Websites:
http://www.sciencepub.net/academia
http://www.sciencepub.net

Emails:
aarena@gmail.com
editor@sciencepub.net

Academia Arena

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

OPTIMIZED WEB CONTENT MINING

K. Thirugnana Sambanthan[1], Dr. S.S. Dhenakaran[2]

[1]Research Scholar, Dept. Computer Science, Alagappa University, Karaikudi, Tamilnadu, India.
[2]Professor, Dept. Computer Science, Alagappa University, Karaikudi, Tamilnadu, India.
E-mail: shivaperuman@gmail.com,ssdarvind@yahoo.com

Abstract: Web content mining is the mining, integration and extraction of valuable data, information and knowledge from Web page content. It is set of information extraction tools brought forward in order to identify and collect content items, such as Text Extraction and Wrapper Induction. In this present era extracting images, links, html source is becoming a tedious task since there is a large growth in the number of websites. The present day crawlers are not answering the different queries asked by the end user. The aim of this paper is to propose a new optimized web searching technique which can give correct relevant search results. The proposed crawler can search a website and display the results regarding the website accurately.
[K. Thirugnana Sambanthan, Dr. S.S. Dhenakaran. **OPTIMIZED WEB CONTENT MINING.** *Academ Arena* 2024;16(3):23-29]. ISSN 1553-992X (print); ISSN 2158-771X (online). http://www.sciencepub.net/academia. 03. doi:10.7537/marsaaj160324.03.

Key Words - Web Data Extraction, Web Content mining, Optimized Web Content Mining, Crawler, Html Crawler, Link Crawler, Image Crawler.

I. INTRODUCTION.

In this era, the web source acts as a resource pool which provides an infinite set of solutions to the users who are looking for useful information on the web. But the fact that information found in web is not relevant and reliable to a great extent is due to large collection of content. Therefore a framework is needed which would consist of structured content from which the user could pick the most relevant and reliable information for usage. A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm on a web crawler.

2. Web Content Mining

Traditional technique of searching the web was via contents. Web Content mining is the extended work performed by search engines [1].Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. [2]. Data is semi structured is inherent structure which appears implicitly on page and varies from one page to another.

2.1. Structured data mining techniques

Structured data extraction is a progress of extracting information from web pages. A program for extracting such data is usually called a wrapper. Structured data are typically the data records retrieved from underlying database and displayed in the web pages following some templates. Sometime, the template is a table. Sometime, it is a form. Extracting such data records is useful because it enables us to obtain and integrate data from multiple sources (Web sites and pages) to provide value-added services, e.g., customizable Web information gathering, comparative shopping, meta-search, etc.

2.2 Web Spiders

Web spiders are prominently known as crawlers which look for the information across the WWW. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, there are two types of Web Crawler which are called as External and Internal Web crawler. Crawlers are computer programs that traverse the hypertext structure in the web. External Crawler crawls through unknown website. Internal crawler crawls through internal pages of the website which are returned by external crawler.[Fig.1]

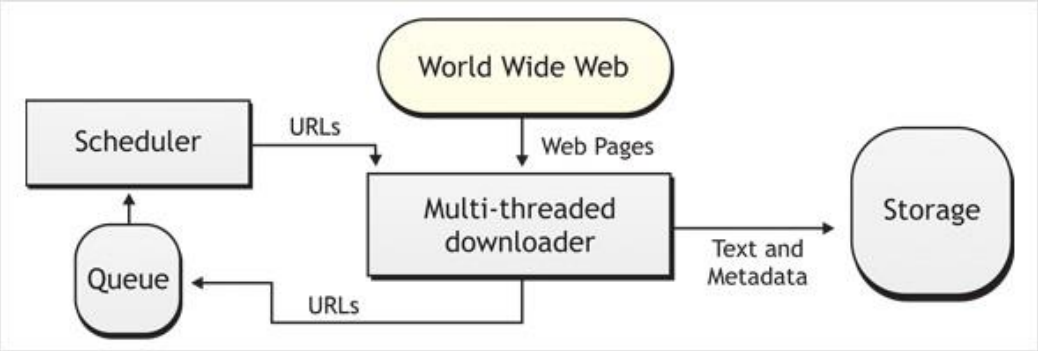


Fig. 1 General Web Crawler Architecture

3. Dhekts Crawler

A new Crawler called Dhekts is created to return Images, Link, Html code from a given website. This Crawler is unique in nature since it can return all the details of a particular website (i.e) Images, Links, Files, details of any website. It crawls through the links in the specified website and it can further crawl to other links in that website. The Dhekts Crawler is designed to crawl images, links, HTML links, Depth and Relevance. Hence, it is named after Image Dhekts Image Crawler, Dhekts Link Crawler, Dhekts HTML Crawler, Dhekts DepthCrawler and Dhekts Relevance Crawler respectively.

3.1. Image Crawler

The Dhekts Image Crawler can be used to browse all images of a website recursively. Dhekts Image Crawler is used to collect a multitude of images from the website. The images can be viewed as thumbnail, the url of the image is also listed beside it. The crawler crawls the website and searches for images (jpg, gif, png etc.) and returns the image with the url. There may be n number of images in the website, this crawler will easily display the images with the url. [Fig 2]. Uniqueness of the Dhekts Image Crawler is, without storing the details in the database it can directly display the results by crawling a particular website.

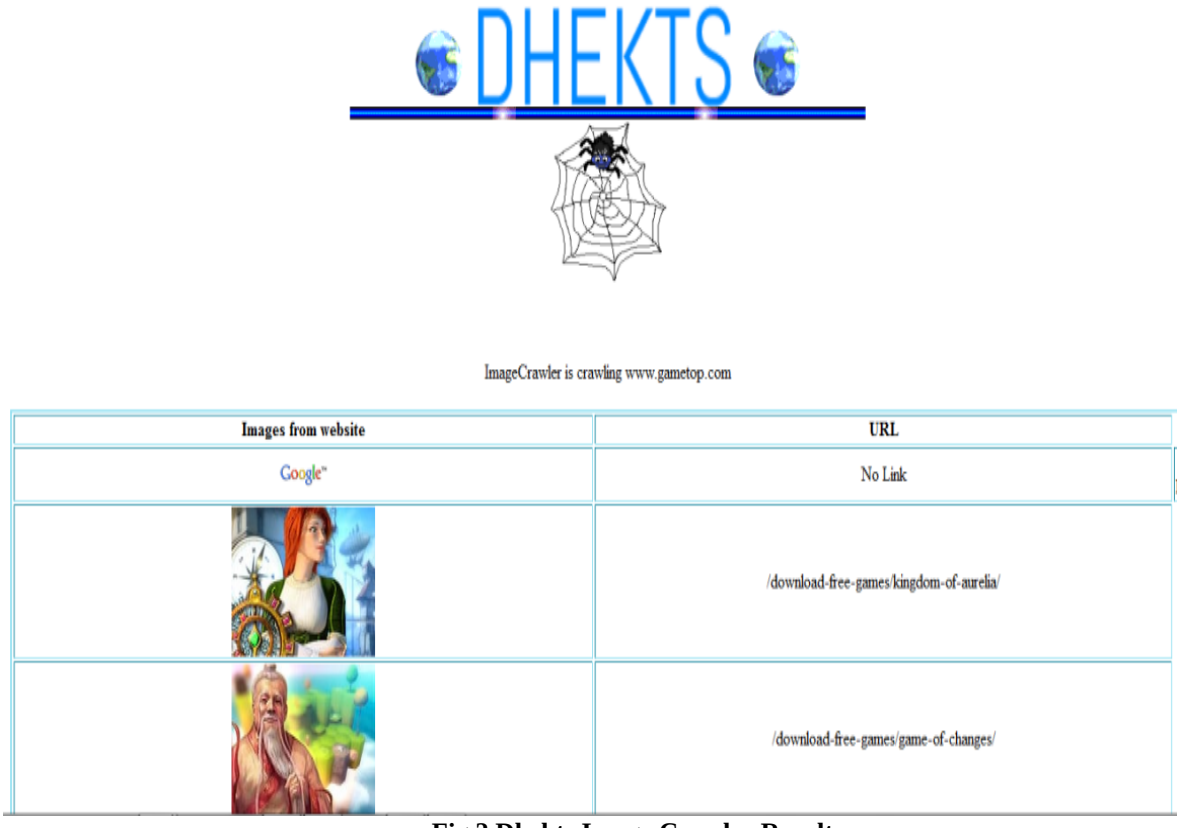
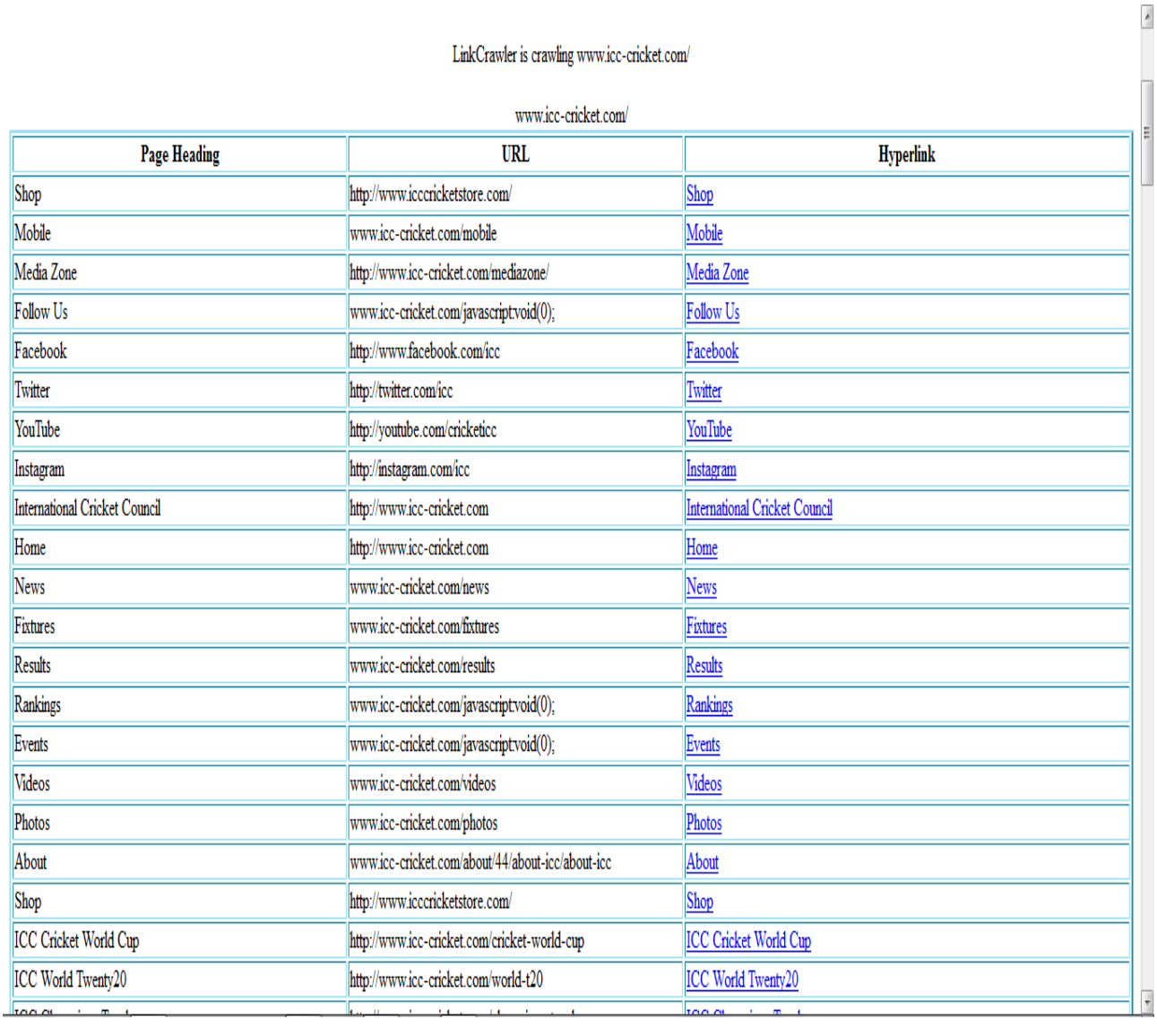


Fig 2 Dhekts Image Crawler Results

3.2 Link Crawler

The Dhekts Link Crawler will crawl all the links that are available in any given websites. It has a unique way of crawling the websites. It crawls a site and gathers information about Internal and External Links. Uniqueness of the Dhekts Link Crawler is that it can crawl a website and list the Page heading, URL, hyperlink available in the entire website. Dhekts Link Crawler acts like a site map provider for any site [Fig 3]. Dhekts Link Crawler can display the results directly without storing in the database. This will save more storage space in the server.



LinkCrawler is crawling www.icc-cricket.com/

www.icc-cricket.com/

Page Heading	URL	Hyperlink
Shop	http://www.icccricketstore.com/	Shop
Mobile	www.icc-cricket.com/mobile	Mobile
Media Zone	http://www.icc-cricket.com/mediazone/	Media Zone
Follow Us	www.icc-cricket.com/javascript:void(0);	Follow Us
Facebook	http://www.facebook.com/icc	Facebook
Twitter	http://twitter.com/icc	Twitter
YouTube	http://youtube.com/cricketicc	YouTube
Instagram	http://instagram.com/icc	Instagram
International Cricket Council	http://www.icc-cricket.com	International Cricket Council
Home	http://www.icc-cricket.com	Home
News	www.icc-cricket.com/news	News
Fixtures	www.icc-cricket.com/fixtures	Fixtures
Results	www.icc-cricket.com/results	Results
Rankings	www.icc-cricket.com/javascript:void(0);	Rankings
Events	www.icc-cricket.com/javascript:void(0);	Events
Videos	www.icc-cricket.com/videos	Videos
Photos	www.icc-cricket.com/photos	Photos
About	www.icc-cricket.com/about/44/about-icc/about-icc	About
Shop	http://www.icccricketstore.com/	Shop
ICC Cricket World Cup	http://www.icc-cricket.com/cricket-world-cup	ICC Cricket World Cup
ICC World Twenty20	http://www.icc-cricket.com/world-t20	ICC World Twenty20

Fig 3 Dhekts Image Crawler Results

3.3 HTML Crawler

The Dhekts Html Crawler will crawl a particular website will list all the html links and it can also give the html coding of the entire website. This crawler can be used to do research with regard to a particular website (i.e) one can analyze the coding techniques used by the website and the structure of the website using this crawler. [Fig 4] Its uniqueness is it can get the html code from any websites which has download or right click restriction.

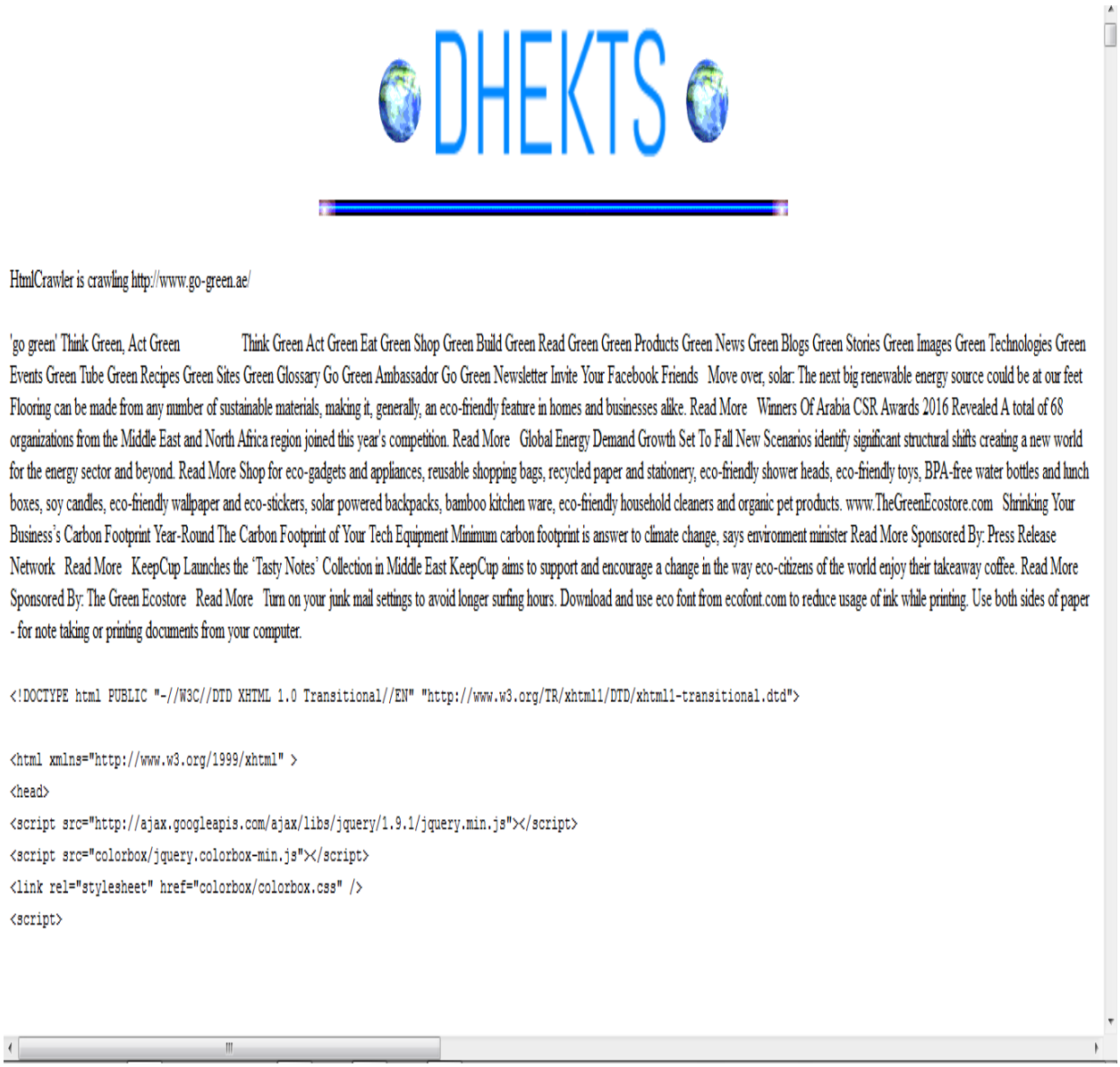


Fig 4 Dhekts HTML Crawler Results

3.4 DepthCrawler

The Dhekts Depth Crawler will crawl the entire website and will continue its crawling to other websites based on the link given in that. Crawl depth is the extent to which a search engine indexes pages within a website. Most sites contain multiple pages, which in turn can contain subpages. The pages and subpages grow deeper in a manner similar to the way folders and subfolders (or directories and subdirectories) grow deeper in computer storage. A Web site's home page has a crawl depth of 0 by default. Pages in the same site that are linked directly (with one click) from within the home page have a crawl depth of 1; pages that are linked directly from within crawl-depth-1 pages have a crawl depth of 2, and so on. Dhekts Depth crawler can crawl accurately and display results till depth of 5. [Fig 5]

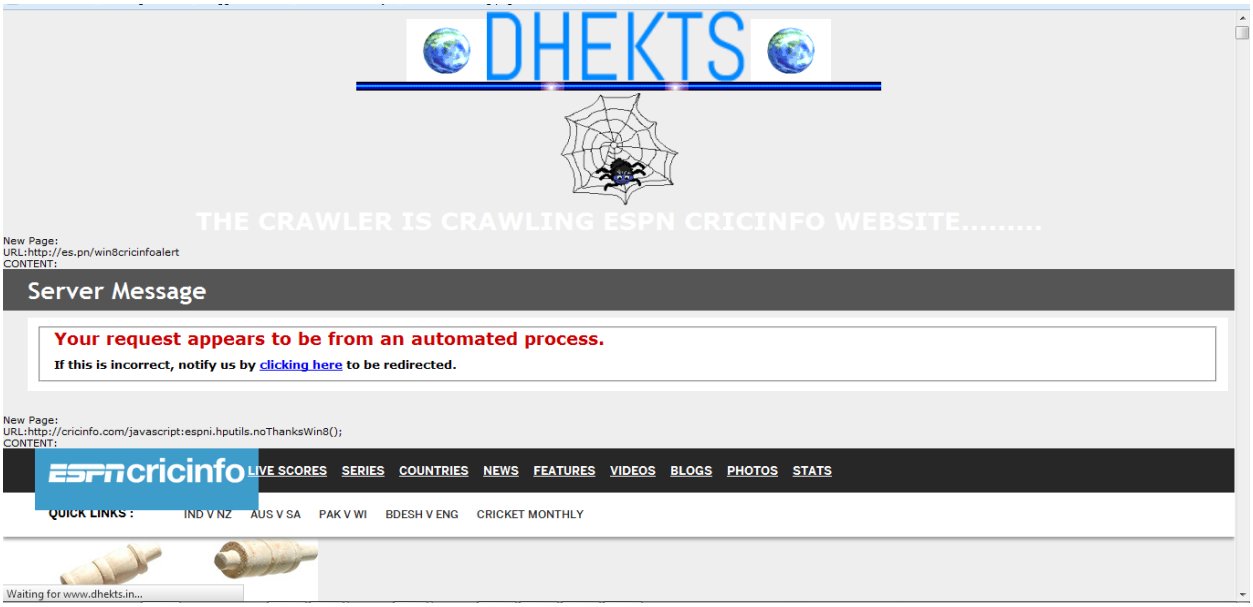


Fig. 5 Dhekts DepthCrawler Results

3.5 Relevance Crawler

Finally, Dhekts Crawler works to bring relevant information from a website which is called Relevance Crawler. This Crawler works based on the keywords search available, no. of keywords present in a particular website [9], User Relevance rating given to the website. This crawler helps in getting the accurate results in searches.

Pseudocode of Relevance Crawler

START
Enter the search query
Check for connection errors
Prepare the query string
Create the new keyword search
Start request to server
If Search is found
{
For each source page
{
Extract the web address and parse the result
For each Source code
{
Check for user rating and the no.of keywords present in the website
}
End For
Connect to database
If no response
Set connection error
Else
{
Prepare database
Check for updates if required
Check for redundancy
Insert URL and disk address
Check for redundancy
Sort based on user rating and no. of keywords
Display the result
}
End If
Check for termination
If the termination condition reached
Exit
Else
Increment the next source page
End if
}
End for
}
Else
Display an error message
End If
END

Fig.6 Relevance Crawler

Conclusion

Web mining is a rapid growing research area. Due to the heterogeneity and the lack of structure of Web data, automated discovery of targeted or unexpected knowledge information still present many challenging research problems. This article is tried to create a optimized web content mining technique which will pave a way for future search engines to give relevant and correct result in a short span of time. The Future Search engines can combine personal, social and local data to show relevant results. Future researchers can follow these crawling procedures and develop Dynamic Relevance Search Engines.

REFERENCES

[1]. A.Mendez-Torreblanca, M.Monte,"A Trend Discovery for Dynamic Web Content Mining", IEEE, Inteligence System, Vol 14, pages.20-22, 2002.

[2]. Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining. World Academy of Science, Engineering and Technology 49.

[3]. Chen Enhog,Wang Sufi "Semi Structure data extraction and Schema Knowledge Mining",EUROMICRO Conference, Proceeding 25 Volume 2 Issue, pages 310-312,1999.

[4]. Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education. Inamdar, S. A. and shinde, G. N. 2010.

[5]. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[6]. Faustina Johnsonm Faustina Johnson Web Content Mining Techniques: A Survey An Approach To Web Content Mining Nita Patil, Chhaya Das, Shreya Patanakar, Kshitija Pol

Department of Computer Engg. Datta Meghe College of Engineering, Airoli , Navi MumbaiAutomation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.

[7]. Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys, Vol. 40,No. 2, April 2008, pp. 5:1 – 5:60.

[8]. Web Content Mining: Its Techniques and Uses- Govind Murari Upadhyay, Kanika Dhingra

[9]. Zhao Li, Wee Keong Ng,"WICCAP: From Semi-Structured Data to Structured Data", In Proceeding of the 11 Th IEEE International Conference and Workshop on the Engineering of Computer-Based System, 2004.

[10]. https://wordcounter.net/

[11]. http://stackoverflow.com/

[12]. http://whatis.techtarget.com/

3/21/2024