ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 1, Issue 8, October 2012

Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining

Arvind Kumar Sharma[1], P.C. Gupta[2]

Abstract—**Web content mining in normal parlance is to download information available on the websites. Such a process involves tremendous stress and time-taking. To augment such a process the software related to web content mining can be used so that a computer can use this software or tools to download the essential information that one would require. It collects the appropriate and perfectly fitting information from websites that one requires. In this paper several tools for web content mining are discussed and their relative merits and demerits are mentioned.**

Index Terms- Web Data Mining, Web Content Mining Techniques, Web Content Mining Tools

I. INTRODUCTION

As websites are a key communication channel not only for companies, but also for private individuals trying to find diverse information, it is important to find ways to make the web more usable. A website is a collection of related web pages containing images, videos or other digital assets [1]. In order to, for example, understand user behaviour or results of search engines it is necessary to analyse the information available on the Web. The field that describes these tasks is called Web Mining. World Wide Web is an evolving system of interlinked files like containing audio, images, videos, and other multimedia. The term Web Data Mining is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest. The Web contains massively information and provides an access to it at any place at any time. Most of the people browsing the internet for retrieving information, but most of the time, they get lots of insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

This paper is organized into six sections. Section-2 contains web data mining, web mining category, and web mining tasks. Section-3 consists of related works. Section-4 includes web content mining and its techniques. Section-5 contains various web mining tools and their functioning. Section-6 includes conclusion while references are shown in the last section.

II. WEB DATA MINING

This technology is popular with many businesses because it allows them to learn more about two important and active areas of current research are data mining and the World Wide Web. Although data mining is a relatively new term, the technology is not. Companies have used powerful computers to sift through volumes of supermarket scanner data and analyze market research reports for years [24]. However, continuous innovations in computer processing power, disk storage, and statistical software are dramatically increasing the accuracy of analysis while driving down the cost. Analysis and discovery of useful information from World Wide Web poses a phenomenal challenge to the researchers in this area. Such type of phenomena of retrieving valuable information by adopting data mining techniques is known as Web mining. Web mining is an application of the data mining techniques to automatically discover and extract knowledge from the Web [4].

A. WEB MINING CATEGORY

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). Web mining can be categorized in the fig.1 shown below.

Manuscript received Sep 15, 2012.
Arvind Kumar Sharma*, Ph.D Scholar, Jaipur National University, Jaipur, India.*
P.C. Gupta*, Asso. Professor, Department of CSI, University of Kota, India.*

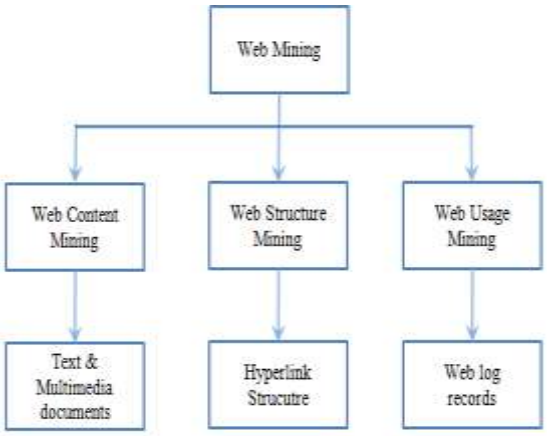

Fig.1: Web Mining Categories

Web content mining though uses data mining techniques; it differs from data mining because Web data are mostly unstructured and/or semi-structured, while data mining deals mainly with structured data. It is associated to text mining because much of the Web contents are texts. Web content mining differs from text mining because of the semi structure quality of the Web, while text mining deals with unstructured texts. Web content mining thus requires inventive applications of text mining and/or data mining techniques and also its own distinct approaches.

B. WEB MINING TASKS

Web usage mining includes the data from server access logs, user registration or profiles, user sessions or transactions, in short, mining the Web log data. Web mining consists of the different essential tasks [2], which are described in a fig.2 below.

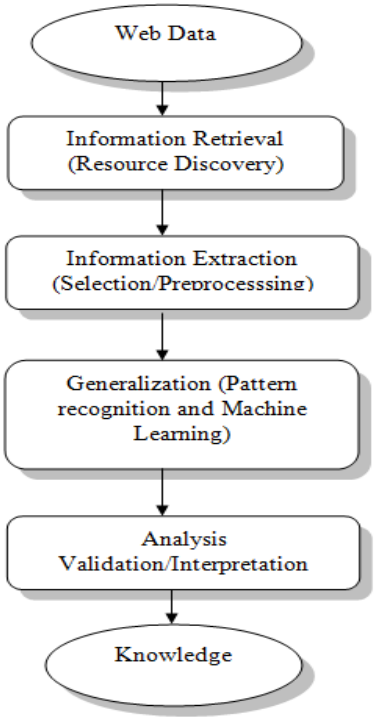

Fig.2: Web Mining Tasks

B.1 Information Retrieval

It is the task of retrieving the intended information from the Web. It locates the unfamiliar documents and services on the Web.

B.2 Pre-processing

It is the task of automatically selecting and pre-processing specific information from retrieved Web resources.

B.3 Pattern Recognition & Machine Learning

It is the task to automatically discover general patterns of individual Web sites as well as across multiple sites.

B.4 Analysis

It is the task of analyzing, validating and interpreting the mined patterns.

III. RELATED WORKS

Various scholars and researches have proposed related work in Web content mining, which are discussed below:

Aidan Finn [19] discusses in his research work "Fact or fiction: Content classification for digital libraries", methods for content extraction from "single-article" sources, where content is supposed to be in a single body. The algorithm tokenizes a page into either words or tags; the page is then sectioned into 3 contiguous regions, placing boundaries to partition the document such that most tags are placed into outside regions and word tokens into the center region. This approach works well for single-body documents, but destroys the structure of the HTML and doesn't produce good results for multi-body documents, i.e., where content is segmented into multiple smaller pieces like we find on WEB Blogs. McKeon [23] in the NLP (Natural Language Processing) group at Columbia University detects the largest body of text on a webpage (by counting the number of words) and classifies that as content. This method works well with simple pages. However, this algorithm produces noisy or inaccurate results handling multi-body documents, especially with random advertisement and image placement. Rahman [20], [23] in first International workshop on "Web Document Analysis" propose another technique that uses structural analysis, contextual analysis, and summarization. The structure of an HTML document is first analyzed and then properly decomposed into smaller subsections. The content of the individual sections is then extracted and summarized. However, this proposal has yet to be implemented. Furthermore, while the paper lays out prerequisites for content extraction, it doesn't actually propose methods to do so. Thus it again proves ineffective in actual content extraction. A variety of approaches have been suggested for formatting web pages to fit on the small screens of cellular phones and PDAs however, they basically end up only reorganizing the content of the webpage to fit on a constrained device and require a user to scroll and hunt for content. The main aim is however to device a method for the generic Web documents accessible on any device. Buyukkokten [21-22] defines "accordion summarization" as

a strategy where a page can be shrunk or expanded much like the instrument. They also discuss a method to transform a web page into a hierarchy of individual content units called Semantic Textual Units, or STUs. First, STUs are built by analyzing syntactic features of an HTML document, such as text contained within paragraph (<P>), table cell (<TD>), and frame component (<FRAME>) tags. These features are then arranged into a hierarchy based on the HTML formatting of each STU.

IV. WEB CONTENT MINING TECHNIQUES

It identifies the useful information from the web contents/data/documents, however, such a data in its broader form has to be further narrowed down to useful information. Web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. Here, the several approaches in web content mining are represented.

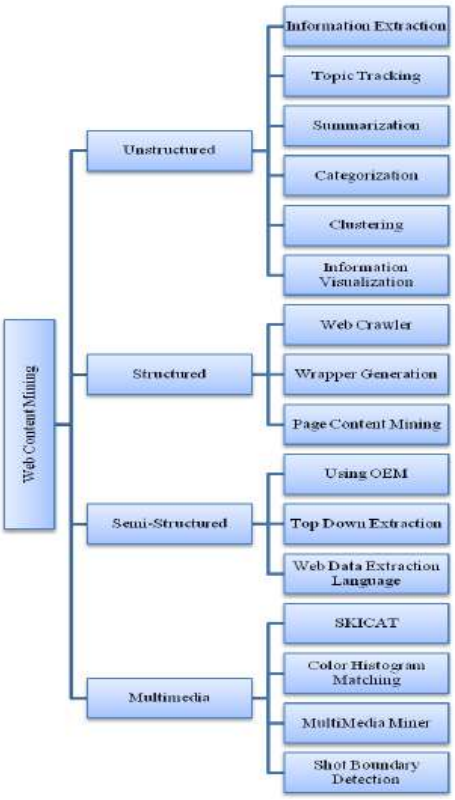

Fig.3: Taxonomy of WCM Techniques

Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

4.1 Unstructured Data Mining Techniques

Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Hence one could consider text mining as an instance of web content mining. To provide effectively exploitable results, preprocessing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques [9]. Content mining has been accomplished on unstructured data such as text. Mining of unstructured data provides unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic content mining is a type of text mining [6]. Some of the useful techniques used in text mining are as follows: Information Extraction, Information Visualization, Topic Tracking, Summarization, Categorization, and Clustering. In the following sections all these techniques are explained briefly.

4.1.1 Information Extraction

The pattern matching technique is used to extract information from unstructured data. In this case, keyword and phrases are traces out and then connections with the keywords are found within the text. This technique is very useful when there is large volume of text. Information Extraction is the basis of many other techniques used for unstructured mining [8]. It can be provided to Knowledge Discovery in Databases (KDD) module because information extraction has to transform unstructured text to more structured data. Firstly the information is mined from the extracted data and then using different types of rules, the missed out information are found out. Information Extraction that makes incorrect predictions on data is discarded [12].

4.1.2 Information Visualization

It utilizes feature extraction and key term indexing to build a graphical representation. The documents having similarity are determined using Information Visualization [12]. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is very useful to find out related topic from a very large amount of documents [8].

4.1.3 Topic Tracking

This technique checks the documents viewed by the user and studies the user profiles. According to each user it predicts the other documents related to users interest. In Topic Tracking applied by Yahoo, user can give a keyword and if anything related to the keyword pops up then it would be informed to the user. Same can be applied in the case of mining unstructured data. An example for topic tracking is that if we select the competitors name then if at any time their name will come up in the news and this information will be passed to the company. Topic tracking can be applied in different areas. Two such areas are medical field and education field. In medical field doctors can easily come to

know latest treatments. In education field topic tracking can be used to find out the latest reference for research related work. Topic tracking helps to track all subsequent stories in the news stream. The demerit of this technique is that when we search for topics we may be provided with information which is not related to our interest. For example, if user sets an alert for *Web Mining* it can provide us with topics related to mineral mining which are not useful for the users [12].

4.1.4 Summarization

It has been used to reduce the length of the document by maintaining the main points. It assists the user to decide whether they should read this topic or not. The time taken by the technique to summarize the document is less than the time taken by the user to read the first paragraph. The challenge in summarization is to teach software to analyze semantics and to interpret the meaning. This software statistically weighs the sentence and then extracts important sentences from the document. To understand the important key points, summarization tool search for headings and sub headings to find out the important points of that document. This tool also give the freedom to the user to select how much percentage of the total text they want extracted as summary. It can work along with other tools such as Topic tracking and Categorization to summarize the document. An example for text Summarization is Micro Soft Word's Auto Summarize [8].

4.1.5 Categorization

This technique is used to identify main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does not process the actual information. It decides the main topic from the counts[9]. It gives rank to the document according to the topics. Documents having majority content on a particular topic are ranked first. This technique has been used in business and industries to provide customer support [8].

4.1.6 Clustering

This technique has been used to group similar documents. Here in clustering, grouping is not done based on predefined topics. It is done based on fly. Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering technique helps the user to easily select the topic of interest. Clustering technology has been used in Management Information Systems [8].

4.2 Structured Data Mining Techniques

The techniques which have been used for mining structured data are referred as Structured Data Mining Techniques. These techniques are explained in the following sections:

4.2.1 Web Crawler

Web Crawlers are computer programs which traverse the hypertext structure in the Web. There are two categories of Web Crawler such as: Internal and External Web Crawler. Internal Crawler crawls through internal pages of the Website which are returned by external crawler. External Crawler crawls through unknown Website.

4.2.2 Page Content Mining

Page Content Mining is structured data mining technique which works on the pages ranked by traditional search engines. By comparing page content rank it classifies the pages [14].

4.2.2 Wrapper Generation

This technique provides information on the capability of sources. Web pages are already ranked by traditional search engines. According to the query web pages are retrieved by using the value of page rank. The wrappers will also provide a variety of Meta Information. i.e. domains, statistics, index look up about the sources.

4.3 Semi-Structured Data Mining Techniques

The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language.

4.3.1 Object Exchange Model

Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange Model (OEM). It helps the user to understand the information structure on the web more accurately. It is best suited for heterogeneous and dynamic environment. A main feature of object exchange model is self describing; there is no need to describe in advance the structure of an object.

4.3.2 Top down Extraction

It extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted.

4.3.3 Web Data Extraction Language

Web data extraction language converts web data to structured data and delivers to end users. It stores data in the form of tables [14].

4.4 Multimedia Data Mining Techniques

Some of the Multimedia Data Mining Techniques are SKICAT, Multimedia Miner, Color Histogram Matching and Shot Boundary Detection.

4.4.1 SKICAT

SKICAT is a Successful Astronomical Data Analysis and Cataloging System that produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set [11].

4.4.2 Multimedia Miner

Multimedia Miner contains four major steps, Image excavator for extraction of image and Video's, a preprocessor for extraction of image features and they are stored in a database. A search kernel is used for matching queries with image and video available in the database. The discovery module performs image information mining routines to trace out the patterns in images [10].

4.4.3 Colour Histogram Matching

It contains Colour Histogram Equalization and Smoothing. Equalization tries to find out correlation between color components. The problem faced by equalization is sparse data problem which is the presence of unwanted artifacts in equalized images. This problem is solved by using smoothening [3].

4.4.4 Shot Boundary Detection

It is a technique which automatically detects boundaries shots in the Videos [5-7].

V. STUDY OF WEB CONTENT MINING TOOLS

Web Content Mining tools are software which helps to download the essential information for users. It collects appropriate and perfectly fitting information. Different types of Web content mining tools are discussed in this section.

5.1 Automation Anywhere:

Automation anywhere is a web data extraction tool used for retrieving web data effortlessly, screen scrape from web pages are use it for web mining. The Intelligent Automation Software, used for automating and scheduling business process and IT tasks in easier way.

5.1.1 Features of Automation Anywhere:

• Intelligent automation is used for business and IT tasks.
• Unique SMART Automation Technology automates complex tasks in the faster way.
• Creating automation tasks takes few minutes, record keyboard and mouse strokes, or use easy point-and-click wizards.
• Distributes tasks to multiple computers easily, using Task to SMART Exe capability
• Web recorder: (Used for extracting multiple Data and to extract Table)
• Use Automation anywhere to automate scripts in disparate formats.
• Powerful task scheduling and auto-login – run scheduled tasks anytime, even when computer is locked.
• 385 plus actions are included: conditional, loop, prompt, file management, database, system, Internet. More great features: fast speeds, automatic email notification, task chaining, hotkey, variables, logging, etc.

5.2 Web Info Extractor:

This tool is helpful in mining web data, extracting web content, and monitoring content update. Thorny template rules are not required to be defined. For mining web data and for content retrieval it is a very powerful tool. It can retrieve unstructured or structured data from web page, reorganize into local file or save to database, place into web server[16]. Difficult template rules are not required to be defined.

5.2.1 Features of Web Info Extractor:

• It is Easy to define extraction task and no need to learn boring and dense template rules.
• Retrieve unstructured data as well as tabular data to file, database.
• Monitor web pages and retrieve new content.
• Deal with any kind of files like, image, text and other link file.
• Unicode support can process web page in all languages.
• Support recursive task definition.
• Can run multi-task at the same time.

***5.3** Web Content Extractor:*

It is the most powerful and easy-to-use data extraction tool for web scraping, data mining or data extraction from the internet. It offers you a friendly, wizard-driven interface that will walk you through the process of building a data extraction pattern and creating crawling rules in a simple point-and-click manner.

It tool allows users to extract data from various websites such as online stores, online auctions, shopping sites, real estate sites, financial sites, business directories, and etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source. This variety of export formats allows you to process and analyze data in your customary format.

5.3.1 Features of Web Content Extractor

- This tool helps businessmen extract and collect the market figures, product pricing data, or real estate data.
- It helps book lovers extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- This tool assists hobbyists and collectors automate extraction of betting and auction information from auction sites.
- This tool assists to Journalists extract news and articles from news sites.
- It extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

- This tool helps single people extract information from dating sites and manage it appropriately. Get married soon!
- It helps people seeking a job extract job postings from online job websites. Find a new job faster and with minimum inconveniences.

5.4 Screen-Scraper:

Screen-scraping is a tool for extracting information from web sites which can be used in other contexts. Like a database, it allows to mine the data of the World Wide Web[18]. It allows mining the content from the web, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements.

5.4.1 Features of Screen-Scraper:

Graphical interface is provided by the Screen-scraper allowing you to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data. Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked[18]. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper. This also facilitates scraping of information at periodic intervals. One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time, after which the results are displayed in a single interface.

5.1.4. Mozenda:

To extract web data easily and to manage it affordably Mozenda is useful. With Mozenda, users can set up agents that regularly extract, store and circulate data to several destinations. Once information is in the Mozenda systems users can repurpose, format, and mash up the data to be used in other online/offline applications or as intelligence.

5.5 Merits & Demerits of WCM Tools

5.5.1 Merits:

- All the tools automate the business task and retrieve the web data in an efficient manner
- All the tools are performed on structured and unstructured web data.

5.5.2 Demerits:

- Screen-scrapper needs prior knowledge of proxy server and some knowledge of HTML and HTTP where as other tools do not require any such knowledge and it need Internet connection to run.
- Automation-AnyWhere5.5 provides a facility for recording of actions; this facility is not provided in the other tools.

5.6 Comparison of WCM Tools

The following table represents the web content mining tools and their respective tasks [17].

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

Table 1: Comparison of WCM Tools

In the above table we have explored some of the popular web content mining tools and provided their comparisons and differences. The analysis results that the Screen Scraper tool is not user friendly among the different web content mining tools discussed. Also we observe that some of these tools seem to be applicable for E-mail Data Mining.

VI. CONCLUSION

Web mining uses various data mining techniques, but it is not an application of traditional data mining due to heterogeneity and unstructured nature of the data available on the World Wide Web. The main uses of web content mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. The mining tools are imperative to scanning the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines, in order of relevance giving more productive results of each search. Detailed study and analysis of each web mining tools have been done in this paper. Future scope of web content mining includes predicting user needs in order to improve the usability, scalability and user retention.

REFERENCES

[1] Arvind Kumar Sharma, P.C. Gupta, "Exploration of efficient methodologies for the improvement in web mining techniques-A survey", International Journal of Research in IT & Management (ISSN 2231-4334) Vol.1, Issue 3, July 2011.

[2] G. Srivastava, K. Sharma, V. Kumar," Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.

[3] Bassiou, N. and Kotropoulos, C. 2006. Color Histogram Equalization using Probability Smoothening. Proceedings of XIV European Signal Processing Conference.

[4] R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[5] Cooper, M., Foote, J., Adcock, J. and Casi, S. 2003. Shot Boundary Detection via Similarity Analysis. In Proceedings of TRECVID 2003 workshop.

[6] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics. Pearson Education.

[7] Smeaton, A. F., Over, P. and Doherty, A. R. 2010. Video Shot Boundary Detection: Seven years of TRECVID Activity. Elsevier, Computer Vision and Image Understanding. Vol. 114, Issue 4. Pp. 411-418

[8] Fan, W., Wallace, L., Rich, S. and Zhang, Z. 2005. Tapping into the Power of Text Mining. Communications of the ACM – Privacy and Security in highly dynamic systems. Vol. 49, Issue-9.

[9] Pol, K., Patil, N., Patankar, S. and Das, C. 2008. A Survey on Web Content Mining and extraction of Structured and Semi structured Data. IEEE First International Conference on Emerging.

[10] Zhang, J., Hsu, W. and Lee, M. L. 2001. Image Mining: Issues, Frame Works and Techniques. In Proceedings of the 2nd International Workshop Multimedia Data Mining. pp. 13-20.

[11] Oh, J. and Bandi, B. 2002.Multimedia Data Mining Framework for Raw video sequences. ACM. Third International Workshop on Multimedia Data Mining. pp.1-10.

[12] Gupta, V. and Lehal, G. S. 2009. A Survey of Text Mining Techniques and Applications. Journal of Emerging Technologies in Web Intelligence. Vol. 1, pp. 60-76.

[13] Inamdar, S. A. and Shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.

[14] Nimgaonkar, S. and Duppala, S. 2012. A Survey on Web Content Mining and extraction of Structured and Semi structured data, IJCA Journal

[15] Kosla, R. and Blockeel, H. 2000. Web Mining Research: A Survey. SIG KDD Explorations; Vol. 2, 1-15.

[16] Web Info Extractor Manual

[17] www.Mozenda.com

[18] www.screen-scraper.com

[19] Aidan Finn, Nicholas Kushmerick and Barry Smyth. "Fact or fiction: Content classification for digital libraries". In Joint DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries (Dublin) - Proceedings - No. 01/W03 18 - 20 June 2001

[20] A. F. R. Rahman, H. Alam and R. Hartono. "Content Extraction from HTML Documents". In 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.

[21] O. Buyukkokten, H. Garcia-Molina and A. Paepcke. "Accordion Summarization for End-Game Browsing on PDAs and Cellular Phones". In Proc. of Conf. on Human Factors in Computing Systems; Pages 213 - 220 2001.

[22] O. Buyukkokten, H, Garcia-Molina and A. Paepcke. "Seeing the Whole in Parts: Text Summarization for Web Browsing on Handheld Devices". In Proceedings of the 10th international conference on World Wide Web Pages 652-662, May 01 - 05, 2001

[23] A. F. R. Rahman, H. Alam and R. Hartono. "Understanding the Flow of Content in Summarizing HTML Documents". In Int. Workshop on Document Layout Interpretation and its Applications, DLIA01, Sep., 2001.

[24] Anjali Singh, "Web Content Extraction to Facilitate Web Mining", International Journal of Electronics and Computer Science Engineering, Volume1, Number 3, 2010.

AUTHOR'S PROFILE

Arvind Kumar Sharma received his Master's Degree in Computer Science from Maharshi Dayanand University, Rohtak and M.Phil Computer Science from Alagappa University, Karaikudi. Currently, he is pursuing Ph.D Computer Science from School of Computer and Systems Sciences, Jaipur National University, Jaipur, Rajasthan, India. His areas of interests include Web Data Mining, Web Usage Mining and Web Applications.

Dr. P.C. Gupta received Ph.D Computer Science from Bundelkhand University, Jhansi. He is working as Associate Professor, Department of Computer Science & Informatics, University of Kota, Rajasthan, India. He has been guiding different Ph.D students. He has published various technical and research papers in National and International Conferences and Journals. His research interest lies in Artificial Intelligence and Neural Networks.