# Integrating Content Search with Structure Analysis for Hypermedia Retrieval and Management

**Wen-Syan Li** and **K. Selçuk Candan**

*C&C Research Laboratories.*, *NEC USA Inc.*

110 Rio Robles, M/S SJ100, San Jose, CA, 95134, USA
{wen,candan}@ccrl.sj.nec.com

**Abstract:** Hypertext and hypermedia have emerged as primary means for structuring documents and for accessing the Web. It has two aspects: content and structural information. It is believed that integration of content search with structure analysis can improve hypermedia document retrieval and management; consequently increasing information utilization. This research topic has emerged as a main focus of many recent conferences and research communities, including computer-human interface, information retrieval, hypertext, databases, and Web. In this paper, we summarize the state-of-the-art techniques and functionalities of integrating content search with structure analysis for Web documents retrieval and management.

Categories and Subject Descriptors: H.3.1 Information storage and retrieval Content Analysis and Indexing [Abstracting methods] H.3.3 Information storage and retrieval Information Search and Retrieval [Information filtering]

General Terms: Hypertext, Web

Additional Key Words and Phrases: Link analysis, topic distillation, organization

# 1. Introduction

Hypertext and hypermedia have emerged as primary means for structuring documents and for accessing the Web. With the explosive growth of the WWW, most searches retrieve a large number of documents. The Web has no aggregate structure which organizes distinct localities. Users have no global view of the entire Web from which to forage for relevant pages. As a solution to these challenging issues, integration of content search and structure analysis has been proposed by many research communities, including computer-human interface [Pitkow 1996], information retrieval [Gudivada 1997], [Khan 1998], and databases [Chaudhuri 1998], [Florescu 1998]. It is believed that such an integration can improve hypermedia document retrieval and classification, as suggested by the study [Chakrabarti 1998], [Gibson 1998]; consequently increasing information utilization. This research topic has emerged as a main focus of many recent conferences on hypertext and WWW, including the first ACM Digital Library Workshop on Organizing Web Space held in Berkeley, California, USA in August 14, 1999. In this paper, we summarize the state-of-the-art techniques and functionalities of integrating content search with structure analysis for hypermedia retrieval and management.

# 2. Organizing Web query space

Although search engines are one of the most popular methods to retrieve information of interest from the Web, they usually return thousands of URLs that match the user-specified query terms. Due to the information overload problem and the mix of useful information and low quality data, it is essential to support filtering and organization functionalities. Many systems and techniques have been proposed to utilize links to determine document associations. Connectivity relationships indicate that two documents are relevant to each other if there are links between them. Co-citation relationships indicate that two documents are relevant to each other if they are linked via a common document. Social filtering relationships indicate that two documents are relevant to each other if they link to a common document. A transitivity relationship can be derived if the user can navigate from one document to another via a sequence of links. Such relationships have been used by many search engines to rank query results. They assume that the quality of a document can be "assured" by the number of links pointing to it.

An interesting approach to organizing Web query results is "topic distillation" proposed by J. Kleinberg [Kleinberg 1998]. This work aims at selecting a small subset of the most "authoritative" pages from a much larger set of query result pages. An authoritative page is a page with many incoming links and a hub page is a page with many outgoing links. Such authoritative pages and hub pages are mutually reinforced: good authoritative pages are linked by a large number of good hub pages and vice versa. This technique organizes topic spaces as a smaller set of hub and authoritative pages and provides an effective means for summarizing query results.

Bharat and Henzinger [Bharat 1998] improved the basic topic distillation algorithm [Chakrabarti 1998a] by adding additional heuristics. The modified topic distillation algorithm considers only

those pages that are in different domains with similar contents for mutual authority/hub reinforcement. Preliminary experiments show improved results by introducing similarity measurements of linked documents.

Another major variation of the basic topic distillation algorithm is proposed by Brin and Page [Brin 1998]. Their algorithm further considers page fan out in propagating scores. Topic distillation has been applied to many search engines, including *Google* [Google Search Engine], NEC NetPlaza [NEC 1999], and IBM Clever [Chakrabarti 1998b]. Many of these basic and modified topic distillation algorithms have been also used to identify latent Web communities [Gibson 1998a], [Kumar 1999].

Links are also useful information for efficient crawling of high quality documents on the Web. Thus, in-degree and out-degree are important parameters for crawlers in determining path exploration strategies. For example, a crawler initiated to find information on `Pokemon` may apply the following heuristics: "the pages linked by more pages whose contents are related to `Pokemon` are more likely to be related to `Pokemon` when compared to those that do not". An advanced crawling technique to find documents as well as patterns to improve crawling strategies is dual iterative pattern relation expansion (DIPRE) [Brin 1998a]. Chakrabarti et al. [Chakrabarti 1999b] proposed a new approach to topic-specific Web resource discovery: *focused crawler*. Provided with example documents, a focused crawler analyzes its crawl boundary to find the links that are likely to be more relevant for the crawl, and avoids irrelevant regions of the Web.

# 3. Facilitating search, navigation, and associating Web pages

A substantial amount of work has been carried out by the database community to augment traditional SQL, the database query language for content search, with path expressions for link traversal. Florescu et al. [Florescu 1998] surveyed existing work and categorized SQL languages/systems with such extended capabilities as the first generation of Web query languages and systems. Today, almost every institution has its home Web site, with links connecting internal pages together. With huge amounts of information to be displayed and maintained, manually authoring and maintaining institution Web sites is almost infeasible. The Web pages of many of large sites are generated by programs, and database systems are used as backend data providers. A second generation of Web query languages and systems, defined by Florescu et al. [Florescu 1998], extends the first generation by providing data organization, link generation, and document/structure authoring capabilities. Representative systems include Strudel [Fernandez 1998] and WebOQL [Arocena 1998].

Another effort by database researchers is to provide more flexible and efficient search on hypermedia. Existing Web search engines return only "physical" pages containing query terms. Since the WWW encourages hypermedia document authoring, authors tend to to create Web documents that are composed of multiple pages connected with hyperlinks. Consequently, a Web

document for a topic may span multiple pages and be authored in many different ways. Li and Wu [Li 1999] introduced the concept of *information unit*, which can be viewed as a logical Web document consisting of multiple physical pages as one atomic retrieval unit. A framework of query relaxation by structure is proposed. In this framework, a set of connected physical pages which, as a whole, contains all query terms can be retrieved. This framework supports desirable progressive processing for Web queries, i.e. it generates the best K results in the order of ranking. Such properties are essential since exploring Web structures is an expensive task. The work by Tajima et al. [Tajima 1999] extends the concept of information units by considering keyword occurrence frequency and distribution. Another effort, by Goldman et al. [Goldman 1998] to support efficient proximity search in an XML document database is to establish so-called hub nodes as a way of indexing for structure-based search.

Another approach to assisting users in navigating the Web is to provide related pages or recommend paths for traversal. Dean and Henzinger [Dean 1999] proposed two algorithms, *companion* and *co-citation* to identify related pages, and compared their algorithms with the Netscape algorithm [Netscape 1999] used to implement the `What's Related?` function. By extending the scope from documents to Web sites, Bharat and Broder [Bharat 1999] conducted a study to compare several algorithms for identifying mirrored hosts on the Web. The algorithms operate on the basis of URL strings and linkage data: the type of information easily available from web proxies and crawlers. Similarly, researchers in the artificial intelligence community have developed Web navigation tour guides, such as WebWatcher [Joachims 1997]. WebWatcher utilizes user access patterns in a particular Web site (i.e. paths) to recommend to users proper navigation paths for a given topic.

# 4. Concluding remarks

Hypermedia has two aspects: content and structural information. Researchers in various fields have been adapting and developing techniques for making hypermedia space, like the World-Wide Web, more usable. The current efforts have been focused on hypermedia modeling, visualization, information retrieval, computer human interaction, and query processing. In this paper, we summarized the recent research and development trends by functionality. We observed that significant technology evolution is by means of integrating content search and structural analysis for hypermedia organization and management. With the introduction of XML [Suciu 1998] and the rapid information growth on the Web, we believe that the demand and the importance of research in this area will continue.

## Bibliography


**[Arocena 1998]**

Gustavo Arocena and Alberto Mendelzon. "WebOQL: Restructuring Documents, Databases, and Webs" in Proceedings of the 14th International Conference on Data Engineering, Orlando, Florida, 24-33, February 1998.

**[Bharat 1998]**

Krishna Bharat and Monika R. Henzinger. "Improved algorithms for topic distillation in a hyperlinked environment" in Proceedings of ACM SIGIR '98, Melbourne, Australia, 104-111, [Online: ftp://ftp.digital.com/pub/DEC/SRC/publications/monika/sigir98.pdf], August 1998.

**[Bharat 1999]**

Krishna Bharat and Andrei Z. Broder. "Mirror, Mirror, on the Web: A Study of Host Pairs with Replicated Content" in Proceedings of the 8th World-Wide Web Conference (Toronto, Canada, May 1999), 1999.

**[Brin 1998]**

Sergey Brin and Lawrence Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine" in Proceedings of World-Wide Web '98 (WWW7), [Online: http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm], April 1998.

**[Brin 1998a]**

Sergey Brin, Rajeev Motwani, Lawrence Page, and Terry Winograd. "What can you do with a Web in your packet?" in Bulletin of the Technical Committee on Data Engineering, 21(2), 37-47, June 1998.

**[Chakrabarti 1998]**

Soumen Chakrabarti, Byron E. Dom, and Piotr Indyk. "Enhanced Hypertext Categorization using Hyperlinks" in Proceedings of ACM SIGMOD '98, [Online: http://www.cs.berkeley.edu/~soumen/sigmod98.ps], 1998.

**[Chakrabarti 1998a]**

Soumen Chakrabarti, Byron E. Dom, Prabhakar Raghavan, Sridhar Rajagopalan, David Gibson, and Jon M. Kleinberg. "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text" in Proceedings of World-Wide Web '98 (WWW7), Brisbane, Australia, 65-74, [Online: http://www7.scu.edu.au/programme/fullpapers/1898/com1898.html], April 1998.

**[Chakrabarti 1999b]**

Soumen Chakrabarti, Martin van den Berg, and Byron E. Dom. "Focused Crawling: A New Approach for Topic-Specific Resource Discovery" in Computer Networks, 31:1623-1640, 1999. First appeared in Proceedings of the Eighth International World Wide Web Conference, Toronto, Canada, [Online: http://www8.org/w8-papers/5a-search-query/crawling/index.html], May 1999.

**[Chaudhuri 1998]**

Surajit Chaudhuri (editor). Special Issue on Databases and the World Wide Web, Bulletin of the IEEE Technical Committee on Data Engineering, 21(2), June 1998.

**[Netscape 1999]**

Netscape Communications Corporation. What's Related web page. Information available at http://home.netscape.com/netscapes/related/faq.html.

**[Dean 1999]**

J. Dean and Monika R. Henzinger. "Finding Related Pages in the World Wide Web" in Proceedings of the Eighth World-Wide Web Conference, Toronto, Canada, May 1999.

**[Fernandez 1998]**

Mary F. Fernandez, Daniela Florescu, Jaewoo Kang, Alon Y. Levy, and Dan Suciu. "Catching the Boat with Strudel: Experiences with a Web-Site Management System" in Proceedings of ACM SIGMOD '98, Seattle, WA, 414-425, June 1998.

**[Florescu 1998]**

Daniela Florescu, Alon Levy and Alberto Mendelzon "Database techniques for the world-wide web: A survey" in SIGMOD Record 27, 3, 59-74, 1998.

**[Gibson 1998]**

David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. "Clustering categorical data: An approach based on dynamic systems" in Proceedings of the 24th International Conference on Vary Large Databases, September, 1998.

**[Gibson 1998a]**

David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. "Inferring Web Communities from Link Topology" in Proceedings of ACM Hypertext '98, Pittsburgh, PA, 225-234, June 1998.

**[Goldman 1998]**

Roy Goldman, Narayanan Shivakumar, Surish Venkatasubramanian, and Hector Garcia-Molina. "Proximity Search in Databases" in Proceedings of the 24th International Conference on Very Large Data Bases (New York City, New York, Aug. 1998), pp. 26-37. VLDB.
Google Search Engine. Information available at http://google.stanford.edu/, 1998.

**[Google Search Engine]**

Google Search Engine. Information available at http://google.stanford.edu/.

**[Gudivada 1997]**

Venkat N. Gudivada, Vijay V. Raghavan, William I. Grosky, and Rajesh Kasanagottu. "Information Retrieval on the World Wide Web" in IEEE Internet Computing 1(5), 58-68, 1997.

**[Joachims 1997]**

Thorsten Joachims, Dayne Freitag, and Tom Mitchell. "Webwatcher: A Tour Guide for the World Wide Web" in Proceedings of the IJCAI '97, August, 1997.

**[Khan 1998]**

Kushal Khan and Craig Locatis. "Searching Through Cyberspace: The Effects of Link Cues and Correspondence on Information Retrieval from Hypertext on the World Wide Web" in Journal of the American Society for Information Science 49, 14, 1248-1253, 1998.

**[Kleinberg 1998]**

Jon M. Kleinberg. "Authoritative sources in a hyperlinked environment" in Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 668-677, [Online: http://www.cs.cornell.edu/home/kleinber/auth.ps], January 1998.

**[Kumar 1999]**

S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. "Trawling the Web for Emerging Cyber-Communities" in Proceedings of the 8th World-Wide Web Conference, Toronto, Canada, May, 1999.

**[Li 1999]**

Wen-Syan Li and Yi-Leh Wu. "Query Relaxation By Structure for Document Retrieval on the Web" in Proceedings of 1998 Advanced Database Symposium, Shinjuku, Japan, December, 1999.

**[NEC 1999]**

NEC Corporation. NetPlaza Search Engine. Information available at http://netplaza.biglobe.ne.jp/ .

**[Pitkow 1996]**

James E. Pitkow and Colleen M. Kehoe. "Emerging Trends in the WWW User Population" in Communications of the ACM (CACM), 39(6) 106-108, June 1996.

**[Suciu 1998]**

Dan Suciu. "Semistructured data and XML" in Proceedings of 5th International Conference

of Foundations of Data Organization (FODO'98), Kobe, Japan, November, 1998.

**[Tajima 1999]**

Keishi Tajima and Kenji Hatano and Takeshi Matsukura and Ryoichi Sano and Katsumi Tanaka. "Discovery and Retrieval of Logical Information Units in the Web" in Proceedings of the 1999 ACM Digital Libraries Workshop on Organizing Web Space, Berkeley, CA, USA, August, 1999.