Searching for Needles in a World of Haystacks

Jamie Callan
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213-3890, USA
callan+@cs.cmu.edu

Abstract

Current Web search engines are based on a single database *model of information retrieval. This paper argues that next generation Web search will be based on a* multi-database *model of information retrieval. The strengths, weaknesses, open research problems, and prospects of several multi-database retrieval models are discussed briefly.*

1 Introduction

Web search engines provide access to HTML documents that are stored on computers around the world. The distributed nature of the document collection provides the illusion of a complex information access task, but the first generation of Web search engines is actually based on a simple retrieval paradigm. Documents are copied from their original locations to a central site, where they are added to a central database, indexed, and made available for search. The first generation of Web search engines is based on a *single database* model of information retrieval.

The single database model of information retrieval was successful because initially i) the Web was not very large, ii) the Web was not a commercial medium, and iii) HTML pages were not very complex. A large, but not *too large*, amount of information was available for copying, freely, in a format that was easy to interpret. As the Web evolved, these conditions were violated. There was too much information for it all to be copied, access to some of it became restricted, and a large amount of information was in searchable databases or was generated dynamically in response to a request or user action. Web search engines today provide access to only a small portion of the information on the Web.

The single database approach to Web search is a poor foundation upon which to build next generation systems. New, multi-database architectures provide a more solid foundation, because they explicitly model the multi-site, multi-resource nature of the Web. Traditional Web search engines based on the single database model can still play a valuable role in a multi-database environment, but it will no longer be the central role.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

2 Multi-Database Retrieval Models

Multi-database retrieval models can be classified as appropriate for *small-scale* or *large-scale* environments. A *small-scale* environment might be one that contains a few hundred text databases, perhaps under the control of a single organization. A *large-scale* environment might be one that contains tens of thousands of text databases, perhaps under the control of multiple organizations.

We consider a multi-database retrieval model appropriate for a small-scale environment if human effort is required to maintain a database hierarchy, a knowledge base [13, 4], or training data [14] used for database selection. The determining factor is the ease with which a newly-discovered text database can be made accessible. We also consider a multi-database retrieval model appropriate for a small-scale environment if each query requires some degree of communication with every database [9].

Multi-database retrieval models for small-scale environments have their advantages. For example, manual database organization or the use of training data can enable very accurate database selection for typical queries. This approach to database selection is most likely to be used by commercial information services and in customer support centers, for example, where there is a premium on accuracy and less concern about the rapid introduction of new text databases on a regular basis.

We consider a multi-database retrieval model appropriate for a large-scale environment, such as the Web, if i) new text databases can be introduced quickly, easily, and automatically; and ii) only limited communication is required to decide which text databases a particular query should be sent to. Retrieval models that satisfy these requirements can be grouped into two classes: message-passing, and centralized selection.

3 Message-passing

A *message-passing* multi-database retrieval model is one in which a search request is repeatedly passed from one node in a computer network to adjacent nodes until either an answer is found or it reaches a search horizon. If each node passes requests to n neighbors, the search horizon at distance h contains n^h nodes, so search requests can be propagated rapidly to many computers in a distributed manner. Matching objects can be passed back along the search path (e.g., Freenet [5]), or a direct connection can be established between the original search node and the matching node(s) (e.g., Gnutella [10]).

Message-passing models are a truly distributed solution to searching across multiple databases. There is no central site that can be attacked or otherwise shut down [3]. Each node decides for itself whether it has content that matches the query, which makes it easy to integrate diverse information sources, dynamic information content, and resources that charge for content. It is also possible to offer a search client complete anonymity in a message-passing model [5], although doing so may entail significant communication costs for every node between the original search client and nodes with matching content.

Some characteristics of message-passing retrieval models are likely to limit their effectiveness as models for next-generation Web search systems. Perhaps the most serious is one of scale. If a message-passing model is applied on a large scale, the amount of additional Internet traffic and unnecessary computation is staggering. For example, a *single query* with fanout 4 and search horizon 10 would involve $\Sigma_{i=1}^{10} 4^i = 1,398,100$ nodes. It is difficult to imagine applying this model to millions of queries per day.

Even on a small-scale, message-passing models are *non-deterministic*, because answers depend upon the topology of the network at a particular time, and *non-optimal*, because only a portion of the network is searched for each query. Message-passing models are also prone to *spoofing*, in which a node claims to have matching information when it does not.

The limitations of the message-passing model are opportunities for interesting research. For example, issues of scale, non-determinism, and non-optimality might all be addressed by less random message-passing, while spoofing might be handled by holding each node responsible for its past actions. Whatever the solution, these problems must be addressed before message passing architectures are a viable, large-scale solution to multi-database text retrieval.

4 Centralized Resource Selection

Perhaps the most fully developed multi-database retrieval models are based on centralized database selection. A central site gathers information about available databases, organizes it into an index, and then provides a resource selection service to search clients. A simple resource selection service might just return a ranked list of resources (searchable text databases) to the client. A more complex service might select a set of resources, submit the query to those resources, and return to the search client a single, merged list of documents obtained from multiple resources.

The best-known models for centralized database selection are based on the GlOSS family of algorithms [8], the CORI algorithm [1], and the Cue Validity Variance (CVV) algorithm [16]. The three algorithms differ in their details, but all are based on matching queries to "bag of words" representations of text databases. The "bag of words" resource representation is created automatically, making it relatively easy to add new resources to an existing resource index. The algorithms are general, relatively robust, and do not necessarily require the cooperation of the resource provider [1]. Unlike message-passing architectures, the centralized selection model is computationally efficient, is deterministic, and considers all resources for all queries.

Centralized services can be attacked, legally or illegally, and do not provide the anonymity that message-passing retrieval models provide. They also raise interesting research issues that are less likely to arise with message-passing retrieval models and retrieval models for small-scale environments.

Resource identification: Techniques for crawling the Web are well-defined, but it can be difficult to identify the type of resource at the other end of a hyperlink. For example, the Common Gateway Interface (CGI) hyperlink type indicates an executable resource, but the resource capabilities are not described. The resource might be a text database with search capabilities, a relational database, or a counter. Text associated with the link might or might not indicate something about its capabilities. A crawler that encounters a link to a CGI resource usually has few clues about the type of resource it points to.

Resource identification can be viewed as a document categorization task, for example assigning documents into categories such as "text database", "relational database", and "other" based on multiple forms of uncertain evidence.

Interoperability: Multi-database models designed for small-scale environments and message-passing retrieval models assume that all resources conform to a common interoperability standard. Multi-database retrieval models for large-scale environments are likely to find this assumption violated more often than satisfied. Rather than excluding such resources, it may be possible learn how to communicate with the resource via its Web page, for example, by probing it and examining its responses. Information can be extracted from returned Web pages using *wrapper induction* techniques [11]. Current wrapper induction algorithms are based on supervised learning, but large-scale interoperability requires wrapper induction algorithms based on unsupervised learning.

Resource representation: Content-based resource selection is based on knowing what subject area(s) each resource covers. For example, if the resource is an archive of Wall Street Journal newspaper articles, it might be described as covering the domains of stocks, bonds, international finance, politics, and similar subjects.

If the resource is a text database, and is controlled by a trusted and cooperative party, the STARTS protocol [7] is an effective method of discovering its contents. The STARTS protocol enables a search service and an information resource to exchange information about the contents and capabilities of the resource. If the resource is controlled by an uncooperative party, the contents of the resource can be discovered by running queries and analyzing the documents that are returned (*query-based sampling*) [1].

Both STARTS and query-based sampling are sufficient for bag-of-words resource representations. More complex resource representations, for example, representations based on statistical language models [15], may require new approaches.

Resource selection: Given a set of information resources, content-based resource selection algorithms rank them by the likelihood that they will satisfy the information need described by a query. The GlOSS family of algorithms [8], CORI [1], and Cue Validity Variance (CVV) [16] are three examples that are based on variations of algorithms for ranking documents. The current generation of algorithms can be effective, but there is also evidence that they are fragile when expectations about query length, document length, and resource description quality are not met [2]. Even under ideal conditions, accuracy is considerably lower than theoretical limits [6].

Result merging: After a set information resources is searched, it may be desirable to merge the results into a single display or representation. This problem is difficult when resources are uncooperative. The state-of-the-art is either to rerank the documents at the search client, to merge the rankings using heuristics [1], or to reverse-engineer the ranking functions of the various search engines [12]. Each of these solutions has weaknesses, hence this remains an important, if often overlooked, research problem.

Robust, large-scale centralized resource selection is arguably the most complex of the multi-database retrieval models discussed above, because it requires solutions to several independent problems. Each of the problems outlined above is interesting and challenging because solutions must be general, automatic, and robust, and because they cannot rely on cooperation from information providers. The lack of cooperation, in particular, makes large-scale centralized resource selection both an appealing solution and an interesting research problem.

5 Forecast

All of the multi-database retrieval models surveyed in this paper are likely to play a role in next generation networked information systems. Models optimized for accuracy in small-scale environments will be used in corporate environments, and on the Web in specific subject areas (e.g., to select among databases covering medical research). Research efforts on this class of multi-database retrieval models will focus on improving accuracy and on reducing the amount of manual intervention necessary.

There are two adoption scenarios for message-passing retrieval models. Message-passing is a good choice for dynamic and diverse information content, because the information provider, which knows its content best, determines what information it can provide for each request. However, this same characteristic causes problems such as spoofing; centralized resource selection may provide adequate access and greater reliability. Message-passing is also a good choice when it is important to present no single point of information access, and when anonymity is important. The most obvious examples are theft of intellectual property such as music and software, and exchange of politically sensitive information.

Centralized resource selection is the retrieval model most likely to be the basis for the next generation of large-scale Web search systems. It scales well, it is effective, it provides consistent results, it supports heterogeneous text resources, and it supports heterogeneous query streams. There are research issues that must be addressed before multi-database retrieval models start becoming common, but they appear to be “near term” research opportunities, rather than “long term” research challenges.

Acknowledgements

This material is based on work supported in part by the National Science Foundation under grants IIS-9873009 and EIA-9983253. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor(s).

References

[1] J. Callan. Distributed information retrieval. In W.B. Croft, editor, *Advances in information retrieval*, chapter 5, pages 127–150. Kluwer Academic Publishers, 2000.

[2] J. Callan, A. L. Powell, J. C. French, and M. Connell. The effects of query-based sampling on automatic database selection algorithms. Technical Report IR-181, Center for Intelligent Information Retrieval, Department of Computer Science, University of Massachusetts, 1999.

[3] A. E. Cha. E-power to the people. In *The Washington Post*, page A01, May 18 2000.

[4] A. S. Chakravarthy and K. B. Haase. NetSerf: Using semantic knowledge to find Internet information archives. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Seattle, 1995. ACM.

[5] I. Clarke. Freenet - Front page. http://freenet.sourceforge.net/, August 2000.

[6] J. French, A. Powell, J. Callan, C. Viles, T. Emmitt, K. Prey, and Y. Mou. Comparing the performance of database selection algorithms. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 238–245. ACM, 1999.

[7] L. Gravano, K. Chang, H. García-Molina, and A. Paepcke. STARTS Stanford proposal for Internet meta-searching. In *Proceedings of the ACM-SIGMOD International Conference on Management of Data*, 1997.

[8] L. Gravano, H. García-Molina, and A. Tomasic. GlOSS: Text-Source Discovery over the Internet. *ACM Transactions on Database Systems*, 24(2):229–264, 1999.

[9] D. Hawking and P. Thistlewaite. Methods for information server selection. *ACM Transactions on Information Systems*, 17(1):40–76, 1999.

[10] Wego.com Incorporated. Welcome to gnutella. http://gnutella.wego.com/, August 2000.

[11] N. Kushmerick, D. S. Weld, and R. Doorenbos. Wrapper induction for information extraction. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI-97)*, 1997.

[12] K. Liu, W. Meng, C. Yu, and N. Rishe. Discovery of similarity computations of search engines. In *Proceedings of the 9th International Conference on Information and Knowledge Management (CIKM)*. ACM, 2000.

[13] R. S. Marcus. An experimental comparison of the effectiveness of computers and humans as search intermediaries. *Journal of the American Society for Information Science*, 34:381–404, 1983.

[14] E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proceedings of the Eighteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 172–179, Seattle, 1995. ACM.

[15] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 254–261, Berkeley, 1999. ACM.

[16] B. Yuwono and D. L. Lee. Search and ranking algorithms for locating resources on the World Wide Web. In S. Y. W. Su, editor, *Proceedings of the 12th International Conference on Data Engineering*, pages 164–171, New Orleans, 1996.