Frontiers in Web Data Management

Junghoo "John" Cho
UCLA Computer Science Department
Los Angeles, CA 90095
cho@cs.ucla.edu

Abstract

In the last decade, the Web has become a primary source of information for many people. Due to its ease and ubiquity, many people first look up pages on the Web whenever they need to look up certain information. The popularity of the Web, however, has brought many interesting challenges. In particular, the ever-expanding size of the Web makes it increasingly difficult to discover, store, organize and retrieve the information on the Web to help users to identify what they are looking for. In this paper, we will briefly go over some of the new challenges in this context and describe our research efforts to address them.

1 Introduction

In the last decade, we have witnessed a tremendous proliferation of information on the World-Wide Web [15, 16]. The World-Wide Web was initially developed to help scientists exchange their new results and findings quickly, but after a decade of exponential growth, it is now being used by millions of world-wide users on a daily basis [3]. As its success and proliferation indicate, the plentiful information of the World-Wide Web is useful for many people. Users access the Web for a variety of purposes, sometimes for casual browsing on a general topic, and sometimes for focused research on a particular problem.

The very success of the Web, however, has also introduced many new challenges that need to be addressed:

- *Information overload*: The success of the Web is partly due to its ease of publication. Any user can publish any information on the Web at any time. This ease has led to the exponential growth of the Web in the last decade, which has simultaneously brought the problem of "information overload." For many topics, there simply are too many Web pages that are potentially relevant, so the users waste significant time and efforts reading irrelevant pages. For example, when we issued the query "database" to Google, one of the major search engines, it returned 46 million Web pages as potentially relevant to the topic. Given the limited time that a user has, the user cannot read all 46 million pages. How can the user identify which page will answer the particular question that the user has? Given the projected growth rate of the Web, this problem will certainly get worse over time.

- *Information transience*: Information on the Internet is inherently ephemeral. It is distributed over millions of servers that are administered by a multitude of organizations. New information is constantly being posted on a server, and existing information is continuously being updated, or even deleted, at an administrator's whim, regardless of how important the information is. This transience of information often causes frustration to many users, since some of the information that they like and frequently access may suddenly disappear. Once the information disappears, users have to waste significant time and efforts looking for similar information from other sources on the Web, or it may be even impossible to access such information any more. There is no guarantee that a particular information will be available for later access.

- *Information bias*: As the Web becomes the primary source information for many people, it starts to introduce a significant bias on people's perception. Since some types of information has more presence on the Web than others, they are discovered and looked at by many people from the Web and introduce a significant influence on how people perceive the world. For example, when we issued the query "Jaguar" to Google the top result was a page on a new version of Mac operation system. Also, the top-10 results were all about either the Mac operating system or Jaguar the car. Does it mean that the original meaning of Jaguar the cat does not have any relevance in the current society?

  Similarly, a large part of the Web is not currently indexed by major search engines due to their technological limitations. This part of the Web is often called the "Hidden Web" because many users rely on search engines to access Web pages, so when the pages are not returned from search engines, they are essentially "hidden" or "inaccessible" to these users. Ho much bias is being introduced because of the inaccessibility of the Hidden Web? Is there a way to reduce this bias?

In our research group, we are conducting various research projects that address some of the above challenges. In the following two sections, we describe our WebArchive project and the Hidden-Web project in more detail.

2 Archiving the Web

In order to alleviate the information transience problem, our WebArchive project tries to build a system that can store and archive the history and evolution of the Web: tracking the changes of the Web, storing multiple versions of Web documents in a concise way, and providing the archived information to users through an intuitive interface. An effective archive system can significantly benefit multiple disciplines in the following ways:

- *Archive of Human Knowledge:* As the Web becomes more popular and widespread, an increasing number of people rely on the Web as their primary source of information. Also, a significant amount of information is available only on the Web in digital form. Therefore, once information disappears from the Web, some of the information may be permanently lost. Unless we archive the constantly changing Web over a long period of time, we may lose information that has taken decades to discover.

- *Web Research Testbed:* Constant changes of Web documents pose many challenges to Web caches, Web crawlers and network routers. Because a Web cache or Web crawler does not know when or how often a Web page changes, it has to download and/or cache the same page multiple times, even if the page has not changed. A large body of research has been conducted to address this challenge and innovative new algorithms and protocols have been developed. Due to the lack of Web change data, however, it has been difficult to validate the algorithms in practice. A central archive of Web history will provide valuable Web change data and will work as a research testbed where researchers can develop and validate new ideas.

- *Study of Knowledge Evolution:* New topics and/or genres grow in popularity and suddenly attract interest from a large number of people. Yet we often don't understand exactly when the topics started or how they became suddenly popular. For example, the Linux project became hugely popular over the last decade. Why has it become so popular? How did the community start? How did people discover the project?

  While answering these questions is not easy, we may get a better understanding by analyzing the history of Web documents. For instance, if we wanted to study the evolution of the Linux project, we could go back to the Web 10 years ago and study what pages were mentioning "Linux" at that time, how the pages were linked to each other, and in what sequence they were created. This analysis might reveal how the community developed over time.

In order to build an effective Web archive system, we are trying to address the following challenges that arise from the distributed nature of the Web:

- *Independent Change:* The information sources on the Internet are updated autonomously and constantly. Since there exist many more changed pages than an archive system can download, the system should "guess" how much and how often the pages are updated, and intelligently decide which updated pages to download. Unless it uses its limited download and storage resources efficiently, it may miss important changes of pages.

- *Scale of Data:* A Web archive system must download and store an enormous amount of data. Textual data on the Web is estimated to be more than 30 terabytes [5, 15, 16], and is constantly being updated [12, 4, 10, 8]. In order to handle the sheer scale of data, the archive system has to employ novel techniques to store, organize and compress the Web history data.

- *Intuitive Access:* The archive system must be intuitive for users to search, browse, and analyze. In addition, it should be able to handle diverse queries that users may pose. For example, a user may simply want to browse multiple versions of a particular Web page, or the user may want to pose a complex query, such as "What are the 10 topics whose popularity (measured in the number of pages mentioning the topic) has increased most rapidly in the last six months?" In order to handle these queries, the system must employ novel indexing and query processing techniques.

Our project tries to address some of the major technical challenges to building an effective Web archive system — a fully scalable and easy-to-use system that can handle the dynamic Web. To attain this goal, we are currently investigating the following three main research issues.

1. *Efficient Change Detection:* We are designing efficient change detection and download algorithms that can identify the change characteristics of Web pages. The new algorithms will efficiently use limited download resources, minimizing the loss of information.

2. *Efficient Storage:* We are developing effective ways to store and organize the Web history data efficiently and compactly. Multiple versions of the same Web page have tremendous potential for efficient storage and compression because changes to a Web page are often minor.

3. *Effective Access:* We are developing appropriate index structures and query processing techniques on the Web history data, so that users can intuitively express their queries and the system can handle the queries efficiently.

As part of the project we are currently building a Web archive prototype; it will store a monthly change history of Web pages related to computer science. While this is a relatively small subset of the whole Web, this prototype will work as a proof-of-concept of the various techniques that we will develop. This prototype will also provide a valuable real history dataset that researchers can test and verify their ideas with. We believe a successful completion of this research will provide both invaluable real Web history data and the technologies that can be useful in managing and archiving any evolving textual database.

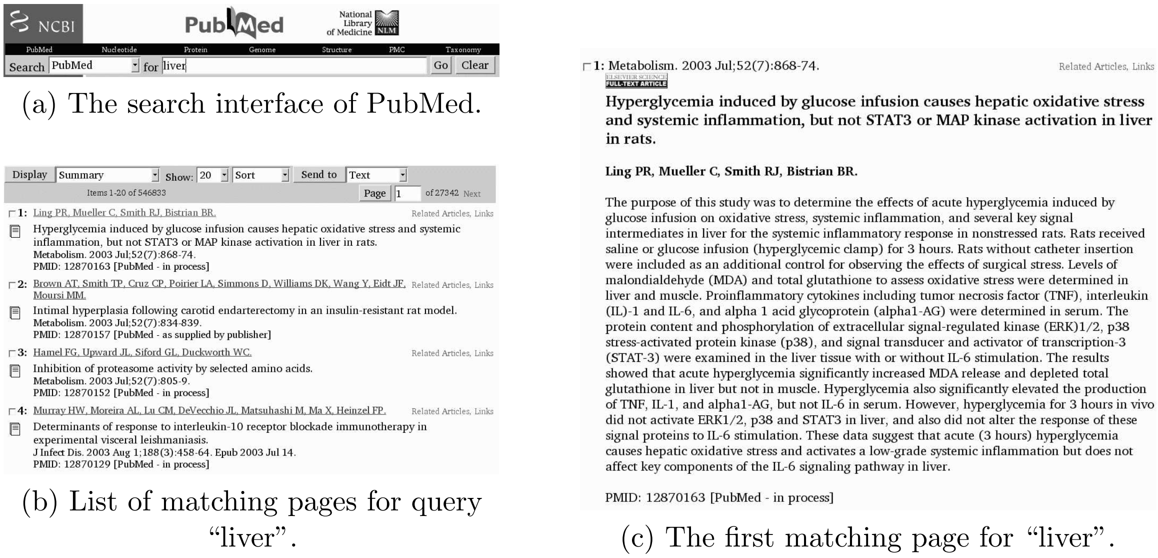
3 Surfacing the Hidden Web

An ever-increasing amount of information on the Web is available only through search interfaces. That is, users have to type a set of keywords, or *queries*, into a search interface in order to access Web pages from certain Web sites: The sites do not provide any static links to their pages. Because Web crawlers[1] simply follow links on the Web to discover pages, search engines cannot download these pages, often referred to as the *Hidden Web* or *Deep Web* [2, 7].

Figure 1 shows an example of the process that a typical user goes through to access pages in the Hidden Web. For every site that the user is interested in, she types a list of keywords to its search interface (an example shown in Figure 1(a)), the site returns a list of potentially relevant pages (Figure 1(b)), and the user clicks on some of the links to retrieve the actual pages (Figure 1(c))

Since the majority of Web users rely on traditional search engines to discover and access information on the Web, the Hidden Web is practically inaccessible to most users and "hidden" from them. When the pages are not returned from major search engines, users simply give up and ignore those pages. Even if users are aware of a certain part of the Hidden

[1] A Web crawler is a program that downloads Web pages for search engines



Search PubMed for liver

(a) The search interface of PubMed.

Summary Show: 20 Sort Send to Text

Items 1-20 of 546833 Page 1 of 27342

1: Ling PR, Mueller C, Smith RJ, Bistrian BR. Related Articles, Links
Hyperglycemia induced by glucose infusion causes hepatic oxidative stress and systemic inflammation, but not STAT3 or MAP kinase activation in liver in rats.
Metabolism. 2003 Jul;52(7):868-74.
PMID: 12870163 [PubMed - in process]

2: Brown AT, Smith TP, Cruz CP, Poirier LA, Simmons D, Williams DK, Wang Y, Eidt JF, Moursi MM. Related Articles, Links
Intimal hyperplasia following carotid endarterectomy in an insulin-resistant rat model.
Metabolism. 2003 Jul;52(7):834-839.
PMID: 12870157 [PubMed - as supplied by publisher]

3: Hamel FG, Upward JL, Siford GL, Duckworth WC. Related Articles, Links
Inhibition of proteasome activity by selected amino acids.
Metabolism. 2003 Jul;52(7):805-9.
PMID: 12870152 [PubMed - in process]

4: Murray HW, Moreira AL, Lu CM, DeVecchio JL, Matsuhashi M, Ma X, Heinzel FP. Related Articles, Links
Determinants of response to interleukin-10 receptor blockade immunotherapy in experimental visceral leishmaniasis.
J Infect Dis. 2003 Aug 1;188(3):458-64. Epub 2003 Jul 14.
PMID: 12870129 [PubMed - in process]

(b) List of matching pages for query "liver".

1: Metabolism. 2003 Jul;52(7):868-74. Related Articles, Links

Hyperglycemia induced by glucose infusion causes hepatic oxidative stress and systemic inflammation, but not STAT3 or MAP kinase activation in liver in rats.

Ling PR, Mueller C, Smith RJ, Bistrian BR.

The purpose of this study was to determine the effects of acute hyperglycemia induced by glucose infusion on oxidative stress, systemic inflammation, and several key signal intermediates in liver for the systemic inflammatory response in nonstressed rats. Rats received saline or glucose infusion (hyperglycemic clamp) for 3 hours. Rats without catheter insertion were included as an additional control for observing the effects of surgical stress. Levels of malondialdehyde (MDA) and total glutathione to assess oxidative stress were determined in liver and muscle. Proinflammatory cytokines including tumor necrosis factor (TNF), interleukin (IL)-1 and IL-6, and alpha 1 acid glycoprotein (alpha1-AG) were determined in serum. The protein content and phosphorylation of extracellular signal-regulated kinase (ERK)1/2, p38 stress-activated protein kinase (p38), and signal transducer and activator of transcription-3 (STAT-3) were examined in the liver tissue with or without IL-6 stimulation. The results showed that acute hyperglycemia significantly increased MDA release and depleted total glutathione in liver but not in muscle. Hyperglycemia also significantly elevated the production of TNF, IL-1, and alpha1-AG, but not IL-6 in serum. However, hyperglycemia for 3 hours in vivo did not activate ERK1/2, p38 and STAT3 in liver, and also did not alter the response of these signal proteins to IL-6 stimulation. These data suggest that acute (3 hours) hyperglycemia causes hepatic oxidative stress and activates a low-grade systemic inflammation but does not affect key components of the IL-6 signaling pathway in liver.

PMID: 12870163 [PubMed - in process]

(c) The first matching page for "liver".

Figure 1: Pages from the PubMed Web site.

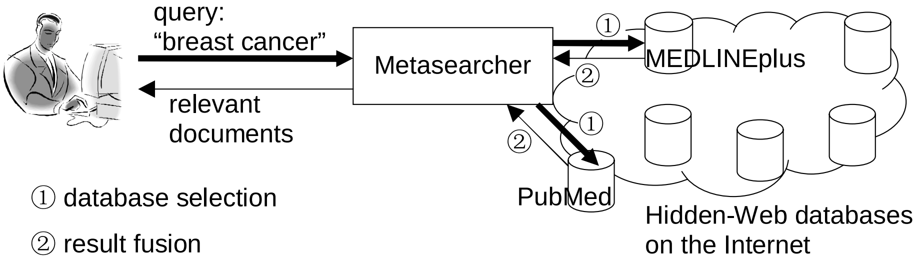


Figure 2: The metasearching process

Web, they need to go through the painful process of issuing queries to all potentially relevant sites with Hidden-Web contents and investigating the results manually. On the other hand, the information in the Hidden Web is estimated to be significantly larger and of higher quality than the "Surface Web" indexed by search engines [2].

In our research group, we are conducting two research projects to make the Hidden Web easy to access by average users: 1) *Hidden-Web metasearcher* and 2) *Hidden-Web crawler* projects. In the following subsections, we briefly describe these two projects.

3.1 Hidden-Web metasearcher

The goal of this project is to build a *metasearcher* or a *mediator* that automatically selects the most relevant Hidden-Web sites to a user's query, so that the users can simply come to our metasearcher and get access to most of the relevant Hidden-Web data without knowing individual sites [1, 6, 14, 13, 17, 18, 19, 20]. Given a user's query (e.g. "breast cancer" as shown in Figure 2), the metasearcher determines which sites are the most likely to be relevant, directs the user's query to those sites and collects the search results back to the user. Given this scenario, an effective metasearcher needs to accomplish two challenging tasks:

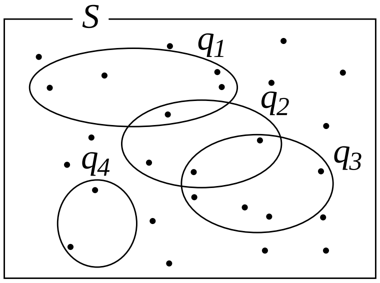


Figure 3: A formal model of the Hidden-Web crawling problem

1. Based on the user's query, the metasearcher has to identify a few sites that are the most relevant, so that it can direct the query to those sites (the arrows labelled ① in Figure 2). This task is often referred to as *database selection* or *database discovery*. In Figure 2, the metasearcher directs the query "breast cancer" to two sites: "PubMed" and "MEDLINEplus.[2]"

2. The metasearcher gathers the query results from the selected sites, and selectively presents the results from multiple sources to the user (The arrows labelled ② in Figure 2). This task is also known as *result merging* or *result fusion*.

So far our research group has mainly focused on the database selection problem and we plan to investigate the result fusion problem in the future. We believe that the success of this research can significantly simplify people's information search on the Hidden Web, as much as internet search engines simplified the search on the Surface Web.

3.2 Hidden-Web crawler

Another approach that we are taking is to build a *Hidden-Web crawler* that can download pages from the Hidden-Web sites automatically, so that we can use well-known indexing technologies to identify the set of pages relevant to a user query.

Since pages in the Hidden Web may be generated dynamically when the user issues a query, the main challenge of a Hidden-Web crawler is how it can discover pages in the Hidden-Web sites. That is, there exist no static links to Hidden Web pages, so the crawler has to automatically generate and issue queries in order to discover such pages. Automatic query generation is clearly hard, because the crawler does not understand the semantics of a query interface.

Note that a Hidden-Web crawler often has limited time and network resources. Thus, a crawler has to carefully select and issue keyword queries, so that it can download the maximum number of pages using minimum resources. If it issues completely random queries that do not return any matching pages, it may waste all of its resources simply issuing queries without retrieving actual pages. Theoretically, this query-selection problem can be formalized as a *minimum-cover problem* in a graph [9]. That is, we assume that a crawler downloads pages from a Web site that has a set of pages S (the rectangle in Figure 3). We represent each Web page in S as a vertex in a graph (dots in Figure 3). We also represent

[2] http://www.nlm.nih.gov/medlineplus/

each potential query q_i that the crawler can issue as a hyperedge in the graph (circles in Figure 3). A hyperedge q_i connects all the vertices (pages) that are returned when the crawler issues q_i to the site. Each hyperedge is also associated with a weight that represents the cost of issuing the query.[3] Under this formalization, our problem is to select the set of hyperedges (queries) that cover the maximum number of vertices (Web pages) with the minimum total weight (cost).

There are two main difficulties in this formalization. First, a Hidden-Web crawler does not know which Web pages will be returned for a query, so the hyperedges of the graph are unknown. Without knowing the hyperedges, the crawler cannot select them. Second, the minimum-cover problem is known to be NP-hard [11], so it is not known whether there exists an efficient algorithm to solve the problem optimally.

We are currently investigating various ideas to address the above issues. Our main idea is to *predict* how many pages will be returned for a future query q_i by analyzing the pages that we downloaded from previous queries $q_1, \ldots, q_{i-1}$. For example, if the keyword "medicine" appears more often than the keyword "violin" in the pages returned from previous queries, we may expect that the site may return more pages if we issue the query "medicine" than "violin." Although our prediction may not be completely accurate, we believe that this approach provide enough clues that the crawler will be able to select significantly better queries.

We recently conducted a preliminary experiment on a real Web site of 40,000 Web pages, and the result was very promising. When we used a relatively straightforward query selection algorithm, we were able to download 95% of the site after issuing fewer than 100 queries. We will continue our study both experimentally and theoretically.

4 Conclusion

In this paper, we briefly went over exciting challenges that the Web has brought us today. We also described our research projects that try to address some of these challenges. Our WebArchive project is developing essential technologies to archive the history of the Web, so that we can still access important information even after it disappears from the Web. The technologies that we develop can be used to archive any type of textual databases, and the pages that we archive will provide a valuable dataset that researchers can investigate. In our Hidden-Web project, we are developing an effective metasearching framework that can provide a single access point for Hidden-Web information. We are also developing a novel Hidden-Web crawler that can download pages from the Hidden Web automatically without user input.

The proliferation of information on the Web has clearly enabled an average person to access an enormous amount of information that was not possible a decade ago. We believe that the success of our research projects will make the Web even closer to become the ideal information source, where any user can access any information at any time through an intuitive interface.

[3] The cost of a query consists of three factors: 1) the cost of issuing the query to the site, 2) the cost of retrieving the answer that contains the list of matching pages and 3) the cost of actually downloading the matching pages.

References

[1] C. Baumgarten. A probabilistic solution to the selection and fusion problem in distributed information retrieval. In *Proc. of ACM SIGIR Conference*, 1999.

[2] M. Bergman. The deep web: Surfacing hidden value, 2000.

[3] T. Berners-Lee. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web*. HarperBusiness, first edition, 2000.

[4] B. E. Brewington and G. Cybenko. How dynamic is the web. In *Proceedings of the International World-Wide Web Conference (WWW)*, May 2000.

[5] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the International World-Wide Web Conference (WWW)*, April 1998.

[6] J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *SIGMOD Conference*, pages 479–490, 1999.

[7] K. C.-C. Chang, B. He, C. Li, and Z. Zhang. Structured databases on the web: Observations and implications. Technical report, University of Illinois, Urbana-Champaign, 2003.

[8] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *Proceedings of the Twenty-sixth International Conference on Very Large Databases (VLDB)*, September 2000.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. McGraw Hill, second edition, 2001.

[10] F. Douglis, A. Feldmann, and B. Krishnamurthy. Rate of change and other metrics: a live study of the world wide web. In *Proceedings of the Second USENIX Symposium on Internetworking Technologies and Systems*, October 1999.

[11] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W H Freeman & Co., 1979.

[12] B. A. Huberman and L. A. Adamic. Growth dynamics of the World-Wide Web. *Nature*, 401(6749):131, September 1999.

[13] P. Ipeirotis and L. Gravano. Distributed search over the hidden web: Hierarchical database sampling and selection. In *Proceedings of the Twenty-eighth International Conference on Very Large Databases*, 2002.

[14] P. G. Ipeirotis, L. Gravano, and M. Sahami. Probe, count, and classify: Categorizing hidden web databases. In *SIGMOD Conference*, 2001.

[15] S. Lawrence and C. L. Giles. Searching the World Wide Web. *Science*, 280(5360):98–100, April 1998.

[16] S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400(6740):107–109, July 1999.

[17] W. Meng, K.-L. Liu, C. T. Yu, X. Wang, Y. Chang, and N. Rishe. Determining text databases to search in the internet. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, pages 14–25, 24–27 1998.

[18] J. Xu and J. Callan. Effective retrieval with distributed collections. In *Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 112–120, 1998.

[19] C. T. Yu, W. Meng, W. Wu, and K.-L. Liu. Efficient and effective metasearch for text databases incorporating linkages among documents. In *SIGMOD Conference*, 2001.

[20] B. Yuwono and D. L. Lee. Server ranking for distributed text retrieval systems on the internet. In *Database Systems for Advanced Applications*, pages 41–50, 1997.