Study of Distributed Web crawler using Web Mining

Akshay S. Kende **Prof. Bhushan V. Choudhari**

Abstract — **As Web is the large information repository, users find resources by following hypertext links. These links connect from one document to another. In the small systems where resources share the same fundamental classification, users can find resources easily and in efficient manner. perhaps Web now encompasses millions of sites with many different information, navigation is difficult. WebCrawler, is the efficient full-text search engine. It is a tool that assists users in their Web surfing by automating the task of link traversal, creating a searchable index of the web, and fulfilling searchers' queries from the index. Conceptually the distributed crawler harnesses the excess bandwidth and computing resources of clients to crawl the web. In this paper we are going to review some basic concepts of distributed web crawling by using mining.**

Key Words — **Distributed web Crawler, Webpage, web mining.**

I. INTRODUCTION

Internet is the shared global computing network. It enables global communications between all connected computing devices. It provides the platform for web services and the World Wide Web. Web is the totality of web pages stored on web servers. In the early days of the Web, manually locating relevant information was reasonably easy due to the limited amount of information that was available. but There is a spectacular growth in web-based information sources and services. It is estimated that, there is approximately doubling of web pages each year.

In other words The World Wide Web is an architectural framework for accessing linked documents spread out over millions of machines all over the Internet. This hypertext pool is dynamically changing due to this reason it is more difficult to find useful information. So Web Crawler for automatic Data and Web Mining is Useful to Us [2].

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. The web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services[4] .

Thus Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data. It implicitly covers the standard process of knowledge discovery in databases (KDD)

II. BACKGROUND

This paper is mostly related to work in distributed web crawling for load balancing in network.

A.*WEB CRAWLING*

A web crawler is a program that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks [1]. They are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries.

- Check for the next page to download. the system keep track pages to download In a queue.
- Check to see if the page is "allowed" to be downloaded - checking a “robots exclusion” file and also reading the header of the page to see if any exclusion instructions were provided do this. Some people don't want their pages archived by search engines [3].
- Download the whole page.
- Extract all links from the page (additional web site and page addresses) and add those to the queue mentioned above to be downloaded later.
- Extract all words, save them to a database associated with this page, and save the order of the words so that people can search for phrases, not just keywords
- Optionally filter for things like adult content, language type for the page, etc.
- Save the summary of the page and update the "last processed" date for the page so that the system knows when it should re-check the page at a later date.

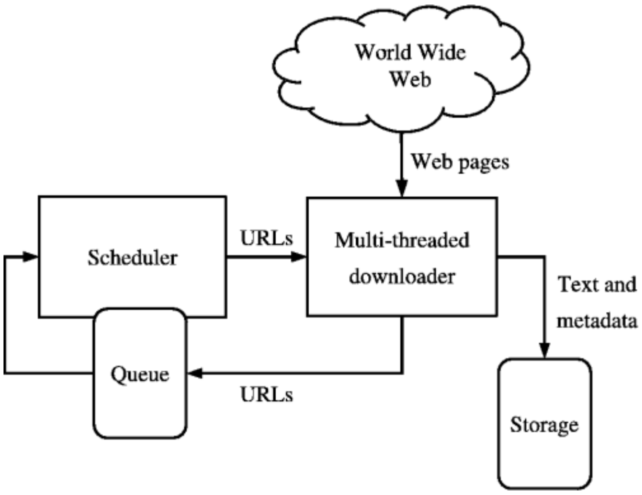

Fig.1 web crawler architecture

B. Distributed Computing

Distributed computing is a field of computer science that studies distributed systems. A *distributed system* is a software system in which components located on networked computers communicate and coordinate their actions by passing messages. The components interact with each other in order to achieve a common goal. Three significant characteristics of distributed systems are: concurrency of components, lack of a global clock, and independent failure of components [6].

III. DISTRIBUTED WEB CRAWLER

Internet search engines now a day's use some advance techniques, one of them is "Distributed web crawling" which is based on distributed computing to index the Internet via web crawling. Such systems may allow for users to voluntarily offer their own computing and bandwidth resources towards crawling web pages. It is costly to maintain a single machine with high configuration for web crawling, so task is divided across number of systems in order to perform load balancing. This technique is efficient and cheaper [5].

This crawler runs on network of workstations. Indexing the web is a very challenging task due to growing and dynamic nature of the web. Due to explosive growth of world wide web, now it is inevitable to use parallelism techniques in order to complete huge crawling tasks in limited time.. A single crawling process even with multithreading will be insufficient for the situation. In that case the process needs to be distributed to multiple processes to make the process scalable. It scales to several hundred pages per second. In distributed web crawler a URL server distributes individual URLs to multiple crawlers, which download web pages in parallel, the crawlers then send the downloaded pages to a central indexer on which links are extracted and sent via the URL server to the crawlers. This distributed nature of crawling process reduces the hardware requirements and increases the overall download speed and reliability along with [5].

Benefits of Distributed web crawling:

1.High crawling throughput and low latency because of spatial locality.

2.Improved network politeness due to distributed servers.

3.Increased availability due to load balancing.

4.Reduced data migration.

IV. PROPOSED WORK

In banking large amount of data is stored on centralized server . These information is use by all branches as well as ATM centers belonging to that bank. The difficulty in these system is that whenever a user want to perform a bank transaction or an ATM transaction his/her data need to be fetch from the centralized server only .these server may be located thousands of miles away From the local branch or local ATM.

For an instance suppose these are 5 lacks transaction occurs for a particular bank or ATM centre. we need to access centralized server for 5 lacks time .these processes involves heavy load on the centralized server as well as heavy traffic on the network bandwidth , also the time take for requesting and retrieving data is considerably high because of physical location of centralized server.

Theses difficulty can be resolved to some extend with the use of distributed web crawling. If we can have small distributed server for each region, that will connect numerous branches and ATM centers in that region. In this case customers of the bank in the target region will transact with the distributed server instead of centralized server. Due to this, load of the centralized server will be divided among all distributed servers. Also network bandwidth will be consumed less than previous case. As the distributed server will be located at the relative nearer location the time for requesting and retrieving information will be reduced considerably.

To maintain the information consistency between centralized and distributed server we need to perform synchronization of data at regular time interval.

Similarly, above distributed web crawling techniques can be used to connect the servers of government offices all across the country.

CONCLUSION

After studying above paper we can conclude that to meet the need of handling huge amount of information distributed web crawlers play an important role because we can reduce cost , time as well as network bandwidth consumption. Also load balancing can be done with the help of distributed web crawlers to reduce the load on centralized server.

REFERENCES

[1] Dr.Ashutosh Dixit,"Web crawler design issue:A review",IJMIE,vol.2,issue 8.ISSN 2249-0558,August 2012

[2] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas, "The web changes every- thing: Understanding the dynamics of web content," in Proceedings of the 2nd International Conference on Web Search and Data Mining, 2009.

[3] P. Boldi, B. Codenotti, , M. Santini, and S. Vigna, "UbiCrawler: A scalable fully distributed web crawler," Software — Practice & Experience, vol. 34, no. 8, pp. 711–726, 2004.

[4] Dhiraj Khurana, Satish Kumar, IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012

[5] Birrell, A.D., Levin, R., Needham, R.M. and Schroeder, M.D. Grapevine: an exercise in distributed computing. Communications of the ACM, 25 (4) 260-274.1992

[6] http://en.wikipedia.org/wiki/Distributed_web_crawling.

AUTHOR'S PROFILE

AKSHAY S. KENDE

A.S.Kende has completed degree of Bachelor of Engineering in Computer Science and Engineering discipline from Sipna college of engineering and technology,Amravati Maharashtra.

BHUSHAN V. CHOUDHARI

B.V Choudhari has completed master's degree in COMPUTER APPLICATION(MCA) from PROF.RAM MEGHE INSTITUTE OF TECHNOLOGY,Amravat.Currently working as assistant professor in P.G.D.C.S.T,D.C.P.E
H.V.P.M college,Amravati