DESIGN OF CATEGORY-WISE FOCUSED WEB CRAWLER

Monika[1], Dr. Jyoti Pruthi[2]

[1]M.tech Scholar, [2]Assistant Professor, Department of Computer Science & Engineering, MRCE, Faridabad, (India)

ABSTRACT

The exponential growth of World Wide Web is rapid and it has scaled to large volume which is difficult to handle, index and search. Due to this increase in the size and diversity of information available on web, it is becoming difficult for traditional crawlers to efficiently crawl the web. Traditional, web crawlers retrieve all the pages that match the query, whether they are relevant for the user or not. So there is a need to develop efficient crawlers. Focused crawler is a crawler which retrieves only the relevant information for the user and discards the information which is irrelevant. In this paper we modify the proposed architecture of category-wise focused crawler which was earlier proposed by us and define the design of Category-Wise Focused Web Crawler.

Keywords: Focused Web Crawler, Relevant page, Web Crawler

I. INTRODUCTION

Web is a fast growing hypermedia database. In 90's the amount of data generated in a year is now generated every day. This availability of data is both structured and unstructured. So, it is very difficult to handle this amount of data. A web crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. With the exponential growth of information on the World Wide Web, there is a great demand for developing efficient and effective methods to organize and retrieve the information available and to develop a web crawler that retrieves most relevant pages. Because of limited computing resources and limited time, focused crawler has been developed.

II. RELATED WORKS

Focused crawling was introduced by Soumen Chackrabarti in 1999[7]. The goal of a focused crawler is to selectively seek out pages that are relevant to a pre-defined set of topics. Fish Search Algorithm for collecting topic-specific pages was proposed by P.M.E DeBra[9]. The algorithm simulates a school of fish, breeding and searching for food. Based on improvement of fish-search algorithm, M.Hersovici et al[8] proposed the shark-search algorithm. Shark-Search is a more aggressive version of Fish-Search. A Generic Framework for Focused Crawler was given by Martin Ester, Matthias Grob, Hans-Peter Kriegel[6]. The framework consists of 2 major components. First component consists of specification of user interest and measuring the relevance of webpage. Second component consists of ordering the links in the crawl frontier. Focused Crawler based on link structure and contents is discussed by Mohsen Jamali, Hassan Sayyadi, Babak Bagheri Hariri and Hassan Abolhassani[5]. They maintain Link Structure of pages and also introduce metric for measuring similarity of a page to a domain. A major problem faced by the above focused crawlers is that it is frequently difficult to learn that some sets of off-topic documents lead reliably to highly relevant documents. For solving this problem Anshika Pal, Deepak

singh tomar[4] proposed a method. For improving the Prediction of Page Relevance of Focused Crawlers Mejdl S. Safran, Abdullah Althagafi[3] proposed a Focused Crawler which uses Naïve Bayesian as the base Prediction Model. Approaches of Focused Web Crawler is discussed by Jay Sampat and Dharmeshkumar Mistry[2].

III. PROBLEM STATEMENT

The objective is to design a crawler that is efficient, fast, user friendly and improves the search based on focused category, so that best matched results will be displayed.

IV. PROPOSED ARCHITECTURE

In this paper we are modifying the proposed architecture of category-wise focused crawler which we have proposed in [1]. The Category-wise Focused Web Crawler crawl a website on the basis of a category and retrieve most relevant pages based on that category. The proposed system architecture has two main modules: Crawl Module & Search Module.

- ***Crawl Module:*** This module will crawl the WebPages using focused approach and store the result in the database. In this module first the user will provide a Domain to crawl, For example- '.com' domain, then user will also provide URL to start crawler. At last the user will choose a category on which crawler will focus while crawling, For example- 'News', 'Education' etc.
- ***Search Module****:* In this module, we will develop a search page. The user can give his/her query under any specified category and best matched results of the query from that category will be retrieved.

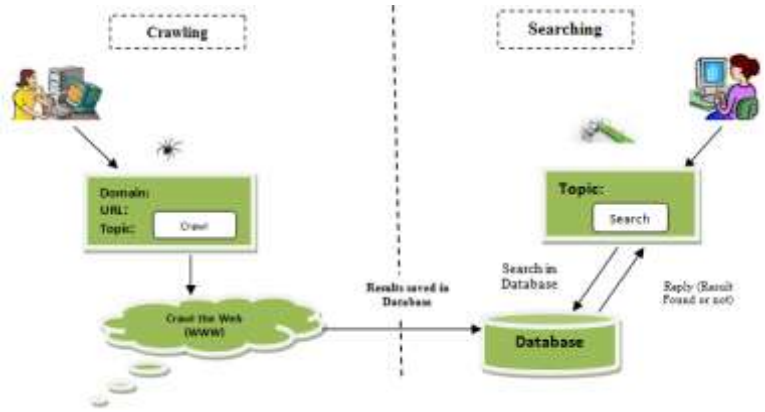

Figure 4.1: Proposed System Architecture

The proposed system crawls the WebPages recursively and stores the relevant data in database. This data includes Title, Meta keywords, Meta title, Meta Description etc of the webpage. When a query is submitted to the Search Engine, it searches its own database in response to it. In Focused Search, a category is also chosen. The WebPages are then retrieved as per the field chosen.

V. DESIGN OF CATEGORY-WISE FOCUSED WEB CRAWLER

The design of the Category-wise Focused Web Crawler is explained with the help of UML Diagram, basic Flowchart and ER Diagram.

5.1 UML Diagram

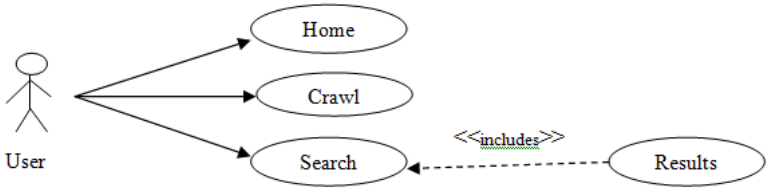

Figure 5.1: UML Diagram

The UML Diagram in Figure 4.2, we have an actor which is the user, the use cases for this actor can be the Home page(which provides basic linking of Crawl and Search Page), Crawl Module and the Search Module. In the Search use case, we have a include relationship with the Results, which specifies that searching includes displaying of results, whether found or not found.

5.2 Flowchart of Category-wise Focused Web Crawler

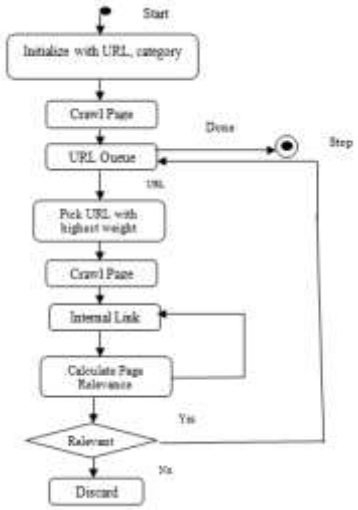

Figure 5.2: Flowchart

The Flowchart in Figure 4.3 explains the basic flow of Category-wise Focused Web Crawler. Here, first we initialize the crawl with Domain, URL and Category. Initially, only single page is crawled and add to the URL Queue. From the URL Queue the webpage with highest Relevance Score is picked and further links of that webpage is explored only. For every internal link Relevance Score or the Page Relevance is calculated. If the webpage is found to be relevant it is stored in database and URL is also added to the URL Queue, otherwise it is discarded. In the next round the URL Queue will have all the relevant internal links of the starting URL. This process will continue until we reach a stopping condition or there are no more webpage's in the URL Queue to crawl.

5.3 ER Diagram

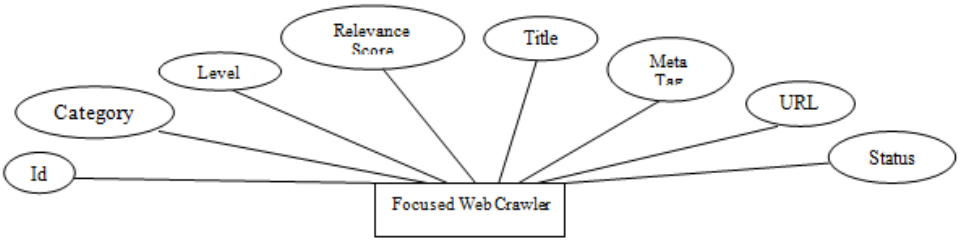

Figure 5.3 ER Diagram

ER Diagram in Figure 4.3, in this diagram we have:

Id: It is a Random Number associated with every crawl. So for every crawl we have a id associated with it. Id will be same for all the URL's of a single crawl.

Category: It is the Area or Field chosen to focus upon while crawling.

Level: It is the depth of the tree constructed by the exploring URL's.

Relevance Score: It is the measure to which a webpage is relevant to a specified category. The Relevance Score is calculated by the counting the frequency of the category in the webpage

Title: Title of the webpage.

Meta Tag: Meta Description of the webpage. In Meta Description, we are storing the keywords of the webpage.

URL: Complete URL of the webpage.

Status: There can be 2 types of Status that we can set:

1. Pending: The URL which is examined, but yet not explored.
2. Done: The URL which is examined and also explored.

VI. CONCLUSION

In today's world of Big Data, we need to prioritize our crawl and search, so that we can more useful data in less time. As there is limited time and resources available there is a need to develop efficient crawlers. So, Focused Crawlers came into existence, which crawl only the relevant page for the user. The modified architecture and also the basic design of the Category-wise Focused Web Crawler are discussed in this paper. We are working on the implementation of this Web Crawler.

VII. ACKNOWLEDGEMENT

I express my sincere and deep gratitude to my guide Dr. Jyoti Pruthi, Assistant Professor, Department of Information Technology, Manav Rachna College of Engineering, Faridabad, for the invaluable guidance, support and encouragement. She provided me all resource and guidance for this work.

REFERENCES

[1] Monika, Dr. Jyoti Pruthi, "Focused Web Crawler: Proposed Architecture", 2nd International Conference on "Innovation and Sustainability: Managing for Change", Jan 2015, 433-437.

[2] Jay Sampat, Anmol Jain, D. Mistry, "Focused Web Crawler and its Approaches", International Journal of Current Engineering and Technology, Volume-4, No.-5, Oct 2014

[3] Mejdl S. Safran, Abdullah Althagafi and Dunren Che, "Improving Relevance Prediction for Focused Web Crawlers", IEEE 11 International Conference on Computer and Information Science, 2012

[4] Anshika Pal, Deepak singh tomar, S.C. Shrivastava, "Efficient Focused Crawling Based on Content and Link Structure analysis", International Journal of computer science and information security, vol. 2, No.-1, June 2009.

[5] Mohsen Jamali, Hassan Sayyadi, Babak Bagheri Hariri and Hassan Abolhassani," A Method for Focused Crawling Using Combination of Link Structure and Content Similarity", IEEE International Conference on Web Intelligence, December 2006, Pages 753-756.

[6] Filippo Menczer, Gautam Pant, Padmini Srinivasan, "Topical Web Crawlers: Evaluating Adaptive Algorithms", ACM Transactions on Internet Technology, Vol. 4, No. 4, November 2004, Pages 378–419.

[7] S. Chakrabarti, M. van den Berg, B. Dom, "Focused crawling: a new approach to topic-specific Web resource discovery," in 8th International WWWConference, May 1999.

[8] Hersovici, M., Jacovi, M., Maarek, Y. S., Pelleg, D., Shtalhaim, M., and UR, S. 1998, " The sharksearch algorithm—An application: TailoredWeb site mapping", 7th International World-Wide Web Conference.

[9] P.M.E. De Bra, R.D.J. Post, "Information Retrieval in the World Wide Web: Making Client-based searching feasible", Computer Networks and ISDN Systems, 27(2) 1994 183-192.