ISSN (e): 2250 – 3005 || Vol, 04 || Issue, 6 || June – 2014 ||
International Journal of Computational Engineering Research (IJCER)

Challenges and Design Issues in Search Engine and Web Crawler

Rahul Mahajan[1], Dr. S.K. Gupta[2], Mr. Rajeev Bedi[3]

[1]*M.Tech Student, Beant College of Engineering & Technology, Gurdaspur, Punjab,India*
[2]*HOD & Associate Professor (Computer Science & Engineering Department), Beant College of Engineering and Technology, Gurdaspur, Punjab, India*
[3]*Assistant Professor (Computer Science & Engineering Department), Beant College of Engineering and Technology, Gurdaspur, Punjab, India.*

Abstract

With the drastic development of number of Internet users and the number of accessible Web pages , it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. To make searching much easier for users, web search engines came into existence. Web Search engines are used to find specific information on the World Wide Web. Without search engines, it would be almost impossible to locate anything on the Web unless or until a specific URL address is known. Hence it is better to know how these search engines actually work and how they present information to the user initiating a search. Web crawling is the process used by search engines to collect pages from the Web. Web crawlers are one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency .This paper discusses the issues and challenges involved in the design of the various types of crawlers and search engine.

Keywords: *Challenges, Design Issues, Web Crawler, Search Engine, Duplicate, Web Search Engine, Spam*

I. INTRODUCTION

With the exponential growth of information on the World Wide Web, there is a great demand for developing efficient and effective methods to organize and retrieve the information available. All the search engines have powerful crawlers that visit the internet time to time for extracting the useful information over the internet. A web-crawler [4] is a program/software or automated script which browses the World Wide Web in a methodical, automated manner. Web crawlers are the programs or software that uses the graphical structure of the Web to move from page to page. Web crawlers are designed to retrieve Web pages and add them or their representations to local repository/databases. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages that will help in fast searches. Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler (sometimes also known as a spider), which is an automated Web browser that follows every link it sees. Search engines [1], [2], [3] operate as a link between web users and web documents. Without search engines, this vast source of information in web pages remain veiled for us. A search engine [6] is a searchable database which collects information from web pages on the Internet, indexes the information and then stores the result in a huge database where from it can be searched quickly.

II. RELATED WORK

Matthew Gray [5] wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996. In 1998, Google introduced its first distributed crawler, which had distinct centralized processes for each task and each central node was a bottleneck. After some time, AltaVista search engine introduced a crawling module named as Mercator [16], which was scalable, for searching the entire Web and extensible. UbiCrawler [14] a distributed crawler by P. Boldi , with multiple crawling agents, each of which run on a different computer. IPMicra [13] by Odysseus a location-aware distributed crawling method, which utilized an IP address hierarchy, crawl links in a near optimal location aware manner. Hammer and Fiddler [7] ,[8] has critically examined the traditional crawling techniques. They purposed web crawling approach based on mobile crawlers powered by mobile agents. The mobile crawlers are able to move to the resources that need to be accessed in order to take advantage of local data access. After accessing a resource, mobile crawlers move on to the next server or to their home machine, carrying the crawling results in their memory.

The following sections describe the various issues and challenges in implementing these different categories of crawlers.

III. CHALLENGES IN SEARCH ENGINE

Web Search Engines are faced with number of difficult problems in maintaining and enhancing the quality of performance. The problems are either unique to particular domain. The various problems are [10] ,[11],[12]:

- Spam
- Content Quality
- Duplicate Hosts

3.1 Spam

The increasing importance of search engines to commercial web sites has given rise to a phenomenon we call “web spam”, that is, web pages that exist only to mislead search engines into mis-leading users to certain web sites. Web spam is a nuisance to users as well as search engines: users have a harder time finding the information they need, and search engines have to cope with an inflated corpus, which in turn causes their cost per query to increase. Therefore, search engines have a strong incentive to weed out spam web pages from their index.

3.2 Content Quality

The web is full of noisy, low quality, unreliable and contradictory content. 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 in web search or information retrieval. The web is so huge, so techniques for judging document quality are essential for generating good search results. The most successful approach to determining the quality on the web is based on link analysis, for instance Page Rank[16,17] and HITS[9] .

3.3 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 finding duplicate or near –duplicate pages in set of crawled pages is well studied [7]. Duplicate hosts 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. Standard check summing techniques can facilitate the easy recognition of documents that are duplicates of each other (as a result of mirroring and plagiarism). Web search engines face considerable problems due to duplicate and near duplicate web pages[15]. These pages enlarge the space required to store the index, either decelerate or amplify the cost of serving results and so exasperate users. The identification of similar or near-duplicate pairs in a large collection is a significant problem with wide-spread applications. In general, predicting whether a page is a duplicate of an already crawled page is very chancy work and lot of work is being done in this field but still it is not able to completely overcome this problem.

IV. WEB CRAWLER DESIGN ISSUES

The web is growing at a very fast rate and moreover the existing pages are changing rapidly in view of these reasons several design issues need to be considered for an efficient web crawler design. Here, some major design issues and corresponding solution are discussed below:-

How should the crawler get relevant pages to query? With the increase in web size, the number of applications for processing data also increases. The goal is to take advantage of the valuable information contain in these pages to perform applications such as querying, searching, data extraction, data mining and feature analysis. For some of these applications, notably for searching, the criteria to determine when a page is to be present in a collection are related to the page contents, e.g., words, phrases, etc.

What pages should the crawler download? In most cases, the crawler cannot download all pages on the Web. Even the 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**.**

How should the crawler refresh pages? Once the crawler has downloaded a significant number of pages, it has to start revisiting the downloaded pages in order to detect changes and refresh the downloaded collection. Because Web pages are changing at very different rates, the crawler needs to carefully decide what page to revisit and what page to skip, because this decision may significantly impact the “freshness" of the downloaded collection. For example, if a certain page rarely changes, the crawler may want to revisit the page less often, in order to visit more frequently changing ones

How should the crawling process be parallelized? 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.

How should crawler get time sensitive information? Time Sensitive Searching is an issue that needs to be addressed to get the time sensitive information from the web. Usually Search engines crawl the web and take vast snapshots of site content. As previous crawls are not archived so search results pertain only to a single, recent instant in time. As a result when users request some pages which require past data then search engines are unable to provide because it is not possible to search files that represents snapshot of the web over time.

V. CONCLUSION

In this paper, I have discussed various issues and challenges faced in the development of the search engine and crawler architectures. I have found the many of the issues and challenges in these architectures are common i.e. reducing the network bandwidth consumption, maintaining the freshness of the database and maintaining the quality of pages etc.

References:

[1] Arvind Arasu, Junghoo Cho, "Searching the Web", ACM Transactions on Internet Technology, August 2001.

[2] Brian E. Brewington, George Cybenko, “How dynamic is the web.”, In Proceedings of the Ninth Internationa l World-Wide Web Conference, Amsterdam, Netherlands, May 2000.

[3] Dirk Lewandowski, “Web searching, search engines an d Information Retrieval, Information Services & Use”, pp: 137-147, IOS Press, May 2005.

[4] Franklin, Curt, “How Internet Search Engines Work”, 2002, www.howstuffworks.com.

[5] Gray M., “Internet Growth and Statistics: Credits and background”, http:www.mit.edu

[6] Heydon A., Najork M., “Mercator: A scalable, extensible Web crawler.”, WWW, vol. 2, no. 4, pp. 219-229, 1999.

[7] J. Fiedler and J. Hammer,” Using the Web Efficiently: Mobile Crawling”, In Proc. Of the 7th Int'l Conf. of the Association of Management (AoM/IAoM) on Computer Science, San Diego, CA, pp. 324-329, August 1999.

[8] J. Fiedler and J. Hammer,” Using Mobile Crawlers to Search the Web efficiently”, International Journal of Computer and Information Science, vol.1, no.1, pp.36-58, 2000.

[9] J. Kleinberg , “ Authoritative sources in hyperlinked environment” , In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, pp:668-677,1998.

[9]. Mark Najork, Allan Heydon, “High- Performance Web Crawling”, September 2001.

[10]. Mike, Burner, “Crawling towards Eternity : Building an archive of the World Wide Web”, Web Techniques Magazine, vol 2 ,no 5, May 1997.

[11]. Niraj Singhal, Ashutosh Dixit, “Retrieving Information from the Web and Search Engine Application”, In Proceedings of National Conference on “Emerging Trends in Software and Network Techniques– ETSNT’09”, Amity University, Noida, India, Apr 2009 .

[12] Odysseas Papapetrou and George Samaras, “*Distributed Location Aware Web Crawling”*, 2004, ACM, New York, USA.

[13] P. Boldi, B. Codenotti, M. Santini and S. Vigna, “UbiCrawler: a Scalable Fully Distributed Web Crawler, Software, Practice and Experience”, Vol. 34, No. 8, pp. 711-726, 2004,

[14] D. Gomes and M.J. Silva, *The Viuva Negra Crawler*, Software, Practice and Experience, , Volume 38, No. 2, 2008

[15] Rahul Mahajan, Dr. S.K. Gupta, Mr. Rajeev Bedi , “A Survey of Duplicate And Near Duplicate Techniques, International Journal of Scientific & Engineering Research, Volume 5, Issue 2, February-2014

[16]. Sergey Brin, Lawrence Page, “The anatomy of a large - scale hyper textual Web search engine”, Proceedings of the Seventh International World Wide Web Conference, pages 107-117, April 1998**.**

[17] S.Brin , L. Page , R. Motwani , T. Winograd , “ What can you do with a Web in your Pocket?”,Bulletin of the Techanical Committee on Data Engineering, pp: 37-47 , 1998