Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 6; April-June, 2015 pp. 546-548

http://www.krishisanskriti.org/acsit.html

An Optimal Focused Crawler

Tannu Priya[1], Prashant Singh[2] and Raj Kumar Singh Rathore[3]

[1,2]Student, Dept. of CSE Galgotia College of Engg. & Technology Greater Noida, UP.
3Dept. of CSE Galgotia College of Engg. & Technology Greater Noida, UP.
E-mail: 1preetypriya.533@gmail.com, 2prashantsingh9173@gmail.com, 3rathore.rajkumar@gmail.com

Abstract— *World wide web has approx. 295 exabytes of data and is growing day by day and till now shows no sign of end, so to get relevant information is very challenging task now a days. Web Crawler is software that accomplishes this cumbersome task of finding out relevant web pages out of all the web pages available on the World Wide Web. A web crawler systematically browses the World Wide Web for web indexing. Web crawler may be called web spider, an ant, an automatic indexer. Web search engines use web crawling to update their web contents. Crawlers apparently gained the name because they crawl through a site a page at a time, following the links to other pages on the site until all pages have been read. Web search engines deployed two types of web crawling strategies namely, "breadth" first search and "best" first search. The "best" first search strategy retrieves only those pages which are pertinent to a given topic. Crawler which uses a "best" first search strategy is identified as a "focused crawler". Focused crawler is a specialized crawler that traverses the web and selects the relevant pages to a defined topic rather than to explore all the regions of the web page. In this paper we would discuss about an optimal approach for focused crawling.*

Index Terms: *Focused crawler, Best first search Crawler, Breadth first search crawler, Automatic indexer, Optimal approach.*

1. INTRODUCTION

The World Wide Web is larger than it looks, with millions and billions of web pages that offer informational content spanning hundreds of domains and many languages across the globe. Due to the colossal size of the WWW [1], search engines have become the imperative tool to search and retrieve information from it [2] but the most prominent challenge with current web crawlers are they cannot download all the pages available

So they need to prioritize the URL's before downloading and parsing them.

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. Focused crawler seeks out only those pages that are relevant to the crawl, 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 Basic 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 technique. In Fig. 1 we show the general crawler crawling activity [2].

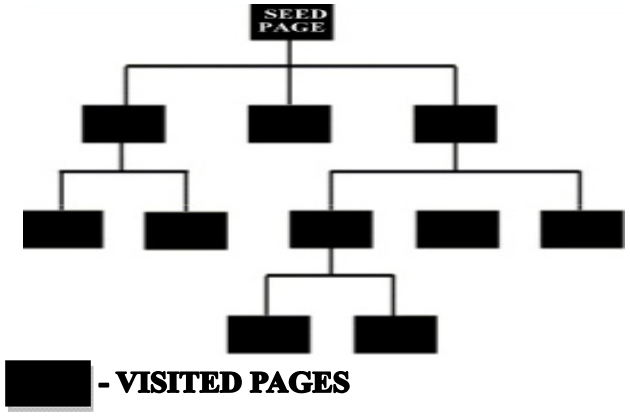

Fig. 1: Standard Crawling [2]

But Focused crawler crawl only those pages that are Domain specific and if they are not domain specific those pages are not crawled. From Fig. 2 we can see that a focused crawler crawls through domain specific pages.

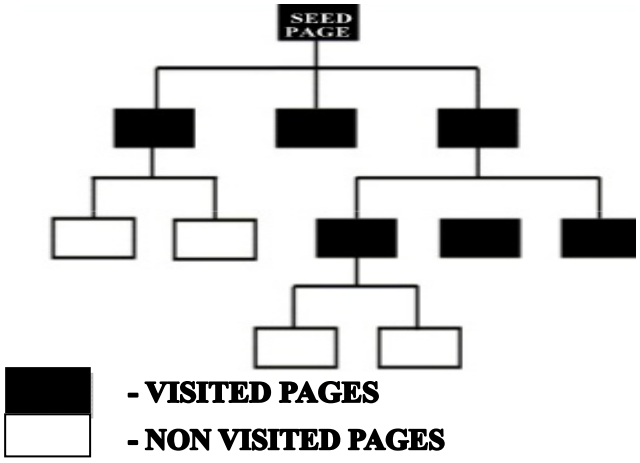

Fig. 2: Focused (Domain Specific) Crawling [2]

Focused crawler has the following two main components: (i) to find the relevancy of a specific web page to the given topic, and (ii) to find how to proceed from seed pages. Our Optimal Focused Crawler aims at providing a simplest alternative for conquering the issue that instantaneous page which are ranked lowly allied to the given topic at hand. By retrieving those pages which are reachable from the initial seeds, a set of candidate pages is obtained. We find the page which has the highest score with respect to the given topic, from the obtainable set of candidate pages. Set of pages again include this page and its relative pages, from which crawling process will get continue and then calculating the relevancy of page on the basis of hits.

2. RELATED WORK

The earlier crawlers [22] on which most of the web search engines are based uses traditional graph algorithms, such as breadth-first or depth-first traversal, to parse the web. A core set of URLs are used as a seed set, and the algorithm recursively follows hyperlinks down to other documents. Document content is paid little attention, as the ultimate goal of the crawler is to traverse the whole web. However, at that time, the web was two to three times smaller than it is today, so those systems did not address the scaling problems inherent in a crawl of today's web.

Depth-first crawling [22] follows each possible path to its conclusion before another path is tried. It works by finding the first link on the first page or the first seed page. It then crawls the page associated with that link, finding the first link on the new page, and so on, until the end of the path has been reached. The process continues until all the branches of all the links have been exhausted.

Breadth-first crawling [2] crawls each link on a page before proceeding on to the next page. Thus, it crawls each link on the first page and then crawls each link on the first page's first link, and so on, until each level of links has been exhausted.

Fish-Search [3] the Web is crawled by a team of crawlers, which are viewed as a school of fish. If the ''fish'' finds a relevant page based on keywords specified in the query, it continues looking by following more links from that page. If the page is not relevant, its child links receive a low preferential value. Shark-Search [4] is a modification of Fish-search which differs in two ways: a child inherits a discounted value of the score of its parent, and this score is combined with a value based on the anchor text that occurs around the link in the Web page.

A focused crawler is computer software used for finding information related to some specific topic from the WWW. However the main goal of focused crawling is that the crawler selects and retrieves pertinent pages only and does not need to gather all web pages. As the crawler is only a computer program, it cannot predict how pertinent a web page is [14]. In an attempt to search pages of a specific type or on a specific topic, focused crawlers aspire to recognize links that are probably to direct to target documents, and pass up links to off topic. Fish Search algorithm and Shark Search algorithm were used previously for crawling with topic keywords mentioned in query.

Naive Best First method exploits the fact that relevant pages possibly link to other relevant pages. Therefore, the relevance of a page a to a topic t, pointed by a page b, is estimated by the relevance of page b to the topic t.

3. TF-IDF RANKING

Term Frequency and Inverse Document Frequency are mathematical term that computes the importance of a word in accordance to a specific document.

Term Frequency (TF): A weight is assigned to each term in a document depending on the number of time that term occurs in the document. This weight is referred to as term frequency.

Inverse Document Frequency (IDF): Term frequency suffers from a significant problem all terms of the document considered equally important when it comes to assessing relevance on a query. In reality, certain terms have little power in characterizing the document. The terms such as full stops, spaces etc. are removed from the document before relevancy estimation.

4. SYSTEM ARCHITECTURE

Fig.3 depicts the system architecture where depending on the input keyword or query, related documents get downloaded from the internet. Then the relevancy of the document is calculated using TF-IDF and work of focused crawler starts by extracting links, finding the most relevant out of them.

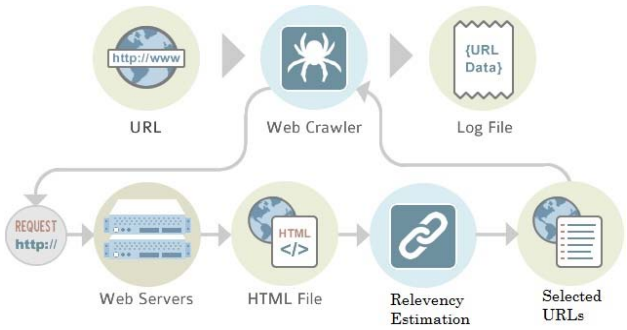

Fig. 2: System Architecture

We will be using various types of URL attributes for measuring that a particular link is relevant for the topic or not.

4.1 Average parent score (APS)

First we need to extract all parent pages of the unvisited link and then we will measure, relevancy of parent pages with those topic keywords. The page with highest Average Parent Score is extracted for further crawling.

Basic Algorithm
The basic Algorithm
{
Pick up the next URL
Connect to the server
GET the URL
When the page arrives, calculate relevancy
REPEAT
}

4.2 Anchor Text Relevancy (ATR)

Anchor Text Relevancy is the relevancy between topic keywords and the anchor text. We find out the synonyms of the word related to anchor text with the help of tool, and find out how much percentage of topic keywords are there in set of related words of topic keywords. The more topic keywords are in set of related words of anchor text, the anchor text is more relevant to topics. This is possible because anchor text describes the some information about URL.

4.3 Number of hits

In this we find number of times particular page is visited. Higher the number of hits and higher will be the priority of web page to be downloaded.

4.4 Estimated Relevancy

To estimate the optimal relevancy of a keyword no of times the link has achieved a hit is multiplied with 100 in order to increase the weightage of that link and then resultant is summed up with Anchor Text Relevancy.

Relevancy = (number of hits*100)+ATR

When estimating relevancy we give priority to the number of hits received by the page. On the basis of relevancy results can be displayed.

5. CONCLUSION

A Basic Crawler and Search engine Crawlers all the link in a document without measuring link relevance, it increases the number of resources required by crawler e.g. storage size, time etc. An Optimal focused Crawler approach calculates the link relevancy, if it relevant to given query then it store link in processing queue and remove irrelevant link. It saves time as well as memory space and also produce more relevant documents. In future work, we would like to select *N* pages (instead of just one) with the highest priority and endorse "Intelligent Crawling" [18].

REFERENCES

[1] S.Lawrence,C.L.Giles,: Searching the World Wide Web. Science, Vol.280, pp. 98-100, (1998) www.sciencemag.or

[2] R.Baeza-Yates, B.Ribeiro-Neto: Modern information retrieval (2nded.). Addison-Wesley-Longman

[3] S. Chakrabarti,M. Van Den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. "Computer Networks, 31(11-16) , pp.1623–1640, 1999.

[4] Aggarwal, C., Al-Garawi, F. & Yu, P., 2001. Intelligent Crawling on the World Wide Web with Arbitrary Predicates, In Proceedings of the 10th international conference on World Wide Web, Hong Kong, Hong Kong, pp. 96 – 105.

[5] Ehrig, M. and Maedche, A., 2003. Ontology-Focused Crawling of Web Documents. In Proceedings of the Symposium on Applied Computing 2003 (SAC 2003), Melbourne, Florida, USA, pp. 1174-

[6] Zhuang, Z., Wagle, R. & Giles, C. L., 2005. What's There and What's Not? Focused Crawling for Missing Documents in Digital Libraries. In Joint Conference on Digital Libraries, (JCDL 2005) pp. 301-310.

[7] Medelyan, O., Schulz, S., Paetzold, J., Poprat, M. & Markó, K., 2006. Language Specific and Topic Focused Web Crawling.In Proceedings of the Language Resources Conference LREC 2006, Genoa, Italy.

[8] McCown, F. and Nelson, M. "Agreeing to Disagree: Search Engines and their Public Interfaces". ACM IEEE Joint Conference on Digital Libraries (JCDL 2007). Vancouver, British Columbia, Canada. pp. 309- 318. June 17-23, 2007.

[9] Bao, S., Li, R., Yu, Y. and Cao, Y. "Competitor Mining with the Web Knowledge". IEEE Transactions on Data Engineering, Volume: 20, Issue: 10, pp. 1297-1310, Oct. 2008.

[10] J. Kleinberg, "Authoritative sources in a hyperlinked environment." Report RJ 10076, IBM, May 1997.

[11] Zhang, T. Zhou, Z.Yu and D.Chen, "URL rule based focusedcrawlers", IEEE International Conference on e-Business Engineering, 2008.

[12] M. Yuvarani, N. Ch. S. N. Iyengar and A. Kannan, "LSCrawler: A Framework for an Enhanced Focused Web Crawler based on Link Semantics" in Proceedings of the IEEEIWIC/ACM International Conference on Web Intelligence, 2006.

[13] Marcus, A. and Maletic, J. I., "Recovering Documentation-to-Source-Code Traceability Links using Latent Semantic Indexing", in *Proceedings 25th IEEE/ACM International Conference on Software Engineering (ICSE'03)*, Portland, OR, May 3-10 2003, pp. 125-137.

[14] Kraft, R. and Stata, R. "Finding buying guides with a web carnivore". *First Latin American Web Congress (LA-WEB'03).* 2003, pages 84–92.

[15] Pant, G., Tsjoutsiouliklis, K., Johnson, J., and Giles, C. L. "Panorama: Extending digital libraries with topical crawlers". *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries.* 2004b, p. 142–150.

[16] De Bra, P., Houben, G., Kornatzky, Y., and Post, R. "Information Retrieval in Distributed Hypertexts". *Proceedings of RIAO'94, Intelligent Multimedia, Information Retrieval Systems and Management, pages 481–491,New York,* 1994.

[17] Aggarwal, C., Al-Garawi, F. and Yu, P. "Intelligent Crawling on the World Wide Web with Arbitrary Predicates.". *Proc. 10th Intl. World Wide Web Conference.* 2001, pp. 96–105..