Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

An Extended model of Topic Driven Focused Crawler using Parallel Crawler

Rohit Kumar
M.Tech. Scholar
Suresh Gyan Vihar Univ.
Jaipur, Rj, India

Virendra Kumar
Assistant Professor
Suresh Gyan Vihar Univ.
Jaipur, Rj, India

Savita Shiwani
Associate Professor
Suresh Gyan Vihar Univ.
Jaipur, Rj, India

Dinesh Goyal
Associate Professor
Suresh Gyan Vihar Univ.
Jaipur, Rj, India

Abstract: ***The World Wide Web today is growing rapidly. It has enabled a publishing explosion of useful internet data, which bring out the fateful side effect of information overload. Now it is challenge for the search engines to make available the relevant information to information seekers. Search engines mainly depend on crawlers to download and index the web. The main task of crawler is to crawl the web for indexing purpose and keeps web pages more up-to-minute, later used this method by search engine to satisfy the end user queries. Major part of the WWW is dynamic therefore a need arises constantly to update the modified web pages continuously. A focused crawler is dedicated to crawl and download a specific part of the WWW. This paper proposes a novel architecture of topic specific crawler which is based on parallel crawling, which makes crawling function more scalable, operative and payload-sharing among the various crawlers which work parallel and download web pages.***

Keywords: Parallel crawler, Focused crawler, WWW, Search Engine, Information Retrieval

I. INTRODUCTION

The program that visits websites to crawl their pages and other data in favour of creating entries for a search engine index is called as crawler. The major search engines on the WWW have corresponding program which is basically known as crawler or sometimes spider. Entire thorough sites or specified pages can be selectively visited or indexed[12]. Crawlers crawl throughout the website and scan a page at a time followed by the links to different pages on the website until all pages present on the website have been scan. Due to huge size of World Wide Web, it is very tough task to gathered useful information, also difficult to design high performance crawling systems. The search engines which works at very large scale frequency updating their index to meet the user requirement and keep themselves up to date. But most of them fail to do so and consume lots of resources and network bandwidth because size of the internet is very large. The basic target of any crawler is to download important web pages first but while working in parallel each crawler may not be aware about the all over collection of the web pages and it downloads the pages collectively at a time. It may increases the downloading ability but it might lead towards poor crawling decisions.

The main objective of a focused crawler or topic specific crawler is to selectively restore pages which are proportional to a pre-defined set of contents. Rather than gathering and making index of all affordable Internet documents to be capable to response all possible ad-hoc queries, a topic specified crawler determine its crawl region to search the links that are probably to be most affording for the crawl, and expel irrelevant area of the Internet. This leads to drastic savings in hardware resources and network overloads, and helps in making the crawl more up-to-minute.

II. RELATED WORK

In the year 1999, Soumen Chakrabarti et al. [1] proposed a new hypertext resource discovery system called Focused Crawler. The main objective of the topic specified crawler is to particularly find out the pages that are applicable to a pre-defined collection of content. The collection of contents are specified and not used the keywords given by the users but basically used the paradigmatic documents. Instead of collecting and indexing all popular web documents to make answerable of all relevant ad-hoc queries of the users.

A topic specific crawler analyses its crawl limit to search the links or hypertext links which is most applicable for the crawl purpose, and avoids inapplicable boundaries of the web [1]. This leads to meaningful savings in system hardware and internet networks resources, hence it results in making crawler more up-to-the-minute. The architecture of focused crawler shown in figure. 1.

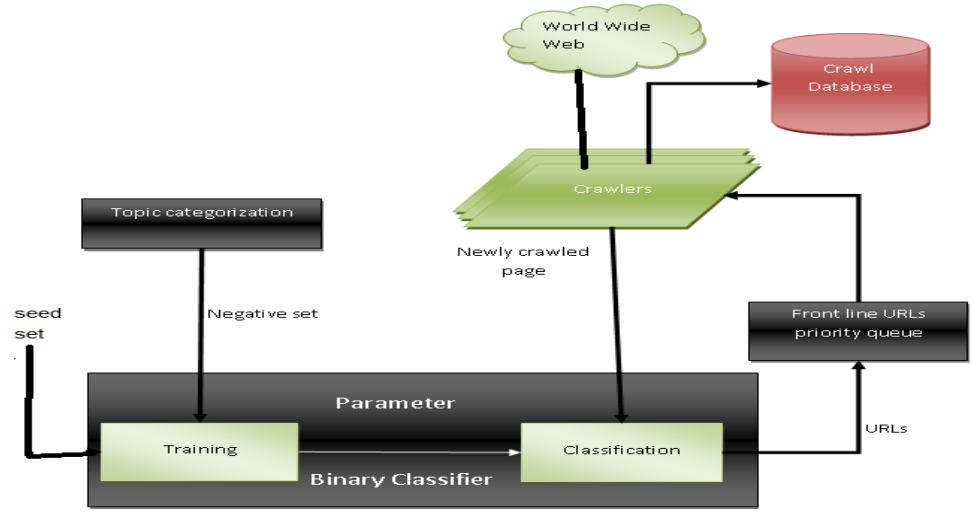

Fig. 1 Architecture of focused crawler

In [11] the year 2002, Junghoo Cho, Hector Garcia – Molina proposed a paper on parallel crawler. This paper basically deals with the design and implementation of a effective parallel crawler. As we know that size of World Wide Web is very huge, it becomes mandatory to parallelize the crawling mechanism, in order to complete the downloading process in the given period of time. In this paper [11] the authors first of all proposed multiple architecture for the mechanism of a parallel crawler and distinguish fundamental issues affiliated to the parallel crawling. The architecture of parallel crawler are shown in figure. 2.

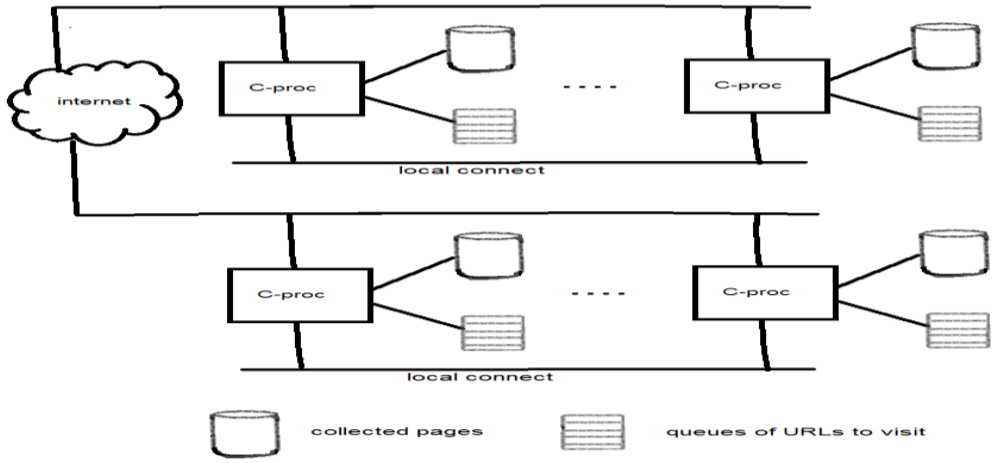

Fig. 2 Architecture of parallel crawler

III. PROPOSED WORK

All In this paper a new architecture for topic specific focused crawler has been proposed. The proposed architecture work in parallel by using C-Procs. The system will start a separate dedicated c-proc for every new topic added in the crawler. The architecture is shown in figure 3.

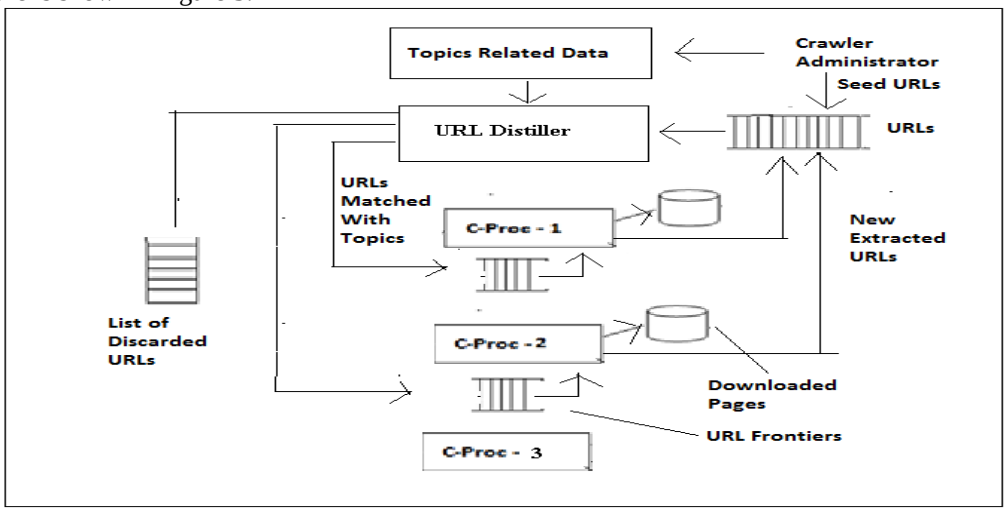

Fig. 3 Proposed Architecture Work in Parallel by using C-Procs

Major components of the proposed architecture are:

URL Distiller: - This module will perform the filtering task for focused crawler. It will fetch URLS from list of URLS and check whether it is related to any of the topic or not. If it is related to any of the topic then distillers will add that URLs to URL-frontier of the corresponding c-proc which is dedicated to that topic. Otherwise distiller will add it in the list of discarded URLs.

Topic related data: - It is a repository that will store the information related to topics. This information is helpful for URLs distiller to decide whether URLs is related to topic or not.

C-Proc: - This module performs all the functions of an individual crawler. The proposed system initiates a separate instance of c-proc for downloading pages for a particular topic

URL frontiers: - It is a list of URLs which are to be downloaded by c-proc. URLs are added in this list by the URL distiller. C-proc fetch URLs from this list, download the web page, extract further out links from that page and add these out links to list of URLs to be distilled by the distiller.

List of Discarded URLs: - It is a list of URLs which are not related to any of the topic.

Downloaded Pages: - It is repository in which downloaded web pages get stored. Pages in this repository will be stored by the c-proc.

Crawler Administrator:- Crawler administrator is the key person to control the proper functioning of the proposed crawler. The administrator initiates the seed URLs and topic related for each and every topic for which crawler is dedicated.

Seed URLs: - it is a list of initial URLs from where crawling process will get started. The administrator provides an initial list of URLs before starting the crawling process.

Procedure/Algorithm for topic driven focused crawler using Parallel Crawler:

1. The Crawler administrator enters the seed URLs for different topics on which crawler is focused. Administrator also provide topic specific data to the URLs distiller which is helpful in relating a URL with given topics. This data contain name of the topics, list of words in the domain of the topics etc.
2. URL Distiller (UD) fetches URLs from a list of URLs and check whether it is related to any of the topic or not. If yes then UD add that URL to the URL frontier of the c-proc dedicated to that topic. If no then UD add that URL to list of discarded URLs which are not related to any of the topics.
3. C-Procs fetches URLs from their specific URL frontiers, download the page, add that page in the repository, extract hyperlinks from it and add these hyperlinks to the list of URLs. Thus c-proc does all the functions of a normal crawler.
4. Steps 2 and 3 are repeated till list of URL is not empty or crawler stopped by the administrator.

IV. CONCLUSION AND FUTURE WORK

The In this paper a new architecture for topic driven focused crawler with the help of parallel crawler has been proposed. The proposed architecture tries to make the crawler scalable using c-procs and efficient by using URLs distiller. So this proposed architecture will start the crawling process from a list of seed URLs for a specified topics and crawl the web with a very fast speed by using C-Proc processes.

Further some work is needed to improve the working of URLs distiller to relate a URL to a topic. The proposed architecture need to be implemented and tested on some topics and results are to be examined.

ACKNOWLEDGMENT

The author specially thanks to the college Suresh Gyan Vihar University for giving such a environment to promote research work and also the faculty member of the college. Without help of all of them this dissertation work can't be completed within time. Among all the faculty member Mr. Virendra Kumar and Mrs. Savita Shiwani specially help a lot.

REFERENCES

[1] S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", *Proc. Of 8th International WWW conference*, Toronto, Canada, May,1999.

[2] Diligenti M., Coetzee F.M., Lawrence S., Giles C.L., Gori M., "Focused Crawling using context graphs", *Proc. International Conference on Very Large Databases (VLDB '00)*,2000,pp. 527-534.

[3] T. Tang, D. Hawking, N. Craswell, and K. Griffiths, "Focused crawling for both topical relevance and quality of medical information," in *Proceedings of CIKM'2005*, Bremen, Germany, 2005.

[4] M. Jamali, H. Sayyadi, B. B. Hariri, and H. Abolhassani. A method for focused crawling using combination of link structure and content similarity. In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, WI '06, pages 753–756, Washington, DC, USA, 2006. IEEE Computer Society.

[5] Naresh Chauhan, A. K. Sharma, " Design of an Agent Based Context Driven Focused Crawler" published in BVICAM'S International Journal of Information Technology, 2008.

[6] S. Mali, B. B. Meshram, "Focused Web Crawler with Revisit Policy" published in International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET, Mumbai, India.

[7] Babaria Rashmin N., "Focused Crawling" thesis report submitted to Computer Science and Automation Indian Institute of Science BANGALORE – 560 012, July 2007.

[8] Rajender Nath, Naresh Kumar, "A Novel Architecture for Parallel Domain Focused Crawler" published in International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 5, September- October 2012, pp.266-269.

[9] M. M. G. Farag, M. S. K. Khan, Gaurav Mishra, P. K. Ganesh, Edvard A. Fox, "Focused Crawler" published in CS5604 Information Storage and Retrieval Fall 2012 Verginia Technology.

[10] Qiuyan HUANG, Qingzhong LI, Zhongmin YAN, Hong FU, "A Novel Incremental Parallel Web Crawler based on Focused Crawling" published in Journal of Computational Information Systems 9: 6 (2013) 2461–2469.

[11] J. Cho and H. Garcia-Molina. Parallel crawlers. In *Proc. of the 11th International World-Wide Web Conference*, 2002.

[12] Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.

