OVERVIEW OF DIFFERENT APPROACHES OF WEB CONTENT MINING TOOLS AND TECHNIQUES

Mrs. Rane Seema Vijay
Assistant professor
Department of Computer Science
Smt.G.G.Khadse College, Muktainagar,
Dist.: Jalgaon, Maharashtra,
Email- seemavijay99@gmail.com

Dr. Vinod M. Patil
Associate Professor & Head
Department of Computer Science
Shri Shivaji College, Akola
IndiaMaharashtra, India
Email- vinmpatil21@yahoo.co.in

ABSTRACT
As the content over the internet (WWW) is rapidly increasing every day, and so web mining is necessary to get a proper judgments form the large amount of data available on web. Web mining is significant area of research which helps to describe the techniques for construe useful information from the web. Web mining is classified in three different types namely web content mining, web structure mining and web usage mining. Web content Mining is the process of finding the useful information from the web documents having text, images, video etc. Web structure mining is the method of discovering the hyperlink structure and document structure of the Web to gain useful information. Web usage mining is also known as web log mining which focuses on the techniques that helps in analyzing the outcomes of conduct of users when they interacting the WWW. This survey paper deals with a brief overview of web mining and its techniques and methods.

Keywords Neural Network, Page Content Mining, Support Vector Machine, Web Crawler, Naive Bayes classifiers

I. INTRODUCTION
The progress in the technology has covered the way for quick Communication and Internet is the powerful medium which provides information in the form of millions of web pages for every searched term but it is not organized information. Due to increase in large amount of information over internet the complexity of web increased. So data mining is required to extract worthful information from the big data set and when it is implemented to the web content it is called as web mining [1]. Web mining is an action of data mining for finding useful knowledge from online data. This data may be web pages which are hyperlinked by other web pages, various inline documents, web logs, and online videos and so forth. At first web mining was introduced by Etizoni[3] in the year 1996. Further web mining is classified into three different techniques i.e. Web Content, Web Usage and Web Structure mining.
The information on the Web are available in three different formats: web content, web structure and web usage. Web content refers to the textual, audio-visual content available on a website. Web content contains any element, for example, text, images, e-mail messages (archived), data, e-services, audio and video files, and so on. Web structure data represents linkage and relationship of web contents to others. Two types of structure namely intra-page and inter-page structure can be considered. Web usage data needs log data collected by web server and application server which is the main source of data. When user interacts with web site, web log data is generated on web server in form of web server log files.

WEB CONTENT MINING
Web Content Mining is a process of fetching useful patent from the data available on Web like text content or multimedia content (Images, audio and video) [4]. It is the process of extraction of useful information from the documents, video, audio, text, structured records (lists and tables) from web. The main resources of web that are mined are individual web pages. Web content mining is mainly related with text mining as most of the web content available is in the form of text [3]. Thus Web content mining needs its own applications of text mining and many other different approaches. It mainly focuses on- Web Text mining and Web Multimedia Mining [3].

For the extraction first step is to collect the data from the different links over the internet and then apply the proper technique depending upon types of `2 data and analysis result will be generated. The extraction of reliable information from the unstructured raw data text of unknown structures is referred to as Web Content Mining. Based on types of data Web Content Mining has mainly three approaches

i. Structured mining
ii. Un-Structured mining
iii. Semi -Structured mining

i. Structured mining:
Structured data is very familiar as it concerns all data which can be stored in database in the form of table i.e. rows and columns. This approach is used when data is fully structured.

Form which consist of specific rows and columns are known as the fully structured data. The techniques used for structured data are:

- Web Crawler
- Wrapper Generation
- Page Content Mining

ii. Un-Structured Mining:

Unstructured data represents the data which often include text and multimedia content. Some examples of this type data include word processing documents, photos, videos, e-mail messages, audio files, presentations, WebPages and many other kinds of business documents.

Unstructured data is everywhere. Many organizations conduct their lives across unstructured data. Same as that with structured data, unstructured data also is either machine generated or human generated. This approach is used when data is un-structured. Images, audio, video etc. are un-structured data. The techniques used for structured data are:

- Multimedia Miner
- Color Histogram Matching
- Shot Boundary Detection

iii. Semi-Structured mining:

Semi-structured data is a sort of structured data that is not same as that of the formal structure of data models linkup with relational databases or other forms of data tables. The data which contains tags are known as the Semi-structured data. [18]

This approach is used when data is partially structured. The techniques used for structured data are:

- OEM (Object Exchange Model)
- Top-Down extraction

WEB CONTENT MINING ALGORITHMS

Web Mining has two common tasks through which useful information can be mined. They are Clustering and Classification. Here various classification algorithms used to get the information are described

(i) Decision Tree:

The decision tree algorithm is one of the powerful classification techniques. Decision trees takes features as its input and output as decision, which denotes the class information. Two widely known algorithms for building decision trees are Classification and Regression Trees and ID3 (Iterative Dichotomiser 3)/C4.5 etc. ID3 (Iterative Dichotomiser 3) originally developed in 1975 by J. Ross Quinlan at the University of Sydney. C4.5 algorithm proposed again by Ross Quinlan in 1993, to overcome the limitations of ID3 algorithm discussed earlier

The ID3 classification algorithm is based on Information Entropy, the basic idea behind the algorithm is that all examples are mapped to various categories according to different values of the condition attribute set. C4.5 is also a well-known algorithm used to generate a decision trees. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also named as a statistical classifier

Splitting of the training data of tree based on the values of the available features to produce a good generalization. This split at each node is based on the features that gives the uttermost information gain. Each leaf node corresponds to a class label. The node attained is viewed as the class mark for that example. The algorithm can naturally manage binary or multiclass classification problems. The leaf nodes can refer to either of the K classes concerned [11].

(ii) k-Nearest Neighbor:

KNN is one of the nonparametric classification algorithms considered among the oldest. To classify an unknown example, the distance (using any distance measure e.g. Euclidean) from that example to each and every training example is measured. The k smallest length are distinguished and the most represented class in these k classes is described the output class label. The value of k is normally determined using a validation set or using cross-validation [11]. This technique is non parametric, which means that it does not make any assumptions on the basic data distribution. This is fairly useful, as in the real world, most of the practical data does not obey the typical theoretical assumptions made (eg Gaussian mixtures, linearly separable etc.). Non parametric algorithms like KNN is used. This algorithm is also a lazy algorithm, which means that it does not use the training data points to do any *generalization* i.e. there is *no explicit training phase*.

KNN algorithm assumes that the data is in a *feature space*. The data can possibly even multi-dimensional vectors. Since the points are in feature space, they have a opinion of the distance.

Each of the training data set consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either positive or negative classes. But KNN, works equally well with arbitrary number of classes.

In the algorithm it is also given a single number "k", which decides how many neighbors (where neighbors is defined based on the distance metric) act upon the classification which is usually an odd number suppose the number of classes is 2. If k=1, then the algorithm is called the nearest neighbor algorithm.

(iii) Naive Bayes:

Based on **Bayes' Theorem,** Naive Bayes classifiers are a collection of classification algorithms which is not an individual algorithm but a category of algorithms where all algorithms share a common principle, i.e. all pairs of features being classified is independent of each other. It is used only when the dimensionality of the inputs is high. The Bayesian Classifier is capable of computing the possible output based on the given input. In this method it is possible to add new raw data at runtime and have a better probabilistic classifier. Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. [12].

(iv) Support Vector Machine:

Support Vector Machines are among one of the most robust and successful classification algorithms which is used for classification and regression analysis. SVM are **supervised learning models** helps to analyze the large amount of data to identify patterns from them. It is a latest 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 [11].

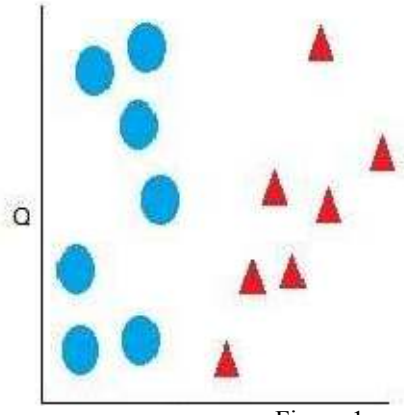

Figure 1

The working of Support Vector Machines is basically understood with a simple example. Let's imagine there are two tags: *red* and *blue*, and the data has two features: *P* and *Q*. input to a classifier is the given a pair of *(P,Q)* coordinates, which outputs if it's either *red* or *blue*. Plotting of such data on a plane is shown in figure 1[20] A support vector machine algorithm returns these data points and outputs the hyperplane, which is simply a line in two dimensions which separates the tags. The line in the figure called as **decision boundary** i.e. one type data set falls to one side for example *blue*, and one type data set falls to one side for example *red* as in Figure 2

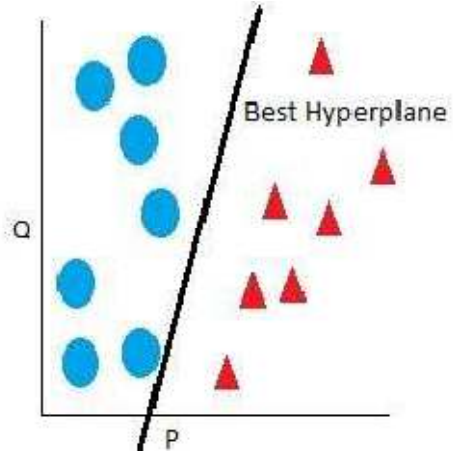

Figure 2

v) Neural Network:

The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. It consist of different layers i.e. input layer, hidden layers and an output layer. Neuron or unit is the basic unit in a neural network. The inputs given to the network represent to the attributes measured for each training tuple. The inputs flows at an identical time into the units making up the input layer. It will be weighted and flows at the same time to a hidden layer. Number of hidden layers is arbitrary, although usually only one. [12]

vii) Cluster Hierarchy Construction Algorithm (CHCA)

This algorithm takes a binary matrix (a table) as input. The rows of the table correspond to the objects that are used for clustering. The columns in the table 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, thus create a reduced table containing only rows with unique attribute patterns (i.e., duplicate rows are removed).The reduced table is used and create a cluster hierarchy by examining each row, starting first with the fewest terms (fewest number of 1's) which will become the most general clusters in our hierarchy.

WEB CONTENT MINING TOOLS

A. Web Content Extractor

For the web scraping, it is considered as the most powerful and easy to use data mining tools [14]. This tool is fundamentally designed for web scraping, data mining, and data extraction. Web Content Extractor will allow users to mine the mark data from a range of Web Pages over the Internet. Web Content Extractor collects data from online stores, company directories, e-commerce web sites, economic web sites, shopping web sites, search engine outcomes, everything you can imagine that is going on the World Wide Web. [15] This tool permits users to take out information from various websites such as online supplies, online public sale, shopping sites, valid domain sites, economic site, trade directory, etc. [14].The collected information can be exported to many
different formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, My SQL script and to any ODBC data source [14]. Features:

- It helps to collects the hierarch data, product pricing data, or valid domain records.
- It helps users to mining the information about books, like book titles, book authors, metaphors, ISBNs, descriptions, and prices, from online booksellers.
- It helps users in automate mining of vendue information beginning auction sites.
- It also helps to the media mining news and articles from report sites.

B. Web Info Extractor **[6][14][15]**

This is a web tool used for web content mining and investigating the contents. This tool also be able to extract structure or unstructured data from web page, alteration into local file or save to database, place to web server. There is no need to define difficult template rules, immediately browse to the web pages which are interesting and hit it off what is actually wished for defining the extraction job, and run it when required, or allow to it run automatically. This is a tool for data mining, extracting Web content, and Web content analysis. [6] Features:

- Monitors Web pages and draw out new content when update.
- This tool deal with text, image and other link file.
- This tool also deal with Web pages in all language.
- Helps to run multi-task at the same time.
- Support recursive task definition.
- This tool extract tabular as well as unstructured data to file, database

C. Web Text Extractor **[15]**

The **"**Web Text Extractor**"** tool is used for extract text from web page and still control label in dialog simply. Users can easily pull out and copy these texts with no selection. This tool help to mine text from web page without text selection, Can mine unselect able text. Even sort out transparent character as well as zero size character automatically. It helps to mine text from windows title, still control and edit control and and handle extracted text.

D. Screen Scraper **[15]**

This Software permits to catch characterbased data from a mainframe repeatedly presented in a green screen and it is very easy to recognize graphical user interface. New Screen Scrapers furnish the information in HTML, thus able to access with a browser. Mozart, lashpoint, Inc, and Intelligent Environments are Top producers. An in recorder presents only click screen scraping. Screen-scraper allows mining the content from the web, like searching a database, SQL server or SQL database interfaces with the software, to achieve the content mining requirements. Screen scraper can also be accessed by using programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP).

Features: [13]

• Some programming languages can also be used to access Screen Scraper.

• Separate spreadsheet is available for Download mine data products.

E. Mozenda **[14]**

This web content mining tool enables users to extract and manage Web data. Users can frame-up agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, he data to be used in other applications or as intelligence. There are two parts of Mozenda's scraper tool:

- Mozenda Web Console: It is a Web application that provides a platform to user to run agents, get data, check the job progress & organize results, and exports the published data extracted.
- For building data extraction project Agent Builder a Windows application is used.
- Features:
  - Easy to use.
  - Platform independency is their. However, Mozenda Agent Builder only runs on Windows.
  - Working place independence: Tune the scraper, manage the scraping process and get scraped data

DATA MINING TECHNIQUES USED IN WEB CONTENT MINING

Web data mining is usually a technique of data mining used to distinguish and extract the information from the documents and services available over internet. So web content mining is effectively manage by different data mining techniques.

1. Association analysis: For finding the association/correlations of recurring pattern among item sets this method is used. This technique is used to discover association between groups of users with specific interests. [15]
2. Clustering: For grouping together items with similar characteristics clustering technique is used which helps in creating users data clusters based on navigational behavior which are derived from web logs. Clustering is also used to content by creating that are related in terms of their content. Clustering. There are various Text based clustering techniques as follows – Partitioned clustering, Hierarchical clustering, Graph based clustering, neural Network based clustering. [15]

3. Sequential Pattern Discovery: this technique combines the concept of association mining along with time sequence. This system uses server logs along with web content mining.[15]

4. Segmentation: This technique is used to group users into various segments by using information from user's profiles and past browsing/purchasing history. [15]

CONCLUSION

The mining of web information still most challenging research problem because the documents available on web own many file format with its knowledge discovery process. In this paper there is discussion on one category used for web mining i.e. web content mining (WCM) and also overviewed the various algorithms like decision tree, SVM, KNN, Naïve Bayes etc. Also focused on web tools which helps to extract data from the web. And techniques used in web content mining to distinguish and extract the information from the documents.

REFERENCES

[1] Faustina Johnson, Santosh Kumar Gupta, "Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.

[2] N. R. Satish, "A Study on Applications, Approaches and Issues of Web Content Mining", International Journal of Trend in Research and Development, Volume 4(6), ISSN: 2394-9333 www.ijtrd.com

[3] Surbhi Sharma, Dinesh Soni, Dr. Arvind K Sharma, "Explorative Study of Web Data Mining Techniques and Tools: A Review", IJCST, Vol. 8, Issue 1, Jan - March 2017.

[4] Prof. Prerak Thakkar, et al., "A SURVEY- WEB MINING TOOLS AND TECHNIQUE",International Journal of Latest Trends in Engineering and Technology, Vol.(7)Issue(4), pp.212-217, DOI: http://dx.doi.org/10.21172/1.74.028 e-ISSN:2278-621X

[5] Dr Eldhose T John, et al., "An Overview of Web Content Mining Tools", Bonfring International Journal of Data Mining, Vol. 6, No. 1, January 2016

[6] Mrs.C.Menaka, et al., "A Survey of Web Content Mining Tools and Future Aspects", International Journal of Advanced Research in Computer Science Engineering and information Technology Volume: 3 Issue: 1 13-Aug- 2014,ISSN_NO: 2321-3337

[7] A. Seshagiri Rao, Dr.P.Venkatraman, "WEB MINING ISSUES IN RESEARCH TRENDS", International Journal of Research in Engineering and Applied Sciences (IMPACT FACTOR – 6.573)

[8] Nidhi Raj, N.K.Singh, "WEB MINING TECHNIQUES IN THE AREA OF THE WEB PRESONALIZATION", International Journal of Advances in Electronics and Computer Science, ISSN: 2393-2835 Volume-4, Issue-5, May-2017

[9] Shipra Saini, Hari Mohan Pandey," Review on Web Content Mining Techniques", International Journal of Computer Applications (0975 – 8887) Volume 118 – No. 18, May 2015.

[10] Anurag kumar, Ravi Kumar Singh, " A Study on Web Content Mining", International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 1 Jan. 2017, Page No. 20003-20006

[11] Darshna 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

[12] Ms. B.Nagarathna1, Dr.M.Moorthi, "A STUDY ON WEB CONTENT MINING AND WEB STRUCTURE MINING", International Journal of Modern Trends in Engineering and Research (IJMTER)Volume 02, Issue 05, [May – 2015] ISSN (Online):2349–9745 ; ISSN (Print):2393-8161

[13] T. Suresh Kumar, M. Arthanari, N. Shanthi, "A Comparative Analysis of Different Web Content Mining Tools", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:8, No:9, 2014

[14] Abdelhakim Herrouz, et al, "Overview of Web Content Mining Tools", The International Journal of Engineering And Science (IJES) ||Volume 2|| ||Issue 6|| ||Pages|| ||2013|| ISSN: 2319 – 1813 ISBN: 2319 – 1805

[15] Deven M. Kene, Dr. Pradeep K. Butey, "Web Content Mining Tools for InformationExtraction in wen Environment", National Conference on "Advanced Technologies in Computing and Networking-ATCON-2015 Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

[16] Manjot Kaur, Prof. Navjot Kaur, "Web ContentMining Techniques: A Survey", IJCST Vol. 4, Iss ue 2, April - June 2013 ISSN: 0976- 8491 (Online) | ISSN: 2229-4333 (Print)

[17] https://www.cs.uic.edu/~liub/WebContentMinin g.html viewed on 5 May 2018

[18]https://www.quora.com/What-are-Structuredsemi-structured-and-unstructured-data-in- Big-Data viewed on 5 May 2018

[19]https://data-flair.training/blogs/svm-supportvector-machine-tutorial/ viewed on 8th Sept. 2018.

[20]https://monkeylearn.com/blog/introduction-tosupport-vector-machines-svm// viewed on 8th Sept. 2018.