Volume 4, Issue 5, May 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Focused Crawler: A Review

Manpreet Kaur
Student, Dept. of CSE
Chandigarh Engineering College (Mohali, India)

Yasmeen Kaur Dhaliwal
Assistant Professor , Dept. of CSE
Chandigarh Engineering College (Mohali,India)

Abstract: ***In this paper, various techniques are discussed which are proposed by researchers to increase the efficiency of the web crawler algorithms. We explore the various developments that have occurred to build crawler that feed the search engines. After symmetric 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 problem remains as fresh as ever in developing algorithm that can have precision value and recall value. Since all search engines take their data fed using crawlers, it is critical to improve its working. Due to size big data Generic Crawlers are no longer applicable in real life. So there is need to develop a domain specific crawler builds on stock of existing algorithms.***

Keywords: Focused crawler, Information retrieval system, and Domain based system.

I. Introduction

There is great demand for developing efficient and effective methods to organize and retrieves web pages because of exponential growth of information on World Wide Web. Focused crawler is an important method for collecting data on, and keeping up with the rapidly expanding internet. A web crawler is a relatively simple automated program or script that methodically scans or 'Crawl' through internet pages to create an index of data. The main goad of focused crawler is to selectively seek out pages that are relevant to pre-defined set of topics [1]. The topics are specified not using keywords, but using exemplary documents. It analyzes the crawl boundary to find links that are likely to be most relevant for crawl, rather than collecting and indexing all accessible web documents to be answer all possible ad-hoc queries.
A standard crawler crawls through all the pages in breadth first strategy. So if we want to crawl through some domain then it will be very inefficient. In Fig-1 we show the general crawler crawling activity [2].

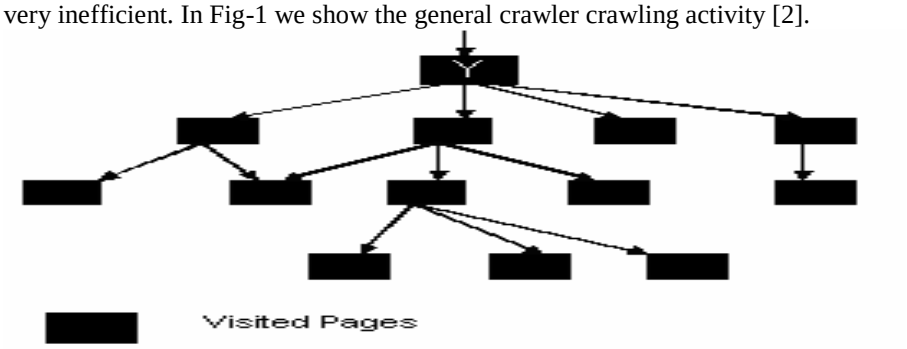

Fig-1. Standard Crawling [2]

If some crawler crawls only through domain specific pages then it is a focused crawler. From Fig- 2 we can see that a focused crawler crawls through domain specific pages. The pages which are not related to the particular domain are not considered.

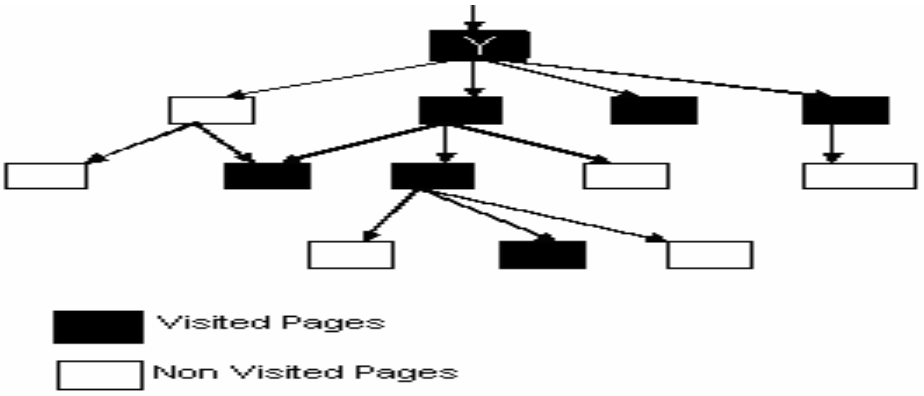

Fig-2. Focused (Domain Specific) Crawling [2]

- **Framework**

Fig-3 shows the architecture of focused crawling system.

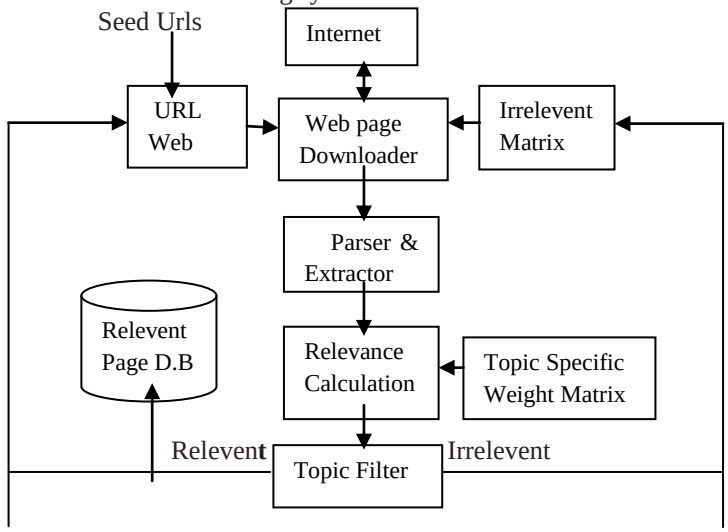

Fig-3. Framework of Focused Crawler [3]

URL Queue contains a list of unvisited URLs maintained by a crawler and is initialized with seed URLS. Web page downloader fetches URLs from URL queue and downloads corresponding pages from the internet [3]. The parser and extractor extracts information such as the terms and the hyperlink URLS from a downloaded page. Relevance calculator calculates relevance of a page with respect to topic, and assigns score to URLs extracted from a page. Topic filter analyzes whether the content of parsed page is related to topic or not. If the page is relevant, the URLs extracted from it will be added to the URL queue, otherwise added to the irrelevant matrix.

- **Use of Focused Crawler**

There are various uses of web crawler , but essentially a web crawler may be used by anyone seeking to collect database out on the internet search engines frequently use web crawlers to collect information about what is available on public web pages. Their primary purpose is to collect data so that search term on their site, then can quickly provide the surfer with relevant web sites. Linguistics may use a web crawler to perform a textual analysis that is, they may comb the internet to determine what words are commonly used today [4].

II. Related Work

Sk.Abdul Nabi et al [5]. addressed domain based information system in which system crawl the information from the web and added all links to the database which are related to specific domain. Because of that searching space and searching time decreases and as a result, it improves the performance of the system. The use pattern matching algorithm in which input is given as rank table of web page and then total rank is calculated. Web is very large and dynamic so searching the required relevant content from web is very difficult. Grouping the collection from the web is always challenging. So there is need to gather from broad range of domains. This work is only limited to no. of collections. So we need to validate large no. of collections from various domains.

Scott Deerwester et al [6]. found a new method for automatic indexing and retrieval. It is designed to overcome a fundamental problem that plagues existing retrieval techniques that try to match words of queries with words of documents. The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic. There are usually many ways to express a given content, so the literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user's query will literally match terms in documents that are not of interest to user. The proposed approach use statistical technique to estimate the latent structure, and get rid of the obscuring "noise". A description of terms and documents based on the latent semantic structure used for indexing and retrieval.

Radhika Gupta et al [7]. In this paper, developed a semi-deterministic algorithm and scoring system that takes benefit of the Latent Semantic Indexing scoring system for crawling web pages that belong to particular domain or is specific to the topic. The proposed algorithm calculates a preference factor in addition to the LSI score to determine which web page needs to preferred for precision values as it builds a queue which is specific to a particular domain/topic which would not have been possible in breadth first algorithm and only LSI based information retrieval systems.

Hong-Wei Hao et al [8]. Developed the improved topic relevance algorithm for focused crawling. Firstly, they implement a prototype system of the focused crawler. Second, experiments on Chinese text corpus show that using latent semantic indexing outperforms using TF-IDF (term frequency-inverse document frequency) for hyperlink topic relevance prediction and pages topic relevance calculation. Third, in real crawling experiments on the prototype system, the crawler using TF-IDF has high performance with the accumulated topic relevance increasing quickly at the beginning of the crawling, however the crawler using LSI can find more related pages and TF-IDF, they proposed TFIDF+LSI performs

the same crawl task and demonstrates the combination advantage of TF-IDF and LSI . However, due to the limitation of LSI and using only anchor text and other factors, topical crawler using TFIDF+LSI may still cause topic drift.

Ahmad Pesaranghader et al [9]. Proposed improved measure called Term frequency-Information Content (TF-IC) to prioritize terms in a multi-term topic accordingly. Through conducted experiments, we compare our measure against both Term frequency-Inverse Document frequency (TF_IDF) and Latent Semantic Indexing (LSI) measures applied in focused crawlers. Experimental results indicate superiority of our measure over TF-IDF and LSI for collecting more relevant web pages of both general and specialized multi-term topics.

M. Diligenti et al [10]. Presented a focused crawling algorithm that builds a model for the context within which topically relevant pages occur on the web. Because the major problem in focused crawling is performing appropriate credit assignment to different documents along a crawl path, such that short-term gains are not pursued at the expense of less-obvious crawl paths that ultimately yield larger sets of valuable pages. This context model can capture typical link hierarchies within which valuable pages occur, as well as model content on documents that frequently co-occur with relevant pages. Their algorithm further leverages the existing capability of large search engines to provide partial reverse crawling capabilities. The algorithm shows significant performance improvements in crawling efficiency over standard focused crawling.

Maintaining currency of search engine indices by exhaustive crawling is rapidly becoming impossible due to the increasing size and dynamic content of the web.

III. Conclusion

Crawling the web is 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 necessity to develop an algorithm which helps in saving computational resources and bandwidth. Hence, the need is to develop an information retrieval system which will have high precision and recall values due to the fact that it has crawled high relevant pages.

References

1. Soumen chakrabarti, Matrin van den Berg, Byron Dom, "*A New Approach to Topic – Specific Web Resource Discovery*" pp. 400 076, 1999.
2. Debajyoti Mukhopadhyay, Arup Biswas, Sukanta Sinha, "*A New Approach to Design Domain Specific Ontology Based Web Crawler*" West Bengal University of Technology, pp.70091.
3. Anshika Pal, Deepak Singh Tomar, S.C. Shrivastava, "Effective *Focused Crawling Based on content and Link Structure Analysis*" Vol. 2, No. 1, June 2009.
4. Gautam Pant, Padmini Srinivasan1, Filippo Menczer, "*Crawling the Web*", Department of Management Sciences, The University of Iowa, Iowa City IA 52242, USA.
5. Sk.Abdul Nabi, Dr. P.Premchand, "Effective *Performance of Information Retrieval by using Domain Based Crawler*", Vol. 4, No.7, 2013.
6. Scott Deerwester, Susan T. Dumais, George W. Furnas and Thomas K. Landauer, Richard Harshman, "*Indexing by Latent Semantic Analysis*" 41(6):391-407, 1990.
7. Radhika Gupta, AP Nidhi, "*Focused Crawling System based in Improved LSI*", Volume 2 Issue 9, September 2013.
8. Hong-Wei Hao, Cui-Xia Mu, Xu-Cheng Yin, Shen Li, Zhi-Bin Wang, "An *Improved Topic Relevance Algorithm for focused Crawling*".
9. Ali Pesaranghader, Ahmad Pesaranghader, Norwati Mustapha, Nurfadhlina Mohd Sharef, "*Improving Multi-term Topics Focused Crawling by Introducing Term Frequencyy-Information Content (TF-IC) Measure*", September 2013.
10. M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles and M. Gori, "*Focused Crawling Using Context Graphs*", NEC Research Institute, Princeton, NJ 08540-6634 USA.

