GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

STUDY ON WEB CONTENT MINING TOOLS, TECHNIQUES AND APPLICATIONS

Manikonda Jhansi Rani[*1], Chenchu Swetha[2] & Prof. T.Venkat Narayana Rao[3]

[*1, 2]Post Graduate Scholars, Computer Science and Engineering Sreenidhi Institute of Science and Technology, Ghatkesar, T.S, India

[3]Professor, Computer Science and Engineering, Sreenidhi Institute of Science and Technology, Ghatkesar, T.S, India

ABSTRACT

With the phenomenal growth of the Web, there is an ever increasing volume of data and information published in numerous Web pages. It is becoming a challenging task to extract and collect the required web pages/information from the web. Data mining is the form of retrieving the knowledge and information which is available in the internet.web mining is extracting information from the web resources and finding interesting patterns and is one of the forms of data mining technique. One of the subfield of Web mining is Web Content Mining. Web content mining is the process of identifying the user specific data from audio, image, text, and video which is already available in the web .This process is also called as web text mining. The web content mining consists of structured data, unstructured and semi-structured data. This paper focus on the web content mining tools , techniques and analyzing the methods for retrieving the data.

Keywords: *Data Mining, Web Mining, Web Content Mining, Mining Tools, Mining Techniques*

I. INTRODUCTION

The advancement in technology paved the way for faster communication. In computer technology a dramatic development is experienced from last ten years, such that with the press of a finger the information about a particular topic appeared in monitors within seconds. As time passed by the complexity of web increased due to enormous large amount of data [1]. The information on the internet is in the form of static and dynamic web pages of various areas from industry, education to every walk of life including blogs. As per the web sites' survey more than 160 million web sites are having inter linked and intra linked web pages. The way the web sites and web pages are accessed, it is useful from the point of business perspective for giving future directions and for decision making. As a result mining became an essential technique to extract valuable information from internet which was named as web mining [3].

Web mining is the application of data mining techniques to find interesting and potentially useful knowledge from web data [6]. Web mining is used to capture relevant information, creating new knowledge out of the relevant information, personalized information, knowing about Consumers or individual users and about several others. Web mining uses data mining techniques to automatically discover and extract information from World Wide Web. Web mining utilizes the automatic way of information extraction from the World Wide Web according to the preferences. The web mining process is shown in the figure 1.

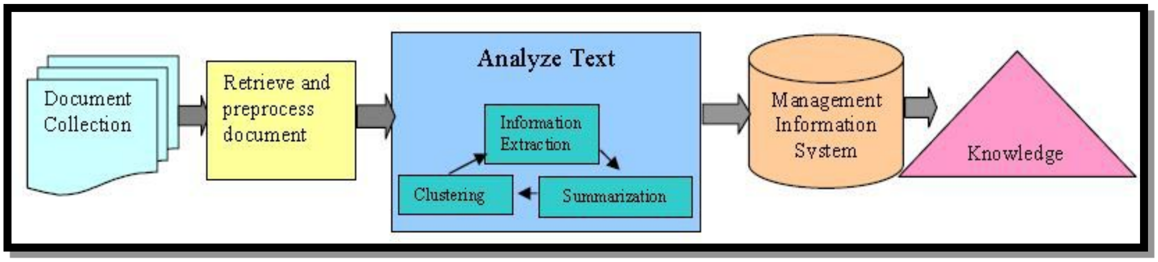

Figure1. web mining process

The three categories used for mining the web

- Web Content Mining

- Web Structure Mining
- Web Usage Mining

A. Web Content Mining

Web content mining is the mining, which is useful for extraction and integration of data, information and knowledge from the content of Web page. It describes the discovery of useful information from the web documents. In web content mining the content may be text, image, audio, video, hyperlinks and metadata etc., Web content mining also distinguishes personal home pages with other web pages. Web content mining contains the creation of wrappers which is a set of extraction rules to extract the data from the web pages, this can be done either manual or automatically [2]. The collection of integrated data may contain texts, images, audios and videos etc., this web content mining includes document tree extraction, classification of data, clustering the data and finally labeling the attributes for results. The research activities are going for computer vision, information retrieval methods and natural language processing.

B. Web Structure Mining

The process of discovering structures information from the web documents are called as web structure mining. It is the process of using graph theory to analyze the node and connection structure of a web site. The information about ranking or authoritativeness and enhance search results of a page through filtering can be provided by graph structure. This mining can be operates on either document level or hyperlink level [6]. A hyperlink is a structural component that connects the web page to a different location and provides clear navigation and point to the pages. This analysis can be done based on knowledge models, scope and properties of analysis and types of algorithms. The other kind is document structure which is using the tree-like structure to analyze and describe the HTML or XML tags within the web page. Web structure mining is classified into two types they are, inter-page structure and intra-page structure. Inter-page structure involves the connection of one page with the other page. Intra-page structure means the existence of links within a page. No separate page will be opened in this case [1].

C. Web Usage Mining

Web usage mining is used to discover the interesting usage patterns form the usage data. This includes server data (IP address), Application server data (web logic), and Application level data (events). This is otherwise a Discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities [5]. The source database is access logs, referrer logs, agent logs, and client-side cookies. It focuses on various data mining techniques to understand and analyze search patterns. The methods that are done in the web usage mining are Data cleaning, Transaction identification, Data integration, Transformation, Pattern Discovery, Pattern Analysis.

Data Mining Techniques – Navigation Patterns
Examples:

- 70% of users who accessed / company/product 2 did so by starting at /company and proceeding through /company/new , /company/products and company/product 1
- 80% of users who accessed the site started from /company/products
- 65% of users left the site after four or less page references

Data Mining Techniques – Sequential Patterns
Examples:

- In Google search, within past week 30% of users who visited /company/product/ had 'camera' as text.
- 60% of users who placed an online order in /company/product 1 also placed an order in /company/product 4 within 15 days

II. WEB CONTENT MINING

Web Content Mining is the process of extracting information and knowledge from web contents. Web Content mining is the Mining. The Search engines are provided with crawlers, to search the web, gathering information and integrating the useful content, indexing techniques to store the information, and to provide the support for query processing to deliver the required information to the users [1][3]. Web Content Mining needs more advanced tools for searching or discovering Web content. It does not give the information about structure of content that we are searching for and no information about various categories of documents that are found. The data may be unstructured (free text) or structured (data from a database) or semi-structured (html) although much of the Web is unstructured. Web content consists of textual, image, audio, video etc and metadata as well as hyperlinks. In the last years the growing of the WWW has overlap any expectations. Today they are several billions of pictures, HTML documents, and multimedia files which are available on the Internet, and their number is continuous increasing. Taking into consideration the huge variety of the web, extracting interesting contents has become a necessity [2].The web content mining techniques is shown in detail in the figure 2.

Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first point of view aims to improve the finding of information and filtering it and could be placed into the following three categories:

- Intelligent Search Agents:

These agents using the domain characteristics search for relevant information and user profiles in order to organize and interpret the discovered information.

- Information Filtering/ Categorization:

To automatically retrieve, filter, and categorize the web documents these agents use information retrieval techniques and characteristics of open hypertext Web documents.

- Personalized Web Agents:

These agents by learning the user preferences they discover Web information based on these preferences, and interests of other users with similar kind of preferences. Modeling the data on the Web into more structured form is targeted in the second approach in order to apply standard database querying mechanism and data mining applications to analyze it. Web query systems and multilevel databases are the two main categories.

Web content Data Structure

- Unstructured data – free text
- Structured data - Table or Database generated HTML pages
- Semi structured data - HTML
- Multimedia data - Receive less attention than text or hypertext

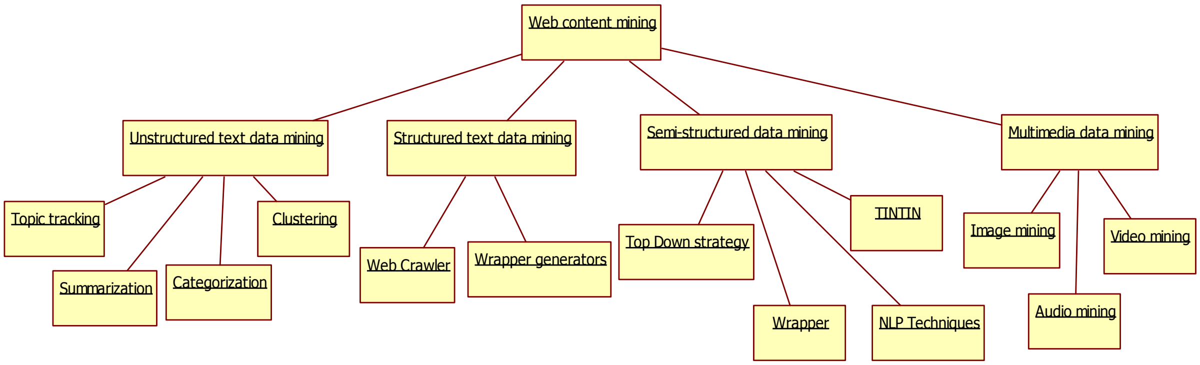

Figure2 .Web content mining techniques

A. Unstructured Text Data Mining

Web content data is much of unstructured text data. Data Mining techniques that are applied to unstructured data is termed as Knowledge Discovery in Text (KDT, or text data mining or simply text mining. Text Mining is a subset of the domain of data mining techniques. We have to use some tools or techniques to get relevant data or information from that data. Next we will discuss techniques used for unstructured data mining [5].

- **Topic Tracking**

Interest of the user is tracked by using this technique. By checking the documents viewed by the user it tries to locate other related documents. Generally used by registered sites use Topic tracking. For e.g. Yahoo uses this technique. The advertisements that are displayed when you login are related to the subject of mails you are receiving [8]. Topic tracking can be beneficial in the field of medicine and research also. Any advancement done anywhere in the world can be notified to the registered user. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest. It can be beneficial in the field of business also by which a company can keep a check on its competitor by analyzing all the news that appear about their competitor. Disadvantage of topic tracking is that sometimes we are not provided by the desired information. We may be provided with off track information.

- **Summarization**

Summarization is the technique used to reduce the length of the document and converts the multiple documents into a short set of words that specifies the meaning of the text or paragraph. It helps the user to decide whether the topic is of his interest or not. For summarization two methods are used they are Extractive and Abstractive. Extractive methods are the methods by which selecting a subset of existing phrases, words, or sentences in the original text is done to form the summary [4]. Abstractive method is closer to what a human might generate first it builds an internal semantic representation and then use natural language generation techniques to create a summary. For summarization Extractive methods are mostly used as compared to that of abstractive methods. Summarization technique can work along with Topic tracking.

- **Categorization**

Categorizing of the document is done by the Categorization technique. By considering the counts which specifies the number of words in a document and from the count it selects the main topic. According to the topic it ranks the document. Document is ranked first if it is having majority content on a particular topic [1][6].

- **Clustering**

From the large unstructured document collection it is very difficult to find out the relevant information. Categorizing of the document is done by the Categorization technique [4]. Same document can appear in different groups. The problem of finding best such grouping can be handled by clustering. A user can easily get the topic of interest from the best relevant clusters. There are various clustering algorithms available

B. Structured Text Data Mining

Structured data are typically the data records retrieved from underlying database and displayed in the web pages. It can be displayed either as tables or forms. Data can be extracted from these sources using structured data extraction techniques. This can be helpful in making value aided services by collecting information from various sources e.g. customized Web information gathering, comparative shopping, meta-search.

Following techniques are used for mining structured data:

- **Web Crawler**

A web crawler is comparatively a simple automated program, or is a script that methodically scans or "crawls" through Internet pages to create an index of the data that the user is searching for; these programs are commonly made which can be used only once, but the programming can be for long-term usage also. There are many uses for the program, possibly the most popular being search engines using it to provide web surfers with relevant websites. Linguists and market researchers are the other users, or anyone trying to search information in an organized manner from the Internet. Alternative names for a web crawler include bot, web robot, crawler, automatic indexer and web spider [4]. On the Internet Crawler programs can be purchased, or from many companies that sell computer software, and the programs can be downloaded to most of the computers.. Search engines frequently use web crawlers to collect information about the data which is available on public web pages. Their initial goal is to collect data so that whenever the Internet surfers enter a search term on their site, within a less time they can provide the surfer with relevant web sites. Web crawler is used by the Linguists may use to perform a textual analysis; that is, they may arrange the Internet to determine what words are commonly used today. Web crawler is used by Market researchers to determine and assess trends in a given market.

- **Wrapper Generators**

On the World Wide Web the effective search is done by the several Meta Search Engines which do not do the search themselves, but take help of the available search engines to find the required information. By the means of Wrappers the Meta Search Engines are connected to search engines [6]. For every search engine there is a wrapper connected to it which translates user's query into native query language and format of the search engine and also extracts the relevant data from the HTML result page of the search engine.

C. Semi Structured Data Mining Techniques

When data is integrated from several heterogeneous sources it does not have a rigid structure on the data it is called Semi structured data. e.g. extracting the Bibliographic data where some books are written by single author and some by two or more authors. Semi structured data mining techniques are required if we have to extract data from the web page and populate it in database. Web pages provide some inherent structure which can be readily recognized but still one web page can differ from other web page significantly so we say the data of web page is semi structure. In case of structured data we can extract data by submitting queries but it becomes difficult to query text data. We need some description of what to extract to get the required content. Following techniques can be applied for extracting data from semi structured data:

- **Top Down Strategy**

Using this strategy the complex objects are extracted by decomposing them into less complex objects until atomic objects have been extracted. Through this technique just a couple of examples are sufficient for extracting hundreds of objects on a new web page. The main goal of this approach is to find the objects identical to the object we are considering. The object that we are considering is supplied by the user and it is very important as whole extraction procedure depends on this example object. Top down strategy works by traversing the structure of example object in preorder form visiting all its components and concatenating them to form new resultant object. Each new object is recognized and extracted in its entirety prior to identification of its component objects.

- **Wrapper**

For representing semi structured data this technique uses OEM (Object Exchange Model). To retrieve the relevant data in OEM format the wrapper is used which in turn uses the extractor, and then executes the query at the wrapper. An OEM answer object is sent to the client, unaware that the data was not stored in a database system [7].

- **NLP Techniques**

Data can be extracted from web sources using NLP (Natural Language Processing). By using NLP techniques relevant fragments are founded and can be extracted from source document.

- **TINTIN(Table Information-based Text Inquiry)**

Based on a purely structured analysis of documents this tool extracts tabular data from unstructured documents.

D. Multimedia Data Mining Techniques

Multimedia data mining (MDM) can be defined as the process of finding interesting patterns from media data such as audio, video, image and also extracting the data cannot be accessed by using queries. MDM is the mining of information, knowledge and high level multimedia database system. Image mining, Text mining, Audio mining, and Video mining come under Multimedia data mining.

- **Image Mining**

Image processing concentrates on retrieving images and also detecting anomalous patterns. Image mining is all about finding unusual patterns. It involves making association between different images present in large database [5].

- **Video Mining**

Mining video data is more complicated than image mining because here we have collection of moving images in the form of animation. Video mining involves finding association between video clips and to find out unusual pattern in video clips.

- **Audio Mining**

Audio data consists of radio, speech or spoken language. To mine audio data one could first convert it into text using speech transcription techniques and then mine the text data. Other way is to directly mine by using audio information processing techniques and then mining selected audio clips.

III. RESEARCH ISSUES ON WEB CONTENT MINING

i. **Data/information extraction**: The main focus of Web content mining will be on extraction of the knowledge from the structured data i.e., from the web pages, such as finding about the products and extracting the results. Extracting such data allows one to provide services.

ii. **Web information integration and schema matching**: Although the Web is a collection of huge amount of data, each web site represents similar information differently. So the main problem is to how we can identify or match semantically similar data with many practical applications.

iii. **Opinion extraction from online sources:** There are many online opinion sources, e.g., customer reviews of blogs, products, chat rooms and forums. Marketing intelligence and product benchmarking can be benefited by mining opinions especially customer mining. More enhanced techniques must be introduced for opinion mining.

iv. **Knowledge synthesis:** In many applications Concept hierarchies or ontology are useful. However, generating them manually is very time consuming. The primary focus is to synthesize and organize the pieces of information on the Web.

v. **Segmenting Web pages and detecting noise:** In many Web applications, the viewer needs only the necessary content of the Web page without any navigation links, copyright notices and advertisements. Automatically segmenting Web page to extract the main content of the pages is one of the interesting problems.

IV. WEB CONTENT TOOLS

As the web consists of huge amount of data, the web content mining tools allows us to extract the required. Some of them are Screen-scraper, Mozenda, Automation Anywhere 6.1, Web Info Extractor, and Web Content Extractor [1].

A. Screen-Scraper

By using Screen-scraping tool the information from websites can be extract/mine. It can be used for searching a SQL database, database or SQL server through which interface is made with the software, for achieving the requirements of content mining. This tool provides a graphical interface which allows the user to delegate URL"s, data elements to be extracted and scripting the logic to traverse pages and work with mined data [4].

Features of Screen-Scraper

- One of the most important usages of this software and services is to mine data on products and download them to a spreadsheet.
- Items have been created, from external languages such as Java, .NET, PHP, and ASP.
- Screen scraper can also be accessed by the programming languages like .NET, Visual Basic ,PHP, Java and Active Server Pages (ASP)

Table1. Applications of Screen-Scraper

Services	Description
Medical	Gather health plan data, Migrate legacy data, Find health professionals.
Financial	Asset Analysis, Aggregate account information, Gather corporate profiles.

Automobile	Generate sales leads, Aggregate inventory, Research industry trends
E-Commerce	Collect product data, Analyze competitors, Integrate with suppliers.
Real Estate	Aggregate Listings, Monitor foreclosures, Collect recorder data

B. Automation Anywhere (AA)

AA is tool of web data extraction used for extracting the web data easily, screen scrape from Web pages or utilizing it for Web mining. It is a unique SMART Automation Technology for fast automation of complex tasks [3]. Some of the applications of this tool are given in the table 2.

Features of Automation Anywhere

- Intelligent automation is used for business and IT tasks.
- Unique SMART Automation Technology automates
- Few minutes are required for creating automation tasks, and to record keyboard and mouse strokes, or use easy point-and-click wizards.
- Distributes tasks to multiple computers easily, using Task to SMART Exe capability
- We can use Automation anywhere to automate scripts in disparate formats.

Table2. Applications of Automation Anywhere

Services	**Description**
Enterprise	With centralized control, distribution and analysis, define process in time.
Small Business	Get more done in less time, Focus on value-added work, Store and share tasks.
Business Users	Use pre-defined automation templates to create an automated task in minutes.
IT & Developers	Enjoy the power of many action wizards, and ability to create a script from scratch.
ERP Environment	A quick ROI, great support and a product built for reliability and ease of use.

C. Web Info Extractor (WIE)

WIE is a tool for mining the data, extracting Web content, and Web content analysis. Through this tool we can extract the structured or unstructured data from Web page and converting into local file or save to database, place into Web server. By this tool we can also deal with image, text, and other link file [6]. Some of the applications of this tool are given in the table 3.

Features of Web Info Extractor

- Extracting tabular data as well as unstructured data to file, database.
- Monitoring the web pages and extracting new content.
- In all languages Unicode support can process web page.
- Running multiple tasks at a time.

Table3. Applications of Web Info Extractor

Services	Description
Medical	Contact Information on Medical Pages with the assistance of IE tools.
Product Catalogues	Extract information about products sold or described online.
Weather Forecasts	Investigate to assist the ontology engineer in reusing existing domain ontology.

D. Mozenda

Mozenda allows the users to extract and manage Web data. Users are able to assign agents that can extract, store, and publish data to multiple destinations. As the information is present in Mozenda systems, users are able to format, repurpose, and mash up the data to be used in other applications or as intelligence [5]. Some of the applications of this tool are given in the table 4.

Features of Mozenda

- Mined data can be exported, accessed online, as well as used throughout an API.
- Mozenda Data Extractor is a tool which does excellent work and performs your scraper within the clouds.
- Working Environment independence.
- Platform independence is maintained.

Table4. Applications of Mozenda

Services	Description
Healthcare	Extract information on disease symptoms from blog posts and forum discussions.
Jobs & Recruiting	Determine which states have the most job openings and for which positions.
Business Intelligence	Analyze competitor recruitments, extract product reviews, performs diligence on candidates.
Web Automation	Gathering and monitoring data, filling forms, submit queries and Performs human functions.
Competitive Pricing	Collecting, monitoring product and pricing information on similar goods sold by competitors.

E. Web Content Extractor (WCE)

WCE is easy and a powerful tool to use data extraction tool for Web scraping, data mining from the Internet. This tool allows users to extract data from various websites such as shopping sites, online auctions, online stores, real estate sites, business directories, financial sites etc. The retrieved data can be converted to a various formats, including Microsoft Excel (CSV), XML, Access, HTML, SQL script, TXT, My SQL script and to any ODBC data source. It helps to retrieve the product pricing data, market figures, or real estate data [3]. Some of the applications of this tool are given in the table 5.

Features of Web Content Extractor

- This tool helps the businessmen to retrieve and collect the product pricing data, market figures, or real estate data.
- It helps book readers extract the information about books, images, authors, descriptions, including their titles, prices and ISBNs from online book sellers.
- This tool assists the Journalists to extract news and articles from news sites.

- The user can extract the online information about vacation and holiday places, including their names, descriptions, addresses, images, and prices, from web sites.

Table5.Applications of Web Content Extractor

Services	Description
Weather System	Regularly download updated web images of weather maps.
Stock Market	Monitor the opening and closing prices of stock portfolio and email them.
Restaurant	Finding all the restaurant information in particular area.
Real Estate	Extract Property information and keeps database up-to-date.
Data Scrape	Scrape unstructured data from the web and transfer it to Excel.
Online System	Scrape data from one online system and transfer it to another online system.

V. APPLICATIONS OF WEB CONTENT MINING

Web content mining is used in various fields of large information maintenance. Cloud users need to extract the information from the cloud provided by web servers can utilize the web mining.

- Online shopping systems use the web mining to extract the information of a product and its specification through web mining.
- Opinion mining is the process of extracting reviews of a customer about the product and its specification using mining techniques.
- Web search makes the user to search over 2 billion data. It maintains the ranks among the pages and advertisement ordering and publish based on the user query. Web wide tracking is effectively done using web mining methodologies.
- Web communities can be maintained such as face book. That is the users of same field of interest can be grouped and they can communicate through the network analyzed.
- Using web mining the customer's behavior can be understood.
- Web page personalization now-a-days are very important to maintain the confidential information. Web mining is used for maintaining personalized data.
- Using web mining techniques Digital library performs automated citation indexing.
- E-services include e-banking, search engines, on-line auctions, personalization, on-line knowledge management, blog analysis, social networking, e-learning and recommendation systems. This can be analyzed for the customers and enable provision to the customers based on their recommendations.

VI. CONCLUSION

The World Wide Web is the embodiment of human knowledge. The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. Thus various Data mining techniques and web content mining tools are used to extract useful information or knowledge from web page contents. This paper is concerned with the study and analysis of web content mining techniques, tools and research issues. 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 fulfill their requirements. There are many concepts available in Web Mining but this paper tried to expose the Web content mining strategy and explore some of the tools, techniques in Web Content mining.

REFERENCES

1. *Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi Overview of Web Content Mining Tools Volume 2 , 2013.*
2. *Xiaoqing Zheng,Yiling Gu,Yinsheng Li,"Data Extraction from Web Pages Based on Structural Semantic Entropy", International World Wide Web conference Committee (IW3C2),April 2012,pp.93-102*
3. *Nimgaonkar, S. and Duppala, S. 2012. A Survey on Web Content Mining and extraction of Structured and Semi structured data, IJCA Journal*
4. *Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).*
5. *Johnson, F., Gupta, S.K., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012)*
6. *Sharma, A.K., Gupta, P.C., Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data mining, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET). Volume 1, Issue 8, October (2012).*
7. *li Ghobadi,Maseud Rahgozar,"An ontology based Semantic Extraction Approach for B2C eCommerce",The International Arab Journal of Information Technology Vol.8, No. 2,April 2011,pp.163-170*
8. *Bharanipriya, V., Prasad, V.K., Web Content Mining Tools: A comparative Study, International Journal of Information Technology and Knowledge Management. Vol. 4, No 1, pp. 211-215 (2011).*