International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

HIT: Web Content Mining Tool

Tripurari Pujan Pratap Singh
M. Phil Computer Science
AISECT University
Bhopal, (M. P) India
chandan_singh954@yahoo.com

Dr. Anurag Seetha
Professor &. Dean
Computer Science &. Engineering
Dr. C. V. Raman. University
Bilaspur, Chhattisgarh India

K. K. Pandey
Asst. Professor
Computer Science &. Engineering
AISECT University
Bhopal, (M. P) India

Abstract **— Among the improvement of information retrieval as well as the expansion of web, websites have turn into an additional and more essential information standard for different associations and community. On other hand, information is distributed worldwide; people have to switch to several different websites also get back the related information consistently. In the changing circumstances World Wide Web became overworked with information and made it durable to mine data according to the requirement. Web mining came as a release for the exceeding difficulty. Web content mining is a section in web mining. This paper compacts amid a study of different techniques as well as prototype of content mining and the parts which has been manipulated by content mining. The web have structured, unstructured, semi structured along with multimedia data. This survey centres on how to pertain content mining going on the hyperlink, image and text.**

Web content mining wants to supply functional information or awareness from Web page contents. separately from conventional jobs of Web page classification and clustering , here are several additional Web content mining assignments, for example, taking out data/information, information incorporation, mining judgments from the user-based content, pulling out the Web to construct perception hierarchies, Web page pre-processing plus cleaning, etc. Here, we will introduce web mining and there types in detail. We will also try to create a simple and user friendly web mining tool "HIT" as well as introduce some other web mining tools and there working with characteristics. We also compare "HIT" with some other web mining tools. In future we will discuss about the coding part of "HIT" and other things.

Keywords **— Web Mining, Web Content Mining, Web Text Mining , Web Image Mining, Web Video Mining, Web structure mining, Web Hyperlink Mining, Web Usage Mining , Comparative Study of web Content Mining Tools, HIT introduction, Working, Conclusion, Proposed Work, References.**

I. INTRODUCTION

The changing scenario has led to the development of more and more advanced technologies in communication and various other emerging fields. People have switched from desktop to laptops and no One cares or even bothers to wait for a second to get the required query in this fast moving time. Explosion of data on the web have created many complicated situations like extracting the most suitable and relevant information as per requirement, learning about the consumers or individual users, searching potentially useful information and data. Further to add, the unstructured format of data had made it extremely difficult to retrieve particular information from the web.

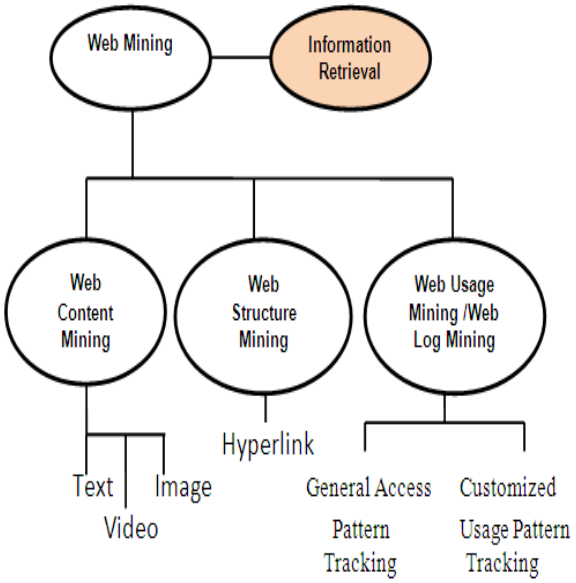


Fig.1. Web Mining

The exponential and rapid growth of data on internet forcing us to adopt some tools that can be useful and efficient to retrieve the valuable information from web. For the data to be conveniently extracted, it has been necessity of our computer to sort the data in any order so as to perform the specified task. Query for the simple data results into lots of unwanted and unspecified matter that has to be neglected for saving time and energy. Although present days working technologies and networking capabilities had made it possible to extract the required information, still the work is time consuming and tedious. The question arises about managing and retrieving these types of data. Looking at the above stated problems web mining came into existence...

II. WEB MINING

Web mining refers to the generally process of discovering potentially valuable and previously unknown information or awareness from the web data. Web mining is the application of data mining techniques to find out patterns from the Web. According to investigation targets, web mining can be separated into three different types, which are Web usage mining, Web content mining and Web structure mining.

A. Web Content Mining

The initial step in data mining appears near is web content mining. It has been become complicated by the crawlers to find the well-organized data from deepness of web for extracting the required subject. Scanning the text also graphics so as to discover the applicable data is what

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

all about content mining. To take out from the strength of web we need habitual tools for query able databases. The first step is structure mining that helps to cluster the web pages also more to take the scanning resulting into answers to the advanced level of application. therefore on the deep web it is promising to find out the results precedence to the search engines according to the highest application of keywords .It has also in use the benefit of semi-structured character of web page, texted consist of the most functional prototype where text is followed by choice list with numeric values. Web content mining requires appearance of new and inventive applications and should also have its own individuality.

A.1. Web Text Mining

Information rescue systems and text handing out systems have been developed are pretty complicated and can regain documents by specifying attributes or key words. On the other hand, in order to be capable to describe at least the meaning of a word within a specified document, generally a vector demonstration is used, where for each word a numerical "meaning" value is stored. The principal approaches based on this proposal are the vector space model [4], the probabilistic model [5] and the logical model [6].

Text mining or else text data mining is the process of finding constructive or attractive patterns, models, guidelines, movements, or rule from unstructured text, is used to explain the purpose of data mining techniques to automated innovation of understanding from text [7]. Text mining has been outlook as a normal extension of data mining [8], sometimes deliberate as a task of applying an equivalent data mining techniques to exact domain [9]. This replicate the fact that starts of text mining relies on the growing field of data mining to a huge degree.

A.2. Web Image Mining

Image mining is an idea used to notice extraordinary patterns and pull out inherent and functional data from images stored in the huge data bases. Thus, we can declare that image mining deals with manufacture associations among different images from huge image databases. Image mining is used in range of fields similar to medical analysis, crop growing, remote sensing, industries, space examine, and also managing hyper phantom images. [2]

A.3. Web Video Mining

Mining video data is yet additional difficult than mining image data. One can observe video to be a group of moving images, much similar to animation. The main areas contain developing query and retrieval techniques for video databases, as well as video indexing, query languages, and optimization approaches.

B. Web Structure Mining

The objective of Web structure mining is to make structural review about the Web site and Web page. Exactly, Web content mining primarily focuses on the structure of internal-document, while Web structure mining try to determine the link structure of hyperlinks at the inter-document stage. Based on the topology of hyperlinks, Web structure mining will classify the Web pages and produce the useful information, for example the similarity and relationship between dissimilar Web sites.

B.1. Web Hyperlink Mining

Web structure mining generally functions on the hyperlink structure of Web pages. Mining focuses on sets of pages, ranging as of a single Web site to the Web as a whole. Web structure mining develops the additional information that is included in the structure of hypertext. Thus, a key application area is the identification of the comparative significance of different pages that show equally applicable when analysed with high opinion to their content in separation.

C. Web Usage Mining

Web usage mining is the innovation of user access prototypes from web server logs, which preserve an explanation of each user browsing activities. Web servers automatically create huge data stored in sever referred as logs containing information regarding the user profile, access model for pages, etc. This can provide information that can be used for resourceful and successful web site administration and the user behaviour. Web usage mining is the method of finding out what users are appearing on the Internet. Several users might be looking at just textual data, but some others might be concerned in multimedia information.

III. WEB CONTENT MINING TOOLS

A. Web Content Extractor

"Web Content Extractor**"** is software designed for web scraping, data mining, and data extraction. Web Content Extractor will permit users to mine the target data from a range of WebPages over the Internet.

Web Content Extractor can collect data from online stores, company directories, e-commerce web sites, economic web sites, shopping web sites, search engine outcomes, everything you can imagine that is going on the World Wide Web.

Web content extractor permits you to export the mined data keen on Excel (CSV), Text (ASCII), and HTML also Microsoft Access, My SQL database.

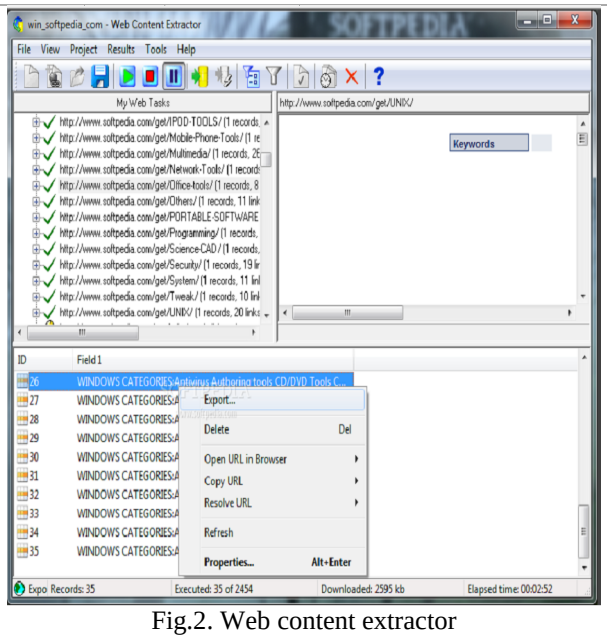


Fig.2. Web content extractor

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Characteristics of web content extractor

· Patterned web information extraction. Extract any Objective data (text and images) from a range of web pages on the Internet;
· Export the extracted data into Access, Excel, Text, HTML, XML, SQL script and My SQL files;
· Personalized Web crawler / web spider. Crawling regulations and multithreaded downloading;
· Assemble data from password sheltered websites.
· Uncomplicated to use design wizard;
· Very easy to use, fast learning arc and exact to the point.

B. Web Info Extractor

"Web Information Extractor**"** is a very powerful tool used for web data mining and content mining, content investigation. It is able to extract structure or unstructured data from web page, alteration into local file or save to database, place to web server. No need to define difficult template rules, immediately browse to the web page you are interesting and hit it off what you wish for define the extraction job, and run it when you want, or allow it run automatically.

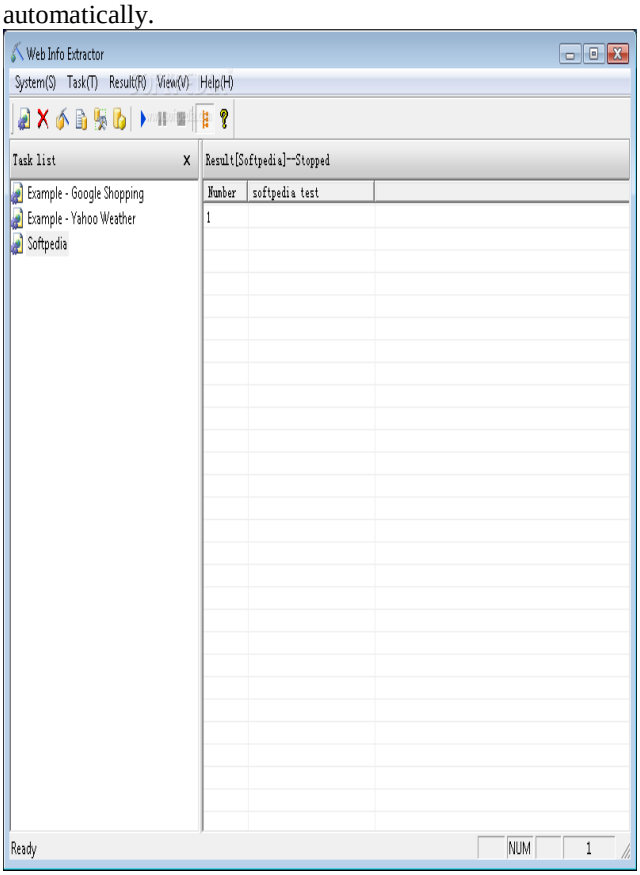


Fig.3. Web Info Extractor

Features

- Simple to classify extraction job, no require learning difficult and dull template rule.
- Mine tabular in addition to unstructured data to file, database.
- Check web pages and extract fresh content when update.
- It can transaction with text, image as well as additional link file.
- Unicode support can compact with web page in all language.
- Carry recursive task (child task) classification.
- Running multiple task at a time.

C. Web Text Extractor

"Web Text Extractor**"** is plan for extract text from web page and still control label in dialog simply. You can pull out and copy these texts with no select them**.**

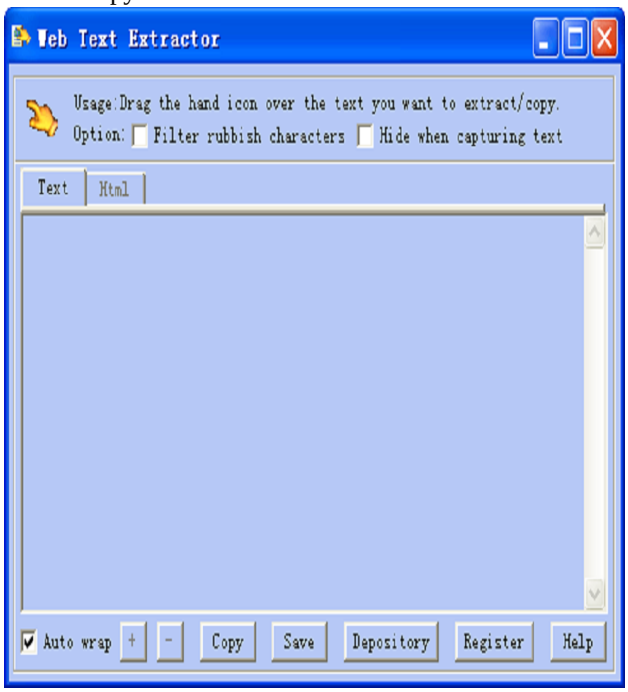


Fig.4. Web Text Extractor

Features

- Mine text from web page, No text selection desired, Can mine unselect able text.
- Sort out transparent character as well as zero size character automatically.
- Mine text from still control, edit control and windows title.
- Handle extracted text for you.

***D.** Screen Scraper*

Software that permits a PC to catch character-based data from a mainframe repeatedly presented in a green screen and it in an easier to recognize graphical user interface. Latest screen scrapers provide the information in HTML, thus it be able to access with a browser. Top producers include Mozart, Flashpoint, Inc, and Intelligent Environments.

An in-built recorder presents only click screen scraping.

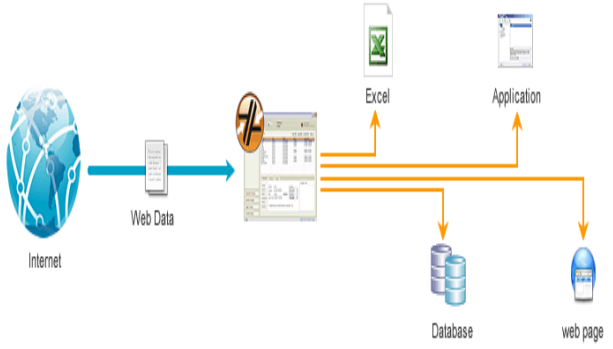


Fig.5. Screen Scraper

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Automation anywhere presents the utilization of its editor through a point-&-click wizard supported tool that can help out you to computerize general screen scraping jobs in minutes.

E. Mozenda

Mozenda is software that permits commercial and non-technical users to simply mine data across web pages. Mozenda now supports logins, paging throughout lists of results, AJAX, frames, with other difficult web sites. Mined data can be accessed online, exported, as well as used throughout an API.

Mozenda Data Extractor is an excellent tool that performs your scraper within the clouds. The circulated character of this web ripper works glowing for large amount scraping and listed and parallel web crop. Mozenda's service used for choosing items as well as appending harvest files fits well for grouping of data from various sources. The out coming export and distributing services (including email notifications) are wonderful characteristics of this screen scraper.

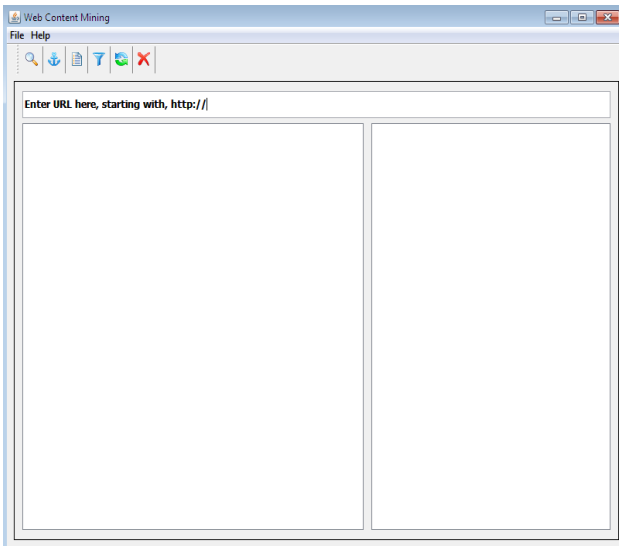
Fig.6. Mozenda

IV. HIT INTRODUCTION

HIT is a java based desktop application developed by us for Hyperlink, Image and Text mining against the given URL's in very easy way. It is very user friendly tool for web mining and covers the web content mining, structure mining and also gives the source code of the given URL's page.

HIT is the client side mining tool and mine web information on user side. It cannot access server side information from example site database or clusters.

Using this tool use can access list of hyperlink, list of images available in web page currently and we can also count any particular word used in web page. Fig.7.1. HIT Main Window



Fi.g.7.1 Hit Main Window

V. HIT FUNCTIONALITY

A. How to access web page source code?

Accessing web page source code is so easy using HIT. Just enter the web page URL With http:// and click on "**Load source Code Button**". It will take some time and access time will depend on the code length of web page. Software captures the source code of the web page and shows the source code in HIT tool window. HIT provide all HTML, CSS & Client Side Scripting code of given URL page.

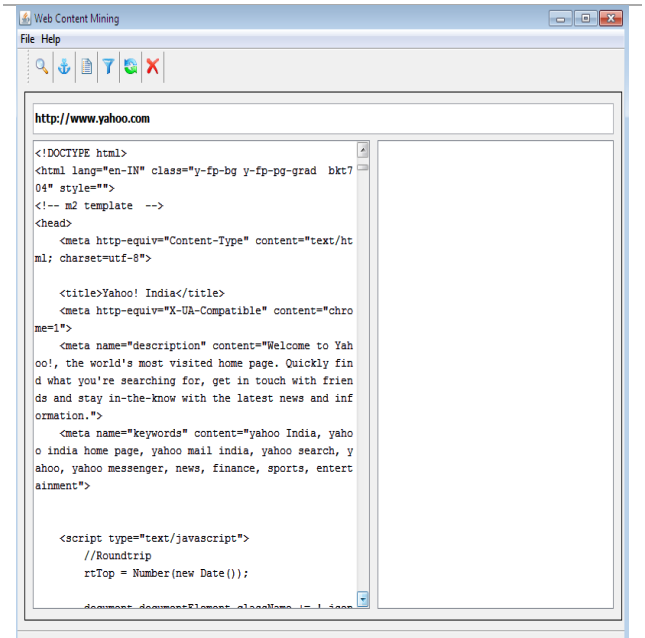


Fig.7.2

***B**. How to mine web page's hyperlinks?*

Put your web page URL With http:// which hyperlink list you want to mine and click on "**Load Hyperlinks Button**". HIT will display all available hyperlinks in right pan window. Result of this operation is URL of all hyperlinks in the form of full URL.

International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 6, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

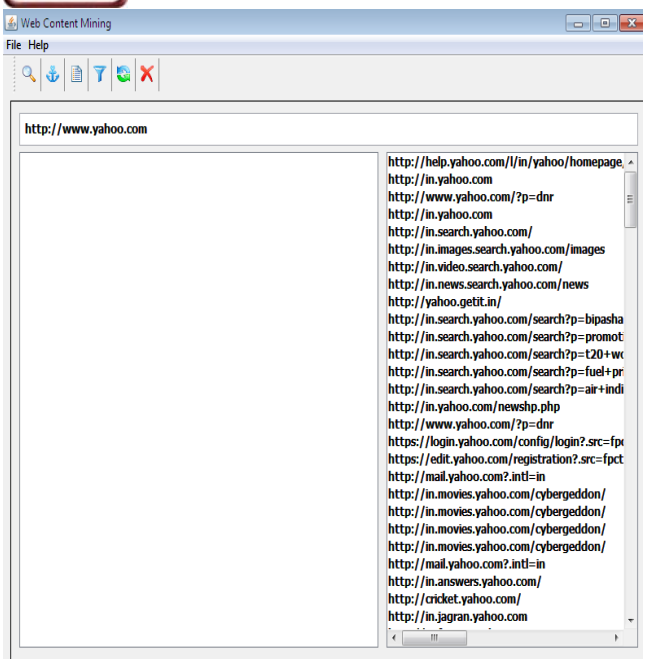


Fig.7.3

C. How to mine image list of web page?

Enter the web page URL must be with http:// and click on "**Load Images Button**". HIT software provides the list of all images name with related information, available in web page currently and will display in list view which is the part of HIT software window. Image mining result will be Image name with extension of images and root of images.

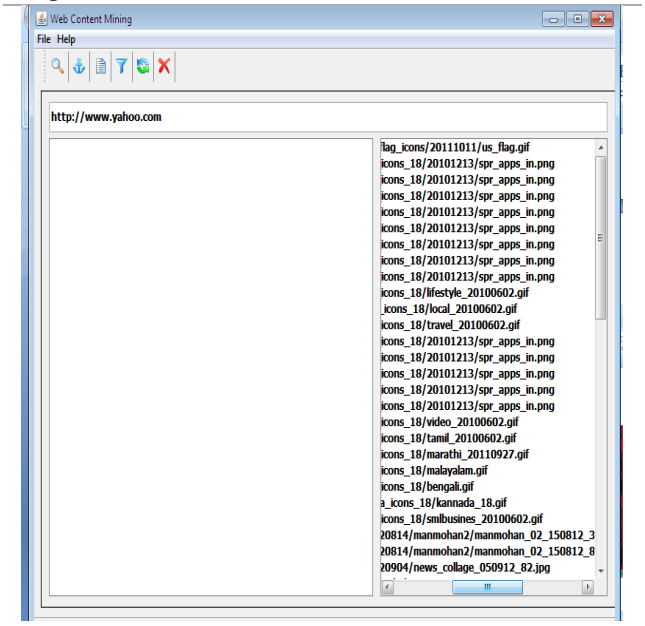


Fig.7.3

D. How to count any text available in web page?

First of all get source code of the Given URL using clicking on "**Load source Code Button**" and after getting source code click on "**Count Text Button**", after clicking then "Count Text Button" an input dialog box will appear and enter your text which you want to count that how many times it is available in that entered URL web page.

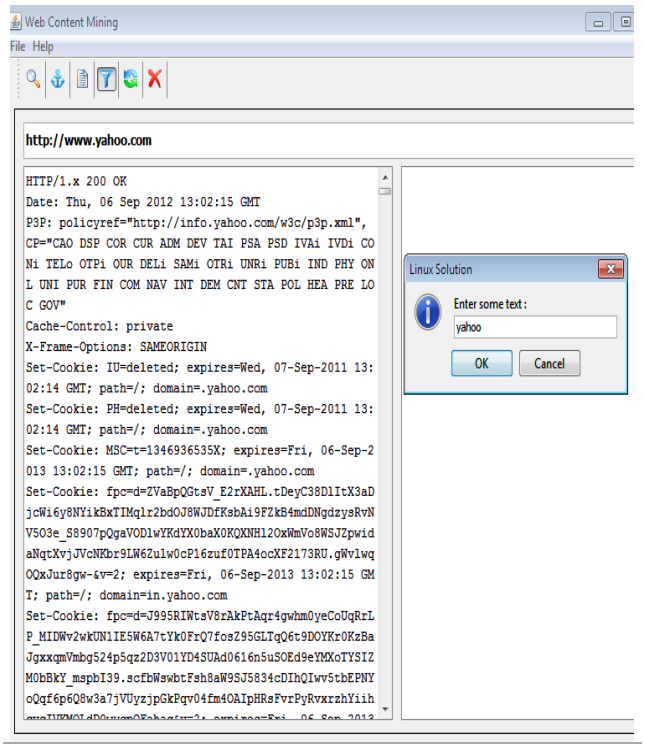


Fig.7.4

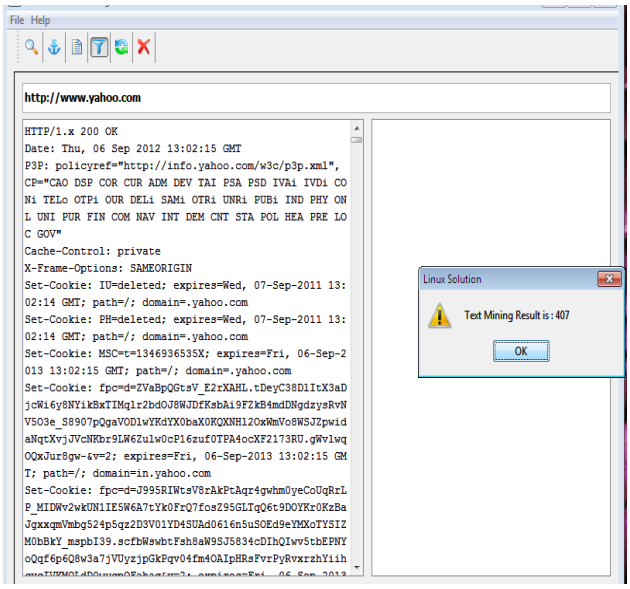


Fig.7.5

VI. Comparative Study of Web Content Mining Tools

Table shows the web content mining tools and the tasks these tools perform.

Tools and their Respective Tasks

In flowing table we present some popular web mining tools and there comparison according to some key points of web mining and also discuss that tool is user friendly on not.

Name of Tool	Tasks			
	Records the Data	Extract Structured data	Extract Unstructured data	User Friendly
Automation Anywhere	Yes	Yes	Yes	Yes
Web Info Extractor	No	Yes	Yes	Yes
Web Content Extractor	No	Yes	Yes	Not for Unstructured data
Screen Scraper	No	Yes	Yes	No
Mozenda	No	Yes	Yes	Yes
HIT	No	Yes	Yes	Yes

Fig.8

Tools and their Price and Platforms

In flowing table we present some popular web mining tools along with their costs that determine that they are compatible with the corresponding operating system. Although windows and linux are the popular O.S. in the market as shown by the statics and HIT supports both of them.

Name of Tool	Price	OS
Automation Anywhere(Server)	$7,000.00	Windows 7 / Vista / XP / 2003 / 2000 / Windows 2008 server
Automation Anywhere(Premier)	$2,495.00	
Automation Anywhere(Standard)	$995.00	
Web Content Extractor	$99	Win
Mozenda	$99 per 5000 pages	Win
HIT	Freeware	Win/Linux

Fig.9

VII. Analysis of Result

Automation Anywhere, Web Info Extractor, Web Content Extractor, Screen Scraper, Mozenda and many more are the latest similar techniques used for web content mining but all of these tools are available with cost and operating system based.HIT is java based tool because of this, it can work on any operation system and any platform. HIT is more simple and user friendly as compare to others. This tool provides very clear and accurate results.HIT has some limitations like it can mine only Image, hyperlink, count the Text and provide the source code of given URL but Video mining, XML mining, semantic web mining ,etc are also the part of web mining.

VIII. Conclusion

This paper talk abuts the methods of web content mining. Web content mining has been proved extremely functional in the business world. The survey moreover talks about the techniques used for mining information as of different natures of data existing in the internet as well as how this mined data can be used for mining functions. Mortally users face some kind of difficulty in getting required information and deciding which information is related to them from common purpose search engines. Web content mining resolves this trouble and facilitates the users to fulfil their requirements. Subject trailing is helpful in guessing the web content associated to user's significance. Summarization assists the user to make a decision whether they ought to read an exacting subject matter or not. Classification is able to use in business as well as industries to supply client support. Web Content mining is also used for distance learning which can be achieved by appalling to business application like mining online sites. Content Mining assists to launch better association with customer by providing accurately what they want. Towards the end paper talk about various tools that provide the web content mining facility as well as about HIT , that is a web content mining tool develop by us and it's working.

IX. Proposed Work

In the future we can develop autonomous agents. These agents analyze the rules that are already discovered and Provide meaningful courses or suggestions to the users.
Future scope of the "Web Content Mining "contains forecasting user needs to improve framing, extendibility, usability, user preservation for web security with the help of web log files.

To manipulate web content for analysis and synthesis by automated system Semantic web is a future revelation.
Internet displays information in human readable format. It is very difficult for the user of internet to understand and interpret HTML and its content. Content explanation, assortment and management are three complicated tasks in the internet. Now a day's humans have to manage these tasks. The stability between human and machine by reducing the above mentioned three tasks can solve by using semantic web and make those tasks automatic.

References

[1] Nasraoui O., Petenes C., "Combining Web Usage Mining and Fuzzy Inference for Website Personalization", In Proc. of Webkdd 2003 – Kdd Workshop On Web Mining As A Premise

to Effective And Intelligent Web Applications, Washington DC, August 2003, P. 37
[2] C. Lakshmi Devasena Et AL. / International Journal on Computer Science and Engineering (IJCSE), ISSN : 0975-3397 , 3 Mar 2011,Vol. 3 No. 1155-1167.
[3] S. Chakrabarti, Mining Theweb, Morgan Kaufmann, San Francisco, CA, 2003.
[4] G. Salton, A. Wong, and C. S. Yang. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11):613–620, (See Also [36] Tr74-218, Cornell University, NY, USA). , 1975
[5] S. E. Robertson. The Probability Ranking Principle. Journal of Documentation, 33:294–304, 1977.
[6] C. J. Van Rijsbergen. A Non-Classical Logic For Information Retrieval. The Computer Journal, 1986, 29(6):481–485.
[7] Chakrabarti S. "Mining the Web: Analysis Of Hypertext And Semi Structured Data", Morgan Kaufmann, San Francisco, CA.
[8] Feldman, R., and Sanger, the Text Mining Handbook. New York: Cambridge University Press, J. (2006). ISBN 978-0-521-83657-9
[9] Dorre J., Gerst P., Seiffert R., "Text Mining: Finding Nuggets In Mountains Of Texture Data", In Proceeding Of 5th International Conf. On KDD-99, PP. 398-401, San Diego, CA, ACM, Short Paper.
[10] Dunham, Data Mining Introductory And Advanced Topics. Pearson Education, M. H. 2003.
[11] Man, L. (2002). Hypertext, Information Retrieval And Data Mining: Behind Link There Is Golden Mine., From http://eww.scholars.nus.edu.sg/cpace/ht/lanman/wm1.htm , Retrieved November 6, 2006
[12] Ding, C., He, X., Husbands, P., Zha, H., & Simon, H. (2001). Pagerank, Hits and a Unified Framework for Link Analysis. Lbnl Tech Report 49372, (Updated Sep 2002).
[13] Cooley, R., Srivastava, J., Mobasher, B.: Web Mining: Information and Pattern Dis-Covery On The World Wide Web. In: Proc. Of The 9th IEEE International Conferenceon Tools With Arti¯Cial Intellegene (ICTAI'97). (1997)
[14] International Journal Of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
[15] Kosla, R. And Blockeel. Web Mining Research: A Survey. Sig Kdd Explorations. , H. 2000, Vol. 2, 1-15.
[16] Torreblanca, A. M., Gomez, M. M. And Lopez, a Trend Discovery System For Dynamic Web Content Mining. Proceedings of the 11th International Conference On Computing, A. L. 2002.
[17] Yang, C. Y., Hsu, H. H. And Hung, A Web Content Suggestion System for Distance Learning. Tamkang Journal of Science and Engineering. , J. C. 2006, Vol. 9, No. 3, 243-254.
[18] [Kosla, R. And Blockeel, Web Mining Research: A Survey. Sig Kdd Explorations. , H. 2000, Vol. 2, 1-15.
[19] Ajoudanian, S. and Jazi M. D., Deep Web Content Mining. World Academy of Science, Engineering and Technology 49, 2009.
[20] Liu, B. And Chiang K. C., Editorial Special Issue on Web Content Mining. Acm. Journal of Machine Learning Research 4, 2004, 177-210.
[21] Singh, B. And Singh, Web Data Mining Research: A Survey. Computational Intelligence and Computing Research (ICCIC).IEEE International Conference, H. K. 2010, 1-10.
[22] Guo, J., Keselj, V. And Gao, Integrating Web Content Clustering Into Web Log Association Rule Mining. Springer Verlag., Q. 2005, Vol. 3501 LNAI, 182-193.
[23] Kazienko, P. And Kiewra, Link Recommendation Method Based On Web Content And Usage Mining. New Trends In Intelligent Information Processing And Web Mining Proc. Of The International IIS: IIPWM '03 Conference. Advances In Soft Computing, Springer Verlag, , M. 2003, 529-534.
[24] Taherizadeh, S. And Moghadam, Integrating Web Content Mining Into Web Usage Mining For Finding Patterns And Predicting User's Behaviors. International Journal Of Information Science And Management, N. 2009, Vol. 7, No. 1.
[25] Gedov, V., Stolz, C., Neuneir, R., Skubacz, M. And Siepel, Matching Web Site Structure Andcontent. Acm. Proceedings Of The 13th International World Wide Web Conference On Alternate Track Papers And Posters, D. 2004.
[26] Pokorny, J. And Smigansky, Page Content Rank: An Approach to the Web Content Mining. In Proceedings of Iadis International Conference Applied Computing. Algarve, Portugal, J. 2005.
[27] Poonkuzhali, G., Thiagarajan, K., Sarukesi, K. And Uma G. V. Signed Approach for Mining Web Content Outliers. World Academy of Science, Engineering and Technology, 2009, 56.
[28] Ahmed, S. S., Halim, Z., Blaig, R. And Bashir, Web Content Mining: A Solution to Consumers Product Hunt. International Journal Of Social And Human Sciences 2, S. 2008 , 6-11.
[29] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. in proceedings of the Acm Sigmod Conference on Management of Data, Pages 207-216, Washington, D. C., May 1993.

AUTHOR'S PROFILE

Tripurari Pujan Pratap Singh

2-1-1984. He is the student of *M. Phil (Computer Science) AISECT University Bhopal, (M. P) India.* Before joining M.Phil he had done PGDCA from MCRPV, Bhopal, India, M. Sc. (IT) from Punjab Technical University Jalandhar India and MCA from Punjab Technical University Jalandhar India.

Anurag Seetha

a Senior Member of IEEE, Computer Society, is presently working as Professor & Dean, Computer Science & Engineering, Dr. C.V. Raman University, Bilaspur (CG), India.
He received his Ph.D. (Computer Applications) degree from Rajiv Gandhi Technical University, Bhopal, India in 2008. Before joining Dr. C.V.Raman University, he was served as Reader, Computer Science & Applications in MCRPV, Bhopal, India. He has been actively involved in the field of Technical Education as Academician, Researcher, Teacher, Planner and Administrator. He has guided several M.Tech./MCA/M.Phil students in Computer Science & Engineering. His research areas are information retrieval, e-learning, educational technology and Web technology.

Krishna Kumar Pandey

is an assistant professor at *AISECT University Bhopal, (M. P) India.* India. He received his B.Tech form aiet UPTU Lucknow and M.Tech from SOIT RGPV Bhopal.