Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

A REVIEW OF TRENDS IN RESEARCH ON WEB CONTENT MINING

Mr. N.SathishKumar
Research Scholar, AVS Arts and Science College, Salem, TN, India

Abstract: Web is a platform for information exchange, as it is simple and easy to publish documents. Searching for information becomes a difficult and time-consuming process as the web grows. Web mining uses various data mining techniques to discover useful knowledge from usage log file from the web. The mining tools are used to scan the HTML documents, images, and text, the results is provided for the search engines. It can assist search engines in providing productive results of each search in order of their relevance. Web mining research relates to several research communities such as Database, information Retrieval and Artificial intelligence, visualization. This paper reviews the research, tools and application issues in web mining besides proving an overall view of Web content mining.

Keywords: web mining, web content mining, web structure mining, web usage mining

I. INTRODUCTION

World Wide Web or Web is the biggest and popular source of information available, reachable and accessible at low cost provides quick response to the users and reduces burden on the users of physical movements. The data on the Web is noisy. The noise comes from two major sources. First, an emblematic Web page contains many pieces of information, e.g., the main content of the page, routing links, advertisements, copyright notices, privacy policies, etc. Second, due to the fact that the Web does not have quality control of information, i.e., one can write almost anything that one likes, a large amount of information on the Web is of low quality, erroneous, or even misleading. Retrieving of the required web page on the web, efficiently and effectively, is becoming a difficult.

With the explosive growth of information sources available on the World Wide Web, it has become increasingly necessary for users to utilize automated tools in order to find, extract, filter, and evaluate the desired information and resources. In addition, with the transformation of the Web into the primary tool for electronic commerce, it is imperative for organizations and companies, who have invested millions in Internet and intranet technologies, to track and analyze user access patterns. These factors give rise to the necessity of creating server-side and client-side intelligent systems that can effectively mine for knowledge both across the Internet and in particular Web localities.

Many organizations and corporations provide information and services on the web such as automated customer support, on-line shopping, and a myriad of resources and applications.web based applications and environments for electronic commerce, distance education, on-line collaboration, news broadcasts etc., are becoming common practice and widespread.

The WWW is becoming ubiquitous and an ordinary tool for everyday activities of common people, from a child sharing music files with friends to a senior receiving photographs and messages from grandchildren across the world. It is typical to see web pages for courses in all fields taught at universities and colleges providing course and related resources even if these courses are delivered in traditional classrooms. It is not surprising that the web is the means of choice to architect modern advanced distance education systems.

II. WEB MINING

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. Although Web mining uses many data mining techniques, as

mentioned above it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Many new mining tasks and algorithms were invented in the past decade. Based on the main kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

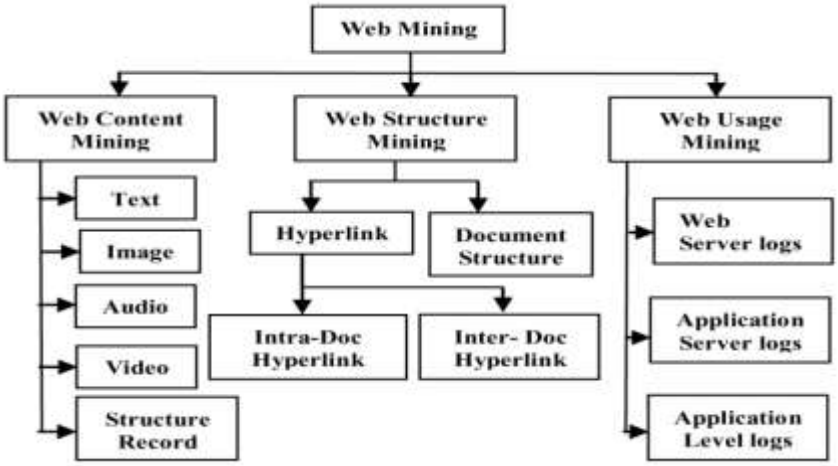

Figure 1. Web Mining Categories

1. Web Content Mining – (find the content of web pages and results of web Searching) The data present on the Web page was designed to provide the users. This usually contained images, graphs, charts, but is not limited to, text and graphics only.

2. Web Structure Mining - (Hyperlink Structure) Data that explains the organization of the content. Intra page (pages from same file) structure information incorporates the arrangement of various HTML or XML tags within a given page. The inter page structure information is hyperlinks connecting pages one by one to another pages.

3. Web Usage Mining - (analyzing user web navigation) Data that gives information about the pattern of Web pages, like IP addresses, page references, and the date and time of data access. Typically, the use of data comes from an Extended Common Log Format (ECLF) Server log files.

III. LITERATURE SURVEY

Web mining is depicted as an insights instrument to help ventures in the powerful rivalry found in ecommerce. The paper exhibited an audit of current Web mining strategies. The creators expressed that Web mining has these fundamental errand, affiliations, arrangement, and consecutive examination. An incredible exchange on the attributes of Web mining is found in [1]. The creators likewise arranged Web mining into three primary classes, Web Content Mining WCM, Web Structured Mining WSM, and Web Usage Mining WUM. WCM is about recovering and mining substance found in the WWW like sight and sound, metadata, hyperlinks, and content. WSM the mining of the structure of WWW, it discovers all the relations with respect to the hyperlinks structure, subsequently we can build a guide of how certain locales are structured, and the motivation behind why a few archives have a bigger number of connections than others. At last, WUM, this is the mining of log documents of web servers, program created logs, treats, bookmarks and parchments. WUM serves to discover the surfing propensities clients and gives experiences on activity of specific destinations.

There have been a few works around substance mining, and structure mining, in light of the exploration of Data mining and Information Retrieval, Information Extraction, and Artificial Intelligence. From the business and requisitions perspective, information acquired from the Web use examples could be straightforwardly connected to effectively oversee exercises identified with e-business, e-administrations, e-training et cetera [2]. A few models like Websift System [4] have additionally been proposed for subtle element investigation of

the web mining courses of action. A model called WHOWEDA (Warehouse of Web Data) has been proposed by Sanjay Madria, Sourav
S Bhowmick [3] in which an exchange has been performed on different issues in web mining region. Different tests have been performed for actualizing web information as a web personalization apparatus [4] in which they have ordered the methodology of web mining in five stages i.e. i) information gathering, ii) information readiness, iii) route design disclosure, iv) design
dissection and visualization, and v) design provisions.
Another part of web mining has been additionally given utilizing two separate perspectives i.e. process-driven perspective which characterized web mining as an arrangement of errands, and data centric- view which characterized web mining regarding the sorts of web information that was being utilized within the mining methodology [5].
Investigates additionally have performed exploration to utilize open Web Apis to comprehend the different parts of web mining [6]. In a review paper Naresh Barsagade has talked about the vitality and future bearings of Web Mining [7].

IV. WEB CONTENT MINING

Web content mining describes the automatic search of information resources available online [6], and involves mining web data contents. In the web mining domain, web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in web documents. The web document usually contains several types of data, such as text, image,audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristic of web data forces the web content mining towards a more complicated approach.
The web content mining is differentiated from two different points of view [8]: Information Retrieval View and Database View.R.kosla et al [9] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in training corpus as features. For the semi structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structures between the documents for document representation.
As for the database view, in order to have the better information management and querying on the web, the mining always tries to infer the structure of the web site of to transform a web site to become a database. Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources.
The various contents of Web Content Mining are

- Web page
- Search page
- Result page

Web Page: A Web page typically contains a mixture of many kinds of information, e.g., main content, advertisements, navigation panels, copyright notices, etc. For a particular application only some part of the information is useful and the rest are noises.
Search Page: A search page is typically used to search a particular Web page of the site, to be accessed numerous times in relevance to search queries. The clustering and organization of Web content in a content database enables effective navigation of the pages by the customer and search engines.
Result page: A result page typically contains the results, the web pages visited and the definition of last accurate result in the result pages of content mining.

V. WEB CONTENT MINING STRATEGIES

Web Content Mining Approaches: Two approaches used in web content mining are Agent based approach and database approach [6]. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, and personalized web agents [10]. 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 predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles [6, 7].

Web content mining has the following approaches to mine data

(1) Unstructured text mining,

(2) structured mining,

(3) Semi-structured text mining, and

(4) Multimedia mining. [9]

i) Unstructured Text Data Mining*:* Most of the Web content data is of unstructured text data. Content mining requires application of data mining and text mining techniques [9]. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are

- Information Extraction,
- Topic Tracking,
- Summarization, Categorization,
- Clustering and
- Information Visualization [8].

ii) Structured Data Mining: The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler,
- Wrapper Generation,
- Page content Mining.[9]

iii) Semi-Structured Data Mining: Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure. The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top Down Extraction, and
- Web Data Extraction language.[9]

iv) Multimedia Data Mining*:* The techniques of Multimedia data mining are;

- SKICAT,
- Color Histogram Matching, Multimedia Miner and
- Shot Boundary Detection.

5.1 Web Content Mining Tools:

Web Content Mining tools are software that helps to download the essential information for users as it collects appropriate and perfectly fitting information. Some of the tools are

i) Web Info Extractor (WIE)*:* This is a tool for data mining, extracting Web content, and web content Analysis and it can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server**[11].**

Features:

- Facilitates to define extraction tools which enable no need of learning boring and complex template rules.
- Extraction of tabular and unstructured data to file or database.
- Extraction of new content while updating and monitoring Web pages.
- Facilitates scraping of information at cyclic intervals.

- Be able to deal with text, image and other link file. Deal with Web page in all language.
- Running multi-task at the similar time. Facilitates recursive task definition.

ii) Mozenda : This is a tool to enable users to extract and manage Web data. The Users can setup agents that normally extract, store, and also publish data to multiple destinations. Previously information is in Mozenda systems, users can format, repurpose, and mash up the data to be used in other applications or as intelligence [12]. There are two parts of Mozenda's scraper tool:

Mozenda Web Console: Mozenda is a Web application that allows user to run agents, view all the results, organize those results, and export the data's extracted.

Agent Builder: Agent Builder is a Windows application used to build data extraction project.

Features:

- Easy to use.
- Platform independency. (Runs only on Windows).
- Working place independence: Tuning the scraper, managing the scraping process and get scraped data from any computer connected to the Web.

iii) Screen-Scrapper*:* This is a tool for extracting/mining information from web sites.

It is used for searching a database, which interfaced with software to attain content mining needs **[12]**. The programming languages such as Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

Features:

- Screen-scraper present a graphical interface allowing the user to allocate 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 ASP, the screen-scraper can be invoked.

iv) Web Content Extractor: WCE is a powerful and easy to use data extraction tool for Web scraping, and data extraction from the Internet. This offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and click manner. This tool permit users to extract data from various websites such as online stores & auctions, shopping, real estate, and economic sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL & MySQL script and to any ODBC data source [13].

Features:

- Helps in the extraction or collection of market figures, product pricing data, or real estate data and to Journalists extract news and articles from news sites.
- Support users to extract the information about books, including their titles, authors, descriptions, ISBNs, images, and prices, from online book sellers.
- Helps users in automate extraction of auction information from auction sites.
- Helps people seeking job postings from online job websites. Finding a new job faster and with minimum inconveniences.

v**) Automation Anywhere 6.1 (AA) :** AA is a Web data extraction tool used in getting web data, screen scratch from Web pages or use it for Web mining [13].

Features:

- Automation Technology for rapid automation of complex tasks.
- Recording keyboard and mouse or use point and click wizards to create automated tasks quickly.
- Web record and Web data extraction.

5.2 Web Content Mining Algorithms

There are two common tasks involved in web mining through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to fetch the information are described

i) Decision Tree: The decision tree is one of the powerful classification techniques. Decision trees

take the input as its features and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3/C4.5. The tree tries to infer a split of the training data based on the values of the available features to produce a good generalization. This split at each node is based on the feature that gives the maximum information gain. Each leaf node corresponds to a class label. The leaf node reached is considered the class label for that example. The algorithm can naturally handle binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [14].

ii) k-Nearest Neighbour: KNN is considered among the oldest nonparametric classification algorithms. To classify an unknown example, the distance (using some distance measure e.g. Euclidean) from that example to every other training example is measured. The k smallest distances are identified, and the most represented class in these k classes is considered the output class label. The value of k is normally determined using a validation set or using cross-validation [14].

iii) Naive Bayes: Naive Bayes is a successful classifier based upon the principle of Maximum A Posteriori (MAP).Given a problem with K classes {C1, . . . , CK} with so called prior probabilities P (C1), . . . , P(CK), can assign the class label c to an unknown example with features such features $x=(x_1,...,x_N)$ such that **$c=\text{argmax}_c P(C=c|x_1.....x_N)$,** is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, that is choosing the class with the maximum a posterior probability given the observed data. This posterior probability observed data. This posterior probability can be formulated,

$$P(C=c|x_1,..,x_N)=p(C=c)P(x_1,....,x_N|C=c)P(x_1.....x_N)$$

As the denominator is the same for all classes, it can be dropped from the comparison. Now, we should compute the so-called class conditional probabilities of the features given the accessible classes.

This may be quite difficult taking into account the dependencies between features. This approach is to assume conditional independence i.e. $x_1,.\ ,\ x_N$ are independent. This simplifies numerator as $P(C = c)\ P\ (x_1|C = c).\ P\ (x_N|C = c)$, and then choosing the class c that maximizes this value over all the classes c = 1. . . K [14].

iv) Support Vector Machine: Support Vector Machines are among the most robust and successful classification algorithms. It is a new classification method for both linear and nonlinear data and uses a nonlinear mapping to transform the original training data into a higher dimension. Among the new dimension, it searches for the linear optimal separating hyper plane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyper plane [14].

***v) Neural Network*:** The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer. The basic unit in a neural network is a neuron or unit. The inputs to the network correspond to the attributes measured for each training tuple. The inputs fed simultaneously into the units making up the input layer. It will be weighted and fed simultaneously to a hidden layer.

Number of hidden layers is arbitrary, although usually only one. Weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction [14]. As network is feed-forward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

vii) Cluster Hierarchy Construction Algorithm (CHCA)

The algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects we are clustering. Here we describing this algorithm with web pages, but the method is applicable to other domains as well. The columns correspond to the possible attributes that the objects may have (terms appearing on the web pages for this particular application). When row i has a value of 1 at column j, it means that the web page corresponding to i contains term j. From this table, which is a binary representation of the presence or absence of terms for each web page, we create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed).

Using the reduced table, we create a cluster hierarchy by examining each row, starting with those with the fewest terms (fewest number of 1's); these will become the most general clusters in our hierarchy. The row becomes a new cluster in the hierarchy, and we determine where in the hierarchy the cluster belongs by checking if any of the clusters we have created so far could be parents of the new cluster.

Potential parents of a cluster are those clusters which contain a subset of the terms of the child cluster. This comes from the notion of inheritance discussed above. If a cluster has no parent clusters, it becomes a base cluster. If it does have a parent or parents, it becomes a child cluster of those clusters which have the most terms in common with it.

This process is repeated until all the rows in the reduced table have been examined or we create a user specified maximum number of clusters, at which point the initial cluster hierarchy has been created. The next step in the algorithm is to assign the web pages to clusters in the hierarchy.

In general there will be some similarity comparison between the terms of each web page (rows in the original table) and the terms associated with each cluster, to determine which cluster is most suitable for each web page.

Once this has been accomplished, the web pages are clustered hierarchically. In the final step we remove any clusters with a number of web pages assigned to them that is below a user defined threshold and re-assign the web pages from those deleted clusters.

VI. SURVEY ON WEB CONTENT MINING

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain [6]. It may consist of text, images, audio, video, or structured records such as lists and tables.

WEB CONTENT MINING		
Author	*Representation*	*Method Used*
(Ahonen, 1998)	Bag of words and word positions	Episode rules
(Billsus & Pazzani, 1999)	Bag of words	TFIDF Naïve Bayes
(Cohen, 1995)	Relational	Propositional rule based system Inductive Logic Programming
(Dumais, 1998)	Bag of words - Phrases	- TFIDF - Decision trees - Naïve Bayes -Bayes nets - Support Vector Machines
(Feldman & Dagan, 1995)	Concept categories	Relative entropy
(Feldman, 1998)	Terms	Association rules
(Frank, 1998)	Phrases and their positions	Naïve Bayes
(Freitag & McCallum, 1999)	Bag of words	Hidden Markov Models
(Hoffmann, 1999)	Bag of words	Unsupervised statistical Method
(Junker, 1999)	Relational	Inductive Logic Programming
(Kargupta, 1999)	Bag of words with n grams	- Unsupervised hierarchical clustering - Decision trees - Statistical analysis
(Nahm & Mooney, 2000)	Bag of words	Decision trees
(Nigam, 1999)	Bag of words	Maximum entropy
(Scott & Matwin, 1999)	- Bag of words - Phrases - Hyponyms and synonyms	Rule based system
(Witten, 1999)	Named entity	Text compression
(Yang, 1999)	Bag if words and phrases	-Clustering algorithms - K-Nearest Neighbor - Decision tree
(Genersereth and Nilsson, 1987)	set of objects	ontology

VII. CONCLUSION

We survey the researches in the area of web mining. Three recognized types of web data mining are introduced generally. Web mining is a rapid growing research area. Web content mining is related but different from data mining and text mining. Web data are mainly semi-structured and/or unstructured. Web content mining requires creative applications of data mining and/or text mining techniques and also its own unique approaches.

REFERENCES

[1] Sankar K. Pal, VarunTalwar, PabitraMitra, (2002) "Web Mining in Soft Computing Framework:Relevance, State of the Art and Future Directions" IEEE Transactions on Neural Networks, Vol. 13, No. 5.
[2] L. Chen, and K. Sycara, WebMate: A Personal Agent for Browsing and Searching, Proceedings of the 2nd International Conference on Autonomous Agents, Minneapolis MN, USA, 1999 132-139.
[3] Sanjay Madria, Sourav s Bhowmick, w. -k ng, e. P. Lim,"Research Issues in Web Data Mining"
[4] A.Jebaraj Ratnakumar,"An Implementation of Web Personalization Using Web Mining Techniques", Journal Of Theoretical And Applied Information Technology", 2005 – 2010 JATIT
[5] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining— Concepts, Applications, and Research Directions", Page 400-417
[6] Hsinchunchen, Xin Li, Michael Chau, Yi-jen Ho, Chunju Tseng, "Using Open Web APIs in teaching web mining" , The University of Arizona, The University of Hong Kong
[7] Chen Ting ,Niu Xiao, Yang Weiping, The Application Of Web Data Mining Technique In Competitive Intelligence System Of Enterprise Based On Xml, Research Paper From IEEE.

[8] R.Cooley,. B.Mobasher,.; J.Srivastava,.; "Web mining: information and pattern discovery on the World Wide Web". In Proceedings of Ninth IEEE International Conference. pp. 558 – 567, 3-8 Nov. 1997.
[9] lohnson., S.K.Gupta,., Web Content Minings Techniques: A Survey, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).
[10] tetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. AAAI/MIT Press, 1991
[11] Zhang,., R.S.Segall,., Web Mining: A Survey of Current Research Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008)
[12] Mozenda,http://www.mozenda.com/web- mining-software Viewed 18 February 2013
[13] Automation Anywhere Manual. AA, http://www.automationanywhere.comViewed 06 February 2013.
[14] Navadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue10,December- 2012 ISSN: 2278-0181