REVIEW PAPER ON WEB CRAWLER

Yaduvir Singh

Associate Professor, BBDIT, Ghaziabad (India)

ABSTRACT

In this paper we have cover about the different types of web crawler exist in today's world. This paper also covers the basic fundamental why the web crawler exists and what is the significance of them. In this paper we also gives you how the web crawler really work and what are the different difficulties in implementing the web crawler and what are the solutions for them.

Keywords: Web crawler, bots, seeds, search engine.

I INTRODUCTION

Web crawler is software or a computer program which will be used for the browsing in World Wide Web in an ordered manner. The methodology used for this type of procedure is known as Web crawling or spidering.The different search engines used for spidering will give you current information. Web crawlers will create the copy of all the visited web pages that is used by the search engine as a reference to speed up the searches.

Web Crawlers can also be used for automation of all the tasks on a Web site, which can be checking links or validating HTML codes. Web Crawlers can be used collect the different information from the Web pages which is not limited to e-mail addresses but its scope is unlimited. It is a type of software agent that starts with a list of URLs that is to be visit and termed as seeds and as the crawler visits these URL's. It identifies all the other links in the page and then it will add them to the list of the URLs to visit, which can be termed as a crawl frontier. The crawler always carries the information in a canonical form.

II WEB CRAWLER AND ITS SIGNIFICANCE

The web crawling is used for numerous Purposes:

i) For site monitoring when there is a updating of information
ii) For copyright legal issues
iii) For checking the syntax of web pages
iv) For checking the semantic of web pages
v) For checking the links with valid syntax and semantic
vi) For increasing the popularity of a site

vii) For increasing the output of an organization

viii) For indexing purpose

Web crawler avoids sticking to the URL that is of large length and in addition to that it also eliminates the URL of non- textual types of data. A web crawler downloads all web pages which are addressed by the URLs and it will extract the hyperlinks contained in the web pages.

III WEB CRAWLER AND ITS WORKING

Web Crawler uses the parsing which starts from a specified web page which may be any hyper page links that is pointed to some other web pages. Crawler starts from a link and move to respective hyperlinks and doing it in repeated manner. It is a set of instructions act as s program which act as an agent for further work and the crawler sends hypertext transfer protocol requests for documents to other computers on the internet .Then the next set of steps are same as that of the web browser used for searching the links and hyper linking the next set of web pages. The web crawler uses the automatic procedure in doing the above mentioned process.

The different search engine uses the same procedure that is used by the crawler for search engine optimization. Web crawler uses the different types of algorithms for implementing the user required scenario in today's world. Seeds also play a very important role in crawling as it uses the basic breadth and depth search strategy for improving the performance in crawling.

The basic principle of web crawling method relays particularly on hypertext links which in fact forms the huge network known as Internet. Hypertext links are represented as URL (Unified Resource Locator) link which contains information about unique location referenced web resource.

The method of web crawling can be represented with the following steps:-

(i) The document fetched from the processing queue

(ii) The document that needs to be downloaded

(iii) The next step is to start the parsing for perfecting of next set of web pages.

(iv) The next task is to store all the web documents that are processed till now.

IV IMPLEMENTATION OF WEB CRAWLER

When we are implementing the concept of web crawler then the type of problem can be categorized in basically two major categories:

(i) Large Web Pages

(ii) The change of rate of web pages.

A greater volume of web page implies that web crawler can only download a fraction of the web pages and hence it is very essential that web crawler should be intelligent enough to prioritize download.

The next problem with world of today is that web pages on the internet change very frequently, as a result, by the time the crawler is downloading the last page from a site, the page may change or a new page has been uploaded/updated to the site.

Implementation part of web crawler not only essential for web but also its behavior and during the selection of the web crawler behavior various points are to be considered:

a. Always avoid the overloaded website means site with too much of traffic
b. The strategy should be properly manage how to revisit the pages for new updates
c. The selection of the web crawler algorithm

V ALGORITHM FOR WEB CRAWLING

The web crawling procedure can efficiently be handle with the different algorithms:

- Focused Crawling
- Path –Ascending Crawling

5.1 Focused Crawling

The main criteria of a page for a crawler can be expressed as a function of the similarity of a page to a given question. During this procedure we can intend web crawler to download pages that are similar to each other, thus it will be called Topical or Focused crawler.

The important question that arises in case of focused crawling in the context of a Web crawler, we would like to be able to predict the similarity of the text of a given page to the query before actually downloading the page. A possible predictor is the anchor text of links; to resolve this problem proposed solution would be to use the complete content of the pages already visited to infer the similarity between the driving query and the pages that have not been visited yet. The performance of a focused crawling depends mostly on the richness of links in the specific topic being searched, and a focused crawling usually relies on a general Web search engine for providing starting points.

5.2 Path –Ascending Crawling

The web crawler tends to download as many resources as possible from a particular Web site. The way in which a crawler would ascend to every path in each URL that it intends to crawl.

The advantage with Path-ascending crawler is that they are very effective in finding isolated resources, or resources for which no inbound link would have been found in regular crawling.

VI CONCLUSION

For efficient building of web crawler is not a difficult problem, but the main point is in choosing the right strategies and an effective architecture which will lead to implementation of highly efficient and intelligent web crawler application. This paper has a lot of future scope to implement in the real scenarios.

REFERENCES

[1]. Gatial E., Z. Balogh Z., Laclavik M., Ciglan M., Hluchy L.: Focused Web Crawling Mechanism based on Page Relevance. In: Proceedings of ITAT 2005 Information Technologies - Applications and Theory, Peter Vojtas (Ed.), Prirodovedecka fakulta Univerzity Pavla Jozefa Safarika v Kosiciach, 2005, pp.41-46. Slovakia, September 2005. ISBN 80-7097-609-8.

[2]. Gatial E., Balogh Z.: Identifying, Retrieving and Determining Relevance of Heterogenous Internet Resources. In: Tools for Acquisition, Organisation and Presenting of Information and Knowledge. P.Navrat et al. (Eds.), Vydavatelstvo STU, Bratislava, 2006, pp.15-21, ISBN 80-227-2468-8. Workshop 26-28 September, Nizke Tatry, Slovakia.

[3]. Advanced Triage (medical term), http://en.wikipedia.org/wiki/Triage#Advanced triage.

[4]. Attributor. http://www.attributor.com

[5]. Z. Bar-Yossef, I. Keidar, and U. Schonfeld, "*Do not crawl in the DUST: Different URLs with similar text,"* in Proceedings of the 16th International World Wide Web Conference, 2007.