



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Information Retrieval Challenges in Web

G.N.K.Suresh Babu¹ S.K.Saravanan² and Dr.S.K.Srivatsa³

¹Associate Professor, Department of Computer Applications, GKM CET, Chennai.

²Assistant Professor [SG], Department of Computer Applications,
Valliammai Engineering College, Chennai, INDIA

³Sr.Professor – St.Joseph's College of Engineering, Chennai, INDIA.

Abstract: Web Information Retrieval (WebIR) is the process of extracting useful information from among the petabytes (10^{15}) of data that make up the WWW. WebIR has a history almost as long as the Web itself. One of the first efforts towards WebIR was the creation of directories of Web sites. An example is yahoo.com. Web directories manually organized contents of the Web into a taxonomy of categories and sub categories. Web site owners who wished to be listed in the directory had to submit their site to the directory for perusal and inclusion. As more information becomes available on the World Wide Web it becomes more difficult to provide effective search tools for information access. Today, people access Web information through two main kinds of search interfaces: Browsers and Query Engines. The first process is tentative and time consuming and the second may not satisfy the user because of many inaccurate and irrelevant results. Better support is needed for expressing one's information need and returning high quality search results by Web search tools. There appears to be a need for systems that do reasoning under uncertainty and are flexible enough to recover from the contradictions, inconsistencies, and irregularities that such reasoning involves. In this paper discusses about the challenges faced by Web and their mitigations.

Keywords: Web, Search Engines, Crawler, IR

I. Introduction

The Web has had very rapid growth in number of pages, number of hosts and number of domain names (e.g. google.com) registered worldwide. There are more than 50 billion Web pages and it is estimated that the number of Web pages will exceed 75 billion by 2013. Almost 10 million pages or 100 Giga bytes of text are added daily, and the average life span of a Web page is about 30 days. To keep up with the changes to Web content, one would need to download about same amount of bytes of information per day, which would mean you would need a connection capable of downloading 10 mega bytes of text per second. The number of internet hosts increases exponentially. The largest search engines have done an impressive job in extending their reach, though Web growth itself has exceeded the crawling ability of search engines. Today Google is probably biggest search engine, and has gathered more than 3 billion pages and covers only about 40 percent of the publicly available Web pages. With the increasing number of users on the Web, and automated systems that query search engines, it is likely that top search engines will handle hundreds of millions of queries per day. Web search engines create and maintain an index of words within documents they find on the Web. They return to a user a ranked list of relevant documents as search results. Few of these results may be valuable to a user. Several ranking methods have been proposed to improve the ranking of resulting documents. For this reason, it may be helpful to use some user information context in returning and ranking results.

II. Information Retrieval on the Web

Retrieving information from the Web can prove to be difficult because of the size and abstractness of data contained on the Web. Approximations for 2013 estimated the Web to be as large as 50 billion Web pages or more. Web retrieval is made increasingly difficult when adding in factors such as word ambiguity (where a single word can take on multiple meanings), and the large amount of typographical errors contained within Web information. It is estimated that one in every two-hundred words, on an average Web site, will contain a textual error.

A. Issues with Information Retrieval

There are several key issues involving information retrieval. These issues are relevance, evaluation, and information needs. However, these are not the only issues involving information retrieval. Other issues such as performance, scalability and occurrences of paging update are other common information retrieval issues. Relevance is the relational value of a given user query to the documents within the database. Relevance of a document is normally based on a document ranking algorithm. These algorithms define how relevant a document is to a user query by using functions that define relations between the query given and the documents collected in the index. The evaluation of the feedback given by the information retrieval system is another issue with information retrieval. The behavior of the system may not meet the expectations of the user or the documents returned from the system may not all be relevant to a query. Depending on the system and the user,

the results of a query should be in a format that most fits the data being searched and returned. Information needs is how the user interacts with the information retrieval system. The data within the system should be able to be accessed easily and in a way that is convenient to the user. Retrieving too much information might be inconvenient in certain systems, also in other systems not returning all relevant information may be unacceptable.

B. Document and Query Management

Managing volumes of information from the Web can be difficult due to the volume of documents the Web contains. This leads to an even larger problem when trying to retrieve the most relevant results to a query. A simple retrieval query can return thousands of documents, many of which are loosely related to the original retrieval criteria. Natural language and word ambiguity can also add distress to document retrieval. To overcome this, an information retrieval system needs to have good query management along with the ability to give weight to documents that are more relevant to the user's query, and present those results first. Development of an efficient document ranking algorithm is crucial to an information retrieval system. There are many methods used to rank documents to a given user query. Common document ranking algorithms are term counting, user relevance feedback, popularity of the document, size of the document, term location within the document, and how often a document is linked. Fig.1 shows the information retrieval process from the Web.

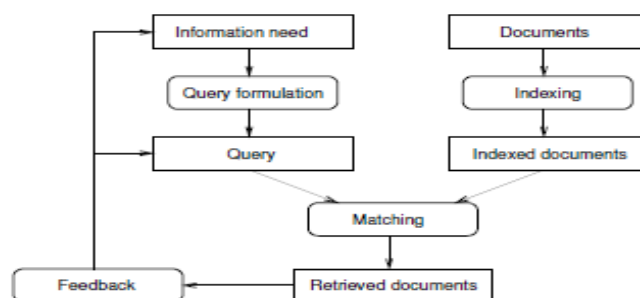


Fig.1 Information Retrieval Process

III. How users search the Web

The users of Web search engines are very heterogeneous and the engines are used by lay persons, as well as by information professionals or experts in certain fields. Apart from studies discussing the common user behaviour, there are some studies that discuss the behaviour of certain user groups. But there are no scientific investigations that discuss how real information professionals in intelligence departments or management consultancies use Web search engines. Instead, most user studies focus on the typical lay user. The main findings of these studies are that the users are not very sophisticated. Only half of the users know about Boolean operators and only slightly more know about advanced search forms. But knowing them does not mean that they are used: Only 14 percent say that they use them. In a laboratory test in the same study, the use of the advanced search forms was even lower. In studies based on transaction log analysis, Spink and Jansen found that Boolean operators are only used in one out of ten queries. Half of the Boolean queries are ill-formed; when plus and minus signs are used (which is generally preferred by the users), the fraction of ill-formed queries rises to two thirds. Users look only seldom at results coming after the first search results page, which means that results which are not among the top 10 are nearly invisible for the general user. There is a tendency that users often only look at the results set that can be seen without scrolling. Within one search session, users look at five documents on average and each document is only shortly examined. Sessions are usually terminated when one suitable document is found. A typical search session lasts less than 15 minutes. Table 1 shows that the differences between Traditional Information Retrieval and Web Information Retrieval.

Table 1 Differences between Traditional IR and Web IR

Traditional IR	Web IR
Databases usually cover only one language or indexing of documents written in different languages with the same vocabulary.	Documents in many different languages. Usually search engines use full text indexing; no additional subject analysis.
Usually all indexed documents have the same format (e.g. PDF) or only bibliographic information is provided.	Several file types, some hard to index because of a lack of textual information.
Document length varies, but not to such a high degree as with the Web documents. Each indexed text is represented with one documentary unit.	Wide range from very short to very long. Longer documents are often divided into parts.
Structured documents allow complex field searching.	HTML documents are semistructures.
Documents are usually not connected. Sometimes citation data is used to determine quality.	Documents are connected heavily. Hyperlink structure can be used to determine quality.

IV. Web Retrieval Challenges

The nature of the Web poses a number of challenges to classic IR systems. Several of these are outlined in this section.

A. Crawling

Web content is distributed across countless Web servers scattered across the Internet, therefore unlike IR collections it is a prerequisite to assemble a snapshot of the Web's content (a crawl) before constructing a representation of it through indexing. Typically snapshots are assembled by automated applications which engage in crawling; the process of recursively fetching documents using a pool of document locations (URLs) which is replenished with discoveries of new URLs referred to in the hyperlinks of fetched documents. Although implementing rudimentary crawlers is relatively straight forward, Google intimate that industry strength crawlers capable of assembling the large crawls typical of major search engines requires a great deal of engineering.

B. Diverse Search Requirements

With developments in Web technology and Web programming, the Web is increasingly functioning as a platform for a growing number of on-line services and Web applications such as Internet banking and Web mail. Changes in the use of the Web induce changes in the intent of Web searchers. Broder presents evidence that informational searches – as formulated in the context of traditional IR account for less than 50% of all searches. The majority of searches are explained by Broder to be either navigational search in which a specific URL such as a corporate homepage is sought or transactional searches in which access to an interactive process (such as on-line shopping) is sought. Broder concludes that search engines are challenged by the need to respond to the different classes of search differently. Although the category of informational searches is common to both IR and Web IR, the abundance of content on the Web demands greater discrimination when returning results for broad-topic searches of this type. A shift in emphasis towards a topic distillation approach to satisfying these queries is advocated by Chakrabarti. By topic distillation, Chakrabarti refers to an approach in which potential search results are evaluated according to how well they represent a topic as opposed to how similar they are to the topic. Chakrabarti experiments with a means to identify this representative quality by analysing topological data. Although numerous techniques for capturing the representative quality of a document through topological analysis have been devised, Marchiori challenges the fairness of these approaches. Instead of topological analysis, he advocates using hyper-information in discerning the added-value of a document, where hyper-information is described as the information that can be obtained through browsing additional content that is hyperlinked.

C. Search Engine Persuasion

Search Engine Persuasion, coined SEP by Marchiori refers to deliberate manipulation of Web search engines in order to boost the ranking of documents in search results. SEP is far more common on the Web than in traditional IR contexts where there is relatively little competition for the attention of collection audiences. Due to the commercial motives of traffic hungry Web site owners, manipulation of this sort ranges from being deceptive to fraudulent. The implicit use of neutral quality judgments in the form of hyperlinks countered the effects of primitive SEP methods such as hidden text. In more advanced SEP hyperlinks are manipulated also. Understandably, efforts made by commercial search engines to maintain the integrity of their search results tend not to be made public.

D. Incorrect Content

Since there are generally no content controls on material published on the Web there is a higher chance that Web documents contain incorrect information than traditional IR collections. Web searchers tend to feel more assured by information that emanates from important sites. The challenge of retrieving correct content is therefore closely tied to that of retrieving authoritative content.

E. Duplication

Duplication of content is far more likely in the context of the Web than it is well controlled collections. Duplication poses a problem for both Search engines and searchers alike. Search engines are computationally burdened by the crawling, indexing and storage of duplicate content and Internet searchers find the presence of duplicates amongst retrieval lists a nuisance. There are generally two approaches to duplicate elimination. Fine-grained duplication elimination concentrates on discovering duplicate pages where as coarse-grained duplicate elimination places an emphasis on identifying duplicate resource directory trees (mirrors).

V. Web IR Implementations

In a technological survey of Web IR systems compiled by Huang, three components are said to be characteristic of Web search engines; an indexer, a crawler and a query server. Huang explains that together the crawler and indexer work to produce a representation of the Web which is optimized for efficient use by the query server. That much is true of IR systems in general. Where Web IR implementations differ from IR systems significantly

is in the variety of information they exploit in retrieval, much of which is unavailable in traditional IR collections. Although the exact details of their systems are generally kept in-house, commercial Web search engines are known to leverage several Web-rich sources of information such as hyperlinks, document structure, document meta-data and usage data from Web servers.

Some consideration was given to document meta-data by Amento, Terveen and Hill in investigating how well a number of measures were able to predict document quality. In their experiments, simplistic meta-data such as number of images and number of documents on site were demonstrated to be effective. A similar study found URL form to be particularly effective for entry page (home page) finding tasks.

Document structure is perhaps more readily available than meta-data. The vast majority of content on the Web is structured in conformance with mark-up languages such as HTML and increasingly XML. Mark-up offers implicit contextual information which facilitates richer modelling of documents than the typical bag-of-words suited to plain text documents. Typically, this representation replaces the standard term frequency meta-data associated with terms occurring in a document with term frequency vectors where each co-ordinate of the vector represents the number of occurrences of the term within designated context classes. Retrieval algorithms can then take this context information into account during relevance evaluations, so that occurrences of terms within certain context classes are more valuable than those occurring within others. Whilst focusing on HTML mark-up, demonstrate that ‘strong’ and ‘anchor’ text are particularly effective descriptors of Web pages.

In XML retrieval, context is important from an addition perspective also. Not only can context aid with retrieval performance, it is also key to meeting a searcher’s requirements. Typically the unit of retrieval in XML retrieval is a particular fragment of an XML document not necessarily the whole document itself. XML queries therefore often feature strict structural constraints so that not only are terms specified in queries but also the required contexts of those terms. The quality of retrieval is consequently not only based on the content resemblance of a fragment to a query but also on the context resemblance where the notion of context resemblance can be expanded in a number of ways. A popular context measure is longest common subsequence which is defined as the how many consecutive components within a context definition match.

McBride’s World Wide Web Worm was the first Web search engine to make use of anchor text. Subsequently anchor text use has proved to be a successful means of improving Web IR systems. In an insight into the architecture of their Web search engine, Google confirm that they index anchor text. Further, structural information pertaining to terms such as font and capitalization are used to enhance their index entries. Terms appearing in URLs and meta-tags are also distinguished within Google’s index structure. From the insight given by Google it is clear that commercial search engines must also concern themselves with optimizing efficiency and eradicating duplication. TREC Web Track participants are more open about their techniques than Web search engine companies and their Web IR research publications are well cited. Participants from the University of Twente are believed to be the first to have published details on the effectiveness of applying URL form evidence in entry page searches, a technique which is thought to have since been adopted in commercial Web search engines.

VI. Conclusion

Web information retrieval is an area open for many research opportunities. The larger problems with Web information retrieval—relevance, evaluation, and information needs—amongst others, are still important topics that require attention. Web Information Retrieval is a pertinent topic of research for the present day. As the Web keeps growing in size, the problem of searching the Web becomes only more complex. There are a number of innovative approaches that have been proposed which hold promise. However, it remains to be seen which approach finally becomes the norm, and to what extent the Web is actually used by users. The kind of WebIR technology that develops over the future would determine whether the Web is destined to be a large storehouse of largely unstructured data, or is actually a huge knowledge network that offers insight to people all over the world.

References

- [1] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener. Graph structure in the Web. *Journal of Computer Networks*, 33(1-6), 309-320. 2000
- [2] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, 30(1--7):65--74. 1998
- [3] N. Craswell, D. Hawking. Overview of the TREC-2004 Web Track, In the Thirteenth Text Retrieval Conference (TREC 2004). 2005
- [4] Lewandowski, D. (2005): *Web Information Retrieval*. Frankfurt am Main, DGI, 2005
- [5] Massimo Marchiori. The quest for correct information on the web. *Comput. Netw. ISDN Syst.*, 29(8-13):1225--123. 1997
- [6] M. De Kunder. “The Size of the World Wide Web (The Internet).” <http://www.worldwidewebsite.com/>. WorldWideWebSize.com. Web. 25 Oct. 2011.
- [7] M. Kobayashi and K. Takeda. “Information Retrieval on the Web.” *ACM Computing Surveys*, 32(2): 144 – 173, June 2000.
- [8] Q. Yang, H. F. Wang, J. R. Wen, G. Zhang, Y. Lu, K. F. Lee and H. J. Zhang. Toward a Next Generation Search engine, *Proceedings of the Sixth Pacific Rim Artificial Intelligence Conference*, Melbourne, Australia, 2000.
- [9] S. Lawrence, Context in web Search, *IEEE Data Engineering Bulletin*, Volume 23, Number 3, pp. 25-32, 2000
- [10] “The Web Robots Pages.” <http://www.robotstxt.org/>. Web. 20 Apr. 2011.
- [11] Tim Berners-Lee, James Hendler, Ora Lassila, *The Semantic Web*. ScientificAmerican, Vol.5, No.1, 2001.
- [12] Monika Henzinger, *Hyperlink Analysis for the Web*, *IEEE Internet Computing*, Jan-Feb 2001.

- [13] Richardson, M., A. Prakash, and E. Brill (2006). Beyond pagerank: machine learning for static ranking. In Proceedings of the 15th international conference on World Wide Web, pp. 707-715. ACM Press.
- [14] Vaughan, L. (2004): New measurements for search engine evaluation proposed and tested. In: Information Processing and Management 40(4), pp. 677-691
- [15] Wu, B.; Davison, B.D.: Identifying Link Farm Spam Pages. Proceedings of WWW 2005, May 10-14, Chiba, Japan, pp. 820-829
- [16] Zhai, C. (2008). Statistical Language Models for Information Retrieval. Foundations and Trends in Information Retrieval 2 (4).