Issues and Challenges with the Web Crawlers

Vishakha

Student, Department of Computer Science and Engineering, Shri Ram College of Engineering and Management, Faridabad, Haryana, India

Abstract: A Web crawler is an Internet bot which helps in indexing the web. It crawls one page at a time through a website until all pages have been indexed. Web crawler helps in collecting information about a website and the links related to them. A web crawler is defined as a mechanized program that accurately scans through Internet pages and downloads any page that can be reached by links. A Web crawler is also known as a Spider, automatic indexer or simply crawler. Now days finding the relevant information over millions bytes of data is very difficult. For this task we need to use efficient crawler so that it can give us best result within minimum time. But there are so many issues associated with each crawler so this paper discusses the issues and challenges involved in the design of the various types of crawlers.

Keywords: Indexer, Website, Scans

1. Introduction

The internet has become the largest unstructured database for accessing information over the documents day by day. [1] It is well known that the information technology has an immense effect on the conduct of the business activities. Internet is now becoming the largest marketplace of the world. [2] With the increasing amount of electronic text from the complex the World Wide Web, more and more knowledge you need is included on the daily basis. But, the huge amount of text also takes so much trouble to people to find useful information. The contributing factor to this volatile growth is the widespread use of microcomputer, increased case of use in computer packages and most importantly incredible opportunities that the web offers to business. For example, the standard Web search engines have very low precision, since typically some Web pages returned resulted information mixed with a large number of irrelevant pages, which is due to the situation that the topic-specific features may occur in different contexts. So, one suitable way of organizing this vast amount of documents is necessary, with increasing hardware and bandwidth resources at their disposal, it is very difficult or almost impossible for search engines to keep up with the growth of the Web. The changes in the web pages on the frequent basis creates the main retrieval challenges for the web crawler. In order to give best result, search engines have to learn about the new data as it become available, the web crawler has to continuously crawl and update the search engine database. Large amount of network bandwidth and other resources by accessing the web resources at a fast speed is consumed by a web crawler. This affects the performance of the server significantly. A huge amount of resources of underlining network are consumed to build a comprehensive full text index of the web. Further crawlers constantly retrieve the pages at a very fast speed to maintain the indices of a search engine updated. Because of the crawling activities of a single engine, there is a daily load of around 60GB to the web. [3] Today this load is not very much harmful to the web crawler but as this load will increase significantly in future as the web grows exponentially in the future.

There are total six types of web crawlers as illustrative below:

- Focused web Crawler
- Distributed web Crawler
- Collaborative Web Crawler
- Parallel web Crawler
- Incremental web Crawler

Here on this paper we basically describes the various issues and challenges in implementing these different kinds of web crawlers.

2. Focused Web Crawler

Focused crawler [4] is such web crawler that downloads the web pages which are correlated with each other. It collects documents which are definite and correlated to the given topic. Therefore, a focused crawler looks for, obtains, indexes and keep up pages on definite set of topics which represents a relevancy narrow segment of the web. It is also called as the Topic Crawler for the cause that is its way of working. The focused crawler ascertains the Relevancy and Way Forward. It ascertains to how much extent the given page is related to the particular topic and how to proceed forward [5] The advantages of focused web crawler are:

a) It is cost-effectively in terms of both hardware and network resources.
b) It can also reduce the network traffic amount and downloads on a website.

3. Issues and Challenges with Focused Web Crawler

There are four issues related to focused crawler as listed below:

Missing Relevant Pages: The main problem with the focused web crawlers is that sometimes they only crawled those pages that are expected to give immediate profit and miss the other relevant pages.

Maintaining Freshness of Database: As the world is changing every second so the web pages containing information gets updated on hourly, daily or weekly basis. The crawler has to crawl the pages frequently to find out the new and updated information out of the web to download those updated pages and updates them into the database to give latest information to the users. If the pages are large in

Volume 8 Issue 3, March 2019
www.ijsr.net
Licensed Under Creative Commons Attribution CC BY

Paper ID: ART20196562 10.21275/ART20196562 1791
number that it need to download on frequent basis then it makes the crawling process slow and puts pressure on the internet traffic as well. Thus, a major issue is to develop a strategy that manages these pages.

**Network Bandwidth and Impact on Web Servers:** The focused crawling techniques download many irrelevant Web pages that lead to consumption of network bandwidth. Focused crawlers used to perform crawling based on deploy multiple crawlers or the polling method to maintain the latest database. A lots of bandwidth get consumed in both the methods, if the crawlers are visiting them for all information then there would be the crippling impact on the performance of the web servers. Thus, this is an another issue to develop a method that only retrieve highly relevant pages so that the resources do not get overloaded.

**Absence of Particular Context:** here the problem is that the crawler basically on a particular topic and find out the related information to that topic but in case if the particular context is not mention then it used to download the large number of irrelevant pages. Due to this it waste the resources in downloading irrelevant information. Thus the challenge is to develop focused crawling techniques that focus on particular context also.

**Distributed Web Crawler**
This crawler [6] sprints on group of workstations. In distributed web crawler, a Uniform Resource Locator (URL) server assigns peculiar URLs to multiple crawlers, which downloads web pages in parallel, the crawlers will then send the downloaded pages to an innermost indexers on which links are retrieved and sent by means of the URL server up to the crawlers. This distributed character of crawling process decreases the hardware requirements and increases the largely download speed and dependability. First distributed crawler was developed which uses client server model that is centralized. But with this model there were problems like obstruction, expensive administration and being a single point of failure. But when it is developed to be fully distributed crawler where there is no central controller because of which problems like scalability, increased autonomy of nodes and failure resilience have been overcome. Although, fault tolerance is considered, then if an error occurs, the whole system will be difficult to maintain, and may be very difficult to ignore the loss of information.

**4. Issues and Challenges with Distributed Web Crawler**
There are six issues related to distributed crawler as listed below.

**Assignment of URL’s among different agents:** the main issue in distributed crawler is how to assign the URLs efficiently and dynamically to download best relevant information among the many crawling agents. It has to consider the various parameters like to minimize the rate of request of web, the crawling agents should be located properly on the network, and the URLs should be exchanged effectively in determining the assignment.

**Priority in Crawling:** The second most challenge is prioritizing the crawling frontier in the dynamically growing web as to provide the best result.

**Partitioning the collection:** The next challenge is to find an effective way of partitioning the data so that when user query something then it provide result best result from the smallest subset of the partitions rather than searching the
result in the whole collection. The chosen subset should be able to provide a high number of relevant documents.

**Load Balancing**: The next challenge is to find an efficient way of balancing the load among the different index servers such that no single server gets the entire load. There must be a good strategy to distribute the data in order to balance the load as much as possible.

**Network bandwidth consumption**: Network bandwidth is a scarce resource and is a big challenge. Therefore, when queries are resolved in a distributed fashion, the servers that should be contacted must be determined efficiently. Also the number of necessary servers should be minimum.

**Efficient cache design**: here the challenge is to design an effective cache that overcomes the network constraints and the hit ratio of finding the relevant information should also high.

**Parallel Web Crawler**

Because the growing amount of data on web, it becomes more difficult to retrieve the whole or a significant portion of the web from the million tons of data by using a single process. Therefore, lots of search engines frequently execute multiple processes in parallel to do the above task, so that download speed is taken full advantage. Multiple crawlers are frequently executed in parallel which are therefore known as Parallel crawlers. Consists of crawling processes called as C-procs.

5. **Issues and Challenges with Parallel Web Crawler**

There are many issues related to focused crawler as listed below:

**Overlapping**: When a number of C-proc’s are working independently it is possible that more than one might download the same page multiple number of times, this is known as an overlap. This is generally happens when one C-proc do not aware about the other C-proc that has already downloaded the same page. Existence of overlaps in a crawl become a challenge and tends to reduce the efficiency of a crawler. Generally, each of the C-proc’s work independently and do not care about overlaps. Each C-proc starts with its own seed and does not communicate with the other C-proc’s. It is assumed that each and every C-proc starts with an independent seed so in the starting the amount of overlap is not very significant.

**Quality of pages**: More important pages are downloaded first; however this could be a difficult task for parallel web crawlers.

**Increased bandwidth Consumption**: To avoid the above two challenges, the one C-proc’s should communicate with other C-proc’s so that both have the complete image of the Web. But this raises another problem of increased bandwidth consumption. If the number of C-proc’s increase then The communication may grow significantly and that need to be resolved to maintain the efficiency of the crawlers, which is a big challenge [4].
Incremental web crawler
Incremental crawler [8] updates current group of downloaded pages as an alternative of restarting the crawl from the beginning each time. This engrosses a way of ascertainment if a page has changed from the last time it was crawled. A crawler will persistently crawl the whole web supported by many group of crawling cycles. An adaptive model is used which uses data from preceding cycles to determine which pages should be tested for updates resulting in huge freshness and achievement of low peak load[9]. So, if the crawler can estimate how frequently pages changes, then the incremental web crawler may re-visit only those pages that have been changed rather than revising the whole collection together.

6. Issues and Challenges with Incremental Web Crawler
There are two issues related to focus crawler as listed below:

**Keep the local collection fresh:** The freshness of a collection can vary widely depending on the strategies used by the crawler to crawl the Web. Thus keeping the collection fresh is a big challenge as best policies need to be used.

**Improve quality of the local collection:** The quality of crawler can be improved by the replacing the less important pages with the more important pages in the local collector. This is necessary because the pages are constantly created and removed and some of the new pages can be more important than existing pages in the collection. Also the importance of existing pages changes over time. Thus improving the quality of collection is another challenge.

Collaborative Web Crawler
A Collaborative Web Crawler is consisting of a group of crawling nodes, where crawling node is responsible for a specific portion of the Web. This section discusses the various issues and challenges involved in implementing the Collaborative Web crawler.

7. Issues and Challenges with the Collaborative Web Crawler
There are two issues related to focused crawler as listed below:

**Overlap:** In the collaborative crawling, it may occur sometimes that different crawling nodes download the same page multiple times. Multiple downloads of the same page are clearly undesirable. Thus the challenge is to develop techniques that reduce or eliminate these overlaps of pages.

**Communication overhead:** In a collaborative crawling, the participating crawling nodes need to exchange URLs to coordinate the overall crawling work. To quantify how much communication is required for this exchange, the communication overhead is defined in terms of the exchanged URLs per downloaded page.

Architecture of Collaborative Web Crawler

8. Conclusions
In this paper, I have discussed the basic function of all crawlers presented by the researchers and the various issues and challenges faced in the development of these crawler architectures. I have found the many of the issues and challenges in these crawlers are common i.e. reducing the network bandwidth consumption, maintaining the freshness of the database and maintaining the quality of pages etc.

References

[7] Satinder Bal Gupta Professor, Department of Computer Science & Applications, Vaish College of Engineering, Rohtak, Haryana, India.
