

A Survey on Web Usage Mining

Nirali H. Panchal¹, Ompriya Kale²

M.E.¹,², Assistant Professor², Computer Engineering Department¹,²,
L J Institute of Engineering and Technology¹,², Ahmadabad, Gujarat, India

Abstract— In Today world Most of the tasks are done by Internet. User spends number of hours over the Internet. So by knowing the behavior of user to access the Internet we can get lots of information about interest of user. Web usage mining is the process to identify the user’s behavior in terms of access the Internet. It is the sub part of Web mining. It focuses on the techniques that discover the usage pattern from web log. It uses the data mining techniques like association rule analysis, clustering, classification and machine learning. Before applying these techniques to web log some data preprocessing steps are needed for prepare the data. After that Pattern discovery techniques are applied to find some interesting pattern. Clustering is one of the Data mining techniques to make a group of user or user session having similar characteristics. Various clustering algorithm are applied to web data. Many proposed algorithm are described here like Gravitational Search algorithm, VPRS approach etc. Comparative study includes various algorithms used for Data preprocessing and Pattern discovery using clustering approach. Each algorithm has their own advantage and limitation in terms of efficiency and scalability.

Keywords— Web log, Data Pre-processing, Clustering

I. INTRODUCTION

Data mining is the process to discover some unknown patterns or knowledge from data. Web mining is part of data mining and Web Usage mining is part of web mining. Web Usage mining include usage characteristics of the user of web on the internet. Data mining techniques are applied to web data to discover web usage pattern. Analyzing such data can help organizations or web site admin to determine the interest level of customer, making some marketing strategies, personalization of web site and find more user friendly logical structure for their web sites. As we know internet is fastest growing area of information gathering. While their navigation web Users leave many records of their action. This tremendous amount of data can be a useful source of knowledge. Some Advance data mining processes are required for this knowledge to be extracted, understood and used. Web Usage Mining processes are specifically designed to carry out this task by analyzing the data representing usage data.

II. WEB USAGE MINING

Web mining refers to the use of data mining techniques to automatically retrieve, extract and analyze information for knowledge discovery from Web documents and services [1]. Web mining contains three sub parts: 1) Web Structure mining. 2) Web Content mining. 3) Web Usage mining. Web Structure Mining is the process of extract knowledge from the World Wide Web organization and links between references and referents in the Web. 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 using graph theory to analyze the node and connection structure of a web site [18]. Web content mining is the process to find out useful information from the content of a page. Mainly, the Web content consists of several types of data like textual, image, audio, video, metadata as well as hyperlinks [18]. Web Usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites[18].

![Fig 1. Web Usage Mining Architecture](image)

A. Types of Data Source

Primary data source in web usage mining is the Server-side log. There are some additional data source are also use for some user and some application which include Client – side log and Proxy-side log.

1) Server-Side Log: These logs usually contain basic information e.g. name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc [2]. This information is generally represented in standard structure like Common Log Format, Extended Log Format, and LogML.

2) Client-Side Log: Usage data can be tracked also on the client side by using Java Script, java applets, or even modified browsers. These techniques avoid the problems of users’ session identification and the problems caused by caching[2].

---

ISSN: 2231-5381  http://www.ijcttjournal.org  Page 177
3) Proxy-Side log: Many internet service providers (ISPs) give to their customer Proxy Server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers [2].

B. Data Pre-processing

Data Preprocessing is the first phase of Web Usage mining. It is the most important task of Web Usage mining. Data that are collected from data source are generally semi structured or unstructured. These logs contain redundancy and some unwanted data. So Data Preprocessing are needed for clean the data. It is main use for reduce the size of data by removing redundancy and irrelevant data. Data Preprocessing contains some technique like: Data Cleaning, User Identification, Session Identification, Pageview Identification and Path Completion.

Data cleaning is use for removing irrelevant and noisy data. Elimination of irrelevant items can be accomplished by checking the suffix of the URL name, which tells one what format these kinds of files are [2]. After cleaning the data next step is to distinguish between different users using User Identification. The aim of session identification is to divide the page accesses of each user into individual sessions. If two accesses from the same user are separated by an interval longer than a threshold they considered as different session. Page view Identification is mostly depended on intra-page structure of the site, page content and domain knowledge of site. Missing references due to caching can be inferred through path completion which is based on the knowledge of site structure and referrer information from server logs.

C. Pattern Discovery

Pattern discovery is the next step after completing data preprocessing step. In this phase using data mining techniques like association rule analysis, clustering, classification and prediction discover some usage pattern. Here some data mining techniques are described that are used in pattern discovery.

1) Association Rule: Association rule generation can be used to find out the relation between pages that are most likely referenced together during a single server session. In the perspective of Web Usage Mining, association rules contains sets of pages that are accessed together with a support value above some specified threshold. These pages may not be directly associated to one another via links.

2) Clustering: It is a technique to group together a set of items having similar characteristics. In the Web Usage mining, there are two types of clusters to be discovered: usage clusters and page clusters. Clustering of users means to establish groups of users that represent similar browsing patterns. In page clusters, clustering of pages will discover groups of pages having correlated content.

3) Classification: Classification is the job of mapping a data item into one of a number of predefined classes. In the Web Usage mining, one is interested in generating a user profile belonging to a particular class or category. Classification uses Supervised learning algorithms in which class label is known. For example, naïve Bayesian classifiers, decision tree classifiers, k-nearest neighbor classifiers etc.

4) Sequential Pattern: The method of sequential pattern discovery attempts to find out some inter-session patterns such that the existence of a set of items is followed by another item in a time-ordered set of sessions or episodes. By using this technique, marketers can predict about the future visit patterns that will be helpful in placing advertisements intended at certain groups of users.

D. Pattern Analysis

Pattern analysis is the final step in the overall Web Usage mining process. The inspiration behind pattern analysis is to filter out uninteresting rules or patterns from the set of pattern found in the pattern discovery phase. The most common type of pattern analysis consists of a knowledge query mechanism like SQL. Another way is to enter usage data into a data cube in order to perform various OLAP operations like roll-up, drill-down etc. Some Visualization techniques, for example, graphing patterns or assignment of colors to various values, can often highlight general patterns or trends in the data.

III. RELATED WORK

Theint Theint Aye[1] have proposed data preprocessing algorithm name as Field extraction and data cleaning algorithm. In field extraction algorithm the log entry

ISSN: 2231-5381  http://www.ijcttjournal.org  Page 178
contains various fields which need to be divide out for the processing. The process of separating field from the single line of the log file is known as field extraction. The server used different characters which work as separators. The most used separator character is ‘or’, ‘space’ character. In data cleaning algorithm errors and inconsistencies will be detected and removed to improve the quality of data.

Shahnaz Parvin Nina, Md. Mahamudur Rahman, Md. Khairul Islam Bhuinyan, Khandakar Entenam Unayes Ahmed [2] provides a clear idea about the pattern discovery of web usage mining. Author proposed various algorithms for data preprocessing like data Preparation, user identification and session identification. After that by applying pattern discovery method they find some result based on mostly used browser, mostly used OS etc.

Uma Maheswari, Dr. P.Sumathi [3] has proposed algorithms for data preprocessing and clustering. Data preprocessing includes reconstruction of sessions and paths are completed by appending missing pages in preprocessing. as well as, the transactions which show the behavior of users are constructed exactly in preprocessing by calculating the Reference Lengths of user access by means of byte rate. Using Web clustering several types of objects can be clustered into different groups for various purposes. By using the theory of distribution in Dempster-Shafer’s theory, the belief function similarity measure in this algorithm adds to the clustering task the ability to capture the uncertainty among Web user’s navigation performance.

Vijayashi Losarwar, Dr. Madhuri Joshi [4] has provided the importance of data preprocessing methods and various steps involved in getting the required content effectively. A complete preprocessing technique is being proposed to preprocess the web log for extraction of user patterns. Data cleaning algorithm removes the irrelevant records from web log and filtering algorithm discards the uninterested attributes from log file. User and sessions are identified by these techniques.

K. Sudheer Reddy, M. Kantha Reddy V. Sitaramulu [5] has presented several data preparation techniques of access stream even before the mining process can be started and these are used to improve the performance of the data preprocessing to identify the unique sessions and unique users. The methods proposed will help to discover meaningful pattern and relationships from the access stream of the user and these are proved to be valid and useful by various research tests.

Xidong Wang, Yiming Ouyang, Xuegang Hu and Yan Zhang[6] has proposed a new algorithm to discover frequent access pattern from web browsing behavior of user. Algorithm name was “FAP-Mining” algorithm. It has two steps. One was Construction of FAP-Tree and second was FAP-Growth. FAP-Tree was constructing according to access path determined from user sessions and FAP-Growth used to find access pattern.

Lin Feng, Baohua Guan[7] has proposed an algorithm based on conjunction matrix. First collect the data from source after that convert that data into formalize expressions. After that establishes topology structure for web site in the form of node and link. Now create a conjunction matrix where column is IP_id and row is User_id. This algorithm can accurately record the time of each visitor visiting various pages. Based on the web diary database they establish a view and apply the SQL inquiry sentence or the multi-dimensional OLAP analysis.

Lin Feng, Baohua Guan[8] has proposed a new navigation approach known as “Web Usage Mining based on Variable Precision Rough Set Model” for web user browsing a website. First, Log data sets are reduced with attribute reduction module by rough set. After that, a reduced Log data set is trained to create a rough classifier. And at last the classification result for identifying Web user is obtained according to rough decision rules.

Ashika Gupta, Rakhi arora, Ranjana sikarwar, Neha Saxena[9] proposed a system for recognize usage pattern from web monitor files of a website. The algorithm known as “new improved fp -tree” Apriori and FP Tree both are well-known algorithms for mining frequent item sets for Boolean association rules. The improved algorithm reduces the database scan and tries to prune the candidate item sets according to the minimum supporting count and get the frequent item sets.

K.Suresh, R.MadanaMohana, A.RamaMohANReddy, A.Subrmanym[10] has proposed a algorithm to find user group with similar characteristics using “Improved FCM algorithm”. In this algorithm they use first information entropy to initialize the cluster centers to find out the number of cluster centers. It can be reduce some errors, and also can improve the algorithm introductions weighting parameters. After that, combine with the merger of ideas, and divide the large chumps into small clusters. Then merge various small clusters according to the merger of the conditions, so that you can solve the irregular datasets clustering.

Mirghani. A. Eltahir, Anour F.A. Dafa-Alla[11] has provided detailed description about web usage mining process. They had included details about Data Source that uses in Web usage mining. Log file format and Information about each field of log file. Author has provided the summarization of web usage mining process which includes Data collection, Data preprocessing, Pattern Discovery and Pattern analysis. Using Deep log analyzer program result has been derived in various figures.

Anna Alphy, S. Prabakaran [12] has proposed cluster optimization technique to ARTI Neural network. Author has
used “Ant Nestmate Approach” which is inspired by social behavior of ants and honeybees. It is belonging to nature inspired algorithm. Author has proposed a cluster optimizing methodology based on ants Nestmate recognition ability and is used for eliminating the data that may occur after the clustering done by the web usage mining methods. For clustering an ARTI-neural network based approach is used.

Nayana Mariya Varghese, Jomina John [13] has provided the cluster optimization technique using the algorithm “Fuzzy Cluster-chase algorithm for cluster optimization“. In this paper author has used “Fuzzy C-means algorithm” to make a group of user having similar characteristics with clustering. After that using “Fuzzy cluster-chase” algorithm optimize the clustering quality by checking the similarity by analyzing fuzziness measures.

Ms. K.Santhisree, Dr. A Damodaram [14] has adopted the clustering algorithm known as “CLIQUE (CLuster in QUEst)” to make a cluster of web session that will be use in web personalization purpose. After that by using various similarity measures of clustering like Euclidean Distance, Jaccard distance, Cosine distance and Fuzzy distance.

V. Diviya Prabha, R. Rathipriya [15] has introduced an algorithm “Gravitational Search Algorithm” to extract highly correlated Bicluster. The correlation based fitness is used to identify the correlated bicluster with large volume. The performance of the proposed algorithm was compared with greedy search. The comparison shows that the proposed algorithm overcomes the shortcomings of greedy biclustering method. It needs minimum number of function to reach the optimal solution. Additionally, the approach can create high quality bicluster with high volume.

jinHuaXu, HongLiu [16] has presented vector analysis and KMeans based algorithms for mining user clusters. They have also applied the proposed algorithms to the real world data. Author has proved that it is feasible and efficient algorithm.

K.Santhisree, Dr A. Damodaram, S.Appaji, D.NagarjunaDevi [17] has presented new “Rough set Dbscan clustering algorithm” which identifies the behavior of the users page visits, order of occurrence of visits . Web data Clusters are produced using the rough set Similarity Upper Approximations. The rough set dbscan clustering algorithm was proved efficient when compared to the rough set agglomerative clustering. As in rough set agglomerative clustering the elements can be present in more than one cluster(soft clustering), where as in proposed algorithm rough set dbscan algorithm(hard clustering), the elements will not occur in other clusters.

<table>
<thead>
<tr>
<th>No.</th>
<th>Algorithm</th>
<th>Advantage</th>
<th>Limitation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Field Extraction and Data Cleaning</td>
<td>Speed Up Data Preprocessing Task</td>
<td>Lacking in Scalability</td>
</tr>
<tr>
<td>2</td>
<td>Greedy Clustering Using Belief Function</td>
<td>Provide Efficient Results Using Dempster–Shafer’s Belief Function</td>
<td>Lacking in Scalability</td>
</tr>
<tr>
<td>3</td>
<td>FAP-Mining Algorithm</td>
<td>Feasible by Extracting Users’ Access Patterns from Users’ Access Paths of Certain Web Site</td>
<td>Need to Improved</td>
</tr>
<tr>
<td>4</td>
<td>Variable Precision Rough Set Approach</td>
<td>Improve Prediction Accuracy Compared to Rough Set Approach</td>
<td>Choice of A Suitable Subset of Attributes For Rule Induction</td>
</tr>
<tr>
<td>5</td>
<td>Improved Frequent Pattern Tree Algorithm</td>
<td>Efficiency Provided In Terms Of Time And Memory Usage Compared to Apriori Algorithm</td>
<td>Lacks in Good Candidate Generation</td>
</tr>
<tr>
<td>6</td>
<td>Improved Fuzzy C Means Algorithm</td>
<td>Able to Identify Initial Cluster and Works on Irregular Datasets</td>
<td>Sensitive to Noise</td>
</tr>
<tr>
<td>7</td>
<td>Cluster Optimization Using Ant-Nestmate Approach</td>
<td>Optimize The Cluster by Increases Precision And Coverage . Thus Accuracy and Completeness of User Profiles Increases</td>
<td>Scalability</td>
</tr>
<tr>
<td>8</td>
<td>Cluster Optimization Using Fuzzy Cluster Chase Algorithm</td>
<td>Less Memory Utilization and Less Run Time</td>
<td>Scalability</td>
</tr>
<tr>
<td>9</td>
<td>CLIQUE (Clustering In Quest) Algorithm</td>
<td>Measure Similarity Between Clustering</td>
<td>More Time and Space Required</td>
</tr>
<tr>
<td>10</td>
<td>K-Means Algorithm</td>
<td>Feasible and Scalable</td>
<td>Sensitive to Initial Parameter K, Unable to Handle Noise</td>
</tr>
<tr>
<td>11</td>
<td>Biclustering Of Web Usage Data Using Gravitational Search Algorithm</td>
<td>Overcomes the Shortcomings of Greedy Biclustering Approach</td>
<td>-</td>
</tr>
</tbody>
</table>
V. CONCLUSIONS

Web Usage mining is a complete process to extract knowledge about browsing behavior of web user from web log. This knowledge is useful in various fields like Website customization, personalization and Recommendation. There are number of data mining techniques are used to mine knowledge from web log. Each technique has its advantage and limitation. From this literature survey, observing various algorithms for Data preprocessing and Pattern discovery task. Data preprocessing algorithms are used to prepare the data for next step. In pattern discovery phase various clustering techniques like K-means, DBSCAN, and FCM are used. Clustering is used to make a group of user having similar browsing behavior. Various algorithms like Ant Nestmate and GSA are used for optimization of cluster.

REFERENCES

ISSN: 2231-5381  http://www.ijcttjournal.org  Page 181