Effective Performance of Information Retrieval by using Domain Based Crawler

Sk.Abdul Nabi[1]
Department of CSE
AVN Inst. Of Engg.& Tech.
Hyderabad, India

Dr. P. Premchand[2]
Dean, Faculty of Engineering
University college of Engineering
Osmania University, Hyderabad, India

***Abstract*—World Wide Web continuously introduces new capabilities and attracts many people [1]. It consists of more than 60 billion pages online. Due to this explosion in size, the information retrieval system or Search Engines are being upgraded day by day and it can be used to access the information effectively and efficiently. In this paper, we have addressed Domain Based Information Retrieval (DBIR) System. In this system we crawl the information from the web and added all links to the data base which are related to a specific domain. It simply ignores which are not related to that domain. Because of that we can save the Storage Space (SS) and Searching Time (ST) and as a result it improves the performance of the system.**

It is an extension of Effective Performance of Web Crawler (EPOW) System [2], in which it has two Crawler modules. The first one is Basic Crawler. It consists of multiple downloaders to achieve parallelization policy. The second one is Master Crawler, which is used to filter the URLs send by the Basic Crawler based on the Domain and sends back to the Basic Crawler to extract the related links. All these related links are collectively stored into the database under a unique domain name.

Keywords—Domain Based Information Retrieval (DBIR); Storage Space (SS); Searching Time (ST); Master Crawler; Basic Crawler; EPOW.

I. INTRODUCTION

The Web crawler [3] is a computer program that downloads data or information from World Wide Web for search engine. Web information is changed or updated rapidly without any information or notice. Web crawler searches the web for updated or new information. Web crawler [4, 5] is a software agent. It can be also called as Spider or Robots, which is the main component of a Search engine. Crawling the whole web is not possible because of its size and growth. Fig .1 represents how fast the amount of internet hosts increased in the last few years. When the crawler has finished with the current state of the network during the time host will have grown a lot larger and the documents which were indexed will have become outdated. Because of this reason web crawlers are being updated day by day and it becomes more popular in Information Retrieval Systems or Search Engines.

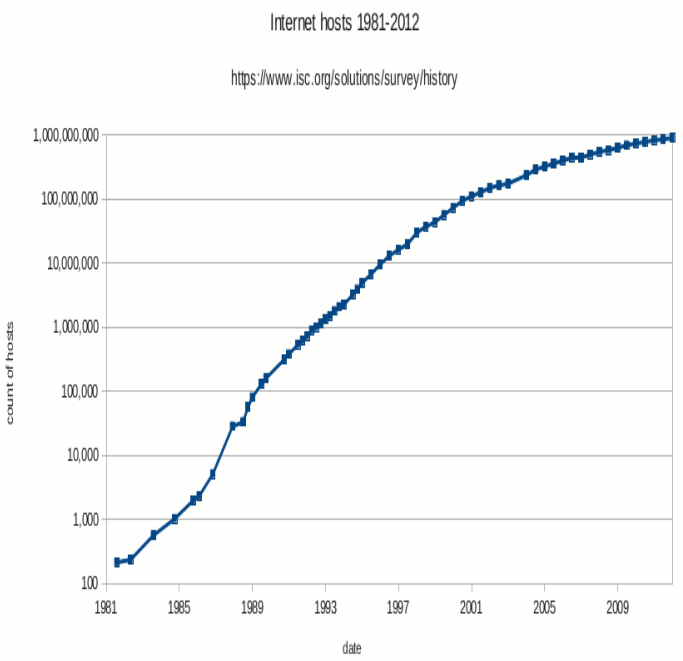

Fig. 1 Internet growth (Total no. of Hosts / Year) (taken from http://en.wikipedia.org/wiki/file)

The advent of the World Wide Web caused a dramatic increase in the usage of the Internet [2, 6]. The World Wide Web is a broadcast medium where a wide range of information can be obtained at a low cost. Information on the World Wide Web is important not only to individual users, but also to the organizations especially when the critical decision making is concerned. Most users obtain World Wide Web information using a combination of search engines [7] and browsers. However these two types of retrieval mechanisms do not necessarily produce all the information needed by the user. When the User tries to search the information hundreds of irrelevant documents return in response to a search query, only less than 18% of web pages are relevant to the user. To overcome this problem we need one effective search engine, which produces maximum relevant information in minimum time and at low cost.

Many academic and industrial researchers have looked at web searching technologies over the last few years, including storage, indexing, ranking techniques, crawling strategies and a significant amount of work on the structural analysis of the web [8]. Thus, highly efficient crawling systems are needed in order to download the hundreds of millions of web pages indexed by the major search engines. In fact, search engines battle against each other primarily based on the size and currency of their primary database, in addition to the quality and response time of their ranking function. Even Popular search engines, such as Google or AltaVista, presently cover up only restricted parts of the web, and a large amount of their data is several months out of date.

Applications involving search are everywhere in the world. The field of computer science which is most involved with R&D for search is Information Retrieval (IR). "[9, 10] Information Retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information." (Salton, 1968). General definition can be applied for many types of information and search applications. Primary focus of IR has been on text and documents since 1950's.

TABLE I. DIMENSIONS OF IR

Content	Applications	Tasks
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

All search engines [11] available on the internet need to traverse web pages and their associated links to copy them and to index them into a web database. The programs associated with the page traversal and recovery is called crawlers. The main decisions associated with the crawlers algorithm are when to visit new sites that have been discovered through links. The parameters to balance are network usage and web server overload against index accuracy and completeness.

II. SCOPE & OBJECTIVES

This proposed system aims at creating a search engine which searches the information in domain form from the web by saving the storage space and searching time and as a result it improves the performances of the system.

The main objective of the system is as follows

- Reduces the Overhead of the user.
- Maintains the Freshness of the page.
- Provides Scalability and Portability.
- Provides High Performance.

III. RELATED WORK

An ultimate goal of web mining is to analyses the structure of Web. This can be achieved by the system that is capable of gathering more relevant information from the web. Web Crawler is defined as "a software component that iteratively collects information from the web, download pages and follows the linked URL's" [12]. Implementation of crawlers stared in early 1990's. But till the Google introduced its distributed crawling module in late 1990's, all previous crawlers were stand alone. Fig .2 represents the Process of a General Crawler. It approaches Breadth First Search or Horizontal search.

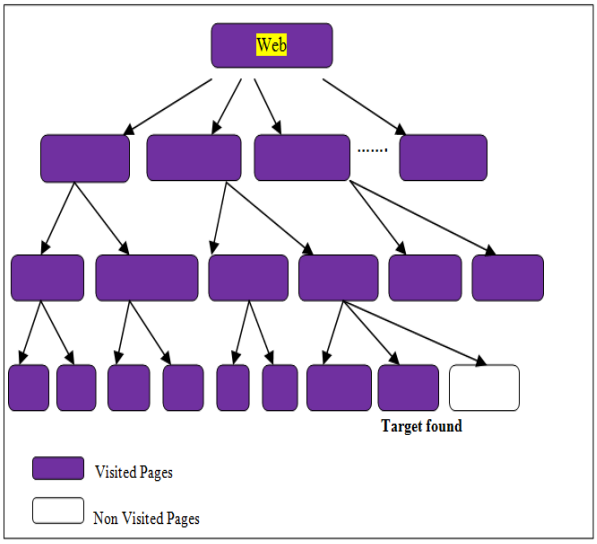

Fig. 2 General Crawler Process.

General crawler crawls all the pages from the web by using breadth first strategy. In this process, when we want to search particular information, it has to search in horizontal manner. Because of this method, it retrieves more non relevant information which is not required to that context. It takes longer time to access the actual information, thus it reduces the precision and recall.

Now days, a major challenge faced by the current web crawlers is to select the important pages for downloading. The crawler cannot download all the pages from the web due to large size. In our DBIR system, crawler will select the pages

and it visits important pages first by prioritizing the URLs in the Queue.

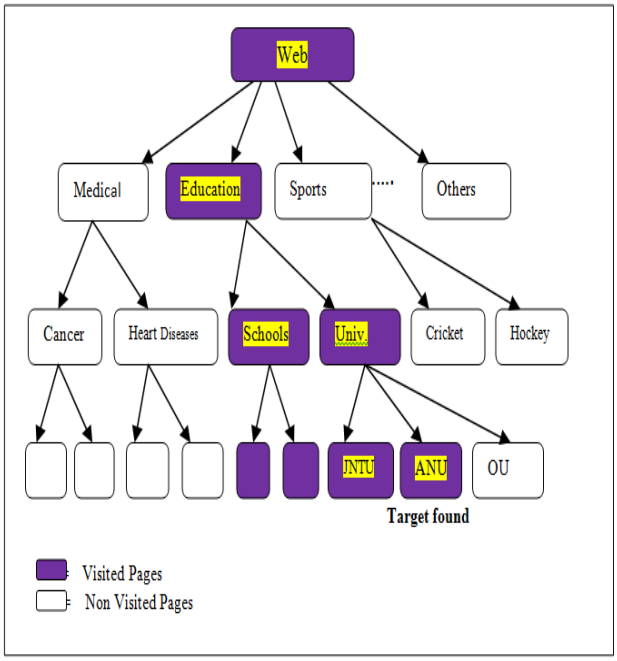

Fig. 3 DBIR Crawler Process.

In our DBIR System, It searches vertically based on the domain. In this process it skips or ignores all non relevant information thus we can improve the precision and recall.

Example to search ANU (Acharya Nagarjuna University) Text and domain name is given as Education in our DBIR System:

The Fig. 3 represents that searching is done in vertically according to the domain name is Education. In each level it skips non relevant pages, thus it improves the performance of the Searching process.

The fig.4 is the example for crawling the web pages which are interconnected through links.

IV. Domain Based Information Retrieval (DBIR) System

It is an extension of our earlier Effective performance of Web crawler (EPOW).In this proposed system we have added Ranking adaption with pattern matching (RAPM) algorithm in Master Crawler(Effective Web Crawler).

In this approach, it has two Crawlers. The first one is Basic Crawler and the second is Effective web crawler (i.e. master crawler). Basic Crawler consists of multiple downloaders. It fetches web pages (documents) from the World Wide Web when provided with a corresponding URL. Effective web crawler receives the URLs which is sent by the basic crawler and decides what page to request next based on the category (Domain) and issues a stream of requests (URLs) to the basic crawler (Downloader).The basic crawler downloads the requested pages and supplies them to the master crawler for analysis and storage.

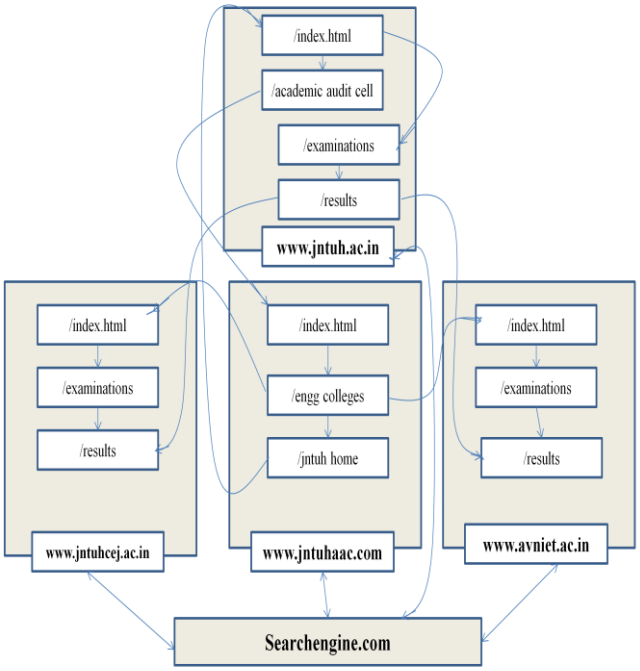

Fig. 4 Example of Crawling the Web

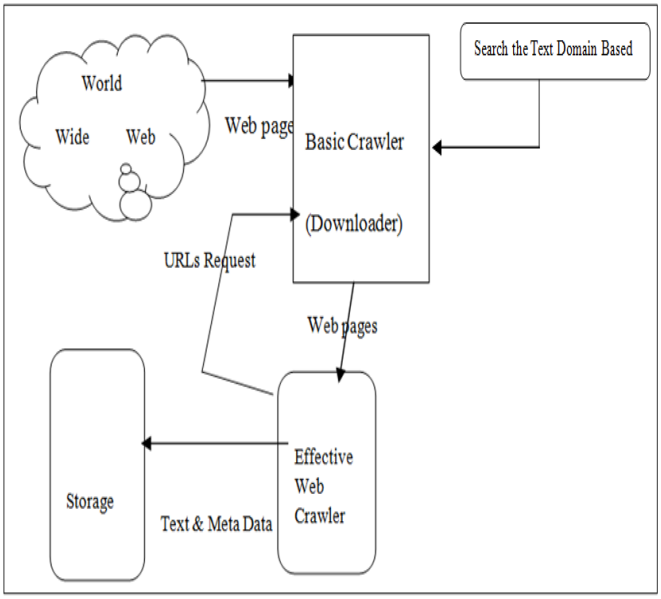

Fig. 5 Domain Based Information Retrieval (DBIR) System.

Master crawler collects all the related information and stored in the database by category wise. This process will be iteratively done till maximum relevant documents are fetched from the web at regular interval.

Once user enters the query, Master crawler analyzes the query and sends multiple URLs list which is relevant to the previous document. The web has a very dynamic nature [13, 14], and crawling a fraction of the web can take weeks or months. By the time a web crawler has finished its crawl, many events can have happened, including creations, updates and

deletions. The pages will remain outdated due to these modifications. The objective of this DBIR system is to keep the average freshness [15] of pages in its collection as high as possible, or to keep the average age of pages as low as possible. For that we have adapted optimal Revisit Policy. This method is for keeping average freshness high by ignoring the pages that change too often, and the optimal for keeping average age low is to access frequencies that monotonically increase with the rate of change of each page.

V. IMPLEMENTATION

A. *Flow of the Data in Domain Based Information Retrieval System*

To implement the crawler in real time we took a GUI application wherein the user has to login to use our DBIR system with a valid id and password. The User can get the id by signing-up into our application which will create a separate profile for the user storing his information, searched terms , sites visited which will provide privacy for each user. The user needs to add the sites in his profile which will be indexed by our DBIR crawlers and maintain the details in the user profile for searching the terms. Adding of sites has been allocated with category wise and admin will see in which category the site to be indexed by the system should. Once a site is indexed, all the details are crawled and maintained by the system in user profile. The system has admin for the application who can control all the activities of the user and system. Admin can upload the domain names, so that whatever indexes crawled related to that domain will be stored in separate storage area.

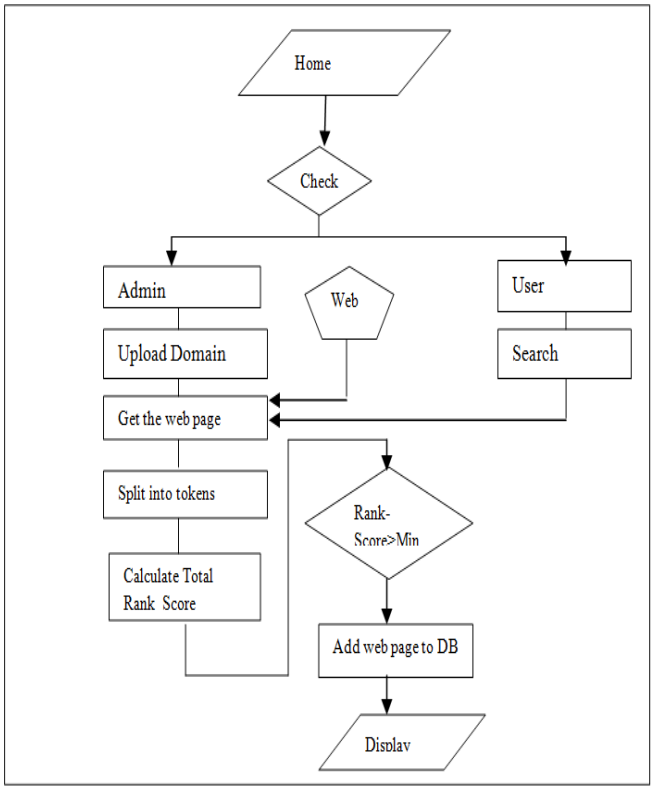

Fig. 6 Flow Chart for Domain Based Information Retrieval System (DBIR)

B. *Rank Adaption with pattern Matching Algorithm*

Input: A webpage (wp), Rank Table

Output: Total Rank of the webpage

Step 1: Initialize input Rank=0;

Step 2: Select the first term (t) and Corresponding rank from Rank table

Step 3: Webpage will be divided into tokens or words (Tokenizes)

Step 4: Compare the first term with webpage (wp).for each matching pattern, the Rank_Score will be added by 1

Step 5: Total Rank_Score = Rank * standard weightage of the page

Step 6: Select next term find the total Rank_Score of that term.

Step7: Repeat until all the terms are compared and find the overall Rank Score of the Webpage.

Step8: If Overall Rank_Score > Min Rank then the web page will be stored in Database.

Step 9: End.

VI. CONCLUSION & FUTURE DETECTIONS

Whenever we are searching for the content from the web, most of the crawlers are not returning the expected results and they give a list of sites to be searched by user which is related to the searched text but not required by the user. Thus it increases the user overhead and consumes time. This system decreases the overhead of the user by providing less in number and most relevant results based on the category of the searched term. To achieve this system uses DBIR Method, unlike other search engines, here two crawlers are used for crawling and indexing sites. The introduction of additional crawler with multiple downloaders increases the speed and performance of the crawling.

The front end of the system enables user to maintain his entire query and search history by providing a login system with Id and password. The validation process provides security and privacy to the users unlike other browsers. The system also has an admin who can control all the users and its activities. User or Admin has to add the sites to be crawled by this DBIR system. User can also share the search results with others using SMS and Email. Overall, the system is designed to decrease the burden on the internet surfers and provide an optimized GUI for safe and easy browsing.

Web is very large and dynamic. Searching the required relevant content from the web is always difficult. Grouping the collections from the web is always challenging. We need to gather from broad range of domains. This project validated limited no of collections. We need to validate a large number of collections from various domains. This Front end application can be implemented for the specific organizations or institutes to provide the privacy to the users to access the web. It can be used for maintaining their history and profile of their employees.

ACKNOWLEDGMENT

We would like to thank everyone who has motivated and supported us for preparing this manuscript.

REFERENCES

[1] Ricardo Baeza-Yates, Felipe Saint-Jean, and Carlos Castillo. Web structure, dynamics and page quality, In Proceedings of String Processing and Information Retrieval (SPIRE), Springer LNCS, 2002.

[2] Sk.Abdul Nabi and Dr. PremChand ,Effective Performance of Information Retrieval, International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012.

[3] S S Vishwakarma , A Jain , A K Sachan, A Novel Web Crawler Algorithm on Query based Approach with Increases Efficiency,International Journal of Computer Applications , Volume 46– No.1, May 2012.

[4] Wenqing Yuan , Research on Prototype Framework of a Multi-Threading Web Crawler for E-Commerce, In Proceedings of the International Conference on Management and Service Science, MASS '09. IEEE Transactions, 2009.

[5] Heydon and Najork. Mercator: "A scalable, extensible Web crawler", World wide web2 (4), 1999.

[6] RM. Vidhyavathy, E.Ramaraj, N. Venkatesan, A Study Of Mining Web Intelligent Information, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 9, September 2012.

[7] H.Vernon Leighton and J. Srivastava, Precision among www search services (Search Engines), 1997.

[8] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's new on the web?: the evolution of the web from a search engine perspective. In Proc. Of 13th conference on World Wide Web, May 2004.

[9] Kowalski, Gerald, Mark T Maybury, Information Retrieval Systems: Theory and Implementation, Kluwer Academic Press, 1997.

[10] Terrence A.Brooks. Web Search : how the web has changed information retrieval. Information Research, April 2003.

[11] H.Vernon Leighton and J. Srivastava. Precision among www search services (search Engines), 1997.

[12] S.Brin , L.Page , " The Anatomy of a large–Scale Hyper Textual Web Search Engine". In Proc. Of the 7th World Wide Web Conference, 1998.

[13] Recardo Baeza – Yates Carlos Castillo and Felipe Saint – Jean., Web Dynamics , Springer , 2004.

[14] Brain E. Brewington and George Cybenko, How Dynamics is the Web. In Proceedings of the Ninth International World – Wide Web Conference , May 2000.

[15] Junghoo Cho and Hector Garcia-Molina. Effective page refresh policies for web crawlers.ACM Transactions on Database Systems, 28(4), December 2003.

AUTHORS

Prof Shaik.Abdul Nabi1 is the Head of the Dept. of CSE, AVN Inst. Of Engg.& Tech, Hyderabad, AP, India. He completed his B.E (Computer Science) from Osmania University, A.P. He received his M.Tech. from JNTU Hyderabad campus and currently he is pursuing Ph.D. in the area of Web Mining from Acharya Nagarjuna University, Guntur, AP, India. He is a certified professional by Microsoft. His expertise areas are Data warehousing and Data Mining, Data Structures & UNIX Networking Programming.

Dr P.Premchand2 is a professor in the Department of Computer Science & Engineering, Osmania University, Hyderabad, A.P,India. He completed his ME (Computer Science) from Andhra University, A.P. He received Ph.D degree from Andhra University, A.P. He has guided many scholars towards the award of Ph.D degree from various Universities. He was a director of AICTE, New Delhi, during 1998-99. He also worked as the Head of the Dept of CSE , Additional Controller of Examinations and chairman of BOS, Faculty of Engineering, OU. Currently he is working as the Dean, Faculty of Engineering, Osmania University, Hyderabad, and AP.