ISSN:0975-9646

Design and Implementation of Web Crawler

Ankita Dangre, Vishakha Wankhede, Priyanka Akre, Puja Kolpyakwar

Dept. of Information Technology
Rajiv Gandhi college of Engineering and Research
Nagpur,India

***Abstract* – As the number of Internet users and the number of accessible Web pages grows, it is becoming increasingly difficult for users to find documents that are relevant to their particular needs. The key factors for the success of the World Wide Web are its large size and the lack of a centralized control over its contents. Users must either browse through a large hierarchy of concepts to find the information for which they are looking or submit a query to a publicly available search engine and wade through hundreds of results, most of them irrelevant.**

Web crawling is the process used by search engines to collect pages from the Web.Web crawlers are one of the most crucial components in search engines and their optimization would have a great effect on improving the searching efficiency. This paper, introduces web crawler that uses a concept of irrelevant pages for improving its crawling performance. Despite their conceptual simplicity, implementing high-performance web crawlers poses major engineering challenges due to the scale of the web. This crawler computes the weights for the pages we come across during the crawling process and hence decide how much a particular page is important to us. Both issues are also the most important source of problems for locating information. The Web is a context in which traditional Information Retrieval methods are challenged, and given the volume of the Web and its speed of change, the coverage of modern search engines is relatively small. Moreover, the distribution of quality is very skewed, and interesting pages are scarce in comparison with the rest of the content.

Keywords:

Web crawler:
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.

Seed:
It is starting URL from where Web Crawler starts traversing World Wide Web.

Frontier:
It is list of unvisited URLs.

Page weight:
Weight of page which is decided on the certain parameters.

Threshold value:
Certain limit which we decide the importance of page.

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. 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. As the Web grows grander and more diverse, search engines also have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated. In Internet data are highly unstructured which makes it extremely difficult to search and retrieve valuable information. Search engines define content by keywords.

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. In addition, with the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and clientside intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities. Many organizations and corporations provide information and services on the web such as automated customer support, on-line shopping, and a myriad of resources and applications.web based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts etc., are becoming common practice and widespread. The WWW is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child sharing music files with friends to a senior receiving photographs and messages from grandchildren across the world.

It is typical to see web pages for courses in all fields taught at universities and colleges providing course and related resources even if these courses are delivered in traditional classrooms. It is not surprising that the web is the means of choice to architect modern advanced distance education systems. There are several important issues, unique to the Web paradigm that comes into play if sophisticated types of analyses are to be done on server side data collections. These include the necessity of integrating various data sources such as server access logs, user registration or profile information; resolving difficulties in the identification of users due to missing unique key attributes in collected data; and the importance of identifying user sessions or transactions from usage data, site topologies, and models of user behavior. We devote the main part of this paper to the discussion of issues and problems that characterize Web usage mining. Web is the totality of web pages stored on web servers. 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. As the Web grows grander and more diverse, search engines have assumed a central role in the World Wide Web's infrastructure as its scale and impact have escalated.

Web crawlers are an important component of web search engines, where they are used to collect the corpus of web pages indexed by the search engine. Web *crawlers* are programs that exploit the graph structure of the Web to move from page to page. In their infancy such programs were also called wanderers, robots, spiders, ¯sh, and

worms, words that are quite evocative of Web imagery. The large size and the dynamic nature of the Web highlight the need for continuous support and updating of Web based information retrieval systems The last key dimension is regarding crawler evaluation strategies necessary to make comparisons and determine circumstances under which one or the other crawlers work best. Crawlers facilitate the process by following the hyperlinks in Web pages to automatically download a partial snapshot of the Web.

MOTIVATION:

Earlier work was based on how the web crawler works, the process of web crawler and how the sequence of accepting the URL, fetching the page, parsing the page, extracting all the hyperlinks is performed. While performing the following sequence, we are downloading the page we need to verify for the evaluation. Hence, while downloading the page we anyways use up the bandwidth. It will be even more beneficial if we utilize the used bandwidth and get more out of it.

Thus implementing the following method, we use the downloaded pages' bandwidth and get the same bandwidth to get the title, body and the number of outgoing links on that particular page.

METHODOLOGY:

We define the factors for which we specify the page importance:

```
weight(page) = weight(URL) + weight(outlinks) + weight(title) + weight(body)
```

where,

```
1)      if ( search string present in URL)
        {
                weight(URL) returns a predefined weight
        }
Else
{
        Return 0
}
```

This will return the weight assigned for the URL occurrence. If the search string is found in the URL, the page acquires certain importance.

```
2)      if ( search string present in title)
        {
                weight(title) returns a predefined weight
        }
Else
{
        Return 0
}
```

This will return the weight assigned for the title occurrence. If the search string is found in the title, the page acquires certain importance.

```
3)      Occurrence of search string in the body
        {
                weight(body)=occurrence*weight for each occurence
        }
```

This will return the weight assigned for the body occurrence. If the search string is found in the body, the page acquires certain importance. When the search string occurs certain number of times in the body, the occurrence is noted and the page importance is calculated using the occurrence count.

```
4)      Number of hyperlinks on the page
        {
        weight(outlinks)=occurrence*weight for each occurence
        }
```

This will return the weight assigned for the out-links occurrence. The number of links linking to the other page has also been assigned some importance.

Giving importance to each component of the parsed page, we have assigned weight to each component and hence acquired the page importance in totality. As we get the page weight, we will compare it with the threshold frequency implicitly provided to the algorithm. Depending on the result of comparison, the links are either added to the output or they may be discarded.

Thus, we get the search more focused to the search string eliminating the least important topic.

PROPOSED ALGORITHM (PSEUDO – CODE):

```
1. Start
2. Initialize frontier with seed URL.
3. While (frontier is not empty)
{
1. Pick URL from frontier
2. Fetch page
3. Parse page
4. calculate weight(page).
}
4. if(weight(page) >  (threshold_value))
{
        Add to output.
}
5. Stop
```

CONCLUSION:

Hence by using the concept of Page Weight, we scan web pages as well as compute the weight of page and hence we can increase efficiency of web crawler as output set of URL generated by this way will always be of better importance than what traditional web crawler is generating.

FUTURE SCOPE:

As we parse the page, we have only extracted the hyperlinks on the page. We can proceed the work by extracting the images, videos and other non textual content and hence carry out the further process.

ACKNOWLEDGMENT:

We would like to express my special thanks of gratitude to our Guide *Prof. Rahul Sathawane* as well as our Co-guide Prof. *Prashant Dahiwale*.

REFERENCE:

1. Rahul Sathawane(Guide), Prashant Dahiwale(Co-guide)
2. Marc Najork "Web Crawler Architecture".
3. Gautam Pant, Padmini Srinivasan and Filippo Menczer "Crawling the Web".
4. From Wikipedia -"Web Crawler".
5. Pinkerton B.,"Finding what people want: Experiences with the WebCrawler", In Proceedings of the First World Wide Web Conference, Geneva, Switzerland , 1994.