UPDATES TECHNIQUES FOR USER BEHAVIOUR BASED ON WEB LOGS

Ms. Mrunmayee R. Joshi[1], Dr. Prakash S. Prasad[2]
M.Tech Scholar, Department of CSE, PIET Nagpur, India

Abstract: Discovering association rules, correlations and other important data mining tasks frequent item set data mining plays an important role in mining. Massive small files are not properly get handle by some frequent item sets mining algorithms. FP-growth algorithm recursively generates large amounts of conditional pattern bases and conditional FP-trees when the dataset is large. In such a case, both the memory usage and computational cost are high; the FP-tree can't meet the memory requirement leading to low efficiency and high cost. We use MapReduce to implement the parallelization of FP-Growth, thereby improving the overall performance of frequent item sets mining. The results show that the IPFP algorithm is better and a higher mining efficiency.
Keywords: Weblogs Dataset, User Behaviour; Apriori Algorithm; KNN Algorithm; FP-Growth Algorithm; IPFP (Improved Parallel FP-Growth algorithm).

I. INTRODUCTION

Analyzing and modeling web navigation behaviour is helpful in understanding demands for online users. To predict User behaviour requires relevant weblog data for processing. Web mining technique is use to extract the relevant data from the vast weblog data. To extract and analyze useful information from web data, uses web mining which is a data mining technique. Web mining can be classified in to three different categories namely web content mining, web usage mining and web structure mining based on the kind of data. The aim of web usage mining is to get into the records of the servers (log files) that store the transactions that are performed in the web in order to find patterns.

1. Web content mining: Discovery of contents from web documents such as text, image, video,audioetc.
2. Web usage mining: Analyses the browsing activity.
3. Web structure mining: Examine the physical link structure of websites.

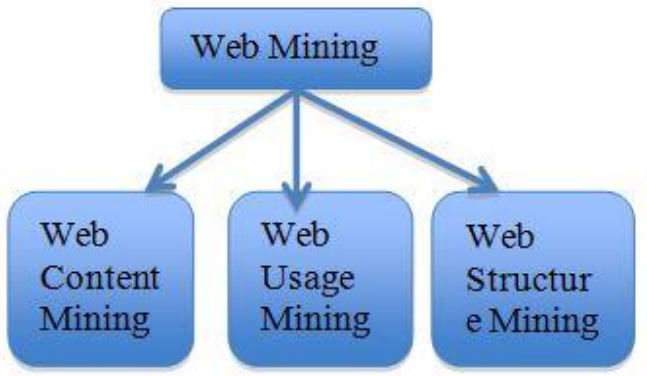


Figure 1: Web Mining Structure

The web usage data consists of data from weblog. The user accessing information from the websites is stored as logs. The log contains series of user transactions which are frequently updated whenever the user accesses the website.

The prediction of user behaviour can be identified only through logs. The weblog contains unstructured format, so convert to raw weblog to processed weblog using data preprocessing. Data Cleaning, User identification, Session Identification is used to get user navigation pattern.

Web log files are the text files, which get generated whenever there is an interaction between user and the web. Each user interaction with web will be recorded as a single record in the web log file. Generally web log file records contains fields such as IP address, URL accessed, time stamp, number of bytes, method used for making request and protocol details. These web log files can be used to understand or study the web user behaviour. The data, which is stored in web log files, will be consisting of huge amount information with some kind of incomplete and unwanted data too. Data mining techniques can be applied to on web log files to remove out unnecessary data and then finding patterns out of preprocessed data for analyzing the data to study web user behaviour.

In web server, web log file is automatically created and maintained. Each time a visitor request any file such as page,image, video, etc. from that website information on their request is appended to a current log file. It contains the information about the user like time span, URL, IP address, etc.

The host is an IP address of the system. A user Id is the unique name, which is used to identify who visit a particular web page. It is displayed when the user would like to make any transactions on the website and URL is a website address.

Objectives

The primary objective of the research is to find out the frequent item dataset using different data mining techniques and performs comparative analysis with the help of execution time and with the help of count.

- We will collect the web logs of the user from Browser history.
- Predicting frequency pattern of user related to its web behaviour (Food habits, video relation, Topic posted liked, wear pattern, political interest and all).
- We will collect the web logs of the user from browser history.
- Predicting frequency pattern of the user related to its web behaviour.

 Copyright 2019.All rights reserved.

Motivation

In the recent times more and more users are using the Internet services for their necessity. This helps to do research in getting useful information and user interest from the web using data mining techniques. To extract the user's interest measure the web logs are one of the most effective features. Since web logs are updated each time the user visits a particular web site, so it is considered as moving data.

Predicting user behaviour on the Web has a wide range of practical applications. Millions of users access web sites. When they access a websites, a huge amount of data generated in log files which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files, which is helpful to find out the user behaviour. Using this behaviour information, we can find out the accurate user next request prediction, which can reduce the browsing time of web pages.

The main motivation of this is to know what research has been done on Web usage mining in future request prediction.

II. PROPOSED ALGORITHM

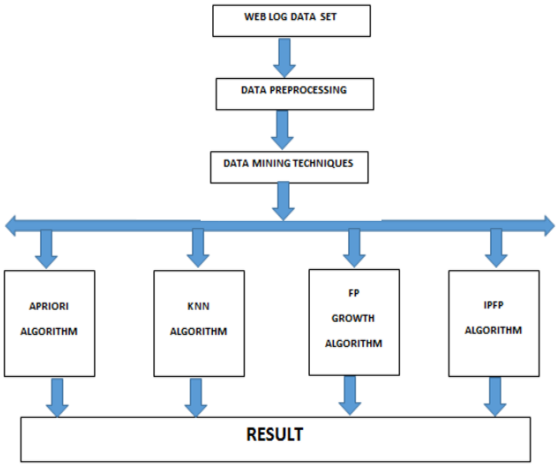


Figure 2: Proposed Methodology

System architecture is a type of conceptual design of a given system that includes system structure and behaviour. This description of architecture gives explanation about system and its properties. The system architecture for web usage analysis is shown here. There are three main modules in the system namely:

- Data Collection and Pre-processing
- Pattern Discovery
- Pattern Analysis

Data Collection and Preprocessing

The dataset used in this work are the web log files. These are the web log files which are generated in accordance with interaction of web user with server or web. Each record line in web log file indicates an interaction between user and server. There are different kinds of web log files, which store these kinds of automatically generated log data. Those are referrer logs, access logs, client-side cookies and error logs. Here for this work the web server logs are collected from the web, where the sample web log data sets are freely available to process on them. These Web logs can be downloaded and used for experimenting.

Preprocessing

This technique involves removing of the undesirable data and splitting of data into a structured format. All log files do not have the right format, so there is a need for preprocessing technique.

The log contains unstructured format of user navigation information, so conversion of unstructured log to structured log is required.

The steps or methods involved here in web data pre-processing are

- Data Cleaning
- User Identification
- Session Identification

Pattern Discovery

Pattern discovery step is performed to discover interesting patterns or knowledge to analyze web user behaviour. Once the pre-processing step is completed the patterns or knowledge can be discovered from the pre-processed web log data in pattern discovery step. Variety of techniques used to discover the web patterns are like "statistical methods" and also "data mining methods" like "Path analysis", "Association rule", "Sequential patterns", "Clustering" and "Classification". These methods are applied on web log data so as to study web user behaviour. These discovered patterns or knowledge could be represented in some form like table, graph and charts etc. The data mining algorithms can be used here to discover the frequent item sets or patterns. Here in this work two data mining techniques "Association rule", "Classification" are used to discover the frequent item sets.

Association Rules:

The generation of association rules out of given web log data using web-mining techniques is done here. In the web usage-mining domain the association rules will refer to the set of web pages, which are accessed together, and also the set of web page patterns, which are accessed very frequently by the web users. It also refers to the pages, which are referenced together.

Classification:

Classification is the data mining technique, which is used for mapping of a given data item into one of many predefined classes. For example, developing the profile of web users belonging to any particular class. It is a supervised learning. The algorithms like "decision trees", "naive Bayesian classifiers", "k-nearest neighbor classifiers", and "Support Vector Machines" can be used here.

Pattern Analysis

The pattern analysis is performed to remove out the uninteresting patterns or rules from frequent item sets, which are found in previous pattern discovery step. One of the common forms of pattern analysis consists of "knowledge query mechanism" like "SQL".

 Copyright 2019.All rights reserved.

In design and implementation we apply different data mining techniques/algorithms over web log dataset.
Data Mining Techniques
In our project we use four different algorithms, which comes under association rule and the classification data mining techniques.

- Apriori Algorithm
- K-Nearest Neighbors Algorithm
- FP-Growth Algorithm
- Improved Parallel FP-Growth Algorithm

Apriori Algorithm, FP-Growth Algorithm comes under association mining rule and K-Nearest Neighbors Algorithm comes under Classification.

i. Apriori Algorithm:
Apriori algorithm follows a "Botton-Up Approach" where frequent item subset are extended one item at a time.
Apriori is designed to operate on database containing transaction. The algorithm is a supervised influential approach for mining frequent item set for Boolean association rules. It takes weblog file of visited and minimum visitor page characterized as segment of input. Apriori algorithm creates all maximum frequent item set F1, F2, F3…Fk as output. The algorithm identifies and the repeated dataset and are considered for creating frequent IP set in the first pass. In the subsequent passes frequent IP sets accepted in the previous pass are extended with another IP to generate frequent item sets. After k passes if no frequent k- item set is found, the algorithm is ended. But unfortunately Apriori algorithm is not as efficient as that of FP-Growth algorithm in discovering association rules.
Apriori algorithm is a step-wise search, the n number of item sets is used to discover an n+1 item sets.
Apriori algorithm utilize a level-wise approach where it will generate patterns containing 1 items, then 2 items, 3 items, etc.

Disadvantages:

- Finds the frequent elements from the database. But the size of the database increases with the number of items.
- Large search space is needed and I/O cost will increase.
- Number of database scan is increased thus candidate generation will increase results in increase in computational cost.

ii. FP Growth Algorithm:
FP-Growth algorithm one of the popular and efficient data mining algorithms which can used for finding the association rules or frequent item sets or interesting patterns. Here in this work FP-Growth algorithm is used to find frequent item sets from given web server log files. This algorithm works in divide and conquers way. FP-Growth algorithm needs to perform only two database scans. In its first database scan algorithm will find frequent item sets which are sorted by frequency and also in descending order. Next in its second database scan the database will be compressed as FP-Tree. Than the algorithm will start finding or generating FP-Tree for each of the item whose support value will be greater than or same as given threshold value. Here in this work the FP-Growth algorithm is used in web usage mining. "Web usage mining" includes web server log data in order tofind the frequent web access patterns or frequently accessed web pages. This will help to analyse the web user behaviour so as to help improve theperformance of web applications and also help improve business. Here thefrequently accessed web page patterns are discovered which will help to analyse web usage. By this result we can pre-fetch and keep frequently accessed web page sequence in cache so as to reduce the page access time in future.

Steps:
Step-1. Built a data structure called the FP-Tree.
Step-2. Extracts frequent item sets.

Drawback:

- FP tree may unable to fit in memory.
- FP tree is costly to build.

iii. K-Nearest Neighbours Algorithm:
K-Nearest Neighbors is one of the most basic but essential classification algorithms. It comes under the supervised learning domain and finds application in pattern recognition, data mining.
This algorithm does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as GMM, which assume a Gaussian distribution of the given data).
We are given some training data, which classifies coordinates into groups identified by an attribute.
KNN model can implement by following the below steps:

- Load the data
- Initialize the value of k
- For getting the predicted class, iterate from 1 to total number of training data points
    - Calculate the distance between test data and the training data. Use Euclidean distance formula.
    - Sort the distances in ascending order based on distance values.
    - Get top rows from the sorted array.
    - Get the most frequent class of these rows.
    - Return the predicted class.

Disadvantages:

- Computationally expensive—because the algorithm stores all of the training data.
- High memory requirement.
- Stores all (or almost all) of the training data.
- Prediction stage might be slow (with big N).
- Sensitive to irrelevant features and the scale of the data.

iv. Improved Parallel FP Algorithm

We introduce a small files processing strategy in the FP-Growth algorithm, to compensate defects of low read/write

 Copyright 2019.All rights reserved.

speed and low processing efficiency for handling the massive small file datasets in Hadoop, and to enhance the access efficiency of HDFS and reduce the additional overhead of MapReduce. On the other hand, we use MapReduce to implement the parallelization of FP-Growth algorithm,
Traditional frequent item sets mining algorithms are unable to handle massive small files datasets effectively, we propose an improved Parallel FP-Growth (IPFP) algorithm.
IPFP algorithm is feasible and good speedup and a higher mining efficiency, and can meet the rapidly growing needs of frequent item sets mining for massive small files.
We can implement an IPFP model by following the below steps:
In this section, we propose the IPFP algorithm for mining frequent item sets in massive small files datasets in detail.
(1) Write a small files processing program—Sequence File. The Sequence File is used to merge all massive small files, which are composed of a large amount of transaction datasets stored in database, into a large transaction data file (transaction database).
(2) Equally divide the transaction database into several sub-transaction databases and then assign them to different nodes in Hadoop cluster. This step is automatically operated by database, when necessary we can use the balance command enabling its file system to achieve load balancing.
(3) Compute support count of each item in the transaction database by MapReduce, and then obtain the set of I_list from support count in descending order.

(4) Divide I_list into M groups, denoted as Group_list (abbreviated as G_list), and assign group_id for each group sequentially and each G_list contains a set of items.

(5) Finish the parallel computing of FP-Growth algorithm by MapReduce.
The Map function compares the item of each transaction in the sub-transaction database with the item in G_list. If they are same, then distribute the corresponding transaction to the machine associated with G_list. Otherwise, compares with the next item in G_list .
The Reduce function recursively computes the independent sub-transaction databases generated in step and then constructs the FP-tree.

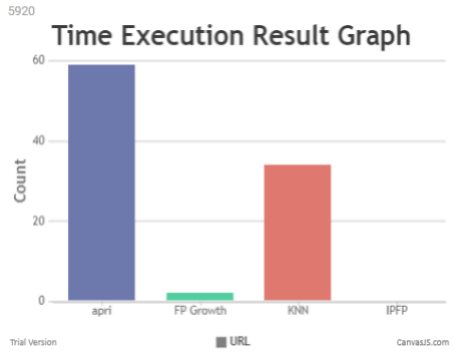
(6) Aggregate the local frequent item sets generated from each node in the cluster by Map Reduce, and finally get the global frequent item sets.

Advantages:

- It eliminates computation dependencies between machines and hence the communication between them.
- Computational time is linear.

III. EXPERIMENT AND RESULT

The data mining algorithms like FP-Growth, KNN, IPFP and Apriori algorithms are used here in this work. The performance comparison of the algorithms in web mining or in generating association rules or finding interesting patterns is done here. Here we have considered two parameters to compare the performance of FP-Growth, KNN, IPFP and Apriori algorithms. One of the two parameters is “time taken” for finding interesting patterns or association rules and other one is the “number of patterns” found. IPFP algorithm will find large number of interesting patterns; KNN takes more time than the fp. And the Fp-Growth algorithm will take less time than Apriori but it will result in very few numbers of interesting patterns. But IPFP is the most efficient data-mining algorithm in respect of number of interesting patterns than FP-Growth, Apriori and KNN as it finds more number of interesting patterns.



IV. CONCLUSION

The web data source used in web usage mining is the web-logs. It contained all the information that the user leaves when accessing the web pages. Web logs are preprocessed into a common log. After, Data mining techniques are applied to discover the interesting characteristics in the hidden patterns. Pattern Analysis is applied to the output of pattern discovery.
Used statistical methods such as classification, association rule mining discovery and statistical correlation analysis, which can find groups of web pages types that are commonly accessed together. Classification is used to map the data item into one of several the predefined classes. The class will belongs into one category such as sport or politics or education etc.
Web Usage Mining of Web Log Files results into interesting “Rules”, “Patterns”, and “Statistics”. In general the goal of web usage mining is to discover interesting patterns or navigational patterns or knowledge about web users. Here gathered information about web users page access patterns can be used for improving the web applications by prefetching and caching web pages. Pre-fetching and caching helps to reduce the page access time or fetching time or latency.

Future Scope
Web usage mining technique is used to mine or extract the knowledge from web logs generated on web server. The applications of web usage mining are like improving website design, improving performance of system, pre-fetching and caching. Web log data mining can done by using R and Python advanced language they can give the prediction of

 Copyright 2019.All rights reserved.

the data and weblog dataset we can use the more advance algorithm for mining data of weblog which will give more efficient result. In an organization administration officer wants to know that on which URL's his employee visit most of the time through out the day. From the data set of his web history with using some advance algorithm we can find out on which URL he or she visited most.

REFERENCE

[1] M. Perkowitz and O. Etzioni. Adaptive websites: an AI challenge. In Proc. 15th Intl. Joint Conf. on Art. Int., 1997.
[2] M. J. Pazzani and D. Billsus. Adaptive web site agentsIn Proc. 3rd Intl. Conf. on Autonomous Agents, 1999.
[3] Adomavicius G and Tuzhilin A. Using Data Mining Methods to Build Customer Profiles, IEEE Feb 2001.
[4] D. KeranaHanirex, Dr. M. A. DoraiRangaswamy International Journal on Computer Science and Engineering, "Efficient Algorithm for Mining Frequent Itemsets Using Clustering Technique, "vol. 3, no. 3, March 2011.
[5] Ketul B. Patel, Dr. A. R. Patel, "Process of web usage mining to find interesting patterns from web usage data, "www. Ijctonline.com vol. 3, no. 1, Aug 2012.
[6] R. Mishra, A. Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Vol 2, 2012.
[7] N. Anand, "Effective prediction of kid's behaviour based on internet use", International Journal of Information and Computation Technology, 2014.
[8] A.P. Singh, R. C. Jain, "A Survey on Different Phases of Web Usage Mining for Anomaly User Behaviour Investigation", IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Vol 3, 2014.
[9] S. Parvatikar and B. Joshi, "Analysis of User Behaviour through Web Usage Mining", ICAST - International Conference on Advances in Science and Technology, 2014.
[10] S. Jagan, and S.P. Rajagopalan, "A survey on web personalization of web usage mining", IRJETInternational Research Journal of Engineering and Technology, 2015.
[11] A. Ladekar, P. Pawar, D. Raikar and J. Chaudhari, "Web Log Based Analysis of User's Browsing Behaviour", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.
[12] Deepa, and P. Raajan, "An efficient preprocessing methodology of log file for Web usage mining", NCRIIAMI - National Conference on Research Issues in Image Analysis and Mining Intelligence, 2015.
[13] Saluja, B. Gour, and L. Singh., "Web Usage Mining Approaches for User's Request Prediction: A Survey", IJCSIT-International Journal of Computer Science and Information Technologies, Vol. 6 (3), 2015.
[14] "An Introduction to Real Time Processing and Streaming of Wireless Network Data", IJARCCE, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 1, ISSN 2319-5940, Jan 2015.
[15] "Strategies in Traffic Controlling Operations in India", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 8, August 2017, ISSN (Online) 2278-1021,ISSN (Print) 2319 5940,pp.210-213.
[16] "Energy Optimization Factors for the Embedded Mobile Processors", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, Issue 3, March 2017, ISSN (Online) 2278-1021, ISSN (Print) 2319 5940,pp.397-399.

 Copyright 2019.All rights reserved.