Intelligent Metasearch Engine for Knowledge Management

Eui-Hong (Sam) Han and George Karypis [*]
iXmatch Inc.
{sam,george}@ixmatch.com

Doug Mewhort and Keith Hatchard
Sagebrush Corp.
{DMewhort,KHatchard}@sagebrushcorp.com

ABSTRACT

The explosive growth of available information sources and the resulting information overload pose several problems for users in many business organizations and educational institutions. First, searching through several information sources, one at a time, is a source of enormous frustration for users. Second, top-ranked documents in search results are frequently irrelevant to what users are interested in. To address these problems, we have developed iXmetafind^{TM}, a powerful metasearch engine that gathers, evaluates, ranks, and reports the most relevant results from multiple information sources, including library catalogs, proprietary databases, intranets, and Web search engines. In addition to basic metasearch capabilities, iXmetafind uses personalization and clustering techniques to find the most relevant results for users. In this paper, we briefly describe technologies used in iXmetafind and present Pinpoint^{TM} from Sagebrush Corporation, the smart research tool^{TM} in the kindergarten through twelfth grade (K-12) school environment. Pinpoint showcases iXmetafind in the knowledge management domain of the K-12 school environment.

Categories and Subject Descriptors

H.3.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Algorithms, Management

Keywords

metasearch, collection fusion, personalization, collaboration, clustering, library automation

1. INTRODUCTION

The explosive growth of available information sources and the resulting information overload pose several problems for users in many business organizations and educational institutions. First, searching through several information sources, one at a time, is a source of enormous frustration for users. Information sources can be difficult to search, and the information a user wants could be in any of them, or pieces of what he wants could be found in several. As a result, search results are often incomplete, leaving the user disappointed. Second, top-ranked documents in search results are frequently irrelevant to what users are interested in. This might be due to limited query capabilities (e.g, lack of boolean query support), the poor ranking mechanism of search engines, a poor choice of keywords, and/or the problems of word synonymy and polysemy.

In response to the first problem, frustration at searching through multiple information sources, several metasearch engines have been developed [6, 11, 9]. Meta-search engines take query keywords, simultaneously transmit the query to several individual search engines, and collate together results from all search engines. In response to the second problem, irrelevant results, document clustering approach has been developed to provide intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters [17, 4]. Another approach has been developed based on personalization of information [1, 2, 12]. Personalized information filtering systems typically try to find pertinent information based on the interest of the user as an individual or as a member of a group.

We have developed iXmetafind, a powerful metasearch engine that gathers, evaluates, ranks, and reports the most relevant results from multiple information sources, including library catalogs, proprietary databases, intranets, and Web search engines. To the best of our knowledge, iXmetafind is the first product of its kind that has all the features discussed above: metasearch capabilities, personalization, and clustering methods.

In this paper, we briefly describe technologies used in iXmetafind to support these features. We also present Pinpoint, the smart research tool in the K-12 school environment from Sagebrush Corporation. Pinpoint showcases iXmetafind in the knowledge management domain of the K-12 school environment.

[*] also with Department of Computer Science, University of Minnesota

CIKM'03, November 3–8, 2003, New Orleans, Louisiana, USA.
Copyright 2003 ACM 1-58113-723-0/03/0011 ...$5.00.

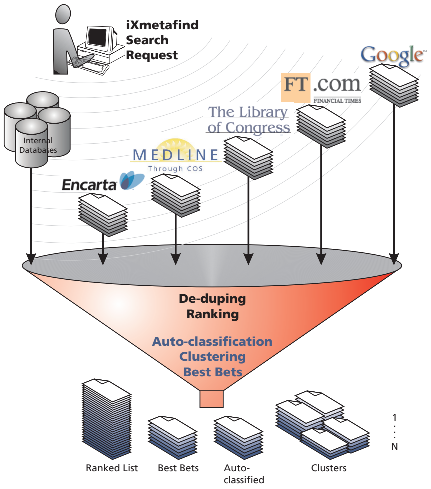

Figure 1: Main components and features of iXmetafind.

2. IXMETAFIND SEARCH AND NAVIGATION TECHNOLOGY

Figure 1 shows the main components and features of iXmetafind. iXmetafind provides flexible mechanisms for collection selection and uses the state-of-the-art collection fusion method for metasearch. In addition to these metasearch capabilities, iXmetafind provides personalization and relevance feedback based on query histories, and clustering and categorization for exploration and browsing.

2.1 The Search Template

Each information source is described by a search template in iXmetafind. The search template defines the interface mechanism (http, z39.50, ODBC, web services, etc.); the query interface format that specifies how to pass in search terms, authentication, selection of databases, and how to parse information for results (HTML page/tag information, XML tag information, etc.).

The search template also maps globally defined (mutually agreed) values into search engine-specific values. For example, the data access security level can be mapped into the specific security level of an individual search interface. Given global security levels of “level 1”, “level 2”, and “level 3”, a particular search template might map “level 1” as “s1”, “level 2” as “s2”, and “level 3” as “s3” in the query interface format. Another search template might map “level 1” as “basic”, “level 2” as “restricted”, and “level 3” as “confidential”.

2.2 Collection Selection Using a Query Profile

iXmetafind defines which information resources (defined as search templates) to use and the relative importance of each source using a query profile. In other words, the dispatch mechanism [5] of metasearch is explicitly controlled by the query profile.

The query profile also defines global values for search. For example, the query profile can set one security level for users of that profile. Several profiles can be maintained in the system and each corresponds to a particular mix of information resources, their importance, and global values. Dynamic profiles can be used as well. In that case, a dynamic profile corresponding to a particular user and query is built and used for the query.

In addition to the information sources and global values, the query profile contains which query history to use for personalization (see Section 2.4).

2.3 Expert Agreement and Content-Based Collection Fusion

A key component of a metasearch engine is the method used to merge the individual lists of documents returned by different engines to produce a ranked list that is presented to the user. The overall quality of this ranking is critical, as users tend to examine the top-ranked documents more than the lower-ranked documents. In iXmetafind, we use Mearf, the state of art fusion mechanism described in [11].

Most of the metasearch engines for which technical details are available ([5], [14], [6]), use a variation of the linear combination of scores scheme (LC) described by Vogt and Cottrell [15]. This scheme requires that a weight be associated with each source (reflecting its importance) as well as a weight associated with each document reflecting how well it matches the query. Then, the LC-based scheme uses a product of the two weights to compute an overall score for each document to be used in ranking. If the weight of each source is unknown or uniform and if the sources only provide a ranked list of documents (no numerical scores) – which is the case for most search engines – then this scheme becomes equivalent to that of interleaving the ranked documents produced by the different sources. No relative scoring is actually being done or considered.

Mearf has four novel methods for merging results from different search engines. The schemes proposed are motivated by the observation that even though the various search engines cover different parts of the web and use different ranking mechanisms, they tend to return results in which the higher-ranked documents are more relevant to the query. Presence of the same documents in the results of different search engines in top ranks can be a good indication about their relevance to the query. In fact, some existing LC-based methods already use this observation to boost the ranks of such documents. The methods used in Mearf take advantage of this observation and also look for common themes present in top-ranked documents to extract a signature that can be used to re-rank other documents. As a result, unlike LC-based methods, the methods in Mearf can boost the ranks of the documents that are similar in content to the top-ranked documents deemed relevant. These new methods that use expert agreement in content to merge and rerank documents have been shown to outperform traditional LC-based methods commonly used in various search engines and Google, a popular and highly regarded search engine.

2.4 Personalization and Relevance Feedback Using Query History

iXmetafind provides mechanisms to create multiple query histories and to record relevance feedback. The query history contains the latest queries and their corresponding “hits” – information content that users selected/liked. The query history is used in two ways in iXmetafind. In the collection fusion process, the relevant hits with respect to a given query are retrieved from the query history. For example, if the query is “pilot training”, then similar queries in the query history are searched and corresponding hits are returned. This set of hits forms one expert opinion for the fusion process and influences the final ranking of the hits. Note that, by using different query histories, the final ranking of the same query can be personalized to a particular set of users corresponding to the query history used. For example, when the query history of aviation authorities is used, hits related to FAA flight training programs might come up at the top of the final hits. With the same query, when the query history of FBI agents is used, hits related to terrorists taking flight training programs might come up at the top of the hits.

The query history is also used to provide “best bets” of the given query. The set of hits from the query history forms a concept, and this concept is used to find the “best bets” from the past query history and from the search results of the current query. Note also that by using different query histories, the “best bets” can be personalized.

2.5 Exploration of Results Utilizing Clustering

Fast and high-quality document clustering algorithms play an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters [17, 4]. In particular, clustering algorithms that build meaningful hierarchies out of large document collections are ideal tools for their interactive visualization and exploration as they provide dataviews that are consistent, predictable, and at different levels of granularity. For example, given set of results coming back from the query “mercury”, clustering algorithms would group the results into topics such as Mercury the planet, Mercury the Greek god, mercury the element, Mercury Theater, Mercury the car, and so on. Or, given the search term “apple”, clustering algorithms would provide one set of results about Apple Computer, one set about recipes, one set about the Big Apple, meaning New York City. Since these clusters represent different contexts of the same word, users can easily navigate through search results by selecting relevant clusters.

iXmetafind provides fast document clustering based on partitional algorithms described in [17]. Given a set of documents from the fusion process, the clustering mechanism finds groups of documents that are similar and provides sets of words that describe the clusters. Users can easily identify clusters that best fit their search/browsing needs and narrow down their search or search with a new query that contains additional words in the clusters they found interesting.

2.6 Categorization/Auto-Classification

Automatic text categorization [16, 10, 8, 7, 3], which is the task of assigning text documents to pre-specified classes (topics or themes) of documents, is an important task that can help people find focused information from huge resources. For example, let’s assume a user has topics of interest such as business, sports, travel, books, and movies, and has collected some articles related to these topics as samples. From the search results, this user would like to collect all the information related to individual topics. Once a classification model is learned from documents representative of these topics, all the search results be classified according to this model. Only the documents closely related to the sample documents of the classification model will be identified and presented to the user.

iXmetafind uses a simple linear-time, centroid-based document classification algorithm [7] that, despite its simplicity, has robust performance. In this algorithm, a centroid vector is computed to represent the documents of each class, and a new document is assigned to the class that corresponds to its most similar centroid vector, as measured by the cosine function. Extensive experiments presented in [7] show that this centroid-based classifier consistently and substantially outperforms other algorithms such as Naive Bayesian [10], k-nearest-neighbors [16], and C4.5 [13], on a wide range of datasets. Our analysis shows that the similarity measure used by the centroid-based scheme allows it to classify a new document based on how closely its behavior matches the behavior of the documents belonging to different classes. This matching allows it to dynamically adjust for classes with different densities and accounts for dependencies between the terms in the different classes. We believe that this feature is the reason why the centroid-based classifier consistently outperforms other classifiers that cannot take into account these density differences and dependencies.

3. PINPOINT, THE SMART RESEARCH TOOL FROM SAGEBRUSH

Knowledge management has become a challenging task in the K-12 school environment. Patrons, mostly students, are faced with an overwhelming variety of information sources that can include the school library catalog, on-line subscription services, Internet resources, and catalogs that institutions such as the Library of Congress have made available on Z39.50 servers. In the past, students have relied heavily on the librarian’s help to choose age-appropriate sources. Even then, students have had to search each source separately, manually compiling information from each source. For many students, especially those in lower grades, this process is too cumbersome. As a result, when conducting research for papers and class assignments, students have seldom taken advantage of even a small portion of the available information.

Sagebrush Corporation undertook to develop a searching tool to make the research process less cumbersome for students and librarians alike. This product had to provide students with easy access to age-appropriate sources, preferably without librarian assistance; provide students with an easy method of quickly searching multiple sources; and combine results from multiple sources into a single search result containing highly relevant and age-appropriate items.

To meet these goals, Sagebrush developed Pinpoint, a searching and research tool that enables students to search multiple sources in a single operation using iXmetafind. During setup, the Pinpoint administrator provides the information the program needs to access the information source, the source’s information type (library items, reference facts, articles and pages, biographies, news, reports, and so on), and the source’s priority for each grade level for each source. This information is mapped into iXmetafind search templates and profiles.

When searching, patrons specify the type of information that they want to find and their grade level (e.g., elementary, middle/junior high, high school, or adult), in addition to the search terms. This information is used to find the right search profile and query history. Note that Pinpoint maintains a separate query history for each grade level.

By employing this strategy, Pinpoint ensures that patrons are presented with the most relevant search results. A high school senior and an elementary student may use exactly the same search terms, yet Pinpoint focuses their searches appropriately and yields different results in each case. The high school student is presented with materials suitable for a high school student; the elementary school student is presented with materials suitable for an elementary student.

Students can also hone in on clusters by clicking the 'Zoom in link' (i.e., 'More Like This' utilizing clustering techniques discussed in Section 2.5) for any item in the search results. In addition, students can choose items from a HotPicks list (i.e., 'Best Bets' discussed in Section 2.4) which is generated based on the group's query history.

4. CONCLUDING REMARKS

iXmetafind incorporates cutting-edge search methods to gather, evaluate, rank, and report the most relevant results from multiple databases, including library catalogs, proprietary databases, intranets, and Web search engines. It also incorporates personalization/collaboration techniques for finding the most relevant information for an individual or a group. Clustering and categorization mechanisms in iXmetafind allows intuitive navigation and browsing. To the best of our knowledge, iXmetafind is the first product of its kind that has metasearch capabilities with personalization and clustering capabilities.

iXmetafind can be used as a library portal, virtual union catalog, or an OPAC (on-line public access catalog). The Pinpoint product for K-12 school environment uses iXmetafind tools for helping K-12 students to find the best information they need for learning. Pinpoint has been well received in the market and demonstrates the technology advantages of iXmetafind in the real world. Pinpoint demonstrates that iXmetafind makes quality research results literally child's play.

5. REFERENCES

[1] D. Boley, M. Gini, R. Gross, E.H. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher, and J. Moore. Document categorization and query generation on the world wide web using WebACE. *AI Review*, 13(5-6), 1999.

[2] Liren Chen and Katia Sycara. WebMate: A personal agent for browsing and searching. In Katia P. Sycara and Michael Wooldridge, editors, *Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98)*, pages 132–139, New York, 9–13, 1998. ACM Press.

[3] David Wai-Lok Cheung, Graham J. Williams, and Qing Li, editors. *Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification*, volume 2035 of *Lecture Notes in Computer Science*. Springer, 2001.

[4] Douglass R. Cutting, Jan O. Pedersen, David Karger, and John W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, 1992.

[5] Daniel Dreilinger and Adele E. Howe. Experiences with selecting search engines using metasearch. *ACM Transactions on Information Systems*, 15(3):195–222, 1997.

[6] S. Gauch, G. Wang, and M. Gomez. ProFusion: Intelligent fusion from multiple, distributed search engines. *J.UCS: Journal of Universal Computer Science*, 2(9):637–??, 1996.

[7] E.H. Han and G. Karypis. Centroid-based document classification: Analysis & experimental results. In *Proc. of The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, 2000.

[8] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the European Conference on Machine Learning*, 1998.

[9] Steve Lawrence and C. Lee Giles. Inquirus, the NECI meta search engine. In *Seventh International World Wide Web Conference*, pages 95–105, Brisbane, Australia, 1998. Elsevier Science.

[10] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, 1998.

[11] B. Uygar Oztekin, G. Karypis, and V. Kumar. Expert agreement and content based reranking in a meta search environment using mearf. In *Proc. 11th WWW Conference*, 2002.

[12] Michael J. Pazzani, Jack Muramatsu, and Daniel Billsus. Syskill webert: Identifying interesting web sites. In *AAAI/IAAI, Vol. 1*, pages 54–61, 1996.

[13] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA, 1993.

[14] E. Selberg and O. Etzioni. Multi-service search and comparison using the MetaCrawler. In *Proceedings of the 4th International World-Wide Web Conference*, Darmstadt, Germany, December 1995.

[15] Christopher C. Vogt and Garrison W. Cottrell. Fusion via a linear combination of scores. *Information Retrieval*, 1(3):151–173, 1999.

[16] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR-99*, 1999.

[17] Ying Zhao and George Karypis. Evaluation of hierarchical clustering algorithms for document datasets. Technical Report TR-02-22, Department of Computer Science, University of Minnesota, Minneapolis, 2002.

APPENDIX

iXmatch is a registered trademark and iXmetafind and ixmatch.com are trademarks of iXmatch Inc. Pinpoint, Sagebrush, and smart research tool are trademarks of Sagebrush Corporation. Apple is a trademark of Apple Computer, Inc., registered in the U.S. and other countries. Google is a trademark of Google Inc. Mercury is a trademark of Ford Motor Company. Mercury Theatre is a registered service mark of Mercury Theatre, Inc.