Hypertext Information Retrieval for the Web

Eric W. Brown
IBM T. J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598 USA
ewb@us.ibm.com

Alan F. Smeaton
School of Computer Applications
Dublin City University
Glasnevin, Dublin 9, IRELAND
asmeaton@compapp.dcu.ie

Abstract

The notion of searching a hypertext corpus has been around for some time, and is an especially important topic given the growth of the World Wide Web and the general dissatisfaction users have with the tools currently available for finding information on the Web. In response to this, a workshop was held as part of SIGIR'98 on *Hypertext Information Retrieval for the Web* and this document presents a brief summary of the papers presented at that workshop, along with a set of themes identified as a result of group discussion and some conclusions on where to go next.

1 Introduction

The notion of searching a hypertext corpus has been around for a while. Early work in the area includes the development of formal search models that attempt to capture both the content and graph structure of a hypertext [13, 4], automatically constructing hypertexts (e.g., via content-based clustering or citation linking) to incorporate navigation and visualization into the search process [2, 1], and extending traditional IR search techniques to exploit the link relationships in a hypertext (both authored and automatically generated links) [14, 11, 7, 6, 5]. Although this work is interesting, much of it was conducted at a time when hypertexts were smaller, not distributed, and somewhat esoteric.

With the advent of the World Wide Web, hypertext and the linking together of related pieces of information has become ubiquitous. Of course, hypertext on the WWW has drifted somewhat from the original ideal. Hypertext "documents" are written by multiple, independent authors who can create links between pages with indiscretion. Navigational links, citation links, reference links, and just plain confusing links are creatively mixed together and scattered throughout the pages. Page lengths range from a few words to thousands of words, producing a remarkable variety of hypertext "nodes." Consistency is rarely found within a website (let alone the WWW), and often only as an afterthought when management of the site has become problematic. Given the current state of affairs, it is unclear how much of the early work in hypertext information retrieval is applicable to the WWW.

At the SIGIR'98 Workshop on Hypertext Information Retrieval for the Web we investigated in depth the application of information retrieval techniques to hypertext and Web documents. In particular, we sought to answer the question, "Can we improve on content-based search results by exploiting the links, meta-data, and other additional information available on the Web?" We explored models, algorithms, heuristics, and systems that attempt to do just that. The goal of this workshop was to bring together practitioners in the area, as well as anyone interested in deploying and using WWW search technologies, to identify the problems, explore general approaches, discuss recent results, and propose future directions for research, evaluation, and possibly standardization.

The workshop included over thirty participants, nine of whom were selected by the organizers to give presentations.[1] The presentations are summarized in Section 2. The organizers encouraged the speakers to give informal presentations with plenty of room for discussion. Following the presentations we held a roundtable session during which we attempted to distill and summarize the key ideas and issues brought out during the presentations. The roundtable discussion produced a number of themes, which we discuss in Section 3. Following that, we offer some concluding remarks in Section 4.

2 Presentations

The workshop opened with a presentation by Eric Brown, which defined the background to the area and summarized the most important work reported to date. Brown described a number of common challenges encountered when searching the Web, identified opportunities for improving search results given the additional information available in a hypertext environment, and reviewed a number of approaches to exploiting this information that have been pursued in the literature.

After the preliminary context presentation, Jamie B. Teevan from Yale University and InfoSeek Corp. reported on work that attempts to marry the respective strengths of keyword searching, which is computationally fast and wide ranging, with manually created and maintained directory pages. In Teevan's work, information seeking on the web is a 2-stage process. An initial search of a range of popular search engines is used to locate *directory pages*, which are defined as having a higher ratio of links to pages on other sites than to pages on the same site. Once the directory pages have been located in the initial search, the pages they in turn link to are used to establish a meta-directory, a kind of virtual amalgamation of the content of already existing directory pages. This approach to information seeking can be seen as leveraging link information from the web for locating information and, in particular, taking advantage of the peculiar WWW phenomenon of *directory pages*—manually maintained pages of meta-information.

A team from IBM Almaden described a system called ARC which uses a technique called *spectral filtering* for resource discovery. This technique uses both the contents of a page and its context to measure the usefulness of a page, where context is defined in terms of the pages it points to, the pages that point to it, and the neighborhood of the web in which it appears. This is a derivative of Kleinberg's seminal work on *hubs* and *authorities* [10] but extends beyond the abstract notions that Kleinberg pursued, exploring the characteristics that are idiosyncrasies of the Web, such as self-promotion of pages through same-site links or the emergence of one hub which encompasses all the links of another. The contribution from Chakrabarti et al. includes a report of a preliminary investigation into the effectiveness of ARC, measuring it favorably against *AltaVista* and *Yahoo!*.

At the BT Laboratories in the UK, John Davies and Scott Stewart are pursuing research where the usefulness of a set of pages in terms of a user's query has been determined manually and metrics such as the Term Frequency score, number of pages linking into and from a page, text used as anchors for such links, the scores of linked pages, the title and its similarity to the query, length of URL and of document, presence/absence of spamming and images, and the kind of presentation formatting used in the page, are all recorded. Once all this information has been computed for pages manually assessed for relevance, this information is used to train a neural network which can combine the values of these measures in subsequent queries.

Einat Amitay has investigated the nature of the text used in web pages, comparing it against text from more conventional documents and noting the significant differences. Her approach to computing the usefulness of a page in terms of a query is to examine not just the documents whose links point to a given page,

[1] Many of the presentations and papers are available at http://lorca.compapp.dcu.ie/SIGIR98-wshop/program.html.

but also the anchor text in those pages pointing to a given page. This novel work is an extension of the now almost conventional approach of just using hypertext links between documents and on into utilizing how that link has been anchored in a document. Like much of the work reported at the workshop this is at the formative stage but seems to have good possibilities.

Sue Dumais from Microsoft Research gave a presentation which recalled previous work she had done over a number of years on systems which had exploited structure and relationships between retrieved items, much as we were trying to do in the workshop. Dumais pointed out that two particular kinds of structure which interested her were link structure, as we have in web documents, and usage patterns at the individual or community level such as we get in collaborative filtering. Dumais did have a strong warning, however, about the importance of interface design in any retrieval interface. Simply adding retrieval features to a retrieval system is not necessarily good design and user testing and user-centered design is almost as important as the development of new techniques.

The workshop on Hypertext Information Retrieval and the Web had the pleasure of an address by William I. Chang, Chief Technology Officer at Infoseek Corp. Chang traced the development of InfoSeek into its current position as a major search engine vendor and provider of enterprise and consumer online services. This divergence into other activities besides search provision is common among such companies and is driven by the stiff competition to attract Internet users (and ultimately Internet advertisers).

Andrea Michalek and Douglas Grundman from Infonautics Corp. highlighted the navigational distinctions between searching and browsing, pointing out that a search function takes nodes and generates links for traversal while the traverse function takes links and generates nodes. An argument put forward is that if links can be used to improve a search, then search information should be exploitable to improve link traversal. Particular manifestations of this would include explaining the reasons why documents are retrieved, highlighting relevant parts of long documents, illuminating related metadata, and propagating search information along subsidiary links. An implementation of this is described.

The final presentation by Daan Velthausz and Henk Eertink explored how the retrieval operation for searching the web may be parameterized by real constraints such as budget and time. Assuming that information providers will eventually have to charge customers for searching and/or retrieving information from their sites, a need will emerge for search paradigms that operate within budgets, such as *retrieve information on X within 1 minute and within a search budget of $5*. The authors describe their approach to developing such search strategies.

3 Themes

The presentations reported on a variety of techniques for exploiting the additional information available on the Web and addressing the end user issues that are peculiar to a Web environment. A number of themes arose out of these presentations and the discussion that ensued.

3.1 Tacit Collaboration

One of the strongest themes that emerged from nearly all of the presentations was an effort to exploit the *tacit collaboration* that takes place on the Web. Tacit collaboration occurs when individuals work independently (i.e., without central coordination) in a distributed, networked environment, and the results of their work can be combined to produce more value than when the individual efforts are considered separately. There are a number of examples of this, of which we describe a few below.

The most common example is when independent web page authors create hypertext links to other pages on the Web. The primary, pedestrian purpose of these links is to let the reader navigate to other pages of

related material. When all of the inbound links to a given page are taken as a whole, however, we can infer a great deal of information about the page. For example, a web page with many inbound links from independently authored pages is likely to be more authoritative than a page on the same subject with very few inbound links. By linking to someone else's page, an author tacitly indicates approval for that page. When many independent authors express the same approval, we have strong evidence about the quality and value of the page.

Another example of tacit collaboration is the simple act of retrieving a page on the Web. When this act is recorded in a web server log, we end up with usage statistics for the page. From these statistics we can make inferences about the popularity of a page, as well as trends in user interests. Of course, web server logs are not always publicly available, so realization of their full potential is limited. Within certain domains, however, such as a corporate intranet, they can be quite valuable.

One last way to exploit implicit collaboration is through sharing and analysis of bookmarks. Users often bookmark high quality web pages and even organize their bookmarks into hierarchical groups of related pages. When bookmark files are shared, users can take advantage of the cataloging efforts of their colleagues, and more sophisticated analysis can be applied to extract communities of interest and connect users to others with similar interests. Again, the application of these techniques may be restricted to intranets where access to user bookmark files can be provided in a controlled, secure fashion.

The assumptions that must be made regarding the collaborators are a significant challenge to exploiting tacit collaboration. For example, hyperlinks are rarely annotated with semantic information even though there are many different kinds of links—a link that navigates to the next page in a document is clearly different from a link that references the authoritative source on a subject. Much of the work in this area involves developing robust techniques for extracting and interpreting examples of tacit collaboration.

3.2 Enhanced User Experience

The second major theme in the workshop was the problem of how to enhance the end user's experience. There was a general consensus that current user interfaces for finding information on the Web leave a great deal of room for improvement. One shortcoming in particular is the lack of rich user task models. For example, information gathering, question answering, and shopping are distinct user search tasks that require different system interactions and different kinds of results. All too often the user is forced to use a hammer (brute force search) when some other tool would be far more useful.

The ultimate goal is to define distinct task models that allow the search system to tailor its interaction with the user and better satisfy the user's information need by focusing on web pages that are more likely to be relevant given the current task model. Although this may seem like a restatement of the basic search problem, the real issue is the tremendous variety of information sources available on the Web and the many different search tasks being performed by users. The current model of trying to satisfy all of these search tasks using search techniques developed for traditional, more homogeneous document collections simply falls short.

Another area where the user experience can be enhanced is in building interfaces that more naturally combine searching, browsing, and web structure. Search results can be much more informative when presented such that the underlying structure of the documents is exposed to the user. This structure could be the hyperlink structure of the documents or an externally imposed structure such as a categorization taxonomy. Similarly, providing the end user with site-level understanding of a web site, either in the context of a search result or through the use of a site map, can facilitate the user's navigation to useful information.

3.3 Metadata/XML

The third theme identified at the workshop was the need for better exploitation of metadata and document structure. In particular, XML represents a significant opportunity in this area. We acknowledged that the metadata community was making considerable progress (e.g., Dublin Core [15], Warwick Framework [12], and XML/RDF [3]) but recognized that adaptation of these standards on the Web at large is slow and much work remains to be done before search tools will fully exploit this new source of information. Better representation and exploitation of metadata will also aid user task modeling and provide significant support for certain end user tasks, such as shopping.

3.4 Measurement and Evaluation

The last theme brought out at the workshop was the issue of how to measure and evaluate the various techniques being applied to the problem of finding information on the Web. Evaluation of information retrieval systems in traditional environments is already difficult and contentious; the Web only confounds the evaluation problem with its size and new definition of "relevance"—the value of a page is no longer measured by just its content, but also by its hyperlink proximity to other relevant pages. Furthermore, since the information finding process is a combination of search and navigation, the user interface plays an even more critical role.

Some efforts are underway to address this problem. Harman and Over [8] sponsored a panel on tools for searching the Web that provides some insight into the complex process of answering questions on the Web. The Very Large Corpus track at TREC [9] is based heavily on Web data and may evolve into a Web track. Much work remains, however.

4 Conclusions

The SIGIR'98 Workshop on Hypertext Information Retrieval for the Web achieved its goal of bringing together researchers and practitioners to discuss current issues and approaches related to searching the Web. In particular, we saw new approaches for finding information on the Web as well as modifications of earlier hypertext approaches enhanced for use on the Web. We recognized that a number of characteristics distinguish the Web from earlier hypertext collections, including size, number of independent authors, distribution, and incredible variety of information. At a minimum, these characteristics force a review of the applicability of early hypertext search techniques, and ultimately have caused the evolution of many new techniques.

The work described here shows both progress and opportunity. We anticipate vigorous activity in this area for some time as these opportunities are pursued.

References

[1] J. Allan. *Automatic Hypertext Construction*. PhD thesis, Cornell University, Ithaca, NY, 1995.

[2] M. Bernstein. An apprentice that discovers hypertext links. In *Proc. European Conf. on Hypertext (ECHT)*, pages 212–223, 1990.

[3] D. Brickley, R. V. Guha, and A. Layman. Resource description framework (RDF) schema specification. W3C Working Draft, Oct. 1998. http://www.w3.org/TR/WD-rdf-schema/.

[4] Y. Chiaramella and A. Kheirbek. *Information Retrieval and Hypertext*, chapter An Integrated Model for Hypermedia and Information Retrieval, pages 139–178. Kluwer Academic Publishers, Boston, 1996.

[5] W. B. Croft and H. R. Turtle. Retrieval strategies for hypertext. *Inf. Process. & Mgmnt.*, 29(3):313–324, 1993.

[6] H. P. Frei and D. Stieger. Making use of hypertext links when retrieving information. In *Proc. ACM European Conf. on Hypertext (ECHT)*, pages 102–111, 1992.

[7] M. E. Frisse. Searching for information in a hypertext medical handbook. *Commun. ACM*, 31(7):880–886, July 1988.

[8] D. Harman and P. Over. Panel: Tools for searching the web. In *Proc. of the 21st Inter. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 332, Melbourne, Aug. 1998.

[9] D. Hawking and P. Thistlewaite. Overview of trec-6 very large collection track. In E. M. Voorhees and D. K. Harman, editors, *The Sixth Text REtrieval Conference (TREC-6)*, pages 93–105, Gaithersburg, MD, 1998. National Institute of Standards and Technology Special Publication 500-240. http://trec.nist.gov/pubs/trec6/papers/vlc_track.ps.

[10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *ACM Symp. on Discrete Algorithms*, 1998. Also appeared as IBM Research Report RJ10076, May 1997.

[11] K. L. Kwok. A probabilistic theory of indexing and similarity measure based on cited and citing documents. *J. Amer. Soc. Inf. Sci.*, 36:342–351, 1985.

[12] C. Lagoze, C. A. Lynch, and J. Ron Daniel. The warwick framework: A container architecture for aggregating sets of metadata, June 1996.
http://www.ifla.org/documents/libraries/cataloging/metadata/tr961593.pdf.

[13] D. Lucarella and A. Zanzi. Information retrieval from hypertext: An approach using plausible inference. *Inf. Process. & Mgmnt.*, 29(3):299–312, 1993.

[14] G. Salton. Associative document retrieval techniques using bibliographic information. *J. ACM*, 10(4):440–457, Oct. 1963.

[15] S. Weibel, J. Godby, and E. Miller. Oclc/ncsa metadata workshop report, 1995.
http://www.oclc.org:5046/conferences/metadata/dublin_core_report.html.