A Study of Web Content Mining

C.Gunasundari[1], N.Eswari[2]

[1].Assistant Professor, [2].Assistant Professor,

CSE, Roever College of Engineering and Technology, Perambalur,

Tamilnadu ,India

Email:gunasundari.cs@gmail.com

Abstract--**For the rapid growth of the web, today WWW plays a vital role in day-to-day life. The website has enormous number of data for government bodies, private organization and people. But the web does not has all data are structured form, it conceives structured, semi structured and un-structured data. Every minute the content of the web are updated and information are newly included regularly so that the amount of information becomes so big. To retrieve the meaningful data from the web is tedious and hectic process. This problem can be solved by web content mining. This paper discuss about introduction of web mining and web content mining and techniques of web content mining like clustering and classification methods.**

Keywords-Web mining, Web content mining, Web usage mining, Clustering, Classification.

I. INTRODUCTION

Day by day the web has evolve more and more. In this modern environment all the activities are based on the WWW. We extract and hire all the information from the ,but this process is not easier to retrieve the valuable information from the web .Nowadays, the web has the multiple types of data like HTML documents, images, video, audio files etc.. on the internet. Data mining is a technique of extraction of hidden predictive information from the large database .Web mining is the application of data mining technique to discover useful information or knowledge from data on the web .Data can come from variety ways and in variety of types that has not only structured ,traditional relational data and have semi-structured and unstructured data. Web mining helps to understand consumer behavior, helps to estimate 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 significant data. Since now in every field of life manual work is replaced by NEW technology. Same happened in the case of internet. As people already admit that internet is really a magic of technology. Web Mining became a benefit 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.

II. OVERVIEW

Web mining can be classified into three types like 1) Web content mining, 2) Web structure mining and 3) Web usage mining.

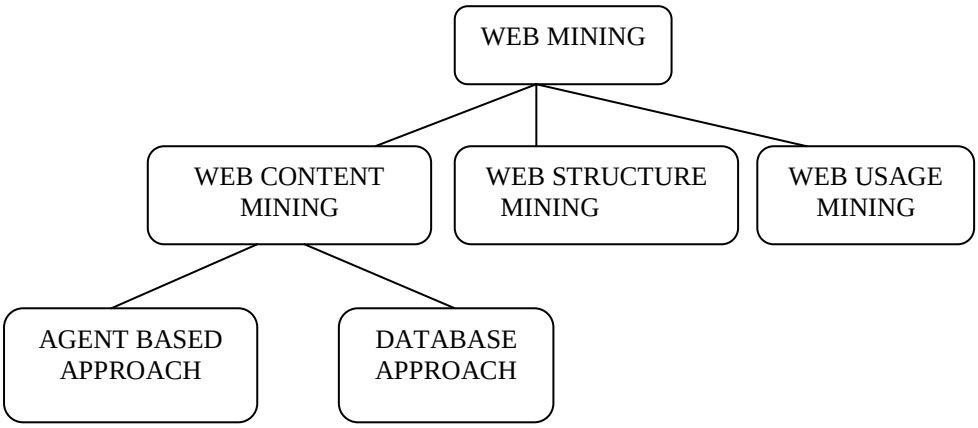

Figure 1: Web Mining types

1. Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents and pages. Data content is the collection of facts a web page is designed to include. It may consist of text, images, audio, video, or structured data such as lists and tables shortly said as multimedia data. Application of text mining to web content has been the most extensively researched. Issues addressed in text mining include topic detection and tracking, extracting association patterns, clustering of web documents and classification of web pages. Research activities on some other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP) is also going on While there exists a significant body of work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to web content mining has been limited[6].

2. Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. The analyzed web resources contain the actual web site, the hyperlinks connecting these sites and the path that online users take on the web to reach a particular site. A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents[6].

3. Web Usage Mining

Web usage mining is the application of data mining techniques to find out interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered: The web usage mining the content of the raw data for web usage mining on the one hand, and the expected knowledge to be derived from it on the other, pose a special challenge[6].

III. WEB CONTENT MINING

A. Approaches used in web content mining

i.Agent based approach

ii.database approach.

In **Agent based approach** There are three types of agents 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 predefine commands. Personalized web agents learn user inclinations and discovers documents related to those user profiles[3].

In **Database approach** it consists of well formed database containing schemas and attributes with defined domains. The algorithm proposed is called Dual Iterative Pattern Relation Extraction for finding the relevant information used by search engines. The content of web page includes no machine readable semantic information. Search engines, subject directories, intelligent agent, cluster analysis and portals are employed to find what a user must look for[3].

B. Views of Web content mining

The study can be done in Web content mining from two different categories of view: i)IR ii)DB views

i)IR view: IR view is mainly to help or to improve the information finding and filtering the information to the users usually based on either incidental or implored user profiles.

ii)**DB View** :DB view mainly tries to represent at the data on the Web and to mix them so that more sophisticated queries other than the keywords based search could be achieved.

C. Unstructured Data Mining Techniques

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.

Some of the techniques used in text mining are

1.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.

2.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.

3.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.

4.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 not process the actual information.

5.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 helps the user to easily select the topic of interest

6.Information Visualization : Visualization utilizes feature extraction and key term indexing to build a graphical representation. Through visualization, documents having similarity are found out.

D. 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

1.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.

2.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.

3.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.

E. 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

1.Object Exchange Model (OEM): Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange model (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.

2.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.

3.Web Data Extraction language: In Web data extraction language it converts web data to structured data and delivers to end users. It stores data in the form of tables.

F. Multimedia Data Mining Techniques

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

1.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 .

2.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 excavator for extraction of image and Video's, 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 .

4.Shot Boundary Detection

It is a technique in which automatically the boundaries are detected between shots in video.

IV. WEB CONTENT MINING TASKS

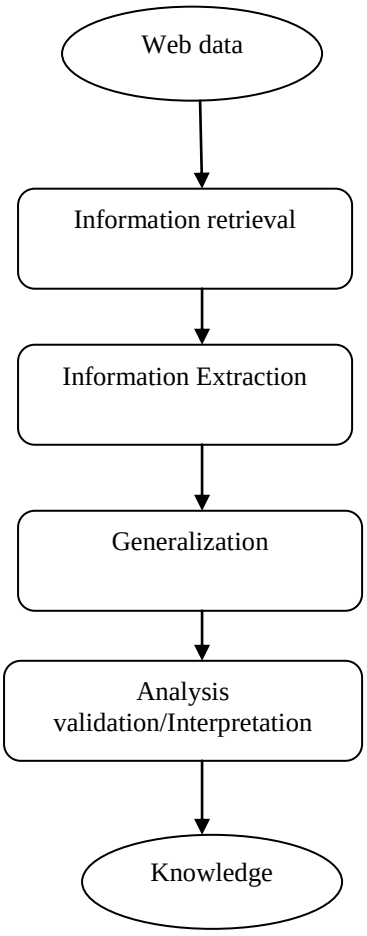

Figure 2:Web content mining tasks

1.Information Retrieval: By resource finding we mean the process of retrieving the data that is either online or offline from the text sources available on the Web such as electronic newsletters, electronic newswire, newsgroups, the text contents of HTML documents obtained by removing HTML tags, and also the manual selection of Web resources. We also include text sources that originally were not accessible from the Web but are accessible now, such as online texts made for research purposes only, text databases, etc.

2.Information Extraction: The information selection and pre-processing step is any kind of transformation processes of the original data retrieved in the IR process. These transformations could be either a kind of pre-processing that are mentioned above such as removing stop words, stemming, etc. or a pre-processing aimed at

obtaining the desired representation such as finding phrases in the training corpus, transforming the representation to relational or first order logic form, etc.

3.Generalization: In step 3 above, machine learning or data mining techniques are typically used for the generalization. We should also note that humans play an important role in the information or knowledge discovery process on the Web since the Web is an interactive medium.

4.Validation: This is especially important for validation and/or interpretation in step 4. Thus, interactive query-triggered knowledge discovery is as important as the more automatic data-triggered knowledge discovery. However, we exclude the knowledge discovery done manually by humans.

V. WEB CONTENT MINING TECHNIQUES

There are two types of web content mining techniques, one is called clustering and other is called classification.

1. Clustering: Clustering is one of the major and most important preprocessing steps in web mining analysis. In this context (Web Usage/Context Mining) items to be studied are web pages. Web page clustering puts together web pages in groups, based on similarity or other relationship measures. Tightly-couple pages, pages in the same cluster, are considered as singular items for following data analysis steps. A complete data mining analysis could be performed by using web pages information as it appears in web logs, but when the number of pages to take into account increases (i.e., in a corporative large scale web server or a server using dynamic web pages) this process could be quite hard or even unbearable. In order to deal with this issue, web page clustering appears as a reasonable solution. These techniques group pages together based on some kind of relationship measure. Pages in the same cluster will be considered as a single item for further data analysis steps

Clustering techniques Web page clustering deal with a set of web pages hosted on a web server to obtain a collection of web page sets (clusters). These clusters are applied in the following steps of the mining process instead of original pages. There are three web clustering criteria: semantic, structure, and usage based. **1.1. Semantic Clustering**:Semantical web page clustering are based on the concept of web page hierarchies. The lowest level leaves in these hierarchies are web pages, that are grouped in higher level nodes based on semantically affinities. For example, product web pages are clustered in several product families that are later grouped in a cluster for all products, beside other clusters of corporative or support information can also be defined. Semantically hierarchies can be defined following many different criteria, depending on the objectives and strategies of this analysis, and, hence, many different collections of clusters can be provided. This web page clustering techniques requires, anyway, some domain information, either from the domain experts or retrieved by any semantic repository. In this later case, there is a range of possible paths, from META-like information provided on the page contents, to Semantic Web principles, including also CMS-based web sites.

1.2. Graph Partitioning : Structure and usage page clustering are both very similar. These two approaches build a web page graph, in which nodes are the different web pages and arcs are the links among these pages. These links can be defined by the actual web links, in the case only web structure is considered or may be weighted by the usage of these transitions. In this last case, web log file is scanned to analyze the frequency of the transitions. In these entire cases web clustering problem is translated in what is called graph partitioning. The graph partitioning problem is NP-hard, and it remains NP-hard even when the number of subsets is 2 or when some unbalancing is allowed . For large graphs (with more than 100 vertices), heuristics algorithms which find suboptimal solutions are the only viable option. Proposed strategies can be classified in combinatorial approaches based on geometric representations , multilevel schemes, evolutionary optimization and genetic algorithms . We can find also a hybrid scheme that combines different approaches. There are a lot of graph partitioning algorithms and, as we cannot describe every single algorithm, we have selected those that we consider more relevant. We will start talking about four graph partitioning heuristics we have used in this study and then we will give a brief description of other clustering algorithms we find interesting.

2. Classification: Classification Techniques: The Classification algorithms are discussed under this section. The need and requirement of the user's of the websites to analyze the user preference become essential due to massive internet usage. Classification techniques are to be applied on the web log data and the performance of these algorithms can be measured. Here, in the following several classifiers are being discussed.

2.1.Decision Tree Classifier: Decision Tree Classifier (DTC) is a simple and widely used classification technique. It is a classifier in the form of a tree structure. In which there is decision node that specifies a test on a single attribute and leaf node that indicates the value of the target attribute. Arc/edge is there for split of one attribute. Path is a disjunction of test to make the final decision. It applies a straight forward idea to solve the classification problem. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online

selection model algorithm. Decision tree classifier has limitation as it is computationally expensive because at each node, each candidate splitting field must be sorted before its best split can be found.

2.2. Naive Baye's Classifier: Naive Bayes classifier (NBC) is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. It can predict class membership probabilities. Naïve Bayes probabilistic classifiers are commonly studied in machine learning. The basic idea in Naive Bayes approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of Naive Bayes methods is the assumption of word independence, i.e. the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption makes the computation of the NB classifiers far more efficient than the exponential complexity of non-naive bayes approaches because it does not use word combinations as predictors.

2.3.Support Vector Machine: Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. SVM constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

2.4. k-Nearest Neighbor: 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.

2.5. Neural Network: The most popular neural network algorithm is back propagation which performs learning on a multilayer feed forward neural network. it consists of 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. Inputs are fed simultaneously into the units making up the input layer. They are then weighted and fed simultaneously to a hidden layer. The number of hidden layers is arbitrary, although usually only one. The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction. The network is feedforward in that none of the weights cycles back to an input unit or to an output unit of a previous layer.

VI. APPLICATIONS OF WEB MINING

With the rapid growth of World Wide Web, Web mining becomes a very hot and popular topic in Web research. E-commerce and E-services are claimed to be the killer applications for Web mining, and Web mining now also plays an important role for E-commerce website and E-services to understand how their websites and services are used and to provide better services for their customers and users.

A few applications are:

E-commerce Customer Behavior Analysis

E-commerce Transaction Analysis

E-commerce Website Design

E-banking

M-commerce

Web Advertisement

Search Engine

Online Auction.

VII. CONCLUSION AND FUTURE WORK

Web has the numerous amount of data to extract the useful information from the web this work is not easy. Web mining is one of solution for the problem. This paper discussed about introduction about web mining and the types like web content mining, web structure mining, web usage mining. The detail study made about Web content mining their approaches and views like IR and Database view features are discussed .The web content mining tasks to extract data from the web is listed out. We described about the various types data available on the net. There are two Web content mining techniques are clustering and another one is classification. Clustering

and classification techniques algorithms are discussed. In future analyze the existing algorithms pros and cons will be made. There are numerous applications in Web content mining like web personalization, online shopping,e-commerce etc.The algorithms and techniques available for the applications discussed later then new techniques will be proposed.

VIII. REFERENCES

[1] Raym Kosala R. & Blockeel H., 'Web Mining Research: A Survey'. Published in ACM SIGKDD, Vol. 2, Issue 1,July 2000.ond Kosala,Hendrik Blockeel,Web Mining Research:A Survey,

[2] R. Cooley, J. Srivastava, and B. Mobasher. "Web mining: Information and pattern discovery on the world wide web". In Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI˜97), 1997.

[3] Faustina Johnson, Santosh Kumar Gupta, Department of Computer Science and Engineering, Krishna Institute of Engineering and Technology, Ghaziabad "Web Content mining techniques: Survey", International Journal of Computer Applications (0975—888) Volume 47-No.11, June 2012.

[4] R. Kosala and H. Blockeel, "Web Mining Research: A Survey,"SSIGKDD Explorations,ACM SIGKDD, July 2000.

[5] Badr Hssina, 2abdelkarim Merbouha,3hanane Ezzikouri, 4mohammed Erritali, , 5belaid Bouikhalene, An Implementation Of Web Content Extraction Using Mining Techniques, JATIT, 517-519, 2013.

[6] Darshna Navadiya, Roshni Patel -Web Content Mining Techniques-A Comprehensive Survey- - (IJERT) December-2012.

[7] Antonio LaTorre, Jos´e M. Pe˜na, V´ıctor Robles, Mar´ıa S. P´erez A Survey in Web Page Clustering Techniques

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

[9] Nimgaonkar, S. and Duppala, S. 2012. A Survey on Web Content Mining and extraction of Structured and Semi structured data, IJCA Journal.

[10] 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).

[11] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "Web Mining -Concepts,Applications & Research Directions",University of Minnesota, Minneapolis,MN 55455, USA