Overview of Search Engine and Crawler

Gaurav Kumar Srivastav [a,*], Irphan Ali [a], Atul Kumar Srivastava [b]

[a] Department of Computer Science & Engineering, NIET, Greater Noida, U.P.T.U., Lucknow, India
[b] Department of Computer Science & Engineering, Amity University, Noida, India

Article Info

Article history:
Received 2 January 2014
Received in revised form
10 January 2014
Accepted 20 January 2014
Available online 1 February 2014

Keywords

World Wide Web,
Search Engine,
Web Crawler,
Parallel Web Crawler

Abstract

Today, Internet is the most important part of human life but growth of internet is major problem of internet user due to internet down loading speed, quality of downloaded web pages and find out the relevant content in the millions number of web pages. Nowadays, internet offering the various services such as business, studies material, ecommerce and search engine on the internet. Due to it is increase the number of web pages in internet.

In this paper we are solve the internet related problem by the help of search engine and improve the Quality of downloaded web pages for internet etc. Search Engine is find out the relevant content for the World Wide Web. We have solve other problem of search engine by the help of web crawler and proposed a working architecture of web crawler. Solve the problem of web crawler by the parallel web crawler.

1. Introduction

Internet is a global system of interconnected between computers by the help of networks. These networks are used the standard Internet protocol (*Transmission Control Protocol/Internet Protocol*) to serve several billion users world wide web. It is a *network of networks* interconnection of computer that consists of millions of private, public, academic, business, and government networks these networks linked by wireless, copper wire and optical networking technologies. The Internet carries an extensive range of information resources and services, such as the web site. Web site are the collection of web pages and every page inter-linked for another web page [8].

Nowadays, the internet are becomes the largest interconnection system of world. Internet is a big network where one can get a large amount of information. World Wide Web is collection of interconnected document and provides the services accessible by the help of internet such as e-mail, file sharing, online shopping, online gaming, search the any keyword on internet etc. Search of any keyword or web document within WWW is performed by the help of program called the web search engine [9].

Search engines are the program that search web document for a specified keywords and return a list of the all web documents where the keywords were found in World Wide Web [2].World Wide Web are combination of the millions of web pages about all type of topic. Web search engines are special web site that is designed to help people find the relevant information for store on these millions of pages.

Search Engine is managed in five steps.

- Crawling: Crawler is program/software that browses the web document on search engine.it starts with a set link Seed URLs and visits the all URL in queue, and downloads the web document. If extract the any URLs in the downloaded document then put the new URLs in the queue. This process is repeated until queue is not empty.
- Repository Management: It is managed and stores a large collection of web document. It collects of full html of every web document and every web document compressed and store in the repository.
- Index: Indexing is contains of individual web sites or web documents. Which is provide a more useful vocabulary for internet search.
- Query: Query means that receiving the relevant web document. It is filling search request from user
- Ranking: The search engine to display the most appropriate result to the user. But ranking define the sorting of result according the need of user.

Corresponding Author,
E-mail address: gauravjau@gmail.com

Web Search Engine normally works as a “bag of word”. In this section user enter any query in terms of keyword then the web search engine appears web document in ranks by their probability of applicability to the query in terms of keyword. These methodologies are effective and require the indexing of web document on World Wide Web. This index is collection of retrieve web document from the World Wide Web, and this process is done by a web crawler [10].

A Web crawler is a program that browses the World Wide Web in a methodical automated manner. This process is called as web crawling. Web crawler is also known as spiders or worms or robots. Web crawler is designed to retrieve web document from World Wide Web and added the retrieve web document in the repository management as temporary. Mainly objective of web crawler is to generate a copy of all visited web documents and store in repository management as a temporary, due to if the user want to further retrieve the visited web document by search engine then search engine retrieve the requested web document from the repository management.

Fig: 1. Processing steps of web crawler

In the fig: 1. web crawler used of the some basics terminology [6]

- Seed page: Seed page or Seed URLs is the initial URL or starting point which is start crawling procedure.
- Frontier: The crawlers start with a given URL, and extract the all un-retrieve URL. This list un-retrieve URL is known as frontier.
- Parser: Parser is to fetched web pages to extract list of un-retrieve URLs from it and return the new un-retrieve URLs to the frontier.

2. Types of Web Crawler

There are various types of web crawler. This is improving the performance and downloading of web crawler.

2.1 Focused Crawler

A focus crawler [chakabarti et. A1.][3] is similar to the crawler but it is returns relevant web pages on given topic from internet. A focused crawling analyzes it’s topic of interest to find the URL that are likely to most relevant web document for the crawl and avoid the irrelevant we document of World Wide Web.

2.2 Incremental Crawler

An incremental crawler [junghoo cho et. A1.][2] is show of the updates an existing set of downloaded pages instead of restarting the crawl from scratch each time and retrieve the updated web document from the World Wide Web.

2.3 Distributed crawler

A distributed crawler [shapenyuk et. Al.][2] distributes the downloading task among the downloading instance i.e. agents. The number of web document crawled per second per agent are independent of the number of agents.

2.4 Parallel Crawler

Parallel crawler are [junghoo cho, 2002][4] run multiple crawler in parallel. It is also process the multiple crawlers at a same time, and it is fetch the more than one web page at a time from a given host. Parallel crawlers increase the download rate of web document from the World Wide Web.

3. Working Of Web Crawler

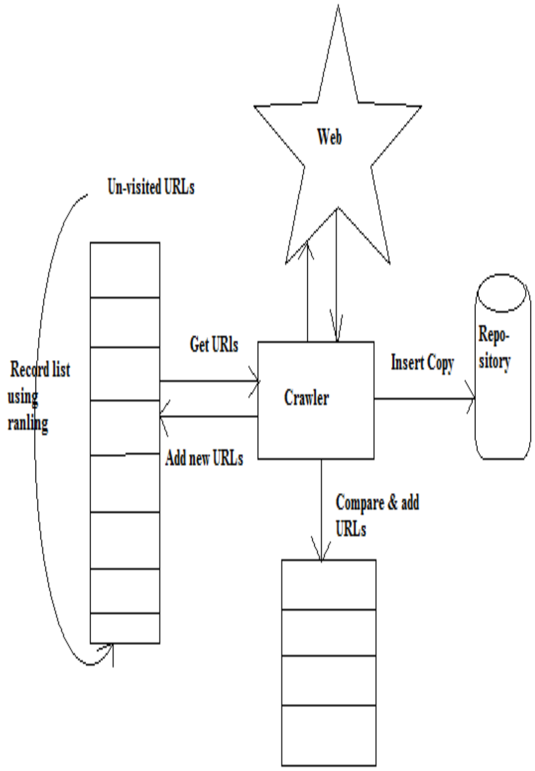

Fig: 2. Working of Crawler

Described the working process of crawler for fig: 2

1. The crawlers start crawling with a set of Seed URLs in queue. Which is known as Seed URLs?

2. The crawler downloads the web document from World Wide Web.
3. After the Above Step that extracts the all un-retrieve URLs from the downloaded web page and added in to a queue. The crawler again retrieves the URLs for downloaded next web document from the queue.
4. All visited web document is saved in the repository management.
5. The above all steps process continues until the crawler stops

4. Issues in Designing of Web Crawler

There are various issues of web crawler designing such as speed, performance and quality of web document.

- How should the crawler get the relevant pages to query?
- How should the crawler refresh web pages?
- How the crawling process should be parallelized?
- How should crawler get time sensitive information?
- The above all steps process continues until the crawler stops.

5. Parallel Web Crawler

Currently, the size of the web documents grow day per day, Due to this reason it becomes too difficult to ret-rieve the relevant web document or significant portion of the web document by the help of single web crawler. So, multiple single crawlers run parallel for the above task. This type of crawler as a parallel crawler [6]. The downloading rate of parallel crawler is increase and it is also improve the quality of the web document.

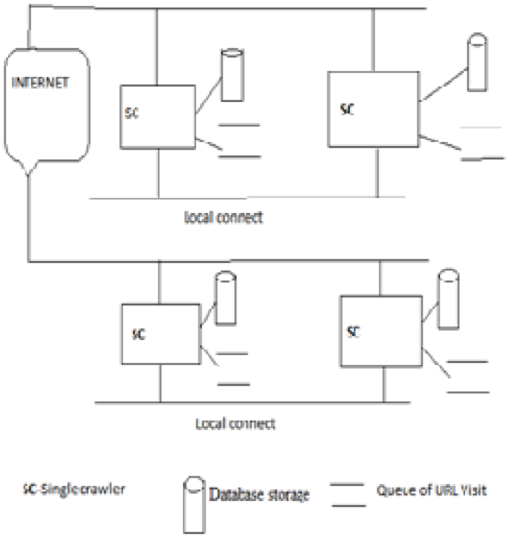

Fig: 3. Parallel Crawler

The parallel crawling architecture [6] is shown in the above Fig: 3. In this architecture each crawler is used to personal database storage and personal queue of the un-visited URLs. Whenever the crawling procedure stop, then added the all collected web pages in repository management.

5.1 Merit of parallel crawler

- Network load dispersion.
- Scalability.
- Reduction of network load.

5.2 Issues of parallel web crawler

- Improve the quality of downloaded web page.
- Reduce the multiple copy of downloading.
- Communication bandwidth.

6. Conclusion And Future Work

In this paper studies about web crawler, working steps of crawler, parallel crawler and reduce the problem of

Single web crawler by the help of parallel web crawler or multiple web crawler. Our future work will be carried out in parallel crawling. We plan to improve the performance of parallel crawler, avoid the spare web pages in parallel web crawling, and improve the quality of downloaded web document from World Wide Web.

References

[1] Junghoo Cho and Hector Garcia-Molina, “Incremental crawler and evolution of web”, Technical Report, Department of Computer Science, Stanford University

[2] V. Shkapenyuk, T. Suel, “Design and implementation of a high performance distributed Web crawler”. In Proceedings of the 18th International Conference on Data Engineering (ICDE'02), San Jose, CA Feb. 26--March 1, pages 357 -368, 2002.

[3] S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: A new approach to topic-specific web resource discovery. In The 8th International World Wide Web Conference, 1999

[4] Junghoo Cho, “Parallel Crawlers” proceedings of www2002, Honolulu, hawaii, USA, May 7-11, 2002. ACM 1-58113-449-5/02/005

[5] Dilip Kumar Sharma, A. K. Sharma,” A Novel Architecture for Deep Web Crawler”, International Journal of Information Technology and Web Engineering, 6(1), 25-48, January-March 2011

[6] Divakar Yadav, A.K Sharma, J.P. Gupta, “ Parallel crawler architecture and web page change detection”, WSEAS transaction on computers, 7 (7), July 2008

[7] Berners-Lee and Daniel Connolly, “Hypertext Markup Language. Internetworking draft”, Published on the World Wide Web at http://www.w3.org/hypertext/WWW/MarkUp/HTML.html, 13 Jul 1993.

[8] Berners-Lee and Daniel Connolly, “Hypertext Markup Language. Internetworking draft”, Published on the WWW at http://www.w3.org/hypertext/WWW/MarkUp/HTML.html.

[9] M. Levene (2005). An Introduction to Search Engine and Web Navigation, Addison Wesley

[10] C. Olston&M. Najork (2010), Web Crawling, Now Publishers Inc