An Efficient Web mining for browsing Corporate Sites.

Bharat Bhushan
Associate Professor & Head
Guru Nanak khalsa College,
Yamuna Nagar (affiliated
to Kurukshetra University,
Kurukshetra, India)

Narender Kumar
Research Scholar
Singhania University ,
V.P.O. - Pacheri Bari, Dist.
Jhunjhunu, Rajasthan -
[INDIA]

ABSTRACT

Web mining in the open web has many business applications. Dependency on internet for web mining is much more especially when one discovers something as per the requirements. No doubt , search engines are useful to fetch the information but for fetching the relevant information from open web really is a tedious & time consuming task. By understanding the user requirements the efficient information retrieval can help to attract more visitors and improve service of e commerce . In this paper an efficient web mining technique is designed for business corporate to browse information using useful Key Attributes.

Keywords

Web crawling, Search Engine, Key Attributes, HTTP request, web pages

.

1. INTRODUCTION

Web mining can be broadly defined as the discovery and analysis of useful information from the World Wide Web. This describes the automatic search of information resources available online i.e. Web content mining and the discovery of user access patterns from Web servers i.e. Web usage mining 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 find the desired information resources and to track and analyze their usage patterns. These factors give rise to the necessity of creating intelligent systems that can effectively mine for knowledge. [1]

In order to submit queries to web search engines for mining ,One have to carefully choose the suitable combination of keywords. Without rich background knowledge about keywords in web documents, it is too difficult to find out invaluable URLs by search engines. Here , The technique is designed for web content mining using the combination of key attributes and anchor text which helps us to find the relevant urls from open web . The information retrieval is more complicated when the depth issue occurs. i.e. upto which level of links one has to retrieve the information to get the valuable information

2. RELATED WORK

For finding the information/content from open web , Bot or crawler are used and the size of web is growing very fast day by day so it is difficult by crawler to cover complete web. So in this respect we will try to search updated web pages & schedule it for crawling. Various algorithms & studies on this topic suggests how web crawler / bot do it efficiently.

Junghoo Cho H et.al. describes several importance metrics, ordering schemes, and performance evaluation measures in what order a crawler should visit the URLs it has seen, in order to obtain more "important" pages first. Obtaining important pages rapidly can be very useful when a crawler cannot visit the entire Web in a reasonable amount of time. Their results show that a crawler with a good ordering scheme can obtain important pages significantly faster than one without.[2]

Marc Najork and Janet L. Wiener Examined the average page quality over time of pages downloaded during a web crawl of 328 million unique pages. They use the connectivity-based metric PageRank to measure the quality of a page. We show that traversing the web graph in breadth-first search order is a good crawling strategy, as it tends to discover high-quality pages early on in the crawl. [3]

Junghoo Cho, Hector Garcia-Molina Proposed architecture to build an effective incremental crawler based on various design choices . The crawler selectively and incrementally updates its index and/or local collection of web pages, instead of periodically refreshing the collection in batch mode. The incremental crawler can improve the “freshness” of the collection significantly and bring in new pages in a more timely manner. [4]

Carlos Castillo et.al. presented a comparative study of strategies for Web crawling. They showed that a combination of breadth first ordering with the largest sites first is a practical alternative since it is fast, simple to implement, and able to retrieve the best ranked pages at a rate that is closer to the optimal than other alternatives. They also explored the effects of large scale parallelism in the page retrieval task and multiple-page requests in a single connection for effective amortization of latency times.[5]

Baeza-Yates, R. and Castillo described a crawling software designed for high-performance, large-scale information discovery and gathering on the Web. This crawler allows the administrator to seek for a balance between the volume of a Web collection and its freshness; and also provides flexibility for defining a quality metric to priorize certain pages.[6]

EelcoHerder studied user's page revisit behavior in resource discovery. Revisit are very common in web navigation, but not as predominant as reported in earlier studies. Backtracking is the most common type of page revisitation and is both used for finding new information and relocating information visited before. Search engines are mainly used for finding new information and users frequently backtrack to result pages. Visits to pages already visited in earlier sessions tend to occur in chunks, but it is not straight forward to create a list of most likely pages that will be revisited. [7]

Proposed Approach

Lots of junk links are crawled by crawler when one tries to fetches information related to business prospects. In today's date lots of information are available & it is very difficult how to fetch relevant contents from the web. The proposed model described here will be helpful to mine the valuable stuff from corporate sites as per the business houses by using key attributes and Anchor Text. For more efficiently use of crawler one should restrict the crawling depth i.e. how much depth one will crawl the web site.

Method:

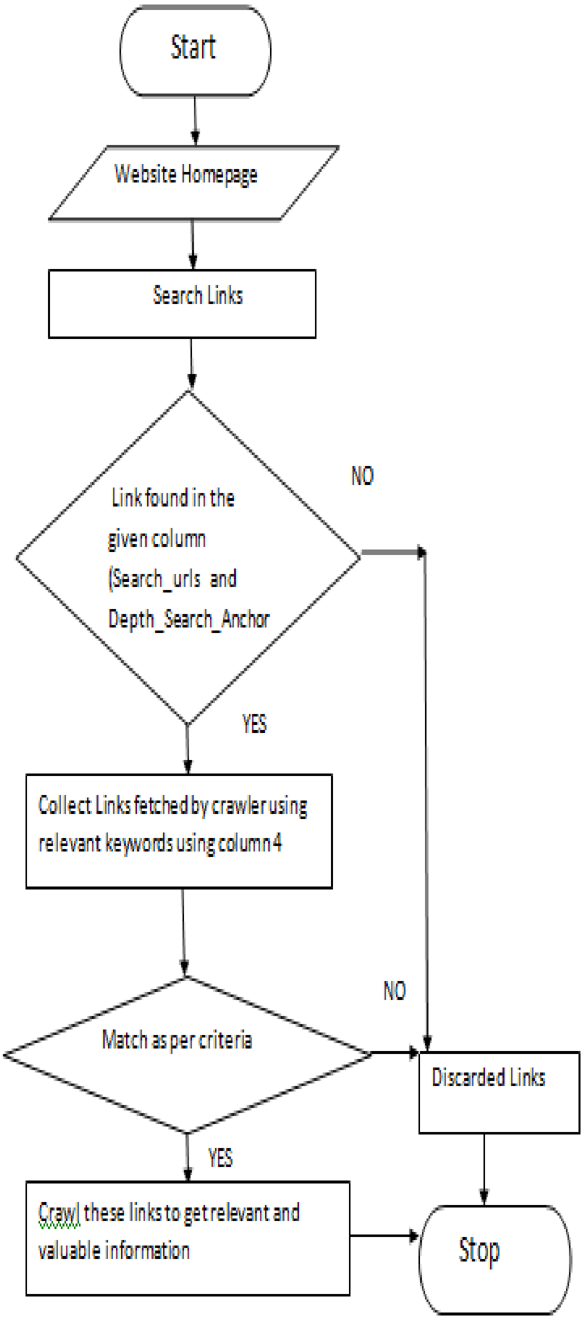
1. Using Sitemap : Sitemap link percentage is very high during browsing the websites. So by using sitemap link one can maximize the crawler as it fetches only information as per pattern keywords/anchor text found under sitemap. Depth should be <=1
2. In few of cases where sitemap not found where one can also send the list of corporate sites to crawler to fetch the links as per pattern keywords / anchor text mention in the list.
   Depth should be <=2
   a. When we open the website and crawl all information from the homepage ,Then the depth is zero. http://www.henderson.com/sites/henderson/home.aspx
   b. If the homepage doesn't contain the information that click on the link available on the homepage and if this link further not giving any valuable information than further click on the link available and soon. Till to get the relevant information from that site. http://www.henderson.com/sites/henderson/aboutus.aspx

The proposed model will work as

- ➢ Website Link : from where we need to search
- ➢ Search_urls : first of all one search the relevant links so crawler search the links given in this column from website [this is Pattern keyword]
- ➢ Depth_Search_Anchor : In the same manner crawler search the links using anchor text.
- ➢ Links that are useful for crawling as per requirement. : By this set one will fetch the list of valuable links & store for further crawling as per need.

As input website link given & using search_url & anchor Column found using crawler & the last column find the key words in particular links & store it in some database. It is the links as one need.

Here one will get the relevant links from corporate sites & if crawl these link with freshness one will get always relevant information in less time & correct data as per need


Start
Website Homepage
Search Links
Link found in the given column (Search_urls and Depth_Search_Anchor
NO
YES
Collect Links fetched by crawler using relevant keywords using column 4
Match as per criteria
NO
YES
Discarded Links
Crawl these links to get relevant and valuable information
Stop

This procedure is tabulated below :

Webiste Link	**Search_urls**	**Depth_Search_Anchor**	**Links that are useful for crawling as per our reqirement.**
http://www.henderson.com/sites/henderson/home.aspx	News	Site Map	News
	Press Release	SiteMap	Press Release
	Events	Sitemap	Events
	Management	AboutUs	Management
	Directors	About Us	Board
	Site Map	About	All Releases
	AboutUs	Investors	Management Team
	About Us	Press	IR Press release
	SEC Filings	Investor	Event Calendar
	Press Room	Media	Executive Profile
	Investor	Company	News & Events
	About	Corporate Governance	Current Press release
	Investors	News &	latest

		Events	Press Release
	Media	Corporate Profile	Directors
	Press Area	Management	Investor
	Company	Executive	News
	Corporate Governance	Board	Directors
	News & Events	News Release	Executive
	Press	Directors	Press kit
	Corporate Profile	PressPass	Company heads
	Sitemap	Leadership	Media
	Executive	Company heads	Profile
	Board	Press kit	Company heads
	Media Centre	News	Announcements
	PressPass	Media releases	News
	About	Profile	Press Release
	News	Site Map	Events
	Press Release	SiteMap	Management
	Events	Sitemap	Board
	Management	AboutUs	All Releases
	Directors	About Us	Management Team
	Site Map	About	IR Press release
	AboutUs	Investors	Event Calendar
	About Us	Press	Executive Profile
	SEC Filings	Investor	
	Press Room	Media	

If one automate this approach & repeat same process again and again (due to dynamically nature of open web) if link expired than new link should be require for that , So one will get another same link after automation. Definitely it will save time, money & effort for fetching the relevant information from open web.

Results :

Id	**Homepage**	**Status**
8487	http://www.integrysgroup.com/	Processed
8488	http://www.transitchicago.com/	Processed
8489	http://e.nikkei.com/e/fr/freetop.aspx	Processed
8490	http://www.bse.hu/	Processed
8491	http://www.wurts.com/	Processed
8492	http://www.state.il.us/srs/gars/home_gars.htm/	Processed
8493	http://www.bsp.gov.ph/	Processed
8494	http://www.defense.gov/	Processed
8495	http://www.fairfaxcounty.gov/retirement/	Processed
8496	http://www.rolls-royce.com/	Processed
8497	http://www.msu.edu/	Processed
8498	http://www.psu.edu/	Processed
8499	http://www.rice.edu/	Processed
8500	http://www.udel.edu/	Processed
8501	http://www.platts.com/	Processed
8502	http://www1.umn.edu/	Processed
8503	http://www.thinkmoney.co.uk/	Processed
8504	http://www.oracle.com/agile/index.html/	Processed
8505	http://www.att.com/	Processed
8506	http://www.polycom.com/	Processed

8507	http://www.harthosp.org/	Processed
8508	http://www.bluestone.com.au/	Processed
8509	http://www.miamiherald.com/	Processed
8510	http://www.nhs.uk/Pages/HomePage.aspx/	Processed
8511	http://www.nlm.nih.gov/	Processed
8512	http://www.nsf.gov/	Processed
8513	http://www.mhra.gov.uk/index.htm	Processed
8514	http://www.salon.com/	Processed
8515	http://www.navy.mil/swf/index.asp/	Processed
8516	http://www.doi.gov/	Processed
8517	http://www.carsales.com.au/	Processed
8518	http://www.selexgalileo.com/	Processed
8519	http://www.dh.gov.uk/en/index.htm	Processed

3. REFERENCES

[1] B.Mobasher , N Jain , E.Han ,and J. Srivastava , "Web Mining : Pattern discovery from world wide web transactions," Technical Report TR 96-0505 , University of Minnesota, Dept. of Computer Science , Minneapolis , 1996.

[2] Cho, J., Gracia-Molina, H. and Page, L. (1998), “Efficient Crawling Through URL Ordering”, Computer Networks and ISDN Systems (0169-7552), Volume 30, No. 1-7, pp 161-172.

[3] Najork, M. and Wiener J. L. (2001), “Breadth-First Search Crawling Yields HighQuality

Pages”, WWW’01, 10th International World Wide Conference, pp. 114-118.

[4] Cho, J. and Gracia-Molina, H. (2000), “The Evolution of the Web and Implications for an Incremental Crawler”, In Proceedings of 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt, pp 200-209.

[5] Castillo, C., Marin, M., Rodrguez A. and Baeza-Yates, R. (2004), “Scheduling Algorithms for Web Crawling”, WebMedia & LA-Web, pp 10-17.

[6] Baeza-Yates, R. and Castillo, C. (2002), “Balancing Volume, Quality and Freshness in Web Crawling”, In Soft Computing Systems – Design, Management and Applications, HIS, Santiago, Chile, pp 565-572.

[7] EelcoHerder Characterizations of User Web Revisit Behavior Eelco Herder Group of Human-Media Interaction, Dept. of Computer Science Twente University, The Netherlands herder@cs.utwente.nl

DISCUSSION & CONCLUSION

The proposed model focuses on reducing the time to fetch the information from web sites using sitemap , anchor text , keywords pattern matching. By using sitemap of a website, one can increase the efficiency of a web crawler to a great extent. The technique used found that while revisiting the websites, if a web crawler can find which web pages have been updated or newly added since last visit, then there is no need to download the complete website every time. With the proposed solution it will be less time consuming for web crawlers to maintain the freshness of downloaded websites used by search engines.

It is also important to mention depth issue. If the crawler has to bring the relevamt information from the maximum depth it will increase the crawling time for that particular link hence we have set the depth <=2. During study 5000 Websites it is found that 90 % of relevant information is available on depth 0,1 or 2.

.