International Journal of New Technology and Research (IJNTR)
ISSN:2454-4116, Volume-4, Issue-2, February 2018 Pages 07-11

Pattern Discovery Using Association Rule Mining on Clustered Data

Htun Zaw Oo, Nang Saing Moon Kham

***Abstract*—Many organizations use Word Wide Web as a multipurpose platform during these days. It is very important to understand how a website is being used by users. Web Usage Mining is the one of the data mining technique that aims to discover interesting usage patterns from web log data. It also helps effective website management, creating adaptive websites, personalization and so on. In this paper, the aim is to find frequent user access pattern from web log entries. Combined effort of clustering and association rule mining is used to apply for pattern discovery.**

***Index Terms*—Apriori, Association Rule Mining, Clustering, DBSCAN, Web Usage Mining.**

I. Introduction

Web mining is the application of data mining techniques to discover patterns from the World Wide Web. As the name proposes, this is information gathered by mining the web. It makes utilization of automated apparatuses to reveal and extract data from servers, and it permits organizations to get to both organized and unstructured information from browser activities, server logs, website and link structure, page content and different sources. Web mining can be divided into three different types – Web usage mining, Web structure mining and Web content mining [9].

A. *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 website. Web usage mining process is divided into three phases: 1) pre-processing, 2) pattern discovery and 3) pattern analysis.

B. *Web Structure Mining*

Web structure mining uses graph theory to analyze the node and connection structure of a website. According to the type of web structural data, web structure mining can be divided into two kinds: i) extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location ii) mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

Htun Zaw Oo, University of Computer Studies, Yangon, Myanmar,
Nang Saing Moon Kham, Faculty of Information Science, University of Computer Studies, Yangon, Myanmar

C. *Web Content Mining*

Web content mining is the mining, scanning and extraction of text, videos, graphs and pictures from web documents. It is also known as text mining. Web content mining analyzes the content of web resources. Most of the data available on the web is unstructured data. Two different points of view of web content mining are: the information retrieval view and the database view. The main goal of content mining from information retrieval view is to improve the filtering and finding of the information to the users. The main goal of database view is to manage the web data [1].

II. Problem Statements

United Nations High Commissioner for Refugees (UNHCR) is a United Nations programme mandated to protect and support refugees at the request of a government or the United Nations itself and assists in their voluntary repatriation, local integration or resettlement to a third country. The agency has been using operational data portal - ODP (http://data.unhcr.org) [5] as a main inter-agency data and information sharing platform since 2012 and a new version of the website (http://data2.unhcr.org) [5] is being developed to provide better service to users. Therefore, it is very important to understand how the current website is used by the users. Discovering frequently used access patterns, the organization can focus on those pages and the reason for "why it is used" can be revealed. On the other hand, the pages mostly not can be improved and the reason behind its non-use can be identified. Web usage mining helps site administrators and web developers re-design the website layout efficiently and make it more user friendly and also customize the pages based on users' interest.

III. Related Works

Market basket analysis is one of the data mining methods focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Market basket analysis determines the products which are bought together and help the super-markets and stores reorganize the placement of the product items, and also support strategic marketing decisions such that products' purchase can be improved. Hence, the Market consumer behaviors need to be analyzed, which can be done through different data mining techniques [2].

An outlier is defined as data point which is very different from the rest of the data based on some measure. Such a point

often contains useful information on abnormal behavior of the system described by data [3]. On the other hand, many data mining algorithms in the literature find outliers as a side product of clustering algorithms. From the viewpoint of a clustering algorithm, outliers are objects not located in clusters of dataset, usually called noise [4]. Outlier detection problem is one of the very interesting problems arising recently in the data mining research [8].

Sequential Pattern Mining is the method of finding interesting sequential patterns among the large databases. It also finds out frequent subsequences as patterns from a sequence database. Enormous amounts of data are continuously being collected and stored in many industries and they are showing interests in mining sequential patterns from their database. Sequential pattern mining has broad applications including web-log analysis, client purchase behavior analysis and medical record analysis. Sequential or sequence pattern mining is also the task of finding patterns which are present in a certain number of instances of data. The identified patterns are expressed in terms of sub-sequences of the data sequences and expressed in an order that is the order of the elements of the pattern should be respected in all instances where it appears [6].

IV. CLUSTERING TECHNIQUES

Clustering often called unsupervised learning is the process of organizing data instances into groups whose members are similar in some way. A cluster is therefore a collection of data instances which are similar to each other and are dissimilar to data instances in other clusters. In the clustering, a data instance is also called an object as the instance may represent an object in the real world. It is also called a data point as it can be seen as a point in an r-dimension space, where r is the number of attributes in the data. Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity. Many different clustering techniques exist in the literature. In this paper, density based spatial clustering applications with noise (DBSCAN) is used for various reasons. It does not require a-priori specification of number of clusters. It is able to identify noise data while clustering and also able to find different size and shaped of clusters.

A. Density Based Spatial Clustering of Applications with Noise (DBSCAN)

It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. It requires two parameters: ε (eps) and the minimum number of points required to form a dense region (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point and hence be made part of a cluster. If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise [10]. DBSCAN can be used with any distance function [11] as well as similarity functions or other predicates [12]. In this paper, Jaccard distance function is used to measure the distance among the users.

V. ASSOCIATION RULE MINING

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Let $I = \{i1, i2, \ldots, im\}$ be a set of items. Let $T = (t1, t2, \ldots, tn)$ be a set of transaction in the database, where each transaction t_i is a set of items such that $t_i \subseteq I$. An association rule is an implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \phi$. X (or Y) is a set of items, called an itemset. A transaction $t_i \in T$ is said to contain an itemset X if X is a subset of t_i. The support count of X in T (denoted by X.count) is the number of transactions in T that contain X. The strength of a rule is measured by its support and confidence. *Support*: The support of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contains $X \cup Y$, and can be seen as an estimate of the probability, $Pr(X \cup Y)$. The rule support thus determines how frequent the rule is applicable in the transaction set T [13]. Let n be the number of transactions in T. The support of the rule $X \rightarrow Y$ is computed as follows:

$$support = \frac{(X \cup Y).count}{n} \quad (1)$$

Support is a useful measure because if it is too low, the rule may just occur due to chance. *Confidence*: The confidence of a rule, $X \rightarrow Y$, is the percentage of transactions in T that contain X also contain Y. It can be seen as an estimate of the conditional probability, $Pr(Y \mid X)$. It is computed as follows:

$$confidence = \frac{(X \cup Y).count}{X.count} \quad (2)$$

Confidence thus determines the predictability of the rule. If the confidence of a rule is too low, one cannot reliably infer or predict Y from X. A rule with low predictability is of limited use. The best known algorithms for mining association rules are Apriori, AprioriTID, STEM, DIC, Partition-Algorithm, Elcat, FPgrow, etc.

A. Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It

proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database [7]. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. In web usage mining, association rules are used to discover pages that are visited together quite often. Knowledge of these associations can be used either in marketing and business or as guidelines to web designers for (re)structuring Websites. Transactions for mining association rules differ from those in market basket analysis as they cannot be represented as easily as in market basket analysis. Association rules are mined from user clusters containing client IP address, user id, and a set of URLs. As a result of mining for association rules the rule, for example: X,Y → Z (c=85%, s=1%) could be generated. This means that visitors who viewed pages X and Y also viewed page Z in 85 % (confidence) of cases, and that this combination makes up 1% of all transactions in preprocessed logs.

VI. ARCHITECTURE OF THE SYSTEM

The proposed system as shown in the below figure will be implemented using the combination of clustering and association rule mining techniques. There are three major steps:

i) Pre-processing of web log data.
ii) Cluster users who share the similar access patterns.
iii) Generate frequently user access patterns from each cluster.

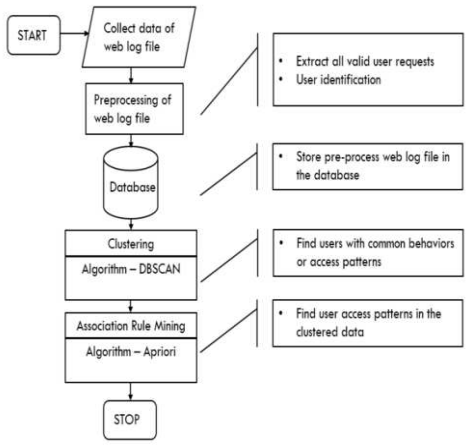

Fig. 1. Proposed system architecture for pattern discovery

As a first step, web log data should be collected from a web server. Normally, the sample web loge file include the below data such as

a) client IP address
b) user identity
c) authenticated user name
d) timestamp
e) requested URL
f) HTTP status code
g) response size
h) referrer of URL
i) user agent.

Once the data is collected, it need to be pre-processed to remove the data which is not necessary for pattern discovery process. The raw web log data may include the log entries that are not explicitly requested by users, for instance java script file, css file, other image files and other records like unsuccessful transactions. These unnecessary data must be removed before user identification process. And also those data which will be feed into the pattern discovery process need to be cleaned by removing query parameters.

Algorithm Name: Data Cleansing of Web Log File
Input: Web Server Log file
Output: Log Database
Step1: Read Log Record from Web Log File
Step2: If(Log Record .url contains(gif, jpeg, jpg, css)) OR (different error like HTTP 404 or more) found then
Remove from web log file.
else
Clean URL.
end if
Step3: Repeat the above two steps until
EOF(Web Log File)
Step4: Stop the process.

After data cleaning, user for each web log entry will be identified using two parameters, client IP address and user agent. If two web log entries have the same client IP address and user agent, it is assumed that they are the same users. If they are not the same, they are different users.

Algorithm Name: User Identification
Input: Processed Web Log File.
Output: Number of Distinct User.
Step1: Read records from Web Log File.
Step2: User's IP addresses of two consecutive entries are compared.
Step3: If (IP address is same) then
check user's browser and operating system
if both are same then consider same user.
else consider new user.
end if
end if
Step 4: Repeat above 2 steps until EOF (Web Log File).

After identifying the user for each log entry, the users are grouped according to their similarity access patterns. To measure the similarity between users, Jaccard distance is used. It is applicable to be used for binary and asymmetric variables. DBSCAN is used to group the users and it needs two parameters, the radius of the cluster and minimum number of points to form a cluster. After clustering, those users who share the similar access patterns will be grouped together and those users who don't have any similarity with other users will be identified as noise.

To discover frequent user access patterns, association rules are generated from each cluster. Apriori algorithm which is one of the popular algorithm to generate association rule mining is used. The user have to provide two parameters: support and confidence. The rules those are conform to

support and confidence are extracted and presented to the user.

VII. IMPLEMENTATION OF THE SYSTEM

The system implementation starts collecting the web log data from data.unhcr.org for 30 days. Each log entry includes nine attributes that was mentioned in section 6.

1. *157.55.39.246 - - [21/Mar/2016:06:34:33 +0100] "GET /car/regional.php/admin/rss/rss/download.php?id=566 HTTP/1.1" 200 33545 "-" "Mozilla/5.0 (compatible; bingbot/2.0; +http://www.bing.com/bingbot.htm)"*
2. *66.249.78.176 - - [21/Mar/2016:06:34:33 +0100] "GET /syrianrefugees/regional.php/js/flood.html HTTP/1.1" 500 61041 "http://data.unhcr.org/syrianrefugees/regional.php/flood.html" "Mozilla/5.0 (compatible; Googlebot/2.1;*

Fig. 2. Sample web log file.

The URLs including java script, css and image files are removed and also the log entries which are flagged as unsuccessful are removed. If HTTP status code is equal to 404 or more, then that transactions are considered as unsuccessful. After cleaning the data, the data indicated that it is reduced to 3.4 million records from 30.8 million which is 89% decreased from the raw web log data.

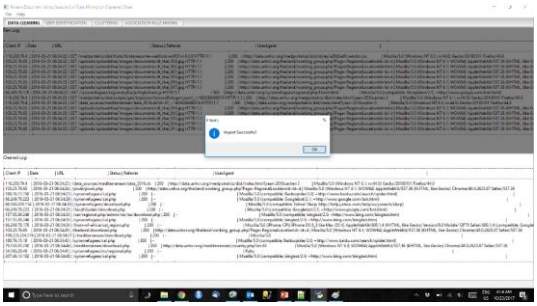

Fig. 3. Raw versus clean log file.

User for each log web entry will be identified using remote IP address and user agent variables. If two web log entries have the same data, then it is considered as the same user. If not, they are different and an identifier for a new user will be used.

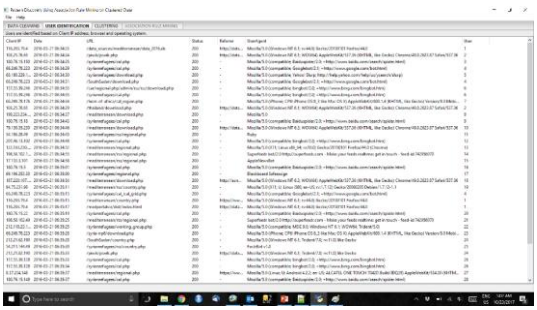

Fig. 4. User identification.

Now, the data is ready for clustering process. DBSCAN uses two parameters. One is for epsilon, the radius of the cluster and another parameter is for the minimum number of points which can be used to form a cluster. User can choose optimal parameters by looking at the purity score that displays after clustering. The output of this process that the number of clusters with users and also identify the noise which doesn't belong to any cluster.

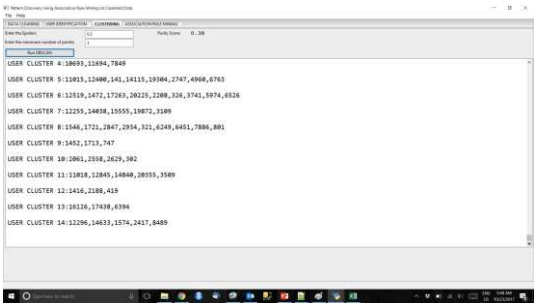

Fig. 5. DBSCAN clustering.

After clustering process, association rule mining function is called to generate more relevant and concise rules for each cluster. Apriori algorithm is used by passing minimum support and confidence parameter, those explain how frequent a rule is and how strong a rule is. If the rules will be generated along with relevant support and confidence variable.

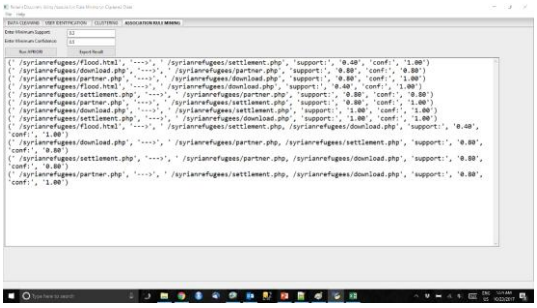

Fig. 6. Association rule mining with Apriori algorithm.

VIII. EVALUATION OF THE SYSTEM

The following evaluation factors has been made based on the system with the following specification – Core i5 processor, 8 GB (RAM), 64bit window OS.

A. Processing Time Evaluation

Using 66,654 sample data set, 0.3 epsilon and minimum number of points (3) for DBSCAN, 0.5 for both minimum support and confidence for Apriori, running time for Apriori on clustered data is 4.7 times faster than that of non-clustered data. When taking into account 0.313 seconds running time for DBSCAN, association rule mining (ARM) on clustered data is 2.6 times faster than mining on non-clustered data.

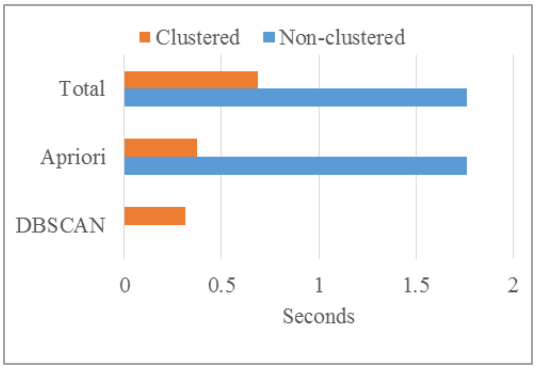

Fig. 7. Processing time evaluation Clustered vs. Non-clustered.

B. *Generating Relevant and Concise Rules*

One of the problem of association rule mining is generating too many rules and users are able to find out which rules are more interesting. This phenomenon is called interesting problem in literature. To overcome interesting problem, the below analysis shows that clustered association rule mining generate less rules to user with high minimum support and confidence where non-clustered association rule mining doesn't generate any rules at all.

Min Support	Min Confidence	Non-Clustered ARM	Clustered ARM	Variation
0.001	0.001	1,487	34	-4,274%
0.001	0.01	1,486	34	-4,271%
0.001	0.2	502	34	-1,376%
0.01	0.01	2	34	94%
0.1	0.1	0	34	100%
0.5	0.5	0	9	100%

Fig. 8. Number of rules generated on Clustered vs. Non-clustered.

IX. Conclusion

Web usage mining techniques are great area of research these days. Providing users easily what they are looking for in websites is the ultimate goal of web usage mining. In this paper, this goal is fulfilled by using association rule mining technique on clustered data. Association rule mining may have drawback of generation of irrelevant rules, generation of too many rules leading to contradictory prediction resulting in reduction of accuracy. Clustering reduces data set for association rule mining and produces relevant frequent access patterns from each cluster. This paper also proved that using association rule mining on small data cluster takes less time than running association rule mining on the whole big data set.

References

[1] Srividya, M., D. Anandhi and M. I. Ahmed. "Web Mining and its categories–a survey." International Journal of Engineering and Computer Science, IJECS 2.4 (2013).
[2] Loraine Charlet Annie M.C. and Ashok Kumar D "Market Basket Analysis for a Supermarket based on Frequent Itemset Mining", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012.
[3] Aggarwal, C. C., Yu, S. P., "An effective and efficient algorithm for high-dimensional outlier detection, The VLDB Journal, 2005, vol. 14, pp. 211–221.
[4] Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying densitybased local outliers.", ACM Conference Proceedings, 2000, pp. 93-104.
[5] Operational Data Portal, The Refugees Operational Portal is a Partners coordination tool for Refugee situations provided by UNHCR.
[6] Ms. Pooja Agrawal, Mr. Suresh kashyap, Mr.Vikas Chandra Pandey and Mr. Suraj Prasad Keshri, An Analytical Study on Sequential Pattern Mining With Progressive Database, International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 3, May 2013.
[7] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", TheMorgan Kaufmann, ISBN 13: 978-1-55860-901-3.
[8] Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad and Mustafa Mat Deris, "A Comparative Study for Outlier Detection Techniques in Data Mining", July 2006.
[9] https://en.wikipedia.org/wiki/Web_mining
[10] https://en.wikipedia.org/wiki/DBSCAN
[11] Schubert, Erich, Sander, Jörg, Ester, Martin, Kriegel, Hans Peter and Xu, Xiaowei (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". ACM Trans. Database Syst. 42 (3), ISSN 0362-5915.
[12] Sander, Jörg, Ester, Martin, Kriegel, Hans-Peter and Xu, Xiaowei (1998). "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications", Data Mining and Knowledge Discovery. Berlin: Springer-Verlag. 2 (2): 169–194.
[13] Bing Liu, "Web Data Mining', Springer, ISBN 978-3-642-19459-7.