IJCSMS International Journal of Computer Science & Management Studies, Special Issue of Vol. 12, June 2012
ISSN (Online): 2231 –5268
www.ijcsms.com

An Improved Approach to perform Crawling and avoid Duplicate Web Pages

Dhiraj Khurana[1], Satish Kumar[2]

[1]Assistant Professor, CSE Department
University Institute of Engineering & Technology
Maharshi Dayanand University, Rohtak(Haryana)
dhirajkhurana23@rediffmail.com

[2]Assistant Professor, CSE Department
Vaish College of Engineering, Rohtak (Haryana)
Krsk23@gmail.com

Abstract

When a web search is performed it includes many duplicate web pages or the websites. It means we can get number of similar pages at different web servers. We are proposing a Web Crawling Approach to Detect and avoid Duplicate or Near Duplicate WebPages. In this proposed work we are presenting a keyword Prioritization based approach to identify the web page over the web. As such pages will be identified it will optimize the web search.

Keywords: Crawler, Optimization, Duplicate, Webpage, Prioritization

1. Introduction

Besides piracy one of the problems on the Internet these days is redundant information, which exist due to replicated pages archived at different locations like mirror sites. As a result, the burden is on Web users to sort through retrieved Web pages to identify non-redundant data, which is a tedious and tiring process. Since the amount of information available on the Internet increases on a daily basis, filtering redundant and similar documents becomes a more difficult task to the user. Due to the rapid growth of electronic documents, redundant information increases on the Web. In order to use the information available on the Web many technologies emerged, information retrieval systems is one of them.

1.1 RESEARCH PROBLEM

The web crawler is the basic requirement to search and download data efficiently from the web. Most of the search engine and downloader uses the same tool to detect and fetch the pages. But today the user requirement is generally very specific such as a researcher only wants to search data on his required topic. For this the topic based web crawler is used.

1.2 PROPOSED GOALS

In this proposed work we are working on topic based incremental crawler. When we perform a topic based search we can find some similar topics also. The proposed work is about to exclude such kind of pages from the list of downloadable pages. For this duplicate page analysis we are proposing a suffix tree based approach that will perform the keyword based matching in optimum time. The proposed work is about to optimize the crawling process by excluding such pages in topic based search.

2. RESEARCH METHODOLOGY

An approach for detecting duplicate web pages in web crawling by use of constructive, analytical and exploratory research design. Constructive research design to get the objectives clearly defined, analytical research design to use facts or information already available, and analyze these to make a critical evaluation for research.

3. RELATED WORK

Akansha Singh performed a work," Faster and Efficient Web Crawling with Parallel Migrating Web Crawler This paper aims at designing and implementing such a parallel migrating crawler in which the work of a crawler is divided amongst a number of independent and parallel crawlers which migrate to different machines to improve network efficiency and speed up the downloading. The migration and parallel working of the proposed

design was experimented and the results were recorded.

Mandlenkosi Victor Gwetu performed a work," The Application of Sampling to the Design of Structural Analysis Web Crawlers". This paper proposes the application of sampling as a selection strategy in the design of structural analysis web crawlers. This has the benefit of alleviating the problems of bandwidth costs to web servers whilst retaining the quality of the data that is mined by crawlers. The initial results of this study are promising and are presented in this paper.

Avanish Kumar Singh performed a work," Novel Architecture of Web Crawler for URL Distribution". In this paper, Author has focused on the problem regarding to performance of crawler, which have been affected because of increasing users load on World Wide Web. Here the two major terms have defined for web pages, firstly the URLs, which hold the address and secondly, URLs size, which have been considered in bit. There is no doubt that because increasing the growth rate of work automation, the work load though internet users are also increasing highly.

Sandeep Pandey performed a work, "User Centric Web Crawling". In this paper Author study how to schedule Web pages for selective (re)downloading into a search engine repository. The scheduling objective is to maximize the quality of the user experience for those who query the search engine. Author begins with a quantitative characterization of the way in which the discrepancy between the content of the repository and the current content of the live Web impacts the quality of the user experience. This characterization leads to a user centric metric of the quality of a search engine's local repository.

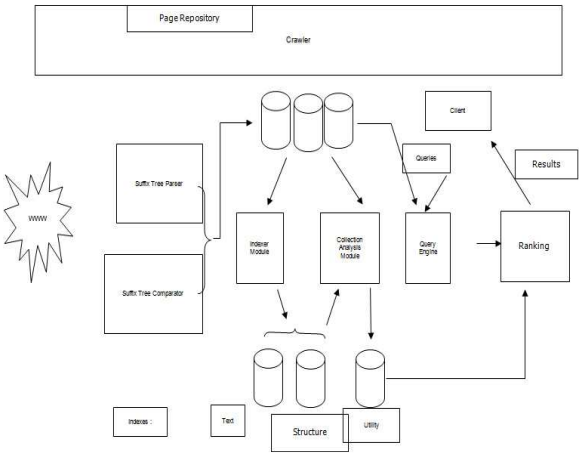
Mike Thelwall performed a work, "A Free Database of University Web Links: Data Collection Issues". This paper describes a free set of databases of the link structures of the university web sites from a selection of countries, as created by a specialist information science web crawler. With the increasing interest in web links by information and computer scientists this is an attempt to make available raw data for research that is not reliant upon the opaque techniques of commercial search engines.

2.1CONCLUSION FROM REVIEW OF LITERATURE & PROBLEM FORMULATION

Duplicate document detection has become a research field. Its purpose is to detect redundant documents to increase search effectiveness and storage efficiency of search engines. Detection of duplicate web pages or documents in a fast way has great importance for users; because users do not want to wait in this process. They want to reach information as quickest as possible and if duplicate detection begins to slow down the access to the information, then they may choose to retrieve duplicate information. The past few years have observed the drastic development of the World Wide Web (WWW). Information is being increasingly accessible on the web. The performance and scalability of the web engines face considerable problems due to the presence of enormous amount of web data. The expansion of internet has resulted in problems for the Search engine owing to the fact that the flooded search results are of less relevance to the users.

4. PROPOSED ARCHITECTURE

The proposed work is about to optimize the

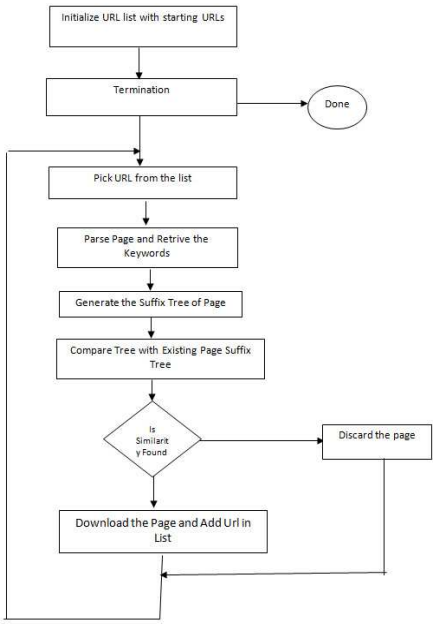


Topic based web crawling process with the concept of exclusion of duplicate pages. For this a new architecture is proposed, this architecture will use the suffix tree based algorithm to detect the duplicate web pages.

As we can see in this proposed architecture the user will interact to the web with his topic based query to retrieve the web pages. As the page is query performed it will perform request to the web and generate the basic URL list. Now it will retrieve the data from the web. For the URL collection it will use some concepts like indexing and the ranking. The indexing will provide a fast access to the web page where as ranking will arrange the list according to the priority.

Now as a web page is fetched, the proposed suffix tree parser will retrieve the keywords form the

document and generate a suffix tree. It will store the tree organization in the respiratory. Now as a new page is retrieved it will generate the suffix tree and perform a suffix tree based comparison to analyze the duplicacy ratio. If the page is duplicate, it will discard the page otherwise retrieve the contents.
The complete process is presented in the form of a flow chart presented



5. ILLUSTRATIVE EXAMPLES

Even worse, different websites often provide conflicting information, as shown in the following examples.
Example (Height of Mount Everest) Suppose a user is interested in how high Mount Everest is and queries Ask.com with What is the height of Mount Everest? Among the top 20 results, one will find the following facts: four websites (including Ask.com itself) say 29,035 feet, five websites say 29,028 feet, one says29, 002 feet, and another one says 29,017 feet. Which answer should the user trust?
Example (Authors of books). When tried to find out who wrote the book Rapid Contextual Design (ISBN: 0123540518). Different sets of authors from different online bookstores are found, and several of them shown in Table 1. From the image of the book cover, it is found that A1 Books provides the most accurate information. In comparison, the information from Powell's books is incomplete, and that from Lakeside books is incorrect.

6. RANKING ALGORITHM

Term Weight age(c1)	Visitor Count(c2)	Like/Dislike(c3)	Page Rank(c4)

We will mark these four things on the scale of 10.

Example: For URL A

1) **Term Weight-age:** In case if term weight-age is 55% for A, then we will count it as 5.5
2) **Visitor Count:** If visitor count is say 4 then, it will be considered as 0.4
3) **Like/Dislike:**

   Like Count: 4

   Dislike Count: 3

   Then, Like-Dislike=1 on the scale of 10, we will mark it as 0.1

   If (Like-Dislike is in negative), then we will use it with a negative sign.

   If Like=3

   Dislike=5

   Then Like-Dislike=-2

   Then we will mark it as -0.2

4) **Page Rank:** If page-rank is 60 then, it will be considered as 6

Now, using these terms we will calculate the rank of URL

x1*c1+x2*c2+x3*c3+x4*c4= Rank of URL A

Where,
x1= 0.3, x2=0.4, x3=0.27, x4= 0.03 (These are tested values)

7. CONCLUSION

Through the web is a huge information store, various features such as the presence of huge volume of unstructured or semi-structured data; their dynamic nature; existence of duplicate and near duplicate documents and the like pose serious difficulties for Information Retrieval.
The voluminous amount of web documents swarming the web have posed a huge challenge to the web search engines making them render results of less relevance to the users. The detection of duplicate and near duplicate web documents in web crawling. The proposed approach has detected the duplicate and near duplicate web pages efficiently based on the keywords extracted from the web pages. Furthermore, reduced memory spaces for web repositories and improved search engine quality have been accomplished through the proposed duplicates detection approach.

8. FURTHER RESEARCH

Further research can be done to avoid unwanted web sites from the extracted links of the web site and to implement this as a browser add-ons.

REFERENCES

[1] Web Crawler Introduction: http://en.wikipedia.org/wiki/Web_crawler
[2] Gautam Pant, Padmini Srinivasan, and Filippo Menczer3, “Crawling the Web”, 2004
[3] Leigh Dodds, “Slug: A Semantic Web Crawler," February 2006.
[4] Junghoo Cho,”Crawling the Web: Discovery and Maintenance of Large Scale Web Data”, Ph.D. thesis submitted in November 2001 at Stanford university, USA.
[5] S.S. Dhenakaran1 and K. Thirugnana Sambanthan2, “Web Crawler – An Overview” International Journal of Computer Science and Communication Vol. 2, No. 1, January-June 2011, pp. 265-267
[6] Sandhya, M. Q. Rafiq ,” Performance Evaluation of Web Crawler” , IJCA Journal, 2011.
[7]Sunil M Kumar and P.Neelima.” "Design and Implementation of Scalable, Fully Distributed Web Crawler for a Web Search Engine”. International Journal of Computer Applications 15(7):8–13, February 2011.
[8] Zhixing GAO, Kunhui LIN, “Design and Implementation of an Efficient Distributed Web Crawler with Scalable Architecture”, Journal of Computational Information Systems5:6 (2009) 1817-1823, 2009
[9] Akansha Singh, “Faster and Efficient Web Crawling with Parallel Migrating Web Crawler”, IJCSI International Journal of Computer Science Issues