ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 2, Issue 12, December 2013

THE STUDY OF WEB MINING - A SURVEY

Ashish Gupta, Anil Khandekar

***Abstract*—** over the year's web mining is the very fast growing research field. Web mining contains two research areas: Data mining and World Wide Web. The Web mining research relates to several research communities such as Database, Information Retrieval and Artificial Intelligence. Web mining - i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage - is the collection of technologies to fulfill this potential . Interest in Web mining has grown rapidly in its short existence, both in the research and practitioner communities. This paper is focus on web mining. This is the review paper which focuses on the study of various techniques used for web mining. Web mining categorize into three areas: Web content mining, Web structure mining, and Web usage mining.

Index Terms— Web Mining, Web Structure Mining, Web Content Mining and Web Usage Mining.

INTRODUCTION

With the rapid development of the World-Wide Web (WWW), the increased popularity and ease of use of its tools, the World Wide Web is becoming the most important media for collecting, sharing and distributing information. 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[1].Web mining is the application of data mining technique to extract knowledge from Web data – including Web documents, hyperlinks between documents, usage logs of web sites, etc. Two different approaches were taken in initially defining Web mining. First was a 'process-centric view, which define Web mining as a sequence of tasks. Second was a data-centric view, which define Web mining in terms of the types of Web data that was being used in the mining process. The second definition has become more acceptable, as is evident from the approach adopted in most recent papers that have addressed the issue [2].

I. WEB MINING

Web mining is the application of data mining techniques to extract knowledge from Web data, i.e. Web Content, Web Structure and Web Usage data. Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks [4].

1. 1. ***Resource finding:*** It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.
2. ***2. Information selection and pre-processing:*** It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in training corpus.

   3. ***Generalization:*** It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization

.

Ashish Gupta, M.E. Scholar CSE, IIST Indore. India 07828498400

Anil Khandekar Asso. Prof., CSE,IIST Indore, India

*4. **Analysis:*** It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web [3].

II. WEB MINING TAXONOMY

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. We provide a brief overview of the three categories.

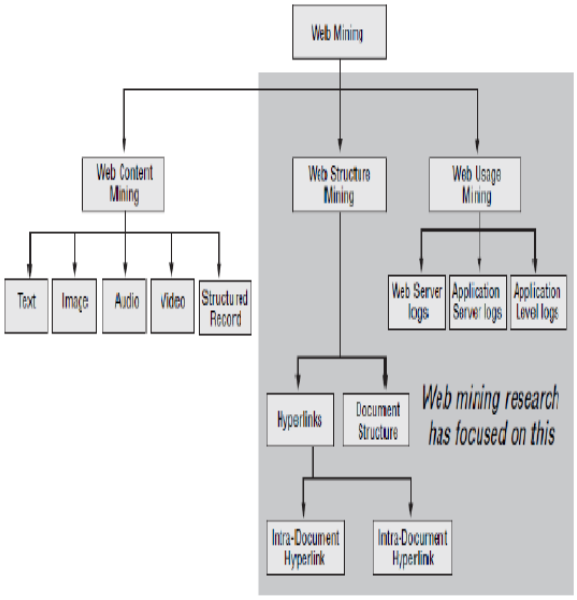

Fig 1. Web Mining Taxonomy

a. Web Content Mining: Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). 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.

b. 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. This can be further divided into two kinds based on the kind of structure information used.

- ***Hyperlinks*:** A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *Intra-Document Hyperlink*, and a hyperlink that connects two different pages is called an *Inter-Document Hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan et al. provide an up-to-date survey.
- ***Document Structure*:** 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 (Wang and Liu 1998, Moh, Lim, and Ng 2000).

c. Web Usage Mining: Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web 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:

- ***Web Server Data:*** The user logs are collected by Web server. Typical data includes IP address, page reference and access time.
- ***Application Server Data:*** Commercial application servers such as web logic story Server have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

- ***Application Level Data:*** New kinds of events can be defined in an application, and logging can be turned on for them - generating histories of these specially defined events [2].

III WEB MINING TECHNIQUE

1. Preprocessing technique - Web Robots

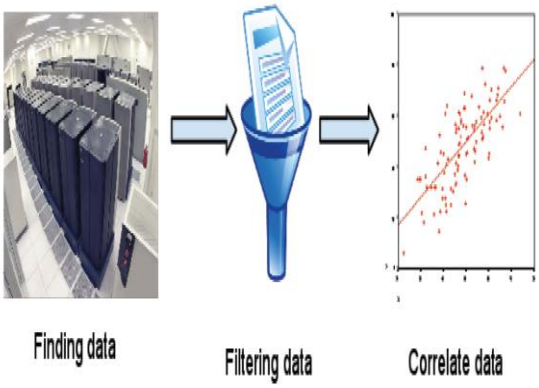

Fig 2. The pipeline of web mining

When attempting to detect web robots from a stream it is desirable to monitor both the Web server log and activity on the client-side. What we are looking for is to distinguish single Web sessions from each other. A Web session is a series of requests to web pages, i.e. visits to web pages. Since the navigation patterns of web robots differs from the navigation patterns of human users the contribution from web robots has to be eliminated before proceeding with any further data mining, i.e. when we are looking into web usage behavior of real users.

(A) Detecting Web Robots - To detect web robots [Tang et al., 2002] uses a technique involving feature classification is used. The classes chosen for evaluation are Temporal Features, Page Features, Communication Features and Path Features. It is desirable to be able to detect the presence of a web robot after as few requests as possible this is of a tradeoff between computational effort and result accuracy.

(B) Avoiding Mislabeled Sessions –

To avoid mislabeling of sessions an ensemble filtering approach [13] is used, where the idea is to instead of just one model for classification, build several models which are used to find classification errors via finding single mislabeled sessions. The set of models acquired are used to classify all sessions respectively. For each session, the amounts of false negative and false positive classifications are counted. A large value of false positive classifications implies that the session is currently assigned to be a non-robot despite being predicted to be a robot in most of the models. A large value of false negative classifications implies that the session might be a non-robot but has the robot classifier.

2. Mining Issue –

(A) Indirect Association

Common association methods often employ patterns that connect objects to each other. Sometimes, on the other hand, it might be valuable to consider indirect association between objects. Indirect association is used to e.g. represent the behavior of distinct user groups. In general, two objects that are indirectly associated have the same path, but are themselves distinct leaf to that path. That is, if one session is {A, B, C} and another is {A, B, D} then C and D are indirectly associated because they share the same traversal path {A, B}, also called "mediator". The algorithm used to discover indirect associations first uses Apriori [5] to distinguish frequent itemsets, i.e. common sessions from single clients. The frequent itemsets are matched against each other in order to discover indirect association candidate triplets, $<a, b, M>$, where a and b are indirectly associated values and M is their mediator. In the matching process a triplet is formed once an itemset L1 and another itemset L2 matches except for one position that is where one has found indirect associated values. Each pair of indirectly associated values is noted in a matrix. The matrix will, after all candidates are considered, contain values combining indirect associated values. The larger a specific matrix value is, the stronger the indirect association. The mediators found for a specific pair of sessions can be considered slightly similar to a pruned tree, i.e. where the leaves are removed.

(B) Clustering-

With the growth of the World Wide Web it can be very time consuming to analyze every web page on its own. Therefore it

is a good idea to cluster web pages based on attributes that can be considered similar to find successful and less successful attributes and patterns. There are many ways to cluster web pages before finding patterns. The most common method is the

K-means algorithm but there are several more like Single pass, Fractionation, Buckshot, Suffix tree and Apriori All, which are described in [6]. In [6] they also measure the execution time of the algorithms. Common ways to gain attributes from web pages are to take specific keywords and comparing their relevance to the rest of the text or excerpts of the web page. Clustering algorithms does not have the responsibility of finding specific web pages but instead making sure that the web pages found are relevant to the users' query. Similarities of two documents are measured by a distance function, which is computed by corresponding term vectors or attributes. The algorithms are arranged in hierarchical and partitional order. Partitional searches are compared to a cluster which yields a score and the pages with the highest score are returned as a result. In a hierarchical algorithm the search moves down a tree, choosing branches with the highest score or when it reaches a predetermined condition. The partitional algorithms are: K-means, Buckshot and Fractionation. Other algorithms are hierarchical. The Suffix tree algorithm starts with the data set as a whole and partitions it into gradually more granular clusters. Each cluster can be seen as a node with branches to smaller clusters. Single-pass uses a bottom up approach and starts at a granular level and analyzes an element of

a web page to determine which cluster it should belong to. This is a highly threshold dependant algorithm where the user determines the threshold. Apriori All studies association rules and learns from the relations of items. A good example is when many users clicks a link and subsequently another link, which creates a relation between the links. K-means are based upon distance calculations between elements, where elements are labeled to their closest centroid. Centroids are randomly placed data points in the data set. After all the elements are assigned to a centroid the centroid is moved to the place which has the shortest summed distance to all its assigned elements. The Buckshot algorithm starts by randomly sampling the data set and then placing elements around

the chosen samples into clusters. It is executed in rectangular time which makes it a fast method, but since it relies on random sampling the clusters can be less than optimal and different executions of the algorithm makes different clusters appear. Fractionation is a more processor demanding algorithm and more thorough. It divides the elements into more granular groups by iterating the clustering algorithm. The downside to Fractionation isthat it is very time consuming.

IV WEB MINING APPLICATION

Web mining is an important tool to gather knowledge of the behavior of Websites visitors and thereby to allow for appropriate adjustments and decisions with respect to Websites' actual users and traffic patterns. Along with a description of the processes involved in Web mining [8] states that Website Modification, System Improvement, Web Personalization and Business Intelligence are four major application areas for Web mining. These are briefly described in the following sections.

1. **Website Modification**

The content and structure of the Website is important to the user xperience/impression of the site and the site's usability. The problem is that different types of users have different preferences, background, knowledge etc. making it difficult (if not impossible) to find a design that is optimal for all users. Web usage mining can then be used to detect which types of users are accessing the website, and their behavior, knowledge which can then be used to manually design/re-design the website, or to automatically change the structure and content based on the profile of the user visiting it. Adaptive Websites are described in more detail in [9].

2. System Improvement

The performance and service of Websites can be improved using knowledge of the Web traffic in order to predict the navigation path of the current user. This may be used e.g. for cashing, load balancing or data distribution to improve the

performance. The path prediction can also be used to detect fraud, break-ins, intrusion etc. [8].

3. Web Personalization

Web Personalization is an attractive application area for Web based companies, allowing for recommendations, marketing campaigns etc. to be specifically customized for different categories of users, and more importantly to do this in real-time, automatically, as the user accesses the Website. For example, [7] and [10] uses association rules and clustering for grouping users and discover the type of user currently accessing the Website (based on the user's pat through the Website), in real-time, to dynamically adapt hyperlinks and content of the Website.

4. Business Intelligence

For Web based companies Web mining is a powerful tool to collect business intelligence to get competitive advantages. Patterns of the customers' activities on the Website can be used as important knowledge in the decision-making process, e.g. predicting customers' future behavior, recruiting new customers and developing new products are beneficial choices. There are many companies providing (among other things) services in the field of Web Mining and Web traffic analysis for extracting business intelligence, e.g.[11] and [12].

V CONCLUSION

Web mining consists of three major parts: collecting the data, preprocessing the data and extracting and analyzing patterns in the data. This paper focuses primarily on web usage data mining. As expected, using Web mining when designing and maintaining Websites is extremely useful for making sure that the Website conforms to the actual usage of the site. The area of Web mining was invented with respect to the needs of web shops, which wanted to be more adaptive to customers. A set of clustering techniques have been listed which significantly speeds up the process of mining data on the Web. The different techniques have a corresponding computation cost and time cost which can determine the technique of choice depending of the size of the data.

REFERENCES

[1] **Web Usage mining for a Better Web-Based Learning Environment** Osmar R. Department of Computing Science University of Alberta Edmonton, Alberta, Canada email: zaianecs.ualberta.ca

[2] Srivastava J, Desikan P and V Kumar, "Web Mining-Accomplishment & Future Directuin" in 2004 Conference

[3] Rekha Jain and Dr G. N Purohit, "Page Ranking Algorithms for Web Mining" International Journal of Computer Applications (0975 – 8887 Volume 13– No.5, January 2011

[4] Srivastava, J., Cooley, R., Deshpande, M., And Tan, P-N. (2000). "Web usage mining: Discovery and applications of usage patterns from web data" SIGKDD Explorations, 1(2), 12-23.H. Poor, An Introduction to Signal Detection and Estimation. New York: Springer-Verlag, 1985, ch.4.

[5] R. Agrawal and R. Skrikant. "Fast algorithms for mining association rules". In Proc. of the 20th VLDB Conference, Santiago, Chile, 1994.

[6] Samuel Sambasivan, Nick Theodosopoulos, "Advanced data clustering methods of mining web documents". 2006

[7] Bamshad Mobasher, Robert Cooley, Jaideep Srivastava, "Creating Adaptive Web Sites through Usage-Based Clustering of URLs". 1999

[8] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data". 1999

[9] Mike Perkowitz, Oren Etzioni, "Adaptive Web Sites: Automatically Synthesizing Web Pages".1998

[10] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, Umeshwar Dayal, "From User Access Patterns to Dynamic Hypertext Linking". 1996

[11]BizIntel, http://www.bizintel.se/ (2011)

[12] webtrends, http://webtrends.com/ (2011)

[13] C. Brodley and M.A. Friedl, "Identifying mislabeled training data", Journal of Artificial Intelligence Research, vol. 11 pp. 131-167, 1999