Towards Context-Based Search Engine Selection

David B. Leake and Ryan Scherle
Computer Science Department, Indiana University
150 S. Woodlawn Avenue, Bloomington, IN 47405
Ph: 812-855-{9756, 8702}. Fax: 812-855-4829
{leake, rscherle}@cs.indiana.edu

ABSTRACT

A well-known problem for web search is targeting search on information that satisfies users' information needs. User queries tend to be short, and hence often ambiguous, which can lead to inappropriate results from general-purpose search engines. This has led to a number of methods for narrowing queries by adding information. This paper presents an alternative approach that aims to improve query results by using knowledge of a user's current activities to select search engines relevant to their information needs, exploiting the proliferation of high-quality special-purpose search services. The paper introduces the PRISM source selection system and describes its approach. It then describes two initial experiments testing the system's methods.

Keywords

Distributed information systems, intelligent web search, just-in-time information access

1. INTRODUCTION

Typical search queries are short—often one or two words. These short queries are often ambiguous, resulting in poor results from general-purpose search engines when off-target results are returned. For the query "home sales," for example, the first page of results for a recent query to AltaVista contained pointers to information on real estate, realtors and mortgages. This is useful information if the user is interested in the mechanics of selling a home, but not for an economist interested in economic indicators. If the context for the query is known to be that the user is writing a document on economics, it is possible to anticipate the type of result that will be useful and refine the query accordingly. A common way to do this is to add additional search terms. Unfortunately, this places an added burden on the user, and it is sometimes difficult even for an expert to select the right query terms for the desired subset of information to be retrieved.

IUI'01, January 14-17, 2001, Santa Fe, New Mexico, USA.

Specialized search resources, on the other hand, can provide coverage that is pre-focused. Many specialized search engines index a carefully-crafted set of resources, often hand-gathered to be relevant to a topic or audience. Thus if the interface can automatically select context-relevant search engines, the focus they provide can decrease the burden of generating focused queries. For example, if the interface can determine that the user is working on a paper in economics, it could use this information to generate a context description [2], and select a context-relevant specialized search engine, such as CNN financial. Sending the "home sales" query there will yield results that an economist might want, such as information on changes in aggregate housing sales trends.

Given sufficient bandwidth, it would be possible to skip the selection process, retrieve documents from many specialized search engines, and then filter the results according to user needs. However, that approach would destroy the benefits of pre-focused document sets, requiring sophisticated on-the-fly filtering to substitute for the careful preselection already done for specialized sources. (Taken to an extreme, this filtering approach would simply provide another general-purpose search engine.) Our approach examines the hypothesis that source selection is a more tractable problem than document filtering: That it is possible to monitor the user's task context, automatically select a small set of on-point sources, and dispatch queries to those sources to provide more useful search results. This paper describes research on the source selection problem in the PRISM system, summarizing the system's methods and initial experimental results.

2. PERSPECTIVE

PRISM's source selection approach relates to research on both distributed searching and "just-in-time" searching. The first distributed information systems grew out of distributed databases (such as [7]), a number of which have been developed for the web [5, 8, 13]. These systems can route queries to a collection of sources, but depend on those sources to cooperate by providing indices or other data to a central distribution system, and require costly updating of central information as database contents change. Another alternative is a metasearch approach, such as first taken by the MetaCrawler [11], to access search engines without explicit cooperation by simply forwarding queries to them and collating the results. These systems ignore the differences in coverage of topics by search engines they index, and bandwidth constraints limit the number of search engines that can be queried. SavvySearch [3], ProFusion [4], and Q-Pilot

[12] use artificial intelligence techniques to select search engines based on the queries they are given. However, these systems are still limited to a relatively small number of search engines, and do not address the problem of ambiguous queries. Systems such as Apple's Sherlock and TheBigHub.com provide categorized lists of specialized search engines, relying on the user to select the right sources.

Recent research has introduced just-in-time information retrieval systems, which attempt to anticipate user information needs, based on a task context inferred from user behavior. The Remembrance Agent monitors what users type in a text editor and sends related queries to local databases and the web [10], the Lumiere project monitors user behavior in Microsoft Office to predict user questions [6], and the Watson system [2] monitors the use of applications to generate context-relevant queries for general search engines.

PRISM combines distributed and just-in-time searching to leverage the advantages of both. Its central claims are that distributed searching can be more effective if it is guided by contextual information, such as the task information gathered by just-in-time retrieval sources, and that just-in-time information systems can benefit from strategic access to a larger selection of information sources.

Research on context-based automatic source selection must address three issues. The first is *context extraction and representation:* how to determine and describe the query context. The second is *source characterization and selection:* how to describe sources to support decisions about source relevance and enable effective access. To enable good coverage, the process for characterizing sources must be simple, quick, and robust, without relying on explicit cooperation from the sources. The third is *selectivity:* how to recognize when the available specialized sources are insufficient. The following sections discuss how these issues are addressed by PRISM.

3. PRISM'S DESIGN

Context extraction: To determine the context of queries, PRISM uses the context-extraction framework in the Watson system (for details on Watson's methods, see [2]). In the combined system of Watson and PRISM, Watson monitors user activities in standard applications such as word processors, uses heuristics to identify relevant content areas, and provides PRISM with context information. PRISM determines appropriate information sources, formulates queries to those sources, sends off those queries, and passes on their results to the user. Typically the query is input by the user, but PRISM can also derive a context description automatically from a user-selected document and search for relevant pages.

Source characterization: PRISM's "specialized search engines" include both pages whose purpose is to search topic-specific material, and medium- to large-sized web sites with internal search services (e.g., the Microsoft web site). PRISM characterizes the "topic" of a search engine with a weighted term vector. This vector is generated from limited quantities of easily-accessible data: Keywords are automatically gathered from a source's "about" or "FAQ" pages; if these are not available the main page for the source is used. This information is augmented with keywords from META tags, if available.

Because each source is different, a wrapper must be used to format queries and interpret the results. Methods for automatic wrapper generation (e.g., [1]) are not yet sufficient to fully automate this process, so some human assistance is needed to create and verify wrappers. However, this burden may be alleviated as services begin to publish wrappers for search engines. For example, Apple's Sherlock searching system uses XML-style plugins to encode wrapper information [9], and MacOS includes plugins for some popular search engines; several independent parties have published lists of Sherlock plugins. PRISM currently represents sources using enhanced Sherlock plugins. Because Sherlock plugins are meant to be used with manual source selection, they do not include the topic and query type information needed for automatic source selection, so PRISM adds several tags to each plugin for this information. PRISM can gather topic information automatically, but information about query types is gathered manually from an "about" or "help" page, or determined by giving the search engine trial queries of different types and lengths to observe results.

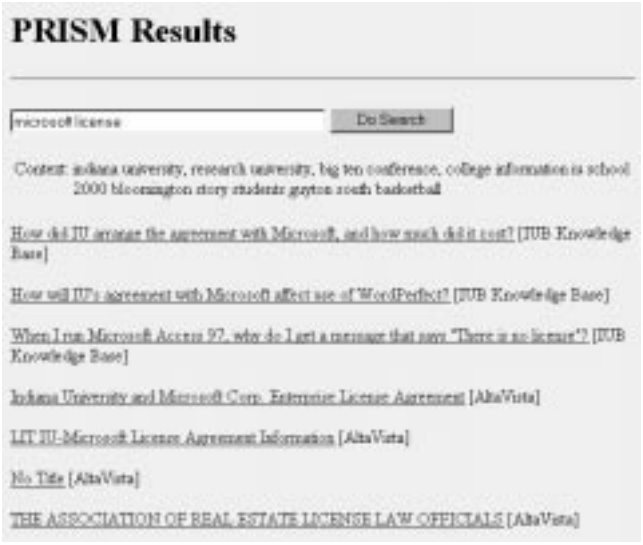

Figure 1: PRISM **recommending pages.**

Source Selection and Selectivity: The search engine representations are stored and accessed by an information retrieval system within PRISM. This system uses the standard information retrieval metrics of TFIDF and cosine similarity. When a query and context are provided to PRISM, the context is used as a query to the internal system, which selects search engines covering similar topics. If no search engines match sufficiently or the query is provided without context, PRISM forwards the query to a fixed set of general-purpose search engines. Figure 1 shows PRISM retrieving suggestions for a query in an automatically derived context.

4. EVALUATION

Our initial tests of the approach focused on two questions: (1) Is source selection sufficiently precise to avoid degrading performance with bad source choices?, and (2) Does use of specialized sources improve results for queries in their content areas?

4.1 Experiment 1

The first experiment studied whether topic-specific sources could be selected automatically with high precision, test-

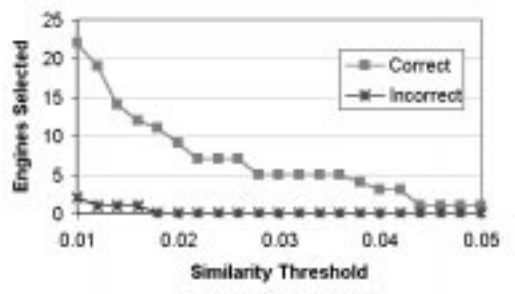

Figure 2: Number of search engines selected for computer science papers with varying similarity thresholds.

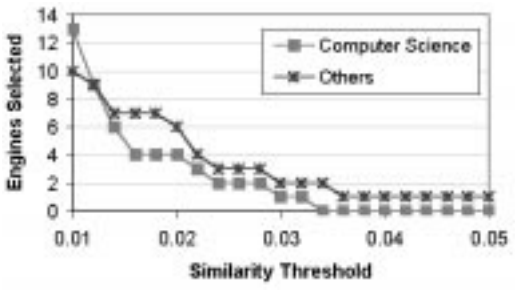

Figure 3: Number of search engines selected for random pages with varying similarity thresholds.

ing the quality of PRISM's internal IR process for search engine representations, based on queries derived automatically from documents.

4.1.1 Method:

PRISM was configured to use 10 search engines, five for the topic of computer science, and five from other topics. 20 papers from computer science and 20 web pages on random topics were given to the system and a tally was kept of the number of times each search engine was selected. The required level of match depended on a similarity threshold, which was varied to determine the effects of making the system more or less likely to pick topic-specific search engines.

4.1.2 Results:

When the context was based on viewing computer science papers with PRISM's default settings, computer science engines were selected 14 times, and a distractor engine was picked only once. When the context was based on viewing random (non-CS) pages, computer science engines were selected six times and other search engines were selected seven times.

Figure 2 shows the number of search engines selected for computer science pages at different similarity levels. The "correct" line indicates computer science engines, and the "incorrect" line indicates distractor engines. Figure 3 shows the number of search engines selected for random pages at various similarity thresholds.

	AV	Google	Spec-1	Spec-2	Spec-3
Average	2.0	2.9	3.0	2.3	2.2
Std. Dev.	1.29	1.67	1.20	1.46	1.22
Subject 5	1.0	1.0	4.5	3.2	1.0

Table 1: Summary of usefulness ratings.

4.1.3 Discussion:

Performance with computer science papers as context is encouraging: For lenient threshold settings, many computer science search engines were selected but few distractors. As the threshold is raised, computer science engines are still selected frequently, but distractor engines cease to be a factor. For non-CS queries, the number of false positives was initially surprising, but because the specialized engines were split evenly between CS and non-CS engines, the false positive rate was actually what would be expected if none of the specific search engines were relevant, and the choice of specific engines were random. As more engines are added to the database, covering a wider range of topics, the TFIDF method used by PRISM to calculate keyword weighting will decrease the weights of common terms such as "system", reducing the similarity levels for unrelated search engines.

4.2 Experiment 2

Experiment 2 examined whether specific sources return better results than general search engines, assuming that correct specific search engines are always chosen. Intuitively, "correct" topic-focused search engines would be expected to give better results than general-purpose search engines, but it is possible that the extremely large databases of general-purpose search engines would counterbalance that advantage.

4.2.1 Method:

The three specific sources selected most frequently in experiment 1 (CiteSeer, Cora, and the Indiana University CS techreport index) were compared to two popular general search engines, Google and AltaVista. Five volunteers submitted computer science papers they had written. Queries were automatically generated from these papers and submitted to each of the five search engines. The responses given by the five search engines were placed in random order and given to the volunteers to rate on a five-point scale for usefulness.

4.2.2 Results:

Effectiveness of the specific search engines varied depending on the subject. For four of the five subjects, at least one specific engine produced better results than the general engines, but the relative scores of the specific engines changed dramatically for different subjects. Google fared surprisingly well, with scores always higher than AltaVista. In the overall average, Google fared better than two of the three specific search engines. Table 1 summarizes the usefulness rankings for each search engine. Google and the first specific engine, CiteSeer, measured significantly better than AltaVista (at the .05 level); the other differences were not statistically significant. For one subject (Subject 5), results from the specific search engines were dramatically better than the general search engines.

4.2.3 Discussion

The quality of search engine responses varied widely in our tests. However, using research papers as context, on average the results of one of the specific search engines, CiteSeer, slightly surpassed Google, the general search engine with the best performance. Two of the subjects commented that they found some very good resources during this experiment, which they will now use as references for future research. Upon investigation, it was found that all of these references came from the specific search engines.

4.3 Overall performance

Experiments 1 and 2 separately examine PRISM's precision in selecting specialized sources, and how specialized sources affect the quality of provided information. The key question concerns their combination: Would PRISM as a whole be successful in the task for experiment 2? To address this question, each of the five papers was given to PRISM as context to examine the usefulness of retrieved results. For each paper, only one search engine matched sufficiently to satisfy the default threshold setting; in four of the five cases the selected engine performed well for that paper, scoring an average of 3.2 on the usefulness measure, which was better than the average for any individual search engine. For the remaining paper, PRISM incorrectly picked a distractor search engine. However, this paper covered the topic of abstract logic, and none of the search engines—general or specific—fared well using it as an input.

5. CONCLUSION

This paper presents ongoing research on context-based selection of specialized web information sources. Our approach aims to provide on-point information by using a description of the user's task context, extracted by monitoring user behavior, to predict the type of information likely to be of interest, and then dispatching search queries to special-purpose search engines tailored towards the user's particular needs. We have conducted two initial experiments to examine the benefits of this approach. The first suggests that PRISM's methods can select specialized search engines with high precision, and the second that the use of specialized search engines can improve results for queries in their content areas. We are now adding additional search engine descriptions to PRISM to increase its range of topic areas in preparation for a larger-scale evaluation.

6. ACKNOWLEDGMENTS

David Leake's research is supported in part by NASA under award No NCC 2-1035. Ryan Scherle's research is supported in part by the Department of Education under award P200A80301-98.

7. REFERENCES

[1] Ashish, N. and Knoblock, C. Wrapper generation for semi-structured internet sources. In *ACM SIGMOD Workshop on Management of Semi-structured Data*, 1997.

[2] Budzik, J. and Hammond, K. User interactions with everyday applications as context for just-in-time information access. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces (IUI2000)* 44–51, 2000.

[3] Dreilinger, D. and Howe, A. E. Experiences with selecting search engines using meta-search. *ACM Transactions on Information Systems*, 15(3), July 1997.

[4] Gauch, S., Wang, G., and Gomez, M. ProFusion: Intelligent fusion from multiple, distributed search engines. *Journal of Universal Computing*, 2(9), September 1996.

[5] Gravano, L., García-Molina, H., and Tomasic, A. Precision and recall of GlOSS estimators for database discovery. In *Proceedings of the third international Conference on Parllel and Distributed Information Systems (PDIS '94)*, 1994.

[6] Horvitz, E. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* 256–265. Morgan Kaufmann, July 1998.

[7] Levy, A. Y., Rajaraman, A., and Ordille, J. J. Querying heterogeneous information sources using source descriptions. In *Proceedings of the 22nd VLDB Conference*, 1996.

[8] Meng, W., Liu, K.-L., Yu, C. T., Wu, W., and Rishe, N. Estimating the usefulness of search engines. In *Proceedings of the 15th International Conference on Data Engineering, 23-26 March 1999, Sydney, Austrialia* 146–153. IEEE Computer Society, 1999.

[9] Montbriand, J. Extending and controlling sherlock. Technical Report 1141, Apple Computer, 1999.

[10] Rhodes, B. J. Margin Notes: Building a contextually aware associative memory. In *Proceedings of the 2000 International Conference on Intelligent User Interfaces (IUI2000)* 219–224, 2000.

[11] Selberg, E. and Etzioni, O. Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World Wide Web Conference (WWW4)*, 1995.

[12] Sugiura, A. and Etzioni, O. Query routing for web search engines: Architechture and experiments. In *Proceedings of the 9th International World Wide Web Conference (WWW9)*, 2000.

[13] Zhu, X., Gauch, S., Gerhard, L., Kral, N., and Pretschner, A. Ontology based web site mapping for information exploration. In *Proceedings of the 8th International Conference on Information and Knowledge Management (CIKM'99)* 188–194, November 1999.