Challenges in Web Search Engines

Monika R. Henzinger* Rajeev Motwani† Craig Silverstein‡

October 17, 2002

Abstract

This article presents a high-level discussion of some problems in information retrieval that are unique to web search engines. The goal is to raise awareness and stimulate research in these areas.

1 Introduction

Web search engines are faced with a number of difficult problems in maintaining or enhancing the quality of their performance. These problems are either unique to this domain, or novel variants of problems that have been studied in the literature. Our goal in writing this article is to raise awareness of several problems that we believe could benefit from increased study by the research community. We deliberately ignore interesting and difficult problems that are already the subject of active research. Thus, we do not consider techniques to improve text-based retrieval, to support natural language queries, to query non-text corpora such as image or audio collections, to search effectively in multiple languages, and so forth.

We begin with a high-level description of the problems that we describe in further detail in the subsequent sections.

Spam. Users of web search engines tend to examine only the first page of search results. Silverstein et al. [17] showed that for 85% of the queries only the first result screen is requested. Thus, inclusion in the first result screen, which usually shows the top 10 results, can lead to an increase in traffic to a web site, while exclusion means that only a small fraction of the users will actually see a link to the web site. For commercially-oriented web sites, whose income depends on their traffic,

*Google Inc., 2400 Bayshore Parkway, Mountain View, CA 94043. *E-mail:* monika@google.com.

†Department of Computer Science, Stanford University, Stanford, CA 94305. *E-mail:* rajeev@cs.stanford.edu.
Part of this work was done while the author was visiting Google Inc. Work also supported in part by NSF Grant IIS-0118173, and research grants from the Okawa Foundation and Veritas.
‡Google Inc., 2400 Bayshore Parkway, Mountain View, CA 94043. *E-mail:* csilvers@google.com.

it is in their interest to be ranked within the top 10 results for a query relevant to the content of the web site.

To achieve this goal, some web authors try to deliberately manipulate their placement in the ranking order of various search engines. The result of this process is commonly called *search engine spam*. In this paper we will simply refer to it as *spam*. To achieve high rankings, authors either use a *text-based* approach, a *link-based* approach, a *cloaking* approach, or a combination thereof. There are web ranking optimization services which, for a fee, claim to place a given web site highly on a given search engine.

Unfortunately, spamming has become so prevalent that every commercial search engine has had to take measures to identify and remove spam. Without such measures, the quality of the rankings suffers severely.

Traditional research in information retrieval (IR) has not had to deal with this problem of "malicious" content in the corpora. Quite certainly, this problem is not present in the benchmark document collections used by IR researchers in the past; indeed, those collections consist exclusively of high-quality content such as newspaper or scientific articles. Similarly, the spam problem is not present in the context of intranets, the web that exists within a corporation. While current research in IR has increased its focus on search and retrieval on the web, there has been relatively little attention paid to this ever-worsening problem of content created to mislead.

Content Quality. Even if the spam problem did not exist, there are many troubling issues concerned with the quality of the content on the web. The web is full of noisy, low-quality, unreliable, and indeed contradictory content. A reasonable approach to IR for relatively high-quality content would be to assume that every document in a collection is authoritative and accurate, design techniques for this context, and then tweak the techniques to incorporate the possibility of low-quality content. However, the democratic nature of content creation on the web leads to a corpus that is fundamentally noisy and of poor quality, and useful information emerges only in a statistical sense. In designing a high-quality search engine, one has to start with the assumption that a typical document cannot be "trusted" in isolation; rather, it is the synthesis of a large number of low-quality documents that provides the best set of results.

As a first step in the direction outlined above, it would be extremely helpful for web search engines to be able to identify the quality of web pages independent of a given user request. There have been link-based approaches, for instance PageRank [5], for estimating the quality of web pages. However, PageRank only uses the link structure of the web to estimate page quality. It seems to us that a better estimate of the quality of a page requires additional sources of information, both within a page (e.g., the reading-level of a page) and across different pages (e.g., correlation of content).

Quality Evaluation. Evaluating the quality of different ranking algorithms is a notoriously difficult problem. Commercial search engines have the benefit of large amounts of user-behavior

data they can use to help evaluate ranking. Users usually will not make the effort to give explicit feedback but nonetheless leave implicit feedback information such as the results on which they clicked. The research issue is to exploit the implicit feedback to evaluate different ranking strategies.

Web Conventions. Most creators of web pages seem to follow simple "rules" without anybody imposing these rules on them. For example, they use the *anchor text* of a link to provide a succinct description of the target page. Since most authors behave this way, we will refer to these rules as *web conventions*, even though there has been no formalization or standardization of such rules.

Search engines rely on these web conventions to improve the quality of their results. Consequently, when webmasters violate these conventions they can confuse search engines. The main issue here is to identify the various conventions that have evolved organically and to develop techniques for accurately determining when the conventions are being violated.

Duplicate Hosts. Web search engines try to avoid crawling and indexing duplicate and near-duplicate pages, as they do not add new information to the search results and clutter up the results. The problem of identifying duplicates within a set of crawled pages is well studied. However, if a search engine can avoid crawling the duplicate content in the first place, the gain is even larger. In general, predicting whether a page will end up being a duplicate of an already-crawled page is chancy work, but the problem becomes more tractable if we limit it to finding duplicate *hosts*, that is, two hostnames that serve the same content. One of the ways that duplicate hosts can arise is via an artifact of the domain name system (DNS) where two hostnames can resolve to the same physical machine. There has only been some preliminary work on the duplicate hosts problem [3].

Vaguely-Structured Data. The degree of structure present in data has had a strong influence on techniques used for search and retrieval. At one extreme, the database community has focused on highly-structured, relational data, while at the other the IR community has been more concerned with essentially unstructured text documents. Of late, there has been some movement toward the middle with the database literature considering the imposition of structure over almost-structured data. In a similar vein, in IR systems, document management systems use accumulated meta-information to introduce more structure. The emergence of XML has led to a flurry of research involving extraction, imposition, or maintenance of partially-structured data.

Web pages in HTML fall into the middle of this continuum of structure in documents, being neither close to free text nor to well-structured data. Instead HTML markup provides limited structural information, typically used to control layout but providing clues about semantic information. Layout information in HTML may seem of limited utility, especially compared to information contained in languages like XML that can be used to tag content, but in fact it is a particularly valuable source of meta-data in unreliable corpora such as the web. The value in layout information stems from the fact that it is visible to the user: Most meta-data which is not user-visible and therefore is particularly susceptible to spam techniques, but layout information is more difficult to use for

spam without affecting the user experience. There has only been some initial, partly related work in this vein [15, 8, 9]. We believe that the exploitation of layout information can lead to direct and dramatic improvement in web search results.

2 Spam

Some web authors try to deliberately manipulate their placement in the rankings of various search engine. The resulting pages are called *spam*. Traditional IR collections did not contain spam. As a result, there has not been much research into making search algorithms resistant to spam techniques. Web search engines, on the other hand, have been consistently developing and improving techniques for detecting and fighting spam. As search engine techniques have developed, new spam techniques have developed in response. Search engines do not publish their anti-spam techniques to avoid helping spammers to circumvent them.

Historical trends indicate that the use and variety of spam will continue to increase. There are challenging research issues involved in both detecting spam and in developing ranking algorithms that are resistant to spam. Current spam falls into following three broad categories: text spam, link spam, and cloaking. A spammer might use one or some combination of them.

2.1 Text Spam

All search engines evaluate the content of a document to determine its ranking for a search query. Text spam techniques are used to modify the text in such a way that the search engine rates the page as being particularly relevant, even though the modifications do not increase perceived relevance to a human reader of a document.

There are two ways to try to improve ranking. One is to concentrate on a small set of keywords and try to improve perceived relevance for that set of keywords. For instance, the document author might repeat those keywords often at the bottom of the document, which it is hoped will not disturb the user. Sometimes the text is presented in small type, or even rendered invisible (e.g., by being written in the page's background color) to accomplish this.

Another technique is to try to increase the number of keywords for which the document is perceived relevant by a search engine. A naïve approach is to include (some subset of) a dictionary at the bottom of the web page, to increase the chances that the page is returned for obscure queries. A less naïve approach is to add text on a different topic to the page to make it appear that this is the main topic of the page. For example, porn sites sometimes add the names of famous personalities to their pages in order to make these pages appear when a user searches for such personalities.

2.2 Link Spam

The advent of link analysis by search engines has been accompanied by an effort by spammers to manipulate link analysis systems. A common approach is for an author to put a *link farm* at the

bottom of every page in a site, where a link farm is a collection of links that points to every other page in that site, or indeed to any site the author controls. The goal is to manipulate systems that use raw counts of incoming links to determine a web page's importance. Since a completely-linked link farm is easy to spot, more sophisticated techniques like pseudo web-rings and random linkage within a member group are now being used.

A problem with link farms is that they distract the reader because they are on pages that also have legitimate content. A more sophisticated form of link farms has been developed, called *doorway pages*. Doorway pages are web pages that consist entirely of links. They are not intended to be viewed by humans; rather, they are constructed in a way that makes it very likely that search engines will discover them. Doorway pages often have thousands of links, often including multiple links to the same page. (There is no text-spam equivalent of doorway pages because text, unlike links, is analyzed by search engines on a per-page basis.)

Both link farms and doorway pages are most effective when the link analysis is sensitive to the absolute number of links. Techniques that concentrate instead on the quality of links, such as PageRank [5, 6],

are not particularly vulnerable to these techniques.

2.3 Cloaking

Cloaking involves serving entirely different content to a search engine crawler than to other users.[1] As a result, the search engine is deceived as to the content of the page and scores the page in ways that, to a human observer, seem rather arbitrary.

Sometimes cloaking is used with the intent to "help" search engines, for instance by giving them an easily digestible, text-only version of a page that is otherwise heavy with multimedia content, or to provide link-based access to a database which is normally only accessible via forms (which search engines cannot yet navigate). Typically, however, cloaking is used to deceive search engines, allowing the author to achieve the benefits of link and text spam without inconveniencing human readers of the web page.

2.4 Defending against Spam

In general, text spam is defended against in a heuristic fashion. For instance, it was once common for sites to "hide" text by writing it in white text on a white background, ensuring that human readers were not affected while search engines were misled about the content. As a result, search engine companies detected such text and ignored it. Such reactive approaches are, obviously, not optimal. Can pro-active approaches succeed? Perhaps these approaches could be combined; it

[1] A search engine crawler is a program that downloads web pages for the purpose of including them in the search engine results. Typically a search engine will download a number of pages using the crawler, then process the pages to create the data structures used to service search requests. These two steps are repeated continuously to ensure the search engine is searching over the most up-to-date content possible.

might be possible for the search engine to notice what pages change in response to the launch of a new anti-spam heuristic, and to consider those pages as potential spam pages.

Typically, link-spam sites have certain patterns of links that are easy to detect, but these patterns can mutate in much the same way as link spam detection techniques. A less heuristic approach to discovering link spam is required. One possibility is, as in the case of text spam, to use a more global analysis of the web instead of merely local page-level or site-level analysis. For example, a cluster of sites that suddenly sprout thousands of new and interlinked webpages is a candidate link-spam site. The work by Ravi Kumar et al. [16] on finding small bipartite clusters in the web is a first step in this direction.

Cloaking can only be discovered by crawling a website twice, once using an HTTP client the cloaker believes is a search engine, and once from a client the cloaker believes is not a search engine. Even this is not good enough, since web pages typically differ between downloads for legitimate reasons, such as changing news headlines.

3 Content Quality

While spams are attempts to deliberately mislead search engines, the web is replete with text that — intentionally or not — misleads its human readers as well. As an example, there is a webpage which claims (falsely!) that Thomas Jefferson was the first president of the United States. Many websites, purposefully or not, contain misleading medical information.[2] Other sites contain information that was once correct but is now out of date; for example, sites giving names of elected officials.

While there has been a great deal of research on determining the relevance of documents, the issue of document quality or accuracy has not been received much attention, whether in web search or other forms of information retrieval. For instance, the TREC conference explicitly states rules for when it considers a document to be relevant, but does not mention the accuracy or reliability of the document at all. This is understandable, since typical research corpora, including the ones used by TREC and found in corporate intranets, consist of document sources that are deemed both reliable and authoritative. The web, of course, is not such a corpus, so techniques for judging document quality is essential for generating good search results. Perhaps the one successful approach to (heuristically) approximating quality on the web is based on link analysis, for instance PageRank [5, 6] and HITS [13]. These techniques are a good start and work well in practice, but there is still ample room for improvement.

One interesting aspect of the problem of document quality is specific to hypertext corpora such as the web: evaluating the quality of *anchor text*. Anchor text is the text, typically displayed underlined and in blue by the web browser, that is used to annotate a hypertext link. Typically, web-based search engines benefit from including anchor-text analysis in their scoring function [11].

[2] One study showed many reputable medical sites contain contradictory information on different pages of their site [2] — a particularly difficult content-quality problem!

However, there has been little research into the perils of anchor-text analysis e.g. due to spam and on methodologies for avoiding the pitfalls.

For instance, for what kinds of low-quality pages might the anchor text still be of high quality? Is it possible to judge the quality of anchor text independently of the quality of the rest of the page? Is it possible to detect anchor text that is intended to be editorial rather than purely descriptive? In addition, many fundamental issues remain open in the application of anchor text to determination of document quality and content. In case of documents with multiple topics, can anchor text analysis be used to identify the themes?

Another promising area of research is to combine established link-analysis quality judgments with text-based judgments. A text-based analysis, for instance, could judge the quality of the Thomas Jefferson page by noting that most references to the first president of the United States in the web corpus attribute the role to George Washington.

4 Quality Evaluation

Search engines cannot easily improve their ranking algorithms without running tests to compare the quality of the new ranking technique with the old. Performing such comparisons with human evaluators is quite work-intensive and runs the danger of not correctly reflecting user needs. Thus, it would be best to have end users perform the evaluation task, as they know their own needs the best.

Users, typically, are very reluctant to give direct feedback. However, web search engines can collect implicit user feedback using log data such as the position of clicks for a search and the time spent on each click. This data is still incomplete. For instance, once the user clicks on a search result, the search engine does not know which pages the user visits until the user returns to the search engine. Also, it is hard to tell whether a user clicking on a page actually ends up finding that page relevant or useful.

Given the incomplete nature of the information, the experimental setup used to college implicit user data becomes particularly important. That is: How should click-through and other data be collected? What metrics should be computed from the data?

One approach is to simply collect the click-through data from a subset of the users — or all users — for two ranking algorithm. The experimenter can then compute metrics such as the percentage of clicks on the top 5 results and the number of clicks per search.

Recently, Joachims [14] suggested another experimental technique which involves merging the results of the two ranking algorithms into a single result set. In this way each user performs a comparison of the two algorithms. Joachims proposes to use the number of clicks as quality metric and shows that, under some weak assumptions, the clickthrough for ranking A is higher than the clickthrough for B if and only if A retrieves more relevant links than B.

5 Web Conventions

As the web has grown and developed, there has been an evolution of conventions for authoring web pages. Search engines assume adherence to these conventions to improve search results. In particular, there are three conventions that are assumed relating to anchor text, hyperlinks, and META tags.

- As discussed in Section 3, the fact that anchor text is meant to be descriptive is a web convention, and this can be exploited in the scoring function of a search engine.

- Search engines typically assume that if a web page author includes a link to another page, it is because the author believes that readers of the source page will find the destination page interesting and relevant. Because of the way people usually construct web pages, this assumption is usually valid. However, there are prominent exceptions: for instance, link exchange programs, in which web page authors agree to reciprocally link in order to improve their connectivity and rankings, and advertisement links. Humans are adept at distinguishing links included primarily for commercial purposes from those included primarily for editorial purposes. Search engines are less so.

  To further complicate matters, the utility of a link is not a binary function. For instance, many pages have links allowing you to download the latest version of Adobe's Acrobat Reader. For visitors that do not have Acrobat Reader, this link is indeed useful, certainly more useful than for those those who have already downloaded the program. Similarly, most sites have a terms of service link at the bottom of every page. When the user first enters the site, this link might well be very useful, but as the user browses other webpages on the site, the link's usefulness immediately decreases.

- A third web convention concerns the use of META tags. These tags are currently the primary way to include metadata within HTML. In theory META tags can include arbitrary content, but conventions have arisen for meaningful content. A META tag of particular importance to search engines is the so-called *Content META tag*, which web page authors use to describe the content of the document. Convention dictates that the content META tag contains either a short textual summary of the page or a brief list of keywords pertaining to the content of the page.

  Abuse of this META tags is common, but even when there is no attempt to deceive, there are those who break the convention, either out of ignorance or overzealousness. For instance, a webpage author might include a summary of their entire site within the META tag, rather than just the individual page. Or, the author might include keywords that are more general than the page warrants, using a META description of "cars for sale" on a web page that only sells a particular model of car.

In general, the correctness of META tags is difficult for search engines to analyze because they are not visible to users and thus are not constrained to being useful to visitors. However, there are many web page authors that use META tags correctly. Thus, if web search engines could correctly judge the usefulness of the text in a given META tag, the search results could potentially be improved significantly. The same applies to other content not normally displayed, such as ALT text associated with the IMAGE tag.

While link analysis has become increasingly important as a technique for web-based information retrieval, there has not been as much research into the different types of links on the web. Such research might try to distinguish commercial from editorial links, or links that relate to meta-information about the site ("This site best viewed with [start link]browser X[end link]") from links that relate to the actual content of the site.

To some extent, existing research on link analysis is helpful, since authors of highly visible web pages are less likely to contravene established web conventions. But clearly this is not sufficient. For instance, highly visible pages are more, rather than less, likely to include advertisements than the average page.

Understanding the nature of links is valuable not only for itself, but also because it enables a more sophisticated treatment of the associated anchor text. A potential approach would be to use text analysis of anchor text, perhaps combined with meta-information such as the URL of the link, in conjunction with information obtained from the web graph.

6 Duplicate Hosts

Web search engines try to avoid crawling and indexing duplicate and near-duplicate pages, since such pages increase the time to crawl and do not contribute new information to the search results. The problem of finding duplicate or near-duplicate pages in a set of crawled pages is well studied [4, 7]. There has also been some research on identifying duplicate or near-duplicate directory trees [10], called *mirrors*.

While mirror detection and individual-page detection try to provide a complete solution to the problem of duplicate pages, a simpler variant can reap most of the benefits while requiring less computational resources. This simpler problem is called *duplicate host detection*. Duplicate hosts ("duphosts") are the single largest source of duplicate pages on the web, so solving the duplicate hosts problem can result in a significantly improved web crawler.

A *host* is merely a name in the domain name system (DNS), and duphosts arise from the fact that two DNS names can resolve to the same IP address.[3] Companies typically reserve more than one name in DNS, both to increase visibility and to protect against domain name "squatters." For

[3]In fact, it's not necessary that they resolve to the same IP address to be duphosts, just that they resolve to the same webserver. Technically even that is not necessary; the minimum requirement is that they resolve to computers that serve the same content for the two hostnames in question.

instance, currently both `bikesport.com` and `bikesportworld.com` resolve to the same IP address, and as a result the sites `http://www.bikesport.com/` and `http://www.bikesportworld.com/` display identical content.

Unfortunately, duplicate IP addresses are neither necessary nor sufficient to identify duplicate hosts. Virtual hosting can result in different sites sharing an IP address, while round-robin DNS can result in a single site having multiple IP addresses.

Merely looking at the content of a small part of the site, such as the homepage, is equally ineffective. Even if two domain names resolve to the same website, their homepages could be different on the two viewings, if for instance the page includes an advertisement or other dynamic content. On the other hand, there are many unrelated sites on the web that have an identical "under construction" home page.

While there has been some work on the duphosts problem [3], it is by no means a solved problem. One difficulty is that the solution needs to be much less expensive than the brute-force approach that compares every pair of hosts. For instance, one approach might be to download every page on two hosts, and then look for a graph isomorphism. However, this defeats the purpose of the project, which is to not have to download pages from both of two sites that are duphosts.

Furthermore, web crawls are never complete, so any link-structure approach would have to be robust against missing pages. Specifically, a transient network problem problem, or server downtime, may keep the crawler from crawling a page in one host of a duphost pair, but not the other. Likewise, due to the increasing amount of dynamic content on the web, text-based approaches cannot check for exact duplicates.

On the other hand, the duphosts problem is simpler than the more general problem of detecting mirrors. Duphosts algorithms can take advantage of the fact that the urls between duphosts are very similar, differing only in the hostname component. Furthermore, they need not worry about content reformatting, which is a common problem with mirror sites.

Finally — and this is not a trivial matter — duphost analysis can benefit from semantic knowledge of DNS. For instance, candidate duphosts `http://foo.com` and `http://foo.co.uk` are, all other things being equal, likely to be duphosts, while candidates `http://foo.com` and `http://bar.com` are not as likely to be duphosts.

7 Vaguely-Structured Data

While IR corpora has tended to be very low in structure, database content is very well structured. This has obviously led to a major difference in how the two fields have evolved over the years. For instance, a practical consequence of this difference is that databases permit a much richer and complex set of queries, while text-based query languages are in general much more restricted.

As database content, or more generally structured data, started being exposed through web interfaces, there developed a third class of data called *semi-structured data*. In the web context, semi-structured data is typically the content of a webpage, or part of a webpage, that contains

structured data but no longer contains unambiguous markup explicating the structure or schema. There has been considerable research on recovering the full structure of semi-structured data, for example, Ahonen, Mannila, and Nikunen [1] and Nestorov, Abiteboul, and Motwani [15].

The three examples above cover three points on the continuum of structured data. However, most web pages do not fall into any of these categories, but instead fall into a fourth category we call *vaguely-structured data*. The information on these web page is not structured in a database sense — typically it's much closer to prose than to data — but it does have some structure, often unintentional, exhibited through the use of HTML markup.

We say that HTML markup provides unintentional structure because it is not typically the intent of the webpage author to describe the document's semantics. Rather, the author uses HTML to control the document's layout, the way the document appears to readers. (It is interesting to note that this subverts the original purpose of HTML, which was meant to be a document description language rather than a page description language.) To give one example, HTML has a tag that is intended to be used to mark up glossary entries. In common browsers, this caused the text to be indented in a particular way, and now the glossary tag is used in any context where the author wants text indented in that manner. Only rarely does this context involve an actual glossary.

Of course, often markup serves both a layout and a semantic purpose. The HTML header tags, for instance, produce large-font, bold text useful for breaking up text, but at the same time they indicate that the text so marked is probably a summary or description of the smaller-font text which follows.

Even when markup provides no reliable semantic information, it can prove valuable to a search engine. To give just one example, users have grown accustomed to ignoring text on the periphery of a web page [12], which in many cases consists of navigational elements or advertisements. Search engines could use positional information, as expressed in the layout code, to adjust the weight given to various sections of text in the document.

In addition, layout can be used to classify pages. For instance, pages with an image in the upper-left of the page are often personal homepages. Pages with a regular markup structure are likely to be lists, which search engines may wish to analyze differently than pages with running text.

Markup can be meta-analyzed as well. It is plausible that pages with many mistakes in the markup are more likely to be of lower quality than pages with no mistakes. Patterns in the markup used may allow a search engine to identify the web authoring tool used to create the page, which in turn might be useful for recovering some amount of structure from the page. Markup might be particularly useful for clustering web pages by author, as authors often use the same template for most of the pages they write.

And, of course, HTML tags can be analyzed for what semantic information can be inferred. In addition to the header tags mentioned above, there are tags that control the font face (bold, italic), size, and color. These can be analyzed to determine which words in the document the author thinks are particularly important.

One advantage of HTML, or any markup language that maps very closely to how the content

is displayed, is that there is less opportunity for abuse: it is difficult to use HTML markup in a way that encourages search engines to think the marked text is important, while to users it appears unimportant. For instance, the fixed meaning of the <H1> tag means that any text in an H1 context will appear prominently on the rendered web page, so it is safe for search engines to weigh this text highly. However, the reliability of HTML markup is decreased by Cascading Style Sheets [18], which separate the names of tags from their representation.

There has been research in extracting information from what structure HTML does possess. For instance, Chakrabarti et al. [8, 9] created a DOM tree of an HTML page and used this information to increase the accuracy of topic distillation, a link-based analysis technique.

However, there has been less research addressing the fact HTML markup is primarily descriptive, that is, that it is usually inserted to affect the way a document appears to a viewer. Such research could benefit from studies of human perception: how people view changes in font size and face as affecting the perceived importance of text, how much more likely people are to pay attention to text at the top of a page than the bottom, and so forth. As newspaper publishers have long known, layout conveys semantic information, but it's not trivial to extract it.

Turning HTML into its markup is also a challenge. It is possible to render the page, of course, but this is computationally expensive. Is there any way to figure out, say, if a given piece of HTML text is in the "middle" of a rendered HTML page without actually rendering it?

Of course, HTML text is only one example of vaguely structured data. What other kinds of content exists that is somewhere between unstructured data and semi-structured data in terms of quantity of annotation? How does it differ from HTML text? For that matter, the continuum of structure is not well-mapped. What techniques appropriate for unstructured data work equally well with vaguely structured data? What techniques work for semi-structured data? How can these techniques be improved as data gets more structured, and is there any way to map the improvements down to less structured forms of data (perhaps by imputing something "structural" to the data, even if that doesn't correspond to any intuitive idea of structure)?

8 Conclusions

In this paper we presented some challenging problems faced by current web search engines. In the last several years, the IR research community has started to work on a proper methodology to evaluate web IR systems. The establishment of the Web TREC conference has greatly contributed to an improved understanding of the issues involved, and an appropriate methodology does not seem far away.

There are other fruitful areas of research related to web search engines we did not touch on. For instance, there are challenging systems issues that arise when hundreds of millions of queries over billions of web pages have to be serviced every day without any down-time and as inexpensively as possible. Furthermore, there are interesting user interface issues: What user interface does not confuse novice users, does not clutter the screen, but still fully empowers the experienced user?

Finally, are there other ways to mine the collection of web pages so as to provide a useful service to the public at large?

9 Resources

Here are two resources for the research community:

- Stanford's WebBase project (`http://www-diglib.stanford.edu/~testbed/doc2/WebBase/`) distributes its content of web pages.
- Web term document frequency is available at Berkeley's Web Term Document Frequency and Rank site (`http://elib.cs.berkeley.edu/docfreq/index.html`.)

References

[1] H. Ahonen, H. Mannila, and E. Nikunen. "Generating grammars for SGML tagged texts lacking DTD." *PODP'94 – Workshop on Principles of Document Processing*, 1994. `http://www.cs.Helsinki.FI/u/hahonen/publications.html`.

[2] G. K. Berland, M. N. Elliott, L. S. Morales, J. I. Algazy, R. L. Kravitz, M. S. Broder, D. E. Kanouse, J. A. Muñoz, J.-A. Puyol, M. Lara, K. E. Watkins, H. Yang, and E. A. McGlynn. "Health Information on the Internet Accessibility, Quality, and Readability in English and Spanish." *Journal of the American Medical Association*, 285(2001): 2612–2621.

[3] K. Bharat, A. Z. Broder, J. Dean, and M. Henzinger. "A comparison of Techniques to Find Mirrored Hosts on the World Wide Web." *Journal of the American Society for Information Science*, 31(2000): 1114–1122.

[4] S. Brin, J. Davis, and H. García-Molina. "Copy detection mechanisms for digital documents." *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1995, pages 398–409.

[5] S. Brin, and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine." In *Proceedings of the 7th International World Wide Web Conference (WWW7)*, 1998, pages 107–117. Also appeared in *Computer Networks* 30(1998): 107–117.

[6] S. Brin, L. Page, R. Motwani, and T. Winograd. "What can you do with a Web in your Pocket?" *Bulletin of the Technical Committee on Data Engineering*, 21(1998): 37-47.

[7] A. Z. Broder. "On the resemblance and containment of documents." In *Proceedings of Compression and Complexity of Sequences*, IEEE Computer Society, 1997, pages 21–29.

[8] S. Chakrabarti. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[9] S. Chakrabarti. Integrating the Document Object Model with hyperlinks for enhanced topic distillation and information extraction. In *Proceedings of the 10th International World Wide Web Conference (WWW10)*, 2001.

[10] J. Cho, N. Shivakumar, and H. Garcia-Molina. "Finding replicated web collections." In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2000, pages 355–366.

[11] N. Craswell, D. Hawking, and S. Robertson. "Effective Site Finding using Link Anchor Information." In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001.

[12] P. Faraday. "Attending to Web Pages." *CHI 2001 Extended Abstracts (Poster)*, 2001, pages 159–160.

[13] J. Kleinberg. "Authoritative sources in a hyperlinked environment." In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1998, pages 668–677.

[14] T. Joachims. "Evaluation Search Engines using Clickthrough Data". To appear, 2002.

[15] S. Nestorov, S. Abiteboul, and R. Motwani. "Extracting Schema from Semistructured Data." In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1998, pages 295–306.

[16] S. Ravi Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins. "Trawling emerging cybercommunities automatically." In *Proceedings of the 8th International World Wide Web Conference (WWW8)*, 1999.

[17] C. Silverstein, M. R. Henzinger, J. Marais, and M. Moricz. "Analysis of a very large AltaVista query log." *SIGIR Forum*, 33(1999): 6–12.

[18] World Wide Web Consortium. "Web Style Sheets." `http://www.w3.org/Style/`.