Abstract:
Data mining is extracting useful information from the web resources and finding interesting patterns that can be useful from ever expanding database of World Wide Web. One of the subfield of web mining is web content mining. The objective of the paper is four folds. Firstly, this paper introduced web mining. Secondly, it tries to explain the interrelationship of Web mining various other areas. Thirdly, it explain various Web Content mining techniques and finally the paper concludes with the analysis of the various techniques and tools.

Keywords: Content mining, Structured Data, Unstructured Data, Semi-structured Data.

1. INTRODUCTION

The advancement in technology paved the way for faster communication. The previous decade experienced a dramatic development in computer technology 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 enormously large amount of data. So extraction of data according to users need became a tedious task. As a result mining became an essential technique to extract valuable information from internet. And this technique was named as web mining. Web mining is further classified into three: They are Web content mining, web structure mining, and web usage mining [5].

Using the objects like text pictures, multimedia etc. content mining is done in the web. In web structure mining, mining is done based on the structure like hyperlinks. In the case of web usage mining, mining is done on web logs which contain the navigational pattern of users. And the study of this navigational pattern will trace out the interest of the users [5]. The structure of the paper is as follows: Section 2 presents the overview of web mining, web content mining, techniques and tools of web content mining. Section 3 deals with literature work. Section 4 shows comparative study of web mining tools and section 5 and 6 are conclusion and future scope respectively.

2. WEB MINING

2.1. Overview

Web mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the web data [15]. Web mining is used to capture relevant information, creating new knowledge out of the relevant data, personalization of the information, learning about Consumers or individual users and several others. Web mining uses data mining techniques to automatically discover and extract information from World Wide Web [2]. Several other techniques like Information retrieval, Information extraction and machine learning have been used in the past to discover the new knowledge from the huge amount of data available in the web. These techniques have been compared with web mining [5].

Information retrieval works by indexing text and then selects useful information [8]. Information Extraction focuses on extracting relevant facts whereas information retrieval selects relevant document. Web mining is now a part of Information retrieval system and Information Extraction system.

It helps in preprocessing phase before web mining. It also helps in indexing which further helps in retrieval. Machine learning is not related to web mining directly but it supports web mining because it improves text classification process better than traditional Information Retrieval process [7].

Web Content mining mines the content like text, image, audio, video, metadata, hyperlinks and extracts useful information. Since Web content mining examines the content of the web as well as the result of the search. Web Content mining mines the content like text, image, audio, video, metadata, hyperlinks and extracts useful information. Web mining helps to understand customer behavior, helps to evaluate the performance of a web site and the research done in web content mining indirectly helps to boost business.

Web content mining examines the search result of search engine. Manually doing things consumes a lot of time. When the data to be analyzed is in large quantities, then it is hard to find out the relevant data. Since now in every field of life manual work is replaced by technology.

Same happened in the case of internet. As people already admit that internet is really a magic of technology. Web Mining became a boon to this magic. In the early stages Web contained few amount of data. So there was no need of web mining tools. As years passed Web got accumulated with large amount of data. Then retrieval of data according to users need became hard task. Web mining came as a rescue for this problem.
Since Web content mining examines the content of the web as well as the result of the search. It can be further classified into Web page content mining and Search result mining. Web page Content mining is a traditional search of web page via content while search result mining is a further search of pages found from previous search [6]. Web Structure mining mines the structures like HTML or XML tags and gets information from the actual organization of the page. It uses interconnections between web pages to give weight to the page. Web usage mining is the application of data mining techniques to understand the web usage patterns. It mines data from logs, user profiles, user sessions, cookies, user queries, bookmarks, mouse click, scrolls etc. The three phases of web usage mining are preprocessing, pattern discovery, and pattern analysis. Web usage mining is classified into two, they are General access pattern tracking and customized usage tracking. The mining using the history of the web page visited by the user is known as General access pattern tracking. It understands access patterns and trends to improve structure. When it is targeted on a specific user it becomes Customized usage tracking. It analyzes access patterns of a user to improve response.

2.2 Web Content Mining
Traditional technique of searching the web was via contents. Web Content mining is the extended work performed by search engines [6]. Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. Two approaches used in web content mining are Agent based approach and database approach. The three types of agents are Intelligent search agents, Information filtering/ Categorizing agent, Personalized web agents. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefined instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well-formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data.

2.2.1 Unstructured Data Mining Techniques
Content mining can be done on unstructured data such as text. Mining of unstructured data give 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. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

Information Extraction
To extract information from unstructured data, pattern matching is used. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. This technique is very useful when there is large volume of text. IE is the basis of many other techniques used for unstructured mining. Information extraction can be provided to KDD module because information extraction has to transform unstructured text to more structured data. First the information is mined from the extracted data and then using different types of rules, the missed out information are found out. IE that makes incorrect predictions on data are discarded.
Topic Tracking
Topic Tracking is a technique in which it 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 will be informed to the user. Same can be applied in the case of mining unstructured data. An example for topic tracking 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 etc. which are not useful for user.

Summarization
Summarization is used to reduce the length of the document by maintaining the main points. It helps 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 weights the sentence and then extracts important sentences from the document. To understand the 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 Microsoft word’s auto summarize [8].

Categorization
Categorization is the technique of identifying main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does hot process the actual information. It decides the main topic from the counts. it ranks the document according to the topics. Documents having majority content on a particular topic are ranked first. Categorization can be used in business and industries to provide customer support [8].

Clustering
Clustering is a technique used to group similar documents. Here in clustering grouping is not done based on predefined topic. 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 technology is useful in management information system [8].

Information Visualization
Visualization utilizes feature extraction and key term indexing to build a graphical representation. Through visualization, documents having similarity are found out [2]. 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 useful to find out related topic from a very large amount of documents [8].

2.2.2. Structured Data Mining Techniques
The techniques used for mining structured data are web crawler, wrapper generation, page content mining.

Web Crawler
There are two types of web crawler which are called as external and internal web crawler. Crawlers are computer programs that traverse the hypertext structure in the web. External Crawler crawls through unknown website. Internal crawler crawls through internal pages of the website which are returned by external crawler.

Wrapper Generation
In Wrapper Generation, it 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 sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, Statistics, Index look up about the sources.

Page Content Mining
Page content mining is structured data extraction technique which works on the pages ranked by traditional search engines. By comparing page content rank it classifies the pages [8].

2.2.3. 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.

Object Exchange Model [OEM]
Relevant information are extracted from semi-structured data and are embeded in a group of useful information and stored in object exchange mode [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.

Top Down Extraction
In 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.

Web Data Extraction Language
In Web Data Extraction Language it converts web data to structure data and delivers to end users. It stores data in the form of tables [8].

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

SKICAT
SKICAT is a successful astronomical data analysis and cataloging system which 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[9].

**Color Histogram Matching**
Color Histogram Matching consists of color 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].

**Multimedia Miner**
Multimedia Miner comprises of four major steps, image extractor 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 [3].

**Shot Boundary Detection**
It is a technique in which automatically the boundaries are detected between shots in video [5, 2].

**2.2.5. Web Content Mining Tools**
Web content mining tools are software that helps to download the essential information for users. It collects appropriate and perfectly fitting information. Some of them are Web Info Extractor, Mozenda, Screen-Scraper, Web Content Extractor, and Automation Anywhere 5.5 [4].

**Web Info Extractor**
This tool is helpful for mining and extracting content and monitoring content update.

**Mozenda**
Users can set up agents that regularly extract, store and circulate data to several destination.

**Screen-scraping**
It searches a database, SQL server or SQL database, which interfaces with the software to achieve content mining requirements.

**Web Content Extractor**
It is powerful and easy tool for web scraping, data mining and data retrieval.

**Automation Anywhere**
It retrieves web data effortlessly, screen scrape from web pages or use it for web mining.

3. LITERATURE WORK
Web content mining can be done by retrieving information from unstructured document such as free text and semi structured document such as hypertext documents. In unstructured documents mining can be done by using word positions in the documents, text classification, event detection and tracking, finding extraction patterns in the text documents. The method used for semi-structured documents are hypertext classification and clustering, learning relations between web documents, learning extraction pattern or rules, and finding patterns in semi-structured data [5]. Web content mining is being used in various different areas like in [7] web content mining is used for mining only news sites. Beyond analyzing the news, they focused on current society interest and measured the social importance finding. For trend analysis they used domain independent statistical analysis. Four stages of dynamic news analysis are Resources identification, preprocessing, Generalization and Analysis. In Resource Identification Phase, dynamic web crawler downloads page from current URL and then filters the downloaded pages and analyze the identified news report. Steps are repeated till the queue of URL’s are empty. In preprocessing stage the news are converted into structured format. Interesting trends among new topics are found in generalization stage. In analysis phase user analyzes the pattern and the process is repeated until interesting news is found. According to this system crawler downloaded 350 web pages each day and only 130 were selected for further analysis during the period of two weeks. It has been shown by experimentation that enough differences on news topics after a period of two weeks have been found. Another area where web content mining has been proved very useful is a web content suggestion system for distance learning and is described in [9]. Two ways of suggestion are collaborative filtering and content based filtering. Collaborative suggestion clusters students into groups with similar behavior. Content based filtering provides web pages to the students who have navigation records. Web page navigation behavior is stored in personal records. Students who are new to attend the course will be having less navigation record so they are asked to poll the interest. Content suggestion system works with the help of six components such as Student Assistant Agents, Student Identification Component, Suggestion Generation Component Suggestion Delivery Component, data ware house. A new algorithm for web content mining succeeded in extracting the information from query interfaces and then matches correlated attribute. First it mines the content of query interfaces and using clustering techniques information is extracted and they are placed in special domains. Query interfaces residing in the domain are matched by the system with the user query and finally query interface nearly similar to the user query are selected. Jaccard measure is used by this algorithm to distinguish positive and negative correlated attribute [2]. Problems faced by web content mining such as extracting information from heterogeneous environment, the redundancy, the linked nature of the web, the dynamic and noisy nature of the web were highlighted. Solutions for some of the above problems were also discussed [6, 3]. In [3] content based image retrieval was discussed in detail. Web usage mining result can be improved by analyzing web content. The system integrates web page clustering into log file association mining and cluster labels are used as web page content indicators. The web page clustering was done using K-means algorithm. The clusters obtained from the web log file and integrated data file were manually summarized. Then A priori association rule mining algorithm was applied. This system utilized web content mining for web usage mining [1]. Integration of web content mining into web usage mining is also possible [4,6]. In [6] the textual content of the web pages are extracted through frequent word sequence. The they are combined with web server logs to study association rule of user’s behavior. The result of the proposed system helps in better recommendation, web personalization, web construction and web user profiling. Connection between web content mining and web structure mining was discussed in [1]. In this approach the web page content is compared with the information defined by the structure of the web site. Each web page is described with a set of keyword. This information is combined with the link structure which generates context.
based description. This comparison helps in finding out semantic information of a web page and its neighborhood. Page content algorithm was created and the aim of the project was to create a better algorithm than page rank algorithm. The importance of page determines the importance of term which the page contains. The importance of the term is calculated based on a given query. For inner classification, page content rank uses neural network [2]. A system was proposed which provides irrelevant data along with the useful data thereby increasing the result of web content mining [2]. A review was done for implementable technique of web content mining and it explained how it can be applied to business field benefitting both the customer and the producer [1].

4. COMPARATIVE STUDY OF WEB CONTENT MINING TOOLS

Table 1 shows the web content mining tools and the tasks these tools perform [4].

<table>
<thead>
<tr>
<th>Name of Tool</th>
<th>Tasks</th>
<th>Records the data</th>
<th>Extract structured data</th>
<th>Extract unstructured data</th>
<th>User friendly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automation anywhere</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Web Info Extractor</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Web content Extractor</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Not for unstructured data</td>
<td></td>
</tr>
<tr>
<td>Screen Scraper</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Mozenda</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

5. CONCLUSION

Data mining techniques used for web information extracts are incredible System and recommended for the maintenance of highly confidential data. This is affluent most intelligent resource extractor, and useful to maintain historical data. Vast Amount of data is maintained the web content mining techniques are used accurately Based on the requirements of the user.

6. REFERENCES


