Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study

[1]R.Malarvizhi, [2]K.Saraswathi
[1]Research scholar, PG & Research Department of Computer Science,
Government Arts College (Autonomous), Coimbatore-641018, TamilNadu, India.
[2]Assistant Professor, PG & Research Department of Computer Science,
Government Arts College (Autonomous), Coimbatore-641018, TamilNadu, India.

Abstract:

Nowadays, the growth of World Wide Web has exceeded a lot with more expectations. Large amount of text documents, multimedia files and images were available in the web and it is still increasing in its forms. Data mining is the form of extracting data's available in the internet. Web mining is a part of data mining which relates to various research communities such as information retrieval, database management systems and Artificial intelligence. The information's in these forms are well structured from the ground principles. This Web mining adopts much of the data mining techniques to discover potentially useful information from web contents. In this paper, the concepts of web mining with its categories were discussed. The paper mainly focused on the Web Content mining tasks along with its techniques and algorithms.

Keywords: Mining tools, techniques, structured data mining.

I. Introduction

The World Wide Web (WWW) is a popular and interactive medium with tremendous growth of amount of data or information available today. The World Wide Web is the collection of documents, text files, images, and other forms of data in structured, semi structured and unstructured form. It is also huge, diverse, and dynamic, hence raises the scalability. The primary aim of web mining is to extract useful information and knowledge from web. The web data store becomes the important source of information for many users in various domains. The web mining becomes the challenging task due to the heterogeneity and lack of structure in web resources. Because of these situations, the web users currently drowning in information and facing information overload [8]. Most of the web users could encounter the following problems, while interaction with the web;

A. Finding Appropriate Information:

When a user wants to find specific information in the web, they input a simple keyword query. The query response will be the list of pages ranked depends on their similarity to the query. Though, today's search tools have some problems such as Low precision (due to the irrelevance of search results) and Low recall (inability to index all the information available).

B. Creation of New Knowledge from the Web:

This problem is a data-triggered process whereas the previous is a query-triggered process. Here the web user has to extract potentially useful information from a collection of available contents.

C. Personalizing Data's:

This is associated with the type and presentation of information, as it is likely that people differ in the contents and presentations they prefer while interacting.

D. Analyzing Individual User Preferences:

This deals with the problem of encountering the needs of web users. This includes personalization of individual user, website design and management, customizing user information etc. The web becomes noisy if it contains various kinds of information. The web mining techniques can be used to solve those issues. The aim of this paper is review and analysis of various techniques, algorithms and tools that are using for web content data mining.

II. Web Mining

A. Overview:

The data mining is defined as the process of discovering useful patterns or knowledge from data repositories such as in the form of databases, texts, images, the Web, etc. The data repositories should be valid, potentially useful, and understandable. With the growth of the text documents, text mining are becoming increasingly important and popular. Web mining is used to capture relevant information, creating new knowledge out of relevant data, personalization of the information and learning about Consumers or individual users and several others. The information will be available from Web are hyperlink structure, page content as well as usage data. Web mining can be divided into three categories depending on the type of data as Web Structure, Web Content and Web Usage Mining. The Web Mining can be decomposed into the following subtasks, namely [13]:

- Resource finding
- Data selection & pre-processing
- Generalization
- Analysis.

B. Web Mining and Information Retrieval:

Information retrieval is the automatic process of retrieving relevant documents. IR has the primary goals of

indexing text and searching for useful documents in a collection and nowadays research in IR includes document classification and categorization, user interfaces, modeling, data visualization, filtering, etc. [3].

C. Web Mining and Information Extraction:

Information Extraction has the goal of transforming a collection of documents, with the help of an IR system, into information that is more analyzed [12]. It aims to extract relevant facts from the documents while information retrieval aims to select relevant documents [10]. While information extraction is interested in the structure or representation of a document, information retrieval views the text in a document just as a bag of unordered words [6]. Thus, in general information extraction works at a finer granularity level than information retrieval does on the documents.

D. Web Mining and Machine Learning Applied on the Web:

The Machine learning techniques support and help Web mining as they could be applied to the processes in Web mining. For example recent research [9] shows that applying machine learning techniques could improve the text classification process compared to the traditional IR techniques. In brief, Web mining intersects with the application of machine learning on the Web.

III. Web Mining Categories

The Web mining analysis relies on three general sets of information: previous usage patterns, degree of shared content [5] and inter-memory associative link structures [26] corresponding to the three subsets in Web mining namely:
(i) Web usage mining,
(ii) Web content mining and
(iii) Web structure mining.

A. Web Usage Mining:

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web. It also uses the secondary data on the web where the activity involves automatic discovery of user access patterns from one or more web servers. It contains four processing stages including data collection, preprocessing, pattern discovery and analysis [102].

i) Data Collection: The data collection is the discovery of hidden information and usage pattern trends, which could aid the Web managers for improving the management, performance and controlling of the Web servers.

ii) Data Preprocessing: The selection of useful data is an important task in the data pre-processing stage. The data's were selected in each data type to generate the cluster models for finding web user access and server usage patterns. The removal of irrelevant and noisy data is an initial step in this task. The most recently accessed data were indexed with higher value of 'time index' while the least recently accessed data were placed at the bottom with lowest value [21]. This becomes the critical step to obtain more precise analysis result due to time dependence characteristics of Web usage data.

iii) Data Clustering: The method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles [28]. The clustering algorithms become the most mining method in websites and the cluster objects include user groups (to describe user actions) and web pages.

iv) Pattern Discovery and Analysis: **U**sing this pattern discovery and pattern analysis, relevant and useful information can be easily predicted based on data analysis and Graph.

Web usages data includes data from web server access, proxy server and browser logs, user profiles, sessions or transactions, queries, registration data, cookies, bookmark data, mouse clicks and scrolls or any other data as result of interaction [15]. Analysis of web access logs for web sites can help understand the user behavior and also its web structure, thus improving the design of this massive collection of resources. There are two tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking [16].

B. Web Structure mining:

Web structure mining is based on the link structures with or without the description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites. The goal of web structure mining is to generate structured summary about websites and web pages. It uses tree-like structure to analyze and describe HTML or XML.

Some algorithms have been proposed to model the Web topology such as HITS [14], PageRank [23] and improvements of HITS by adding content information to the links structure [7] and by using outlier filtering [22]. These models are mainly applied as a method to calculate the quality rank or relevancy of each Web page. Some examples are the Clever system [7] and Google 16]. Some other applications of the models include Web pages categorization [11] and discovering micro communities on the Web [25].

C. Web Content Mining:

The Web content mining refers to the discovery of useful information from web contents which include text, image, audio, video, etc. The mining of link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining [27]. It includes

extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [1].

IV. Web Content Mining Strategies

A. Web Content Mining Approaches:

Two approaches used in web content mining are Agent based approach and database approach [13, 14]. The three types of agents are intelligent search agents, Information filtering/Categorizing agent, personalized web agents [19]. 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 [13, 14]. In Database approach it consists of well formed database containing schemas and attributes with defined domains.

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. [17]

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 [24]. 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 [17].

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.[17]

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.[17]

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

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

B. 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) [20]: 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.

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.
- 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 [29]: 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. 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 [31]: 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. 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.
- Facilitates scraping of information at cyclic intervals. The common purpose of this software and its services is to mine data on products and download them to a spreadsheet.
- A classifier example would be a metasearch 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.

iv) Web Content Extractor [30]: 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.

Features:

- Helps in the extraction or collection of market figures, product pricing data, or real estate data.
- 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.
- Help to Journalists extract news and articles from news sites.
- Helps people seeking job postings from online job websites. Finding a new job faster and with minimum inconveniences.
- Extracting online information about vacation and holiday places, including their detailed descriptions from web sites.

v) Automation Anywhere 6.1 (AA) [32]: AA is a Web data extraction tool used in getting web data, screen scratch from Web pages or use it for Web mining.

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.
- This has 305 plus actions were included: Internet, conditional, loop, prompt, file management, database and system, automatic email notifications, task chaining, etc.

C. Web Content Mining Algorithms in Classification:

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: [24]: 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.

ii) k-Nearest Neighbor [24]: 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.

iii) Naive Bayes [24]: 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 $x = (x1, \ldots, xN)$ such that $c = argmaxcP(C = ckx1, \ldots, xN)$, that is choose the class with the maximum a posterior probability given the observed data. This posterior probability can be formulated, using Bayes theorem, as follows:

$$P(C = ckx1, \ldots, xN) = P(C = c)P(x1, \ldots, xNkC = c)P(x1, \ldots.., xN).$$

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. x1, . . ,xN are independent. This simplifies numerator as P(C = c)P(x1kC = c) . . P(xNkC = c),and then choosing the class c that maximizes this value over all the classes c = 1, . . . ,K.

iv) Support Vector Machine [24]: 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 hyperplane (i.e., "decision boundary"). With an appropriate nonlinear mapping to a adequately high dimension, data from two classes can be partitioned by a hyperplane. The SVM finds this using support vectors ("essential" training tuples) and margins (defined by the support vectors)

v) Neural Network [24]: The most popular neural network algorithm is backpropagation which performs learning on a multilayer feedforward 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. 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.

IV. Conclusion

The importance of web mining continues to increase due to the increasing tendency of web documents. The mining of web data still be present as a challenging research problem in the future. Because the web documents possess numerous file formats along with its knowledge discovery process. There are many concepts available in Web Mining but this paper tried to expose the Web content mining strategy and explore some of the techniques, tools in Web Content mining.

References:

[1] Han, J., Kamber, M. Kamber. "*Data mining: concepts and techniques*". Morgan Kaufmann Publishers, 2000.

[2] Chang G, Healey MJ, McHugh JAM, Wang JTL. Web minig. In Mining the *World Wide Web—An Information Search Approach*, Dordetch: Kluwer; 2001.

[3] R. Baeza-Yates and e. Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Company, 1999.

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

[5] Boley D, Gross R, Gini ML, Han EH, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J. *Document categorization and query generation on the world wide web using WebACE*. JArtif Intell Rev 1999;13(5-6): 365–91.

[6] Y. Wilks. *Information Extraction as a core language technology*, volume 1299 of Lecture Notes in Computer Science, chapter In M-T. Pazienza (ed.), Information Extraction, pages 1–9. Springer, 1997.

[7] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins wid. *Mining the link structure of the world e web*. IEEE Computer, 32(8):60–67, 1999.

[8] P. Maes. *Agents that reduce work and information overload*. Communications of the ACM, 37(7):30–40, 1994.

[9] D. Mladenic. *Text-learning and related intelligent agents*. IEEE Intelligent Systems, 14(4):44–54, 1999.

[10] M. T. Pazienza, editor. *Information Extraction: A multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of Lecture Notes in Computer Science. International Summer School, SCIE-97, Frascati (Rome), Springer, 1997.

[11] S. Chakrabarti, B. Dom, and P. Indyk. *Enhanced hypertext categorization using hyperlinks*. In SIGMOD 1998, Proceedings ACM SIGMOD International Conference on Management of Data, pages 307–318. ACM Press, 1998.

[12] J. Cowie and W. Lehnert. *Information extraction*. Communications of the ACM, 39(1):80–91, 1996.

[13] O. Etzioni. *The world wide web: Quagmire or gold mine*. Communications of the ACM, 39(11):65–68, 1996.

[14] J. M. Kleinberg. *Authoritative sources in a hyperlinked environment*. In Proc. of ACM- SIAM Symposium on Discrete Algorithms, pages 668–677, 1998.

[15] Cooley, R.; Mobasher, B.; Srivastava, J.; "*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.

[16] J. Srivastava, R. Cooley, M. Deshpande, Pag-Ning Tan, *"Web Usage Mining: Discovery and Applications of Usage Patterns from WebData"* in proceedings of ACMSIGKDD Explorations NewsletterVol.1Issue2,January 2000.

[17] Johnson, F., Gupta, S.K., *Web Content Minings Techniques: A Survey*, International Journal of Computer Application. Volume 47 – No.11, p44, June (2012).

[18] Bharanipriya, V. and Prasad, K. 2011. *Web content Mining Tools: A Comparative study*. International Journal of Information Technology and Knowledge Management. Vol. 4. No 1,211- 215.

[19] 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.

[20] Zhang, Q., Segall, R.S., *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).

[21] Aggarwal C, Wolf JL, Yu PS. *Caching on the world wide web*. IEEE Trans Knowledge Data Engg 1999;11(1): 94–107.

[22] K. Bharat and M. R. Henzinger. *Improved algorithms for topic distillation in a hyperlinked environment*. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pages 104–111, 1998.

[23] S. Brin and L. Page. *The anatomy of a large-scale hypertextual Web search engine*. In 7th International WWW Conference, 1998.

[24] Darshna Navadiya, Roshni Patel, *Web Content Mining Techniques-A Comprehensive Survey*, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181

[25] S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. *Trawling the web for emerging cybercommunities.* In Proceedings of the Eighth World Wide Web Conference (WWW8), 1999.

[26] Pirolli P, Pitkow J, Rao R. Silk from a sow's ear: *extracting usable structures from the web*. In Proceedings of Conference on Human Factors in Computing Systems (CHI(96), Vancouver, British Columbia, Canada 1996;1996:118–25.

[27] 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

[28] Wang X, Abraham A, Smith KA. *Web traffic mining using a concurrent neuro-fuzzy approach.* In Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Computing Systems: Design, Management and Applications, Santiago, Chile 2002;2002:853–62.

[29] Mozenda, *http://www.mozenda.com/web-mining-software Viewed 18 February 2013*.

[30] Web Content Extractor help. WCE, *http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013*.

[31] Screen-scraper, *http://www.screen-scraper.com* Viewed 19 February 2013.

[32] Automation Anywhere Manual. AA, *http://www.automationanywhere.com* Viewed 06 February 2013.