International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 02 | Feb 2019 www.irjet.net p-ISSN: 2395-0072

Enhancing Prediction of User Behavior on the basic of Web Logs

Ms. Mrunmayee R. Joshi[1], Dr. Prakash S. Prasad[2]

[1]*M.Tech Scholar, Dept. of CSE, PIET College, Nagpur, India*
[2]*Head of Department, Dept. of CSE, PIET College, Nagpur, India*

Abstract - *As an important part of discovering association rules, frequent item sets data mining plays an important role in mining associations, correlations and other important data mining tasks. Some traditional frequent item sets mining algorithms are unable to handle massive small files. 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 propose an improved Parallel FP-Growth algorithm. In particular, we introduce a small files processing strategy for massive small files datasets to compensate defects of low R/W speed and low processing efficiency in Hadoop. We use MapReduce to implement the parallelization of FP-Growth, thereby improving the overall performance of frequent item sets mining. The experimental results show that the IPFP algorithm is feasible and a higher mining efficiency, and can meet the rapidly growing needs of frequent item sets mining for massive small files datasets.*

Key Words: **Weblogs Dataset, User Behavior, Apriori Algorithm, KNN Algorithm, FP-Growth Algorithm, IPFP (Improved Parallel FP-Growth algorithm).**

1. INTRODUCTION

Un-stoppable growth of knowledge available on Internet makes the users to access the information day by day. It becomes much more difficult for users to access appropriate information efficiently. Analyzing and modeling web navigation behavior is helpful in understanding demands for online users. To predict User behavior, require relevant weblog data for processing. Web mining technique is use to extract the relevant data from the vast weblog data. Web mining is one of the techniques of data mining, which is used to extract and analyze useful information from web data. The categories of web mining are web content mining, web usage mining and web structure mining. Web content mining is the discovery of contents from web documents such as image, text, audio, video etc. Web structure mining focus on detecting the physical link structure of websites. Web usage mining examines the browsing activity.

The user accessing information from the websites is stored as diary of information also called as logs. The log contains series of user transactions which are often updated whenever the user accesses the website. The prediction of user behavior can be identified only through logs. The weblog contains unstructured format, so convert to raw weblog to processed weblog using data preprocessing, the data preprocessing includes Data Cleaning, User identification, Session Identification, content retrieval and path completion to get user navigation pattern.

Prediction of logs is the main activity. The prediction is analyzed by using the attributes and categories in weblog, which is based on user preference. The two types of prediction process one is online and another is offline with respect to web server activity. Offline data analyze historical data, such as log file or weblogs, which are captured from server. Weblog used in online whenever that user comes online for next time. Based on previous user navigation we can identify user behavior that is taken from weblogs.

1.1 Objectives

- We will collect the web logs of the user from browser history.
- Predicting frequency pattern of user related to its web behavior.
- We can recommend user products and services on the basis of that.
- High chances of conversion as we able to predict what he want.

2. RELATED WORK

- In 1997, M. Perkowitz and O. Etzioni [1] suggested that the user would visit the web server by current and past sessions. The algorithm suggest that first top m pages based on user's most visited current sessions, the system should analyze the max-path and min-path of navigated sessions, then only the visitor should easily navigate the path from the list of max list of path, by selecting min path, the algorithm select the best path from the list of frequent path.
- In 1999, M. J. Pazzani and D. Billsus [2] showed the examples of visitors suggest links of each user. The user visit the web server based on user interest, by the next time the browser will show the user suggest website based on user interest. The browser will follow the frequent viewed items, whenever the user will visit some other pages that

will show frequently of webserver links, and then only user can follow the frequently visited links.

- In 2001, Adomavicius and Tuzhilin[3] used the data mining method to mine the access records of the individual users, excavate the association rules and the user registration of the personal information constitute the user model. Sofia Stamou and Alexandros Ntoulas proposed a method of user interest modeling by analyzing query terms and web page subject information. Researchers such as Paul conducted user interest modeling through ODP classification system and data information.
- In 2011, D. Kerana Hanirex and Dr. M. A. Dorai Rangaswamy [4] showed classification of frequent item sets, semi-frequent item sets and In-frequent item sets and database transactions into clusters. Clusters are grouped based on the similar traversal pattern. Then it finds the frequent item set. Semi-frequent item sets and In-frequent item sets, the clusters reduce the number of transaction in the database and efficiency is improved.
- In 2012, Dr. A. R. Patel and Renata Ivancsy [5] Web usage mining predict the user navigation behavior based on the preferences in website, User navigation technique uses data preprocessing suggested that the weblog contains raw log format, so convert to unprocessed weblog to processed weblog using data preprocessing, the data preprocessing technique contains Data Cleaning, User identification, Session Identification, content retrieval and path completion to get user navigation pattern. After getting the processed log, the given log is converted in to sequence of patterns based on user's pattern.
- In 2012, Mishra R. and Choubey R. [6] describe the FP-growth algorithm is obtaining a most frequently access paths and pages from the web log data and providing valuable information to user behavior.
- In 2014, Anand N. [7] describes an Internet usage details and provides them with the tools to understand the online behavior of their teenage children.
- In 2014, Singh A.P. and Jain R.C. [8] Different kinds of web usage mining techniques with their basic models and concepts are provided.
- In 2014, Parvatikar S. and Joshi B. [9] this paper focused on Web Usage Mining is the user navigation patterns and their use of web resources. The different stages involved in this mining process and with the comparative analysis between the pattern discovery algorithms Apriori and FP-growth algorithm.
- In 2015, S.Jagan, and S.P. Rajagopalan [10] describe the web usage mining and algorithms used for providing personalization on the web. In this paper focused the data preprocessing and pattern analysis on the web and using the association rule mining algorithms.
- In 2015, Ladekar A. Pawar A. et al. [11] describe a web-mining algorithm that aims at amending the interpretations of the draft's output of association rule mining. This algorithm is being extremely used in web mining. The results obtained prove the robustness of the algorithm proposed in this paper.
- In 2015, Deepa and Raajan [12] implemented the preprocessing techniques to convert the log file into user sessions, which are suitable for mining and reduce the size of the session file by filtering the least requested pages using the preprocessing technique.
- In 2015, Avneet Saluja et al. [13] in their work is user future request prediction using web log records and user information. The purpose of the effort is to provide a benchmark for evaluating a various methods used in the past, a present and which can be used in a future to reduce the search time of a user on the network.

3. PROPOSED WORK FLOW

1. Web-log data.
2. Data Preprocessing
    a. Data Cleaning
    b. User Identification
    c. Session Identification
3. Apriori Algorithm
4. KNN Algorithm
5. FP-Growth (FP) algorithm
6. Improved Parallel FP-Growth Algorithm (IPFP)
7. Result

4. PROPOSED METHODOLOGY

1. Weblog Data

A Weblog is a Web site that consists of a series of entries arranged in reverse chronological order, repeatedly updated on frequently with new information about particular topics.

2. Data Preprocessing

This technique involves removing of the unwanted data and splitting of data into a structured file format. All

server log files do not have the right format, so there is a essential for preprocessing technique.

a. Data Cleaning - It is the method of recognizing and correcting imprecise records from tables, datasets.

b. User Identification - The user gets identified on the basis of client IP address, user name, requested URL, date, time, server IP address etc. Log file format- Internet Information Service (IIS) records the data.

c. Session Identification - The important operation of navigation pattern mining is to cluster the session. The navigation pattern mining arranges the user based on the session.

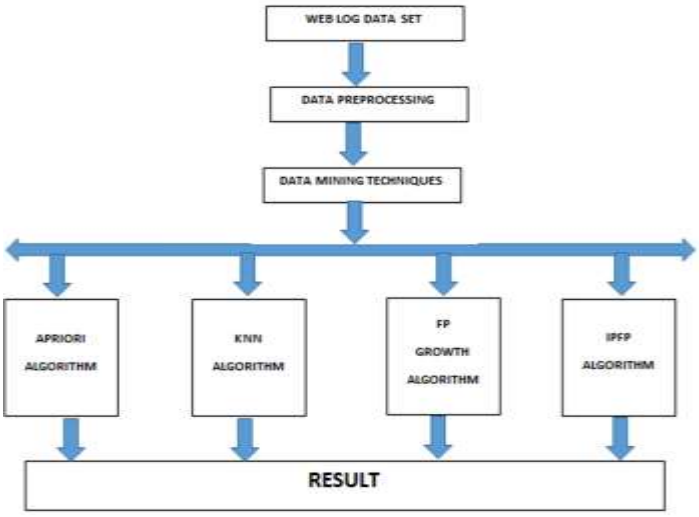

Fig -1: Flow of Design Mode

3. Apriori Algorithm

Apriori algorithm developed by Agrawal and Srikand is the first algorithm to create all frequent item sets and confident association rules. Presently after that, the method was corrected and retiled Apriori by Agrawal. Apriori algorithm follows a "Bottom-Up Approach" where frequent item subset are extended one item at a time.

Apriori is designed to operate on database containing transaction (For ex. Collections of items bought by customer or details of website frequentation).

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.

4. KNN Algorithm

K-Nearest Neighbors is one of the most basic yet important classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds powerful application in pattern detection, data mining and intrusion recognition. It is commonly disposable in real-life scenarios since it is non-parametric.

Its purpose is to use a database in which the data points are separated into some classes to predict the classification of a new data points.

5. FP-Growth Algorithm

The FP-growth algorithm creates numerous data sets from FP Tree by navigating in a bottom up approach.

The algorithm finds frequent item sets that can be used to extract association rules. This is done using the support count of an item set. This frequent information allows for the effective discovery of numerous data sets.

The main idea behind the algorithm is follows a divide and conquer strategy such as it compress the database which provides the frequent sets and then divide this compressed database into a set of conditional databases, each associated with a often set and apply data mining on every database.

5. PROPOSED SYSTEM

IPFP ALGORITHM

IPFP algorithm for mining numerous itemsets in enormous small files datasets.

1) Write a small files processing program—Sequence File. The Sequence File is used to merge all enormous small files, which are composed of a large amount of transaction datasets stored in HDFS, into a large transaction data file (transaction database).
2) Similarly divide the transaction database into several sub transaction databases and then assign them to different nodes in Hadoop cluster. This step is automatically divided by HDFS, when necessary we can use the balance command enabling its file system to achieve load balancing.
3) Calculate 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 successively and each G_list contains a set of items.
5) Complete the parallel calculating of FP-Growth algorithm by MapReduce. The Map function relates 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. Finally, the independent sub-transaction databases corresponded to G list will be produced. The Reduce function recursively calculates the independent sub-transaction databases generated in step , and then constructs the FP-tree. This step is like the process of traditional FP-tree generation, but the difference is a size K maxheap HP that stores frequent pattern of each item.
6) Aggregate the local numerous itemsets generated from each node in the cluster by MapReduce, and finally get the global numerous itemsets.

6. CONCLUSIONS

- In this paper, owing to the small files processing strategy, the IPFP algorithm can reduce memory cost greatly and improve the productivity of data access, thus avoids memory overflow and reduces I/O overhead.

- Meanwhile, the IPFP algorithm is migrated to the MapReduce environment, which can complete frequent item sets mining efficiently and thus enhance the overall performance of FP-Growth algorithm.

IPFP algorithm can make a breakthrough where PFP algorithm has its shortcomings in handling massive small files datasets, and has a good speed up and a higher mining efficiency.

REFERENCES

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. Kerana Hanirex, Dr. M. A. Dorai Rangaswamy 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 behavior 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 Behavior Investigation", IJETTCS - International Journal of Emerging Trends & Technology in Computer Science, Vol 3, 2014.

9) S. Parvatikar and B. Joshi, "Analysis of User Behavior 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 Behavior", IJCSIT - International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015.

12) A. 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) A. 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.

BIOGRAPHIES

Miss. Mrunmayee R. Joshi is a M.Tech Student in the Department of Computer Science & Engineering, Priyadarshini Institute of Engineering and Technology, Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur.

Dr. Prakash S. Prasad is working as Professor and Head of Computer Science and Engineering at Priyadarshini Institute of Engineering and Technology, Nagpur . He is having 20 Years of experience in the field of teaching to engineering students. He completed his engineering in the year 1997, master of engineering in 2006 and PhD in 2014.