IJARCCE

ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 10, October 2021

DOI: 10.17148/IJARCCE.2021.101030

Role of Web Crawler for Network Load Reduction

Dr. Kompal

Govt. College, Panchkula

Abstract: In today's world, the rising popularity of the Web has drastically expanded the probability of sharing significant information and knowledge on a large scale certainly not seen before. Owing to the availability of the bulk amount of data, the search services on the World Wide Web (WWW) are causing more demand among users. Regardless of beneficial part by conventional term-based search engines, precise filtering and retrieving relevant data from the Web to reduce network load is considered as a challenging task.

I. INTRODUCTION

Page relevance is the fundamental aspect for the Web search, as it supports the current and novel search engines, indexing, crawling, and ranking. Traditionally, the computation of page relevancy is performed using the link analysis methods, which uses the hyperlink graph of the Web. These methods acquire the connection from one Webpage to another as an endorsement of the linking page and presume that the more links pointed to a page; the more liable it is significant [1]

II. NETWORK LOAD REDUCTION

Mobile crawlers, specifically, are designed to operate on mobile devices, which typically have limited resources and bandwidth compared to desktop computers [2].

Mobile crawlers identifies the pages which are modified at the remote sites without downloading these pages but it downloads only those pages which are actually amended since last crawl run. This reduces the internet traffic and burden on remote sites significantly [3]. This system can be executed with the help of java aglets. Mobility with respect to Web Crawling is the capability of a crawler to shift itself to each Webserver of its interest before saving pages on that server. Once it completes the saving process on a particular server, the crawler with the saved data moves to the next server or to its home system.

Mobile crawlers are managed by a crawler manager, which supplies each crawler with a list of target Web sites and monitors the location of each crawler. This is necessary to intervene in case one or more crawlers happen to interfere with each other (i.e., crawl the same Web space). However, the crawling strategy and path taken are controlled separately by each crawler through the crawling algorithm. In addition, the crawler manager provides the necessary functionality for extracting the collected data from the crawler for use by the indexer.

Traditional crawlers always download data more than it effectively use (worst case is the whole web). A mobile crawler is directed to each Web source which is expected to contain relevant information for a local preselection of pages.

Firstly, the crawler gets a list of target locations from the crawler manager. These addresses are stated as seed URLs since they indicate the start of the crawling process. In addition, the crawler manager also uploads the crawling strategy into the crawler in form of a program. This program tells the crawler which pages are considered relevant and should be collected. In addition, it also generates the crawler path through the Web site.Before the actual crawling begins, the crawler must migrate to a specific remote site using one of the seed URL's as the target address. After the crawler successfully migrated to the remote host, the crawling algorithm is executed.

The ability of a mobile web crawler to decide carefully about the pages to be transferred through the network, results in significant reduction of load on the resources of the web server and underlying network

The major advantages of a mobile crawler over the traditional crawler are as follows:

i) Mobile crawlers access web pages on local server's results in saving network bandwidth by reducing request/response messages used for data collection.

DOI: 10.17148/IJARCCE.2021.101030

ii) The mobile crawlers are able to collect only the relevant pages before transmitting the relevant pages over the network and results in saving the bandwidth by removing irrelevant information directly at the data source.

iii) The mobile crawlers decrease the contents of web pages before it transmits the contents over the network. This saves network bandwidth by removing irrelevant portions of the retrieved pages.

iv) The mobile crawlers compress the contents of web pages before it transmits them over the network. This saves network bandwidth by decreasing the size of the retrieved data.

Issues/problems with existing mobile crawling technique:

i) The mobile crawlers which always reside in the memory of remote system do occupy a large portion of it.
ii) When number of mobile crawlers from different search engines exist and all mobile crawlers will reside in the memory of the remote system and will consume larger part of the memory which could have been utilized for some other activities.
iii) It is also possible that remote system might not allow mobile crawlers to stay permanently in its memory due to security concerns.
iv) The frequent changes in pages require that the mobile crawlers immediately assess the changed page and direct it to search engine to update the index. This result in excess consumption of network bandwidth and CPU cycles etc.
v) The studies of [4], proposed distributed and parallel crawling systems to enhance the coverage and to reduce the bandwidth consumption but these systems only distribute and localized the load but did not helped in reducing the load. The studies of [5], proposed web crawling approach based on mobile crawlers powered by mobile agents. These mobile crawlers can exploit the information about the pages being crawled in order to reduce the amount of data that needs to be transmitted to the search engine. These mobile crawlers move for accessing the resources. After accessing a resource, mobile crawlers move on to the next server or to their home machine, carrying the crawling results in their memory. The main advantage of mobile crawling is localized data access, remote page selection, filtering and compression.
vi) There are significant security issues due to crawler migration and remote execution of code because a mobile crawler might have harmful codes. Thus the challenge is to explore some methods so that mobile crawlers can be separated from harmful codes.
vii) Integration of the mobile crawler virtual machine into the Web. The mobile crawling will be effective only where the mobile crawler virtual machine exists on most of the machines. This integration can be achieved through Java Servlets.

III. CONCLUSION

The strategies to achieve network load reduction using a mobile crawler are User-Agent and Request Headers, Crawl Rate and Frequency, Content Filtering, Data Compression, Efficient Storage, Battery Optimization. Network load reduction should not compromise the effectiveness of mobile crawler in gathering the desired data.

REFERENCES

[1]. S. Brin and L. Page “ The Anatomy of a Large Scale Hypertextual web search engine” Proc. 7th International world wide web conference, 1998, pp. 107-117.
[2]. Divakar Yadav, AK Sharma, and J. P. Gupta “Parallel Crawler Architecture and Web Page Change Detection”, WSEAS Transactions On Computers, Issue 7, Volume 7, July 2008, pp929-940.
[3]. Stefan Buttcher, Charles L. A. Clarke, and Gordon V Cormack, “Information Retrieval” MIT Press, 2010, ch I.2, pp. 40-50.
[4]. Mohammed Khan “Search Engine Optimization – what do you need to know” SEO expert, 2008, pp. 22.
[5]. Scott Counts, Karen E. Fisher “Mobile Social Networking: An Information Grounds Perspective” Proc 41st Hawaii International Conference on System Sciences – 2008, pp. 1-10.