryserver by AcadeME. This comparison emphasized the contribution of AcadeME to the relevance and the variety of the results.

This research can contribute to the fields of search patterns on the Web, search engines and selecting problem algorithms, especially in light of the paucity of research pertaining to the academic community in these fields. In addition to its theoretical contribution to knowledge in these fields, its practical benefit to the academic community lies in its ability to locate free academic materials on the Web. The results of this study can be further applied to the entire academic community. Future studies to overcome the limits of this research or to enlarge the subject, in light of the current findings, are proposed.

System No.
583769