Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Analysis of Web Content Mining Tools

M.Karpagam *
Assistant Professor, Department of Information Tech.
K.S.R College of Technology
Tiruchengode, India

R.Sasikala
Professor and Head, Department of Information Tech.
K.S.R College of Technology
Tiruchengode, India

Abstract— Web Mining is a process of accessing data publicly. The Web mining is used to get information from structured, unstructured and semi structured form of web pages, which includes fetching information from wide database. The search engine is a very important tool for people to obtain information on Internet, but the low-precision and low-recall exist widely in current search engines [1]. Web mining works on On-Line Data which is stored in server database and web log. This paper compares diverse mining tools as well as archetype of web content mining. This analyzes and focuses on how to relate content mining concept with on-going hyperlink, image and text.

Keywords— Web Mining, Content Extractor, Mining Tools, WISE, Information extractor

I. INTRODUCTION

Web mining - is the appliance of data mining techniques to discover patterns from the Web. Web mining helps to provide solution in discovering the way that users using Web sites. It involves log analysis and the steps that typically have to be gone through to get meaningful data from Web logs - data collection, pre-processing, data enrichment and pattern analysis and discovery. Escalation of web data have created many complicated situations like extracting the most suitable and relevant information as per requirement. World Wide Web is the interactive technique to distribute information today. Web Mining depends on knowledge discovery from web. It is the mining of the knowledge framework which represents in a proper way. Web mining is used to extract the information, image, text, audio, video, documents and multimedia. Web mining can easily extract all features and information about multimedia .Searching a topic from web is difficult to get accurate topic information but Now a days it is easy to get the proper and relevant information due to web mining. Web data mining is based on data mining technique which discovers the hidden data in web log.

Mining data records in Web pages is useful because they typically present their host pages' essential information, such as lists of products and services. Extracting these structured data objects enables one to integrate data/information from multiple Web pages to provide value-added services, e.g., comparative shopping, meta-querying and search.

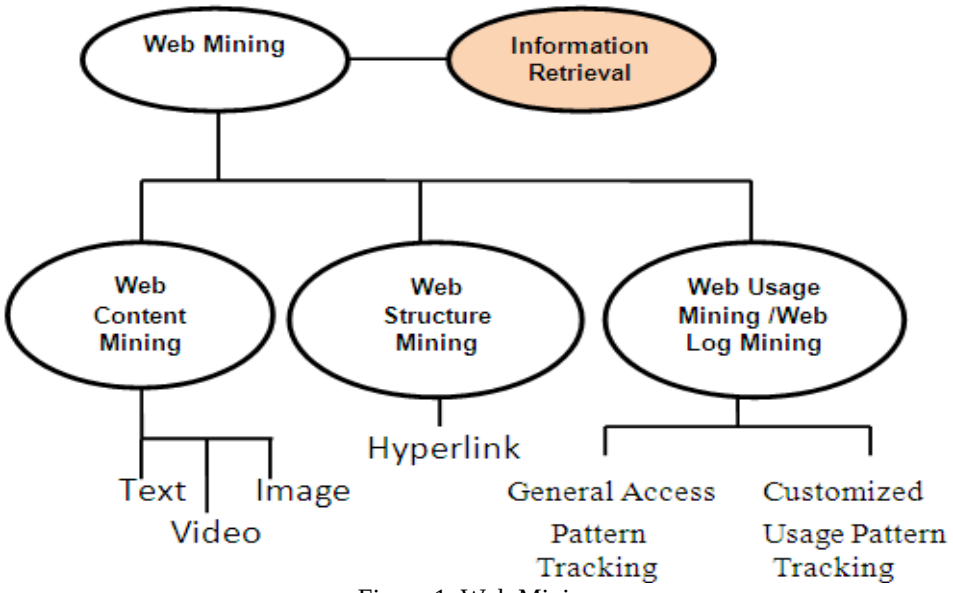

Figure 1. Web Mining

Various data mining methods are used to discover the hidden and useful information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. New approaches should be used which better fit the

properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area[1].The rest of the paper is structured as follows: Section 2 summaries the various categories of web mining. Section 3 provides a general description of the tools and software under test. Section 4 compares the results of the different software and tools used. Ultimately, we close this paper with a conclusion.

II. WEBMINING CATEGORIES

Web Usage mining

Web log mining is the progression of extracting valuable information from server logs. It is used to analyze the activities of website users. It is the relevance of data mining techniques to discover interesting usage patterns from Web data in order to recognize and better serve the needs of Web-based applications[2]. Traditional data captures the identity or origin of Web users along with their browsing behavior at a Web site. Web usage mining itself can be classified depending on the kind of convention data considered[3]:

- Web Server Data: The user logs are serene by the Web server. Distinctive data includes IP address, page reference and access time.
- Application Server Data: A key aspect is the facility to trail various kinds of business events and log them in application server logs.
- Application Level Data: Recent events can be providing in an application, and logging can be twisted on from generating histories of these events.

Web Content Mining

Web content mining is the procedure of extract and integrates of valuable data, information and knowledge from Web page content[4]. Web content mining is differentiating from two diverse points of view: Information Retrieval View and Database View.

Scanning the text and also graphics so as to discern the pertinent data is all about web content mining. Web content mining requires facade of inventive applications and should also have its own eccentricity. For example, we can automatically classify and cluster Web pages according to their topics[2].

Web Text Mining

Liberating information systems and text handing out systems which have been developed are pretty complicated and regain documents by mentioning attributes or key words[5]. Text mining or else text data mining is the procedure to search productive or impressive patterns, models, strategy, schedule, or statute from unstructured text, is used to explain the purpose of data mining techniques to automated improvement of understanding from text .Text mining has been outlook as a normal extension of data mining, sometimes deliberate as a task of applying an equivalent data mining techniques to exact domain .This reproduce the fact that starts of text mining relies on the growing field of data mining to a huge degree.

Web Image Mining

Image mining is an inspiration to recognize extraordinary patterns and pull out inherent and functional data from images stored in the huge data bases[6]. Thus, proclaim 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.

Web Video Mining

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

Web Structure mining

Web structure mining is the procedure of using graph theory to examine the node and connection structure of a web site. Based on the category of web structural data. The web pages are represented as nodes and Hyperlinks as edges. Basically it defines the relationship between user and web[5]. The intention of web structure mining is to generate structured summaries about information on web pages. It shows the link from one web page to another web page.

III. WEB CONTENT MINING TOOLS

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[8]. 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.

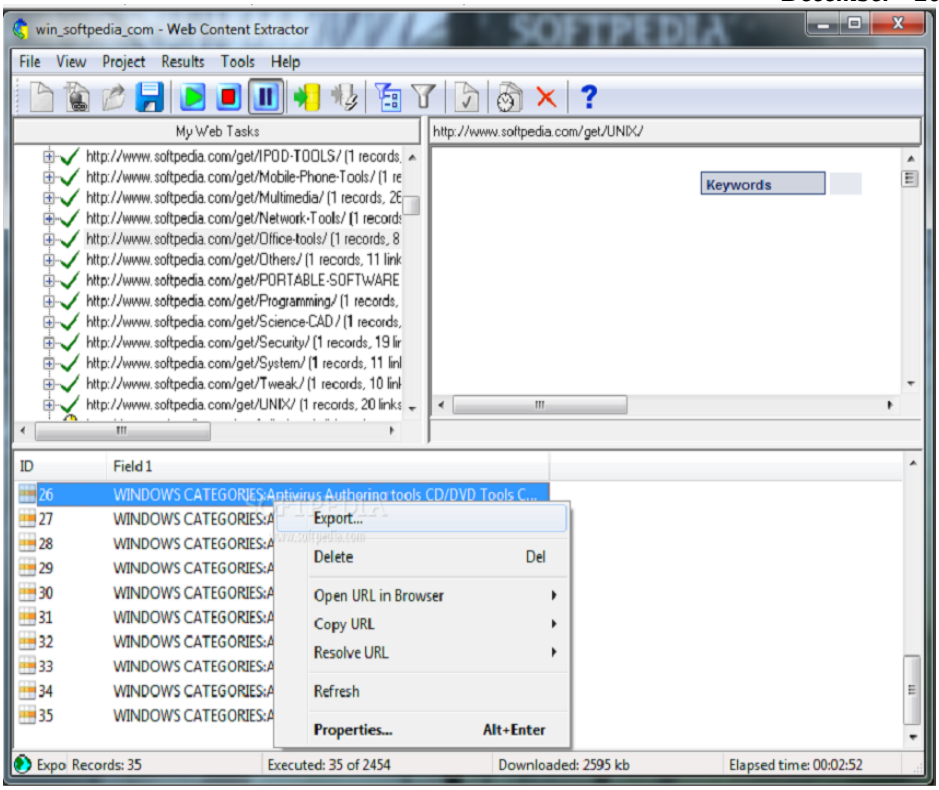

Figure 2. Web Content Extractor

Characteristics of web content extractor

Attractive 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.

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[7]. 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.

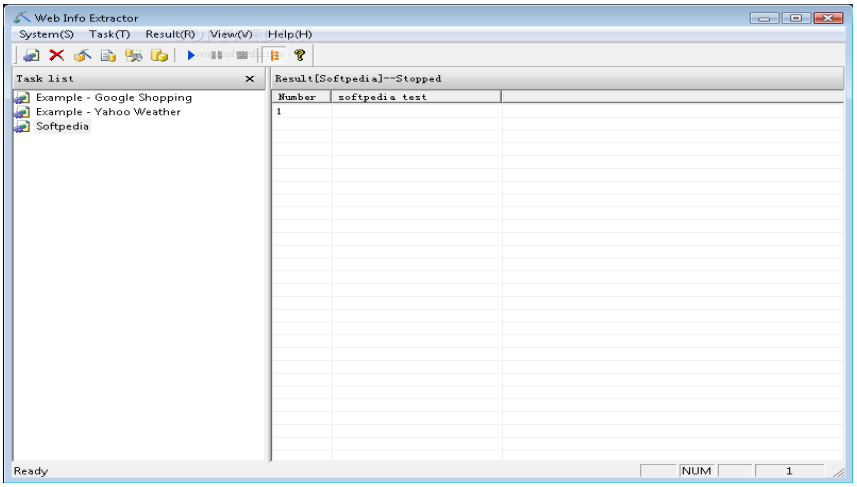

Figure 3. Web Info Extractor

Web Text Extractor

"Web Text Extractor" is plan for extract text from web page and still control label in dialog simply. This tool allows pulling out and copying these texts without selecting them[7].

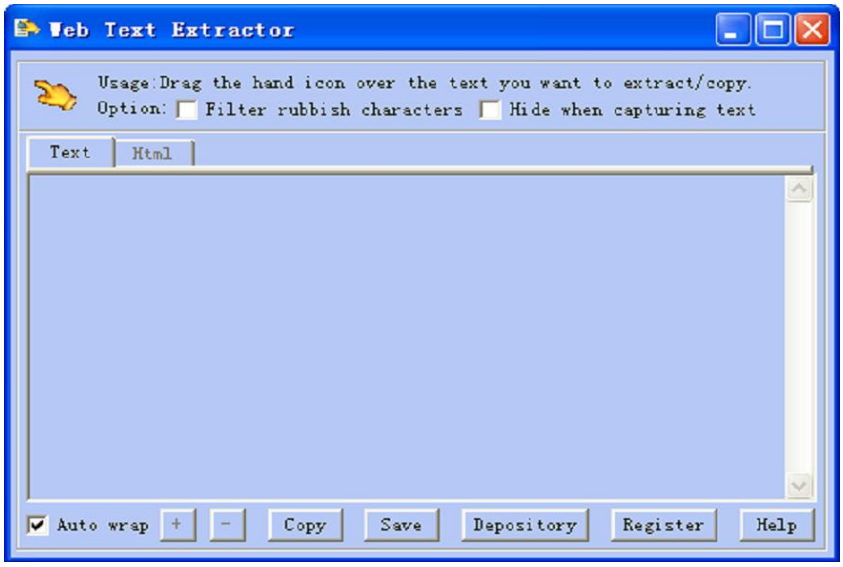

Figure 4. Web Text Extractor

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[7]. Top producers include Mozart, Flashpoint, Inc, and Intelligent Environments. An in-built recorder presents only click screen scraping.

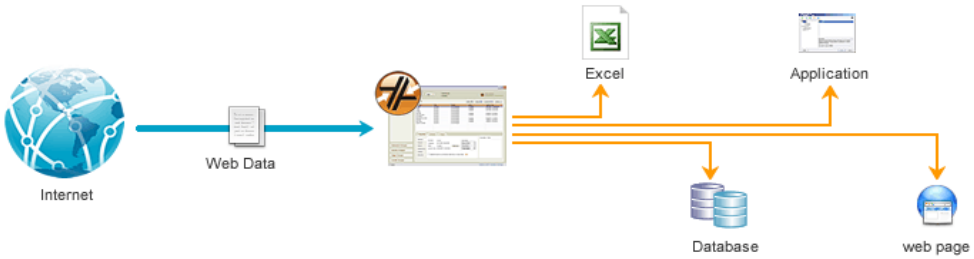

Figure 5. Screen Scrapper

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

Mozenda

Mozenda is software that permits commercial and nontechnical 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[7]. 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.

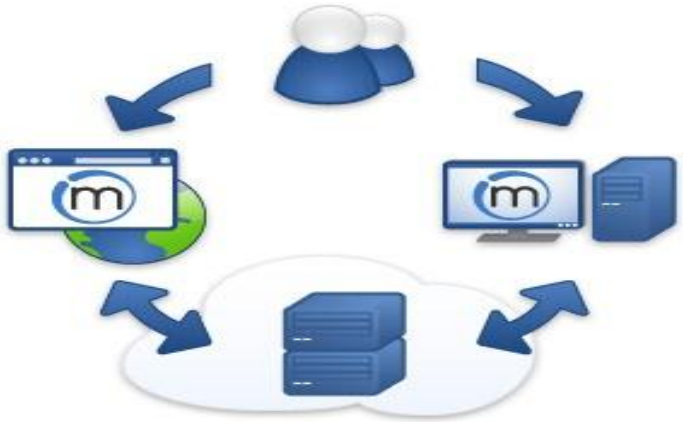

Figure 6. Mozenda

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.

HIT

HIT is a java based desktop application developed 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 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[7]. 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 up any particular word used in web page.

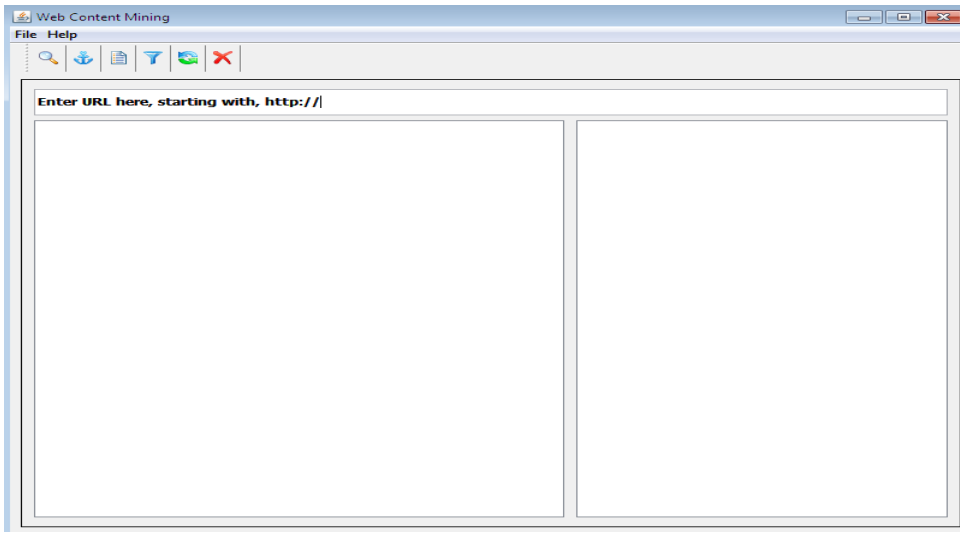

Figure 7. HIT Main Window

Using this tool use can access list of hyperlink, list of images available in web page currently and also used to count any particular word used in web page[7].

WISE-Integrator

WISE-Integrator is developed using JDK1.4 and is now operational [8]. To utilize WordNet, WordNet1.6 is embedded into the system through APIs based on the C language. The system has two components, one for search interface extraction and the other for interface integration. The search interface extraction component is implemented by WISE-iExtractor which can be used alone or embedded into WISE Integrator as a sub-system. WISE-iExtractor takes as input one or more HTML pages containing search interfaces of ESEs. In general, a Web page might contain multiple search forms for different products .Our system requires that all the search forms on a Web page be in the same product domain. After each input interface is extracted, the extractor shows its extracted search interface.

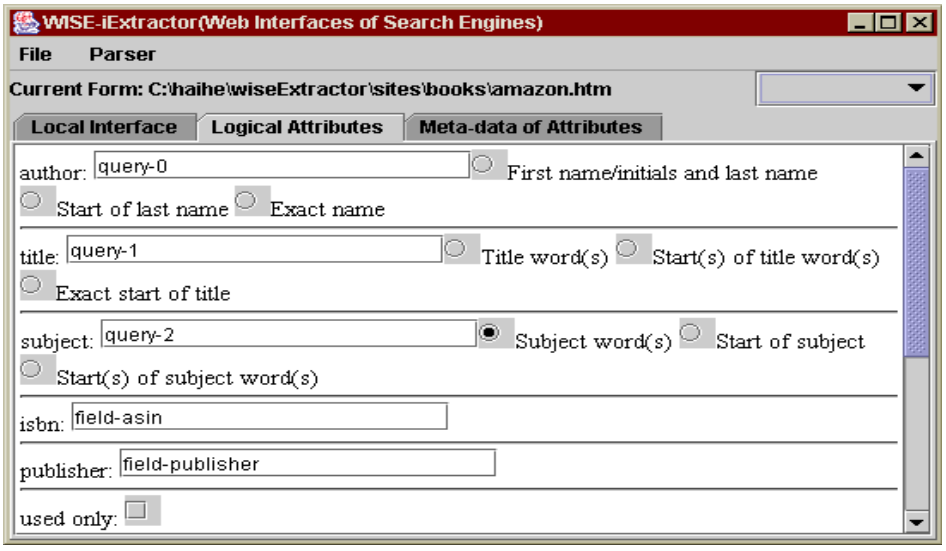

Figure 8(a). A screen-shot of the interface integration component of WISE-Integrator

A special feature of WISE-Integrator is that users can remove an existing interface from or add a new interface to the existing unified interface at any time on the fly, and WISE-Integrator will generate the new unified interface without starting from scratch (i.e., incremental maintenance is implemented)[8]. In addition, users can choose a parameter value to trim some less important attributes from the unified interface to make it more user-friendly.

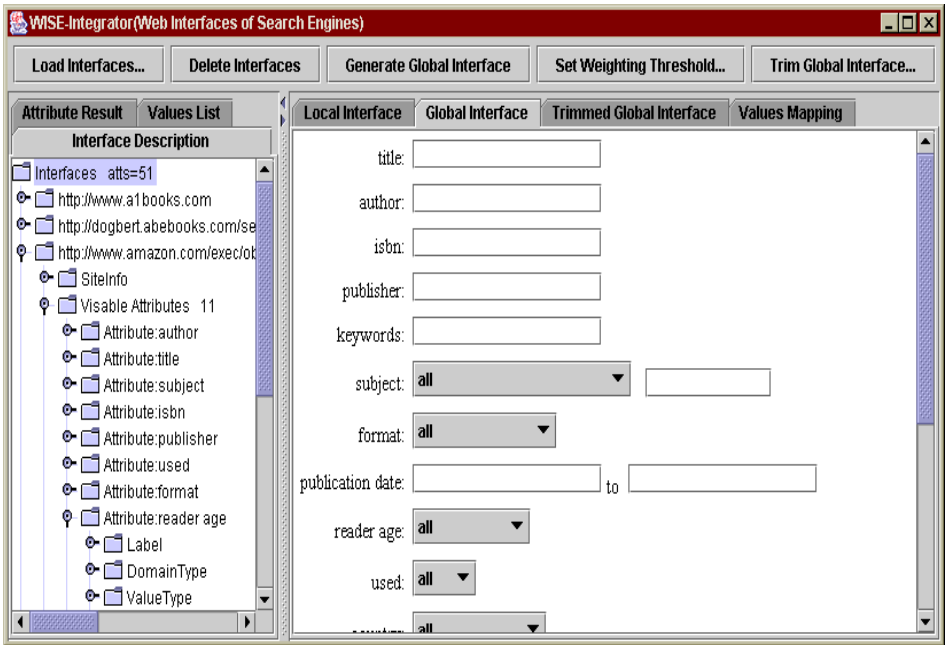

Figure 8(b). Another screen-shot of the interface integration component of WISE-Integrator

Comparative Study Of Web Content Mining Tools

Table shows the web content mining tools and the tasks these tools perform [3]. 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.

Table 1. Comparison of web mining tools

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 structured data
Screen Scraper	No	Yes	Yes	No
Mozenda	No	Yes	Yes	Yes
HIT	No	Yes	Yes	Yes
WISE Tool	Yes-meta data	Yes	Yes	Yes

Analysis of Result

Automation Anywhere, Web Info Extractor, Web Content Extractor, Screen Scraper, Mozenda ,HIT 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[8].WISE is java based tool because of this, it can work on any operation system and any platform. WISE is simpler and user friendly as compare to others. The tool WISE-Integrator explores a rich set of special meta-information that exists in Web search interfaces, and employs the information to identify matching attributes from different search interfaces for integration [8].WISE-Integrator deals with automatic interface integration. WISE-*i*Extractor is also designed to obtain/derive meta-information.

IV. CONCLUSIONS

This paper talks about the methods of web content mining. The survey focuses on the techniques used for mining information on different natures of data existing in the internet. Incurably 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. On the way to the end, this paper talks about various tools that provide the web content mining facility and additionally it includes a WISE-Integrator – tool which automatically integrates the Web Interfaces in Search Engines [8].

REFERENCES

[1] Aishwarya Rastogi, Smita Gupta, Srishti Agarwal, Nimisha Agarwal ," *WEB MINING: A COMPARATIVE STUDY*", /International Journal Of Computational Engineering Researc"h / ISSN: 2250–3005
[2] Abdelhakim Herrouz,Chabane Khentout Mahieddine Djoudi," *Overview Of Web Content Mining Tools*", The International Journal Of Engineering And Science (IJES) ||Volume||2 ||Issue|| 6 ||Pages|| 106-110||2013|| ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805
[3] http://en.wikipedia.org/wiki/Web_mining
[4] http://en.wikipedia.org/wiki/Web_content_mining
[5] Monika Yadav, Mr. Pradeep Mittal," *Web Mining: An Introduction*", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 3, March 2013.
[6] 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
[7] Tripurari Pujan Pratap Singh, Dr. Anurag Seetha, K. K. Pandey, "*HIT: Web Content Mining Tool"*, Morgan Kaufmann, International Journal of Electronics Communication and Computer Engineering (IJECCE), ISSN: 2278–4209.
[8] Hai He1, Weiyi Meng1, Clement Yu2, Zonghuan Wu3, Automatic Integration of Web Search Interfaces with WISE-Integrator.
[9] Yanhong Zhai, Bing Liu," *Web Data Extraction Based on Partial Tree Alignment*", International World Wide conference Committee.
[10] S. E. Robertson. The Probability Ranking Principle. Journal of Documentation, 33:294–304, 1977.
[11] C. J. Van Rijsbergen. "*A Non-Classical Logic For Information Retrieval"*. The Computer Journal, 1986, 29(6):481–485.
[12] Chakrabarti S. "*Mining the Web: Analysis Of Hypertext And Semi Structured Data*", Morgan Kaufmann, San Francisco, CA.
[13] Feldman, R., and Sanger, the Text Mining Handbook. New York: Cambridge University Press, J. (2006). ISBN 978-0-521-83657-9
[14] 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.
[15] Dunham, Data Mining Introductory And Advanced Topics.Pearson Education, M. H. 2003.

