Information Retrieval Issues on the World Wide Web

Ashraf Ali
Department of Computer Science, Singhania University, Pacheri Bari, Rajasthan
aali1979@rediffmail.com

Dr. Israr Ahmad
Department of Computer Science, Jamia Millia Islamia University, New Delhi
israr_ahmad@rediffmail.com

Abstract

The World Wide Web (Web) is the largest information repository containing billions of interconnected documents (called the web pages) which are authored by billions of people and organizations. The Web is huge, diverse, unstructured or semi structured, dynamic contents, and multilingual nature; make the effectively and efficiently searching information on the Web a challenging research problem. In this paper we briefly explore the issues related to finding relevant information on the Web such as crawling, indexing and ranking the Web.

Keywords: Web Information Retrieval, Crawling, Indexing, Ranking

1. Introduction

As the Web is growing with explosive speed and changing rapidly, finding information relevant to what we are seeking is becoming important. Among users looking for information on the Web, 85% submit information requests to various Internet search engines [9]. Given a few search keywords, the search engine would response by supplying thousand of web pages. To be retrieved and presented to the user by search engine, a web page may have passed through three ‘Obstacles’ [4]. These obstacles are crawling, indexing and ranking. First search engines try to find any useful and relevant web pages through crawler that crawl through the Web and rank sufficiently well in crawler prioritization; otherwise it will never make it into index. Second if the system reduces it searches by removing all frequent and non-significant words (such as “the”, “are”, “of”) and uses a global ordering of pages, which is one technique for efficient query processing on very large indexes, the page must be high enough in the index to avoid being reduced. Finally, having been crawled and not reduced, the page must rank highly enough in the result list that the user sees it.

2. Web Information Retrieval (WebIR)

The growth of the Web as a popular communication medium has fostered the development of the field of WebIR. WebIR can be defined as the application and theories and methodologies from Information Retrieval (IR) to the Web. However, compared with classic IR, the WebIR faces several different Challenges. The following are the main differences between IR and WebIR.

Size: The information base on the Web is huge with billions of web pages.

Distributed Data: Documents are spread over millions of different web servers.

Structure: Links between documents exhibit unique patterns on the Web. There are millions of small communities scattered through the Web.

Dynamics: The information base changes over time and changes take place rapidly. Also the Web exhibits a very dynamic behavior. Significant changes to the link structure occur in small periods of times (e.g. week), and URL and content have a very low half-life.

Quality of Data: There is no editorial control, false information, poor quality writing etc.

Heterogeneity: The Web is very heterogeneous environment. Multiple types of document format coexist in this environment, including texts, HTML, PDF, images, and multimedia. The Web also hosts document written in multiple languages.
**Duplication:** Several studies indicate that nearly 30% of Web’s content is duplicated, mainly due to mirroring.

**Users:** Search engines deal with all type of users, generally performing short ill-formed queries. Web information seeking behaviors also have specific characteristics. For example, users rarely pass the first screen of results and rarely rewrite their original query.

All of these characteristics and the nature of the Web require new approaches to the problem of searching the Web.

Open research problem and development in IR field can be witnessed through various research papers [2, 6, 13].

Baeza-Yates et al. [2], determined a set of research directions, such as retrieval of higher quality, combining several evidential sources to improve relevance judgments, Understanding the criteria by which users determine if retrieved information meets their information needs.

Henzinger et al. [6], determined several problems; such as pro-active approach to detect and avoid the spam, combine link-analysis quality judgments with text-based judgments to improve answer quality, quality evaluation.

Sahami [13], referred to high quality search results, dealing with spam and search evaluation, such as identifying the pages of high quality and relevance to a user’s query, Linked-based methods for ranking web pages, Adversarial classification, detecting spam, Evaluating the efficacy of web search engines.

The ultimate challenge of WebIR research is to provide improved systems that retrieve the most relevant information available on the Web to better satisfy users information needs.

2.1. WebIR Components

To address the challenges found in WebIR, the web search system needs very specialized architecture [10, 14]. Figure 1 shows the basic components of a WebIR System. Overall search engines have to address all these aspects and combine them into unique ranking. Below is a brief description of the main components of such systems.

**Crawler:** This includes the crawlers that fetch pages. Typically multiple and distributed crawlers operate simultaneously. Current crawlers continuously harvest the Web, scheduling operation based on web sites profiles.

**Repository:** The Fetched web documents are stored in a specialized database allowing high concurrent access and fast read. Full HTML texts are stored here.

**Indexes:** An indexing engine build several indices optimized for very fast reads. Several types of indices might exist, Including inverted indices, forward indices, hit lists, and lexicons. Documents are parsed for content and link analysis. Previously unknown links are feed to the crawler.

**Ranking:** For each query, ranks the result combining multiple criteria. A rank value is attributed to each document.

**Presentation:** Sorts and presents ranked documents.

All of the above aspects have contributed to the emergence of WebIR as an active field of research.

![Figure 1: Basic components of a Web Information Retrieval System](image)
have changed frequently over the years due to the emergence of new fields. Below is a summary of the main tasks and also of the new or emerging one.

**Ad-hoc**: This ranks documents using non-constrained queries in fixed collection. This is the standard retrieval task in IR.

**Filtering**: This selects documents using a fixed query in a dynamic collection. For example, “retrieve all documents related to ‘Research in India’ from a continuous feed”.

**Topic Distillation**: This finds short lists of good entry points to a broad topic. For example, “Find relevant pages on the topic of Indian History”.

**Homepage Finding**: This finds the URL of a named entity. For example, “Find the URL of the Indian High Commission homepage”.

**Adversarial Web IR**: This develops the methods to identify and address the problem of web spam, namely link spamming that affect the ranking of results.

**Summarization**: This produces a relevant summary of a single or multiple documents.

**Visualization**: This develops methods to present and interact with results.

**Question Answering**: This retrieves small snippets of text that contain an answer for open-domain or closed-domain questions.

**Categorization / Clustering**: This grouping the documents into pre-defined classes/adaptive clusters.

Sahami [13], identified several open research problems and applications, including stemming, link spam detection, adversarial classification and automated evaluation of search results. According to these authors, WebIR is still a fertile ground for research.

3. Web Crawling

A web crawler is a software program that browses and stores web pages in methodical and automated way [9]. Typically a web crawler starts with a list of URLs to visit, called the seeds. As the crawler visit these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are respectively visited according to the set of policies.

The behavior of a web crawler is the outcome of a combination of policies [3]:

- A selection policy that states which page to download
- A re-visit policy that states when to check for changes to the pages
- A politeness policy that states how to avoid overloading web sites, and
- A parallelization policy that states how to coordinate distributed web crawlers

3.1 Issues in the Web Crawling

The crawler module retrieves pages from the Web for later analysis by the indexing module. Given a set of seed Uniform Resource Locators (URLs), a crawler downloads all the web pages addressed by the URLs, extracts the hyperlinks contained in the pages, and iteratively downloads the web pages addressed by these hyperlinks.

Given the enormous size and the change rate of the Web, web crawling has many issues [1, 12] such as:

- Web pages are changing at very different rates. Crawlers that seek broad coverage and good freshness must achieve extremely high throughput, which poses many difficult engineering problems. Modern search engine companies employ thousands of computers and dozens of high-speed network links.
- Most comprehensive search engine currently indexes a small fraction of the entire Web. Given this fact, it is important for the crawler to carefully select the pages and to visit “important” pages first by prioritizing the URLs in the queue properly, so that the fraction of the Web that is visited (and kept up-to-date) is more meaningful.
- Some content providers seek to inject useless or misleading content into the corpus assembled by the crawler. Such behavior is often motivated by financial incentives.
- Due to the enormous size of the Web, crawlers often run on multiple machines and download pages in parallel. This parallelization is often necessary in order to download a large number of pages in a reasonable amount of time. Clearly these parallel crawlers should be coordinated properly, so that different crawlers do not visit the same web site multiple times, and the adopted crawling policy should be strictly enforced. The coordination can incur significant communication overhead, limiting the number of simultaneous crawlers.
The ultimate size of the Web and the impossibility of getting a perfect snapshot led to the development of web Crawler's ability to choose a useful subset of the Web to index.

The design of effective crawlers for facing the information growth problem can be witnessed through various papers [3, 7, 16].

4. Indexing:

The documents crawled by the search engine are stored in an index for efficient retrieval. The purpose of storing index is to optimize speed in finding relevant documents for a search query. Without an index, the search engine would scan every document in the corpus, which would require considerable time and computing power.

There are various methods that have been developed to support efficient search and retrieval over text document collections [2]. The inverted index, which has been shown superior to most other indexing scheme, is a popular one. It is perhaps the most important indexing method used in search engines. This indexing scheme not only allows efficient retrieval of documents that contain query terms, but also very fast to build.

4.1 Issues in Indexing the Web:

As of November 02, 2011, the indexed Web is estimated to contain at least 12.08 billion pages (http://www.worldwideWebsize.com). Due to the dynamic generation of web pages, estimated size of the Web is much longer than the actual size. The responsibility of the web search engine is to retrieve this vast amount of content and store it in an efficiently searchable form. Commercial search engines are estimated to process hundreds of millions of queries daily on their index of the Web. The perfect search engine would give the complete and comprehensive representation of the Web. Indeed, such search engine does not possible.

Indexing the Web holds significant challenges such as selecting which document to index, calculating index term weight, maintaining index integrity as well as retrieval from independent but related index.

Recent work on the challenges in indexing the Web includes the following problems [5, 8, 15]:

- Size of the database, search engine should not index the entire Web. An ideal search should know all the pages of the Web, but there are contents such as duplicates or spam pages that should not be indexed. So the size of its index alone is not a good indicator for the overall quality of a search engine.
- Keeping the index fresh and complete, including hidden content.
- Web Coverage, due to the dynamic environment no one knows the exact index size of the Web. Therefore it is too difficult to determine the Web coverage of search engines.
- Identifying and modifying malicious content and linking.
- Search engine facing the problem in keeping up-to-date with the entire Web, and because of its enormous size and the different cycles of individual websites.
- The Invisible Web is defined as the parts of the Web that general purpose search engine do not index.

5. Ranking the Web:

When the user gives a query, the index is consulted to get the document most relevant to the query. The relevance document then ranked according to their relevance, importance, factor etc.

5.1 Issues in Ranking the Web:

Given a search keyword on the Web, search engine estimates too many relevant documents for almost any query. For example, using “Web Information Retrieval” as the query, the search engine Google estimated that there were 46,500,000 relevant pages. Therefore, the issue is how to rank pages and present the user the “most” relevant page at the top.

There are several approaches to address the problem of presenting the “most” relevant page first. The currently most popular method to address the problem is by ordering the search result and presenting most relevant page first. This method is known as page ranking which is one of the important factors that makes Google currently the most successful search engine. Google uses over 100 factors in their methods to rank the search results [17].

Ranking algorithm is the heart of the search engines. PageRank, proposed by Brin et al. [14], is one of the most significant algorithms based on link analysis. It is used by the Google search engine to rank web results. The algorithm produced the final rank for each web page, its PageRank value. PageRank is more paradigm than a specific algorithm since there are multiple variations on same concepts [11].

One of the main problems with the PageRank paradigm is its weakness to direct manipulation. The
article is widely known as link spamming and its
detection is open research problem [13]. Different
implementation page rank have tried to overcome this
limitation

There are many Issues that affect ranking the Web
such as:

- Quality of the pages, anyone can publish
  anything, so there is no quality control.
- Duplicate content, mainly due to the mirror,
  meaning that identical documents appear on
  the Web with different URLs.
- Spam, Spamming refers to actions that do no
  increase the information value of a page, but
  dramatically increase its rank position by
  misleading search algorithm to rank it high
  than they deserve.

6. Conclusion:
The rapid growth of sources of information,
heterogeneity, dynamic and multilingual nature of the
Web generates the new challenges for the IR research
community, including crawling the Web in order to
find the appropriate web pages to index, indexing the
Web in order to support efficient retrieval for a given
search query, and ranking the Web in order to present
to the user most relevant web page at top.

This paper presents the short overview of some
main issues related to crawling, indexing and ranking
the Web and concludes for the listing research issues.

References:
[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina,
Andreas Paepcke, and Sriram Raghavan “Searching the
Web.” ACM Transactions on Internet Technology, 1(1):
August 2001
Information Retrieval, ACM Press
[3] Carlos Castillo (2005), Effective Web crawling, SIGIR
Retrieval, in Information Retrieval: Searching in the 21st
Century (eds A. G{oker and J. Davies), John Wiley &
Sons, Ltd, Chichester, UK
More than 11.5 billion pages. Proceedings of the Special
interest tracks and posters of the 16th international
conference on World Wide Web, May 10-14, 2005,
Chiba, Japan. pp. 902-903.
Challenges in Web Search Engines, 18th International
Joint Conference on Artificial Intelligence.
locating hidden-Web entry points,” in Proceedings of the
engines and Information Retrieval. In: Information
Services and Use 18(2005)3, 137-147.
Retrieval on the Web, ACM Comput. Surv, 32(2):
144(173), June 2000.
information retrieval paradigm, Journal of Theoretical
Ranking the Web frontier, In WWW 04; Proceedings of the
13th international Conference on World Wide Web,
Press.
Foundations and Trends in Information Retrieval Vol. 4,
the Web information retrieval, Pacific Rim International
Conference on Artificial Intelligence: 3157, p.3-12
large-scale hyper textual Web search engine”, Computer
Networks and, ISDN Systems, 30(1-7):107{117, April
1998.
Guardian Unlimited 6,9,2001, www.guardian.co.uk/
online/story/0,3605,547140,00.html
[16] Soumen Chakrabarti, Martin van den Berg, and Byron
Dom. Focused Crawling: A New Approach to Topic-
Specific Web Resource Discovery. In Proceedings of
the 8th World Wide Web Conference, Toronto, May
1999.
/googleranking-factors.htm"