Volume 3, Issue 6, June 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Review of Domain Based Crawling System

Radhika Gupta,
Department of Computer
Science", Swami ViVekanand
Institute of Engineering
& Technology, Punjab Technical University, Jalandhar, India

AP Gurpinder Kaur
Department of Computer
Science", Swami ViVekanand
Institute of Engineering
& Technology, Punjab Technical University, Jalandhar, India

Abstract- In this research paper we explore the various developments that have occurred to build crawler that feed the search engines. After systematic literature review of algorithms related to information retrieval, we have found that most of the search engines became irrelevant in terms of their results as internet grew, and the challenge remains as fresh as ever in developing algorithm that can have high precision and recall values. Since all search engines take their data fed using crawlers, it is critical to improve its working. Now, due to size Big Data Generic Crawlers are no longer applicable in real life. So there is an urgent need to develop a domain specific crawler built on stock of existing algorithms like LSI so that they become relevant again, the paper proposes such domain specific crawler algorithm.

Keywords: Web crawlers, Latent Semantic Index, Domain based crawler, Focused crawler, Recall, Precision.

I. Introduction

The exponential growth of the World-Wide Web has transformed it into an ecology of knowledge in which highly diverse information is linked in an extremely complex and arbitrary manner. Between 2005 and 2010, the number of Web users doubled, and was expected to surpass two billion in 2010 [1]. Early studies in 1998 and 1999 estimating the size of the web using capture/recapture methods showed that much of the web was not indexed by search engines and the web was much larger than expected [2][3]. According to a 2001 study, there were a massive number, over 550 billion, of documents on the Web, mostly in the invisible Web, or Deep Web [4]. A 2002 survey of 2,024 million Web pages [5] determined that by far the most Web content was in the English language: 56.4%; next were pages in German (7.7%), French (5.6%), and Japanese (4.9%). A more recent study, which used Web searches in 75 different languages to sample the Web, determined that there were over 11.5 billion Web pages in the publicly index able Web as of the end of January 2005 [6]. As of March 2009, the index able web contains at least 25.21 billion pages [7]. On 25 July 2008, Google software engineers Jesse Alpert and Nissan Hajaj announced that Google Search had discovered one trillion unique URLs [8]. As of May 2009, over 109.5 million domains operated [9]. Web search engines work by storing information about many web pages, which they retrieve from the World Wide Web. These pages are retrieved by a Web Crawler (sometimes also known as a spider). A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. The contents of each page are then analyzed to determine how it should be indexed (for example, words can be extracted from the titles, page content, headings, or special fields called Meta tags). Data about web pages are stored in an index database for use in later queries. A query can be a single word. The index helps information be found as quickly as possible [10]. The Current Challenge: Today Big Data is the name of the game, According to a new infographic posted by Intel, it's a staggering 277,000 logins every minute, even as six million Facebook pages are getting viewed in that same period. The infographic show below shows that 639,800 GB is transferred from one destination to another in one minute [11]; therefore, a search engine must keep a pace with the data produced and data crawlered fresh with optimal utilization of resources for doing so.

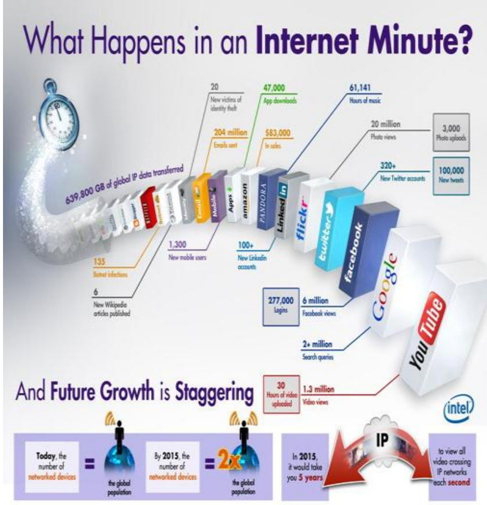

Fig1: Amount of data generated in one min

II. Related Work

The grandfather of all search engines was Archie, created in 1990 by Alan Emtage, a student at McGill University in Montreal [13]. The name stands for "archive" without the "v". The program downloaded the directory listings of all the files located on public anonymous FTP (File Transfer Protocol) sites, creating a searchable database of file names; however, Archie did not index the contents of these sites since the amount of data was so limited it could be readily searched manually. Veronica (*V*ery *E*asy Rodent-*O*riented *N*et-wide *I*ndex to *C*omputerized *A*rchives) was created as a type of searching device similar to Archie but for Gopher files [13]. Another Gopher search service, called Jughead (Jonzy's Universal Gopher Hierarchy Excavation and Display), appeared a little later [13]. Matthew Gray's World Wide Web Wanderer [13]. Was the first robot on the web and was designed to track the web's growth. Initially, it counted only Web servers, but shortly after its introduction, it started to capture URLs as it went along. The database of captured URLs became the Wandex, the first web database. In response to the Wanderer, Martijn Koster [13] created Archie-Like Indexing of the Web, or ALIWEB, in October 1993. As the name implies, ALIWEB was the HTTP equivalent of Archie, and because of this, it is still unique in many ways. Matthew Gray's Wanderer inspired a number of programmers to follow up on the idea of web robots, or spiders. These programs systematically scour the web for pages by exploring all of the links on a starter site, which is a page that contains many links to other pages. By December 1993, three search engines powered by robots had made their debut: Jump Station, the World Wide Web Worm, and the Repository-Based Software Engineering (RBSE) spider [13]. Jump Station's web bot gathered information about the title and header from Web pages and used a very simple search and retrieval system for its web interface. The system searched a database linearly, matching keywords as it went. Needless to say, as the web grew larger, Jump Station became slower and slower. The WWW Worm indexed only the titles and URLs of the pages it visited. It used regular expressions to search the index. Results from Jump Station and the Worm came out in the order that the search found them, meaning that the order of the results was completely irrelevant. The RSBE spider was the first to improve on this process by implementing a ranking system based on relevance to the keyword string. Unfortunately, these spiders all lacked the intelligence to understand what it was that they were indexing. Therefore, if you didn't specifically know what it was that you were looking for, it was unlikely that you'd find it. This deficiency prompted the creation of EINet Galaxy, now known as the Tradewave Galaxy, which is the oldest browsable/searchable web directory. Because it is a directory, Galaxy links are organized into hierarchical categories.

In April 1994, two Stanford University Ph.D. candidates, David Filo and Jerry Yang [13], created some pages that became rather popular. They called the collection of pages Yahoo Lycos and Infoseek followed Web Crawler, however, it was Digital Equipment Corporation's (DEC) AltaVista [13] which had a number of innovative features that quickly catapulted it to the top. The least of the features was its speed. AltaVista was the first to use natural language queries, meaning a user could type in a sentence like "What is the weather like in Tokyo?" and not get a million pages containing the word "What." Additionally, it was the first to implement advanced searching techniques, such as the use of Boolean operators (AND, OR, NOT, etc.). Furthermore, a user could search newsgroup articles and retrieve them via the web as well as specifically search for text in image names, titles, Java applets, and ActiveX objects. Additionally, AltaVista claims to be the first search engine to allow users to add to and delete their own URLs from the index, placing them online within 24 hours. HotBot from Inktomi was introduced in the year 1996. HotBot could index 10 million pages per day. Additionally, HotBot makes extensive use of cookie technology to store personal search preference information. The META engines forward search queries to all of the major web engines at once. The first of these engines was MetaCrawler. MetaCrawler searches Lycos, AltaVista, Yahoo!, Excite, WebCrawler, and Infoseek simultaneously. MetaCrawler was developed in 1995 by Eric Selburg, a Masters student at the University of Washington. Colorado State University also has a tool called Savvy Search that searches up to 20 engines at once, including a number of topic-specific directories such as Four11 (e-mail addresses), FTPSearch95 (files on the Net), and DejaNews (UseNet database). It's faster but less reliable than MetaCrawler.

III. Methodology

As mentioned earlier, there is an urgent need to make existing crawling algorithms to work more efficiently in terms of their recall and precision values while not ignoring the facts related to resources(bandwidth, time, CPU, utilization etc.) involved in running crawling algorithm. Therefore, we propose the following model:

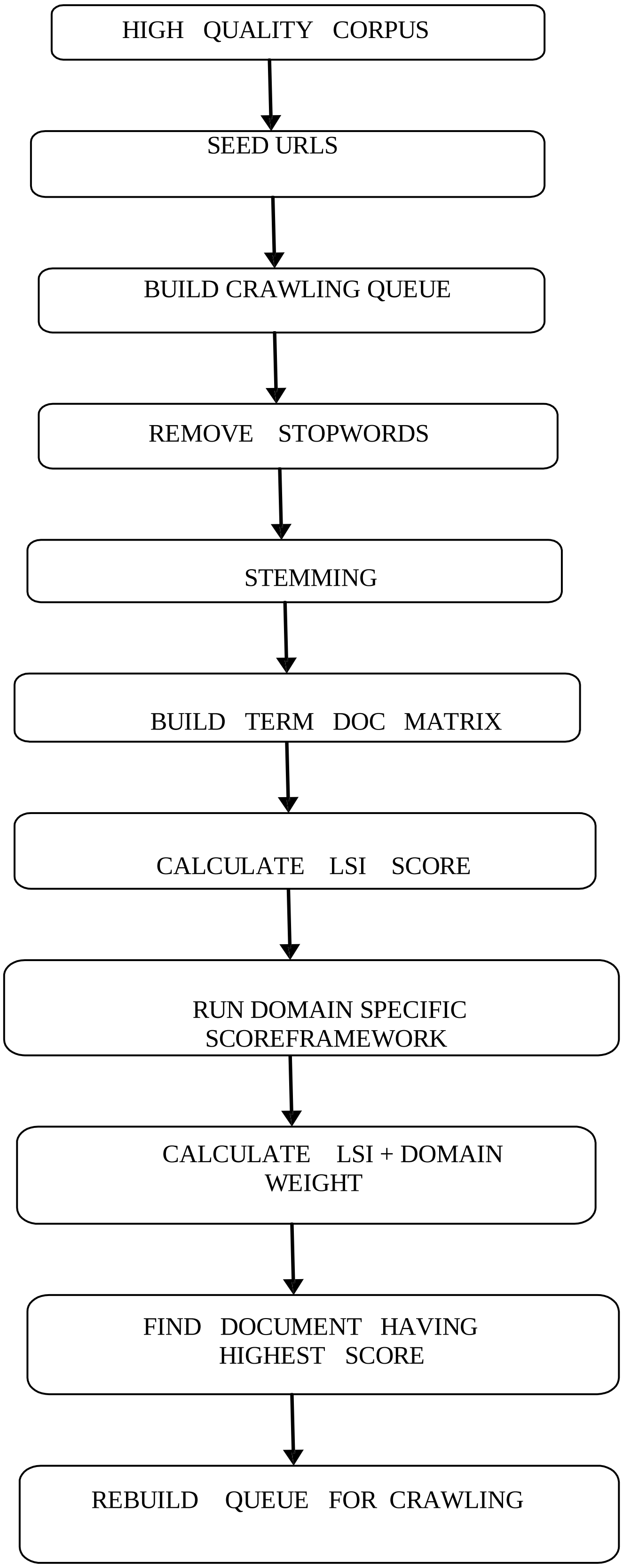
HIGH QUALITY CORPUS
SEED URLS
BUILD CRAWLING QUEUE
REMOVE STOPWORDS
STEMMING
BUILD TERM DOC MATRIX
CALCULATE LSI SCORE
RUN DOMAIN SPECIFIC SCOREFRAMEWORK
CALCULATE LSI + DOMAIN WEIGHT
FIND DOCUMENT HAVING HIGHEST SCORE
REBUILD QUEUE FOR CRAWLING

Fig2. Basic Flow of the research process

IV. Conclusion

Crawling the web is a highly resource intensive task which requires coordination of multiple threads and large spectrum of bandwidth. Secondly, crawling is semi-undeterministic approach for indexing and getting information, therefore, it is a necessity to develop an algorithm which helps in saving computational resources and bandwidth. Hence, the need for focused crawlers. These focused crawlers may be domain specific or knowledge specific in nature, which helps to develop an information retrieval system which will have high precision and recall values due to the fact that it has crawled highly relevant pages. The challenge is to develop algorithm which work on the principal of calculating score on the basis of context of the knowledge domain on which we are working on and the websites which are being crawled. Latent Semantic Indexing (LSI) is one of the promising models to do so and it can further be improvised and improved to give better results.

V. Future Scope

These days many information retrieval systems are being created based on taxonomies, ontologies, knowledge bases. The users want information based on particular domains which would help them save time and effort and would help them retrieve more relevant and useful results. However there is still lot to do in the field of domain specific crawlers. Creation of more domain based crawlers is suggested in various areas such as chemistry, biology, medicine, etc. We can also add other machine learning algorithms like probabilistic algorithms, neural network etc which may result in even better precision.

Acknowledgement

I am thankful to AP GURPINDER KAUR, ASST.PROF, SWAMI VIVEKANAND INSTITUTE OF ENGG. TECHNOLOGY, BANUR, for providing constent guidance and encouragement for this research work.

References

[1] Lynn, Jonathan (19 October 2010). "*Internet users to exceed 2 billion ...” Reuters*. Retrieved 9 January 2013. [2] S. Lawrence, C.L. Giles, "*Searching the World Wide Web,"* Science, 280(5360), 98–100, 1998.

[3] S. Lawrence, C.L. Giles, *"Accessibility of Information on the Web,"* Nature, 400, 107–109, 1999. [4] "*The 'Deep' Web: Surfacing Hidden Value*". Brightplanet.com. Retrieved December 21, 2012.

[5] Netz-tipp.de. Retrieved *"Distribution of languages on the Internet"*. December 27, 2012. [6] Alessio Signorini. *"Indexable Web Size"*. Cs.uiowa.edu. Retrieved December 27, 2012.

[7] *"The size of the World Wide Web"*. Worldwidewebsize.com. Retrieved December 27, 2012.

[8] Alpert, Jesse; Hajaj, Nissan (25 July 2008). *"We knew the web was big...". The Official Google Blog*. [9] "*Domain Counts & Internet Statistics*". Name Intelligence. Retrieved December 27, 2012.

[10] Jawadekar, Waman S (2011), "*8. Knowledge Management: Tools and Technology*", Knowledge Management: Text & Cases, New Delhi: Tata McGraw- Hill Education Private Ltd, p. 278, ISBN 978-0-07-07-0086-4, retrieved January 07 2012

[11] http://readwrite.com/2013/03/20/a-lot-can-happen-in-an-internet-minute#feed=/search?keyword=what can happen on internet in 1 min

[12] http://en.wikipedia.org/wiki/Web_crawler#cite_note-5 retrieved on January 21, 2013

[13] http://www.wiley.com/legacy/compbooks/sonnenreich/history.html retrieved on January 22, 2013

[14] Wenlei Mao, Wesley W. Chu, *"The phrase-based vector space model for automatic retrieval of free-text medical documents*”, Data & Knowledge Engineering Volume 61 Issue 1, April, 2007 Pages 76-92 (Elsevier).

[15] Sergey Brin, Lawrence Page, *"The anatomy of a large-scale hypertextual Web search engine*", Computer Networks and ISDN Systems 30 (1998) 107- 117 (Elsevier).

[16] J. Kleinberg, Authoritative sources in a hyperlinked environment, in: Proceedings of the Ninth Annual ACM-SIAM Symposium, Discrete Algorithms, January 1998, pp. 668-677.

[17] Yiming Yang, Xin Liu, *“A re-examination of text categorization methods"*, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval 1999.

[18] M,Deligenti, F Coetzel , S Lawrence , C Leegiles , M Gori , *“Focused Crawling using Context Graphs*”, presented in 26th International Conference on Very Large Databases,Cairo , Egypt,2000.

[19] Fan Wu, Ching-Chi Hsu, *"Topic-specific crawling on the Web with the measurements of the relevancy context graph*", Information Systems Volume 31 Issue 4-5, June, 2006. (Elsevier).

[20] Joon ho lee, myoung ho Kim, and yoon joon lee, *"Ranking documents in thesaurus-based boolean retrieval systems*", Information Processing and Management Vol. 30, No. 1. PP. 79-91. 1994.