ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

User Data Classification with Web Access Behavior Analysis Using Wi-Fi History

D. Priyadarshini[1], Sree Dhanya C[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[1]

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, K.G.Chavadi, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: In recent trends, mobile devices are very popular and the growth of those mobile devices is tremendous in nature. Due to this marvelous growth, data analysis and user preference and interest detectingon are very new and interesting. The Proposed System helps to detect anomaly and user preferences by using Wi-Fi logs from mobile devices. The analysis of user behavior from Wi-Fi logs are much complicated because the Wi-Fi logs contains many auxiliary information and noises. There is an enormous challenge to perform the elimination of auxiliary information's and performing user summary from Wi-Fi logs. In this proposal, there are 3 algorithms are used after data cleaning for enabling the user preference and anomaly detection. The **Access Pattern Analysis (APA)** analysis Algorithm is proposed to identify user and session are very important for identifying behavioral patterns. The time taken for identifying user and session are considerably reduced due to the effect of APA. Improved Expectation Maximization (IEM) clustering algorithm is proposed to help in identifying very relevant similar groups. The similarities between user access and their behaviors are identified using IEM. The proposed **Multi Label PCA** strategy is used for finding and pruning unwanted SSID and log details. And it also helps to detect every pattern from APA. The experiments show the Proposed System can effectively find the location based user preferences and anomalies with Wi-Fi and Web logs.

KEYWORDS: Web Mining, WebLog Mining, Wi-Fi Log Mining.

I. INTRODUCTION

With more than two billion pages being created by millions of Web page authors and organizations, the World Wide Web (WWW) is growing tremendously as a rich knowledge base. Even from the unique characteristics of Web like hyperlink structure and its content and languages the knowledge can be extracted. Analysis of these characteristics often exposes interesting patterns and new knowledge. Such knowledge can be used to enhance users' efficiency and usefulness in searching for information on the Web and also for using them in applications unrelated to the Web, like support for decision making or business management [1]. The intention of Web Mining is to detect useful information or knowledge from the Web page content, hyperlink structure along with usage data. Various Data Mining techniques used by Web Mining includes Supervised Learning (or classification), Unsupervised Learning (or clustering), association rule mining and sequential pattern mining [2].Web data are data that can be gathered and used in the framework of Web personalization. Such data are grouped in four categories based on Content data, Structure data, Usage data,and User profile data[3].The growth of World Wide Web is tremendous in the past few years with its usage in small research communities to the biggest and most popular way of communication and propagator of information. The Web grows daily on an average of a million electronic pages, being added to the hundreds of millions already available. The Web also acts as a platform for exchanging different kinds of information that include software, research papers, educational content and multimedia content. The uninterrupted expansion in the size and use of the WWW insists for new methods which could process these massive amounts of data effectively. Because of its swift and muddled expansion, the network of information that is resulting from it is a deficit of organization and structure. Besides that, the content is also published in various distinct formats. This fact forces the users to feel disoriented sometimes, lost in the information overload that continues to enlarge over a period of time[4][5].

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

II.PROBLEM DEFINITION

There are several previous works for finding user preference from Web log. But there is an only one paper, which handles Wi-Fi logs to detect user preferences [6]. Analyzing and preprocessing the Wi-Fi logs is very challenging task. Dimensionalityproblem: arise when analyzing and organizing data in high-dimensional spaces. The Maximum coverage problem: computational complexity and this type of process are more suitable in the theoretical way and it needs more resource when they implementing practically. Additionally, Many Wi-Fi access points are named by default settings or without any valid names and meanings. However, the human behaviors are not random, e.g., people visit restaurants around noon, go for work in the daytime, and stay at home at night. This situation may vary if you take passenger data. Namely, this can make use of the visiting patterns of the users to a place to infer the type of the place. Without the proper user Wi-Fi logs this can't annotate the types of the places from the SSIDs without semantics. And this also creates set cover problem at the time deployment.

III.PROPOSED SYSTEM

In the current situation, Wi-Fi based Web access is very useful for almost every activity. So there is a rapid development of mobile and internet in its volume of traffic and the size and complexity of Web sites. The task of Web Mining is the application of Data Mining process, and so on to the Web data and traces user access behaviors and extracts their preference using usage history.

Contributions of the Proposed System

A Wi-Fi server usually registers a history entry for every access of a mobile user across the network. First, raw Wi-Fi history data needs to be cleaned, condensed and transformed in order to retrieve and analyze significant and useful information. Second, pattern mining can be performed on history records to find association patterns, sequential patterns and trends of Web accessing. The followings are the contributions of the Proposed System.

- The **Access Pattern Analysis (APA)** analysis Algorithm is proposed to identify user and session are very important for identifying behavioral patterns. The time taken for identifying user and session are considerably reduced due to the effect of APA.
- IEM (**Improved Expectation Maximization)** clustering algorithm is proposed to help in identifying very relevant similar groups. The similarities between user access and their behaviors are identified using IEM.
- The proposed **PCA** strategy is used for finding and pruning unwanted Service set ID and history details. And it also helps to detect every pattern from APA.

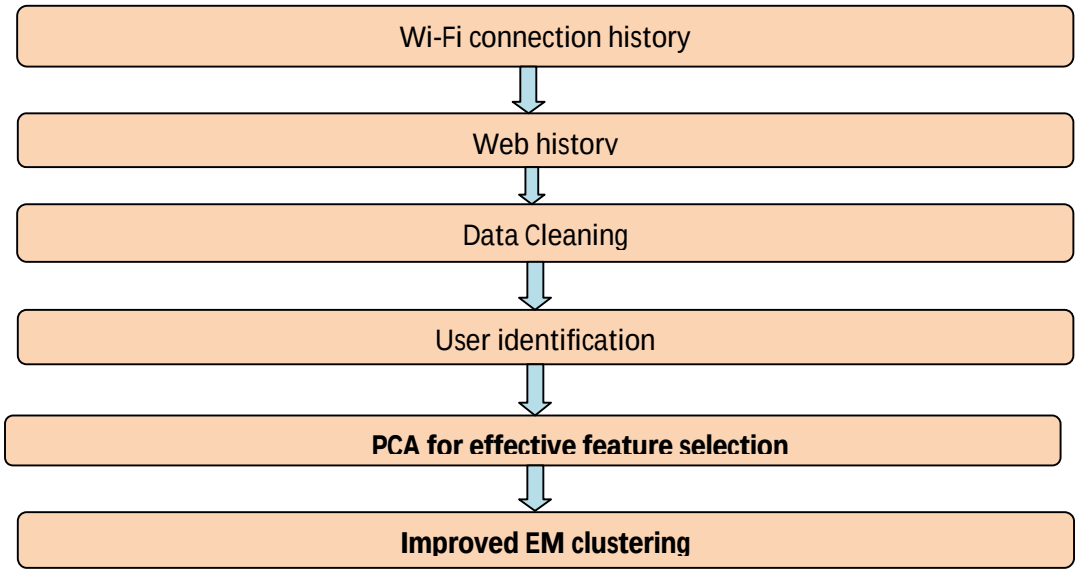

Figure 1.0 The overall process involved with the Proposed System

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

A. Wi-Fi history

Once the dataset was collected from the Wi-Fi router and Web server the raw history file has to undergo some preprocessing works. The history file has to be cleaned by removing the duplicate, improper and missing value entries. The Service set ID information also be included in the history file, those entries are to be removed. The images and the style sheet entries are also removed. Then the user identification based on the IP address is performed. The session identification based on the duration of entry and exit time is identified. In the last stage of preprocessing technique is the feature extraction process in which the feature sets are extracted using the methods

- Service set ID extraction
- Service set ID service details extraction

B.Preprocessing Phase

Data Pre-processing is an important step to filter and organize the appropriate data before applying any algorithm. Wi-Fi data Pre-processing increases the quality of available data by reducing the log file size[7]. The primary use of data Pre-processing is to improve data quality and increase mining accuracy. Pre-processing consist of following steps

- Field Extraction
- Data Cleansing
- User Identification
- Session Identification
- Semantic SSID detection

In this proposed, it will provide an overview of the Wi-Fi log and its feature. main task is to ”clean” the raw Web log files [8] and apply the clustering technique to find the quality of log file data and pattern analysis. So the main steps of this phase are:

- Extract the Wi-Fi logs that collect the data on the Wi-Fi routers.
- Clean the Wi-Fi logs and eliminate the redundant information and inappropriate SSID.
- Applying different Data Mining Techniques for finding the description of SSID

The raw Web log data after Pre-processing and cleansing could be used for pattern discovery, moving pattern analysis, data usage analysis, and generating user profiles.

a) Field Extraction

Each connection and Web log entry is represented as a single line of the log file. The log entry contains many fields as discussed in the earlier section which have to be separated taken for the preprocessing step. The filed extraction is the phase of separating the field from the single line of the server log file. The server used different characters which work as separators. The most used separator character is '/', ‘tab’ and ‘‘character.

b) Delimiter based Field Extraction algorithm:

Input: Wi-Fi connection and Web Log Files, delimiter L

Output: preprocessed data file (PDF)

Steps:

1. Open Log File
2. Read all fields contain in Server Log Files
3. Separate out the Attribute using the delimiter L
4. Extract all fields and Add into the PDF
5. Close and end phase.

In Data Cleansing, log entry involves irrelevant references to the objects like multimedia files which may not be

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

necessary for analyzing purposes. Therefore such kind of useless entries has to be eliminated from log files before performing any analysis phase. By performing Data cleansing phase, errors and inconsistency will be discovered and eliminated to improve the quality of data.

Algorithm: Data Cleansing at the time of SSID and URL extraction

1. Read Entries in PDF (Pre-processed Data File)
   a. For each Entry in PDF
   b. Read fields (ssid)
2. If ssid !=‘ ’and mainilink!= ‘’ Then
   a. Get IP_address and URL_link
   b. If suffix.URL_Link= {*.gif,*.jpg,*.css} Then
   c. Remove suffix.URL_link
   d. SSID,Ip address and URL_Link
   e. End if
3. Else
   a. Next Entry
4. End if

Algorithm1:Data Cleansing at the time of SSID and URL extraction

C.SSID Semantic Enrichment

An SSID (Service Set Identifier) is typically a very short string, which can be extracted from the Wi-Fi log. It is a small set of string, which denotes the Wi-Fi hotspot. A fundamental idea to the study is that it makes use of the conservapedia API, to enrich the semantics of the SSID. If the SSID description not retrieved by the above API then the semantic data is extracted from the Web. With the help of conservapedia API, it can readily expand the meaning of a given SSID and its semantic features. For example, if the SSID “nthu” is input into the API, it can obtain Web documents regarding National TsingHua University, and if the SSID “McDonalds” is emanated, it obtains Web documents of McDonalds descriptions and finds the category of the given SSID. Therefore, with the employment of the Web search conservapedia API, a short, abbreviated SSID string can be expanded into Web documents, which should be more informative than the original SSIDs.

Technique: SSID data extraction and filtering

Input: SSID

Output: extracted SSID description and semantic keywords

Steps:

1. Read the SSID Q.
2. Get the citations Ct
3. Read the page source
4. Extract the page source Ps. And store into a temp location
5. Find auxiliary data. Au from Ps.
6. Use the training dataset Ts to extract desired data from the Ps.
7. Get semantic labels and links
8. Update the results.

Algorithm2:SSID data extraction and filtering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

D .Clustering Models

The similar user behaviors are grouped and categorized based on their SSID preference this phenomenon is known as clustering process. It is an unsupervised learning technique, thus does not need any labeled training samples. The successive clustering methods can be found by density based algorithms using neighbor details. These algorithms are commonly follows the neighbor details and calculate the distance to perform the clustering. The Proposed System used improved I-IEM (**Improved Expectation Maximization)** clustering.

The IEM algorithm is mainly divided into three steps:

1. Extraction of the structure information from the Weblog dataset:
   - Create a neighborhood graph to connect each object to its n neighbors in the cluster and in the given datasets D;
     a. Based on the IP, thus neighborhood graph will be constructed.
   - Calculate a density for each object based on its neighbors to its KNN; number of neighbors for each user based on IP and query.
   - Users and Objects are to be classified into 3 types:
     a. Cluster Supporting Object (CSO): object with density higher than all its neighbors; Based on the Weblog, the system finds highest frequency link.
     b. Cluster Outliers: object with density lower than all its neighbors, and lesser than a predefined threshold;
     c. the rest are grouped as the other category

2. Local/Neighborhood approximation of Principle object memberships:
   - Initialization of Principle object membership:
     a. Each CSO is to be assigned with fixed and full membership to itself to represent one cluster;
     b. All outliers are assigned with the fixed and full membership to the outlier group;
     c. The rest are assigned with the equal memberships to all clusters and the outlier group;
   - Then the Principle object memberships of all type 3 objects are updated by a converging iterative procedure calledNeighborhood/Local Approximation of Principle object Memberships, in which the Principle object membership of each object is updated by a linear combination of the Principle object memberships of its nearest neighbors.
3. Cluster construction from Principle object memberships in two possible ways:
   - One-to-one object-cluster assignment, to assign the each object to the cluster in which it has the highest membership;
   - One-to-multiple object-clusters assignment, to assign the each object to the cluster in which it has a membership higher than a threshold.

IV.IMPLEMENTATION AND RESULTS

A.DATASET

To evaluate the Proposed System there implementation dynamic synthetic datasets have been used. The proposed experiment shows the difference between the existing clustering algorithm and proposed clustering algorithm in Web personalization techniques. The first synthetic data set used in the experiments is the dataset1 which collected from the Website and different Wi-Fi users.Based on the real world dataset shown in figure 2.0, 3.0.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

Time	ssid	wlevel	bssid	sid
12/14/2...	Edimax	-72	d8:c7:c8:79:cb:d2	1
12/14/2...	Dominos	-72	5c:63:bf:c9:84:9a	2
12/14/2...	McDonalds	-74	64:66:b3:4c:6b:80	3
12/14/2...	Dominos	-73	74:d0:2b:88:6d:1c	4
12/14/2...	wenshan	-94	00:13:f7:1b:c8:63	5
12/14/2...	Edimax	-90	f8:d1:11:75:54:5a	6

Figure 2.0 Wi-Fi Connection dataset sample

sno	sid	mainlink	sublinks	ipadd	datetime	uname
1	1	https://in.yahoo.c...	../https://in.yaho...	192.168.1.86	11/13/...	kalai
33	2	http://www.code...	.//	192.168.1.32	11/14/...	hema
71	9	https://in.yahoo.c...	http://movies.ndt...	192.168.1.96	11/13/...	hema
129	5	https://in.yahoo.c...	../https://in.movi...	192.168.1.47	11/13/...	hema
182	27	https://in.yahoo.c...	../https://in.yaho...	192.168.1.2	11/13/...	hema
241	24	https://in.yahoo.c...	http://portal.kipli...	192.168.1.7	11/13/...	hema
343	2	http://www.code...	.//search.aspx?sb...	192.168.1.23	11/14/...	hema
442	8	http://www.code...	.//search.aspx?sb...	192.168.1.40	11/14/...	hema
543	16	https://in.yahoo.c...	../https://cricket....	192.168.1.32	11/14/...	hema

Figure 3.0 Web Log sample

EXPERIMENTAL RESULTS

In this study for conducting experiment the experimental phase used two different data sets, they are dataset 1, dataset 2files, the dataset1 is collected from the proposed Website and the second dataset2 has been used with synthetic dataset. The performance study of the proposed method APA-IEMis compared with two different existing techniques namely K-Means clustering and Expectation Maximization (EM).

Table 1.0 Performance analyses in terms of accuracy

Techniques	Data set 1	Data set 2	Data set 3
K-MEANS	86.03	89.75	91.26
EM	90.08	91.14	93.09
APA-IEM	92.74	94.81	95.96

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 5, Issue 10, October 2017

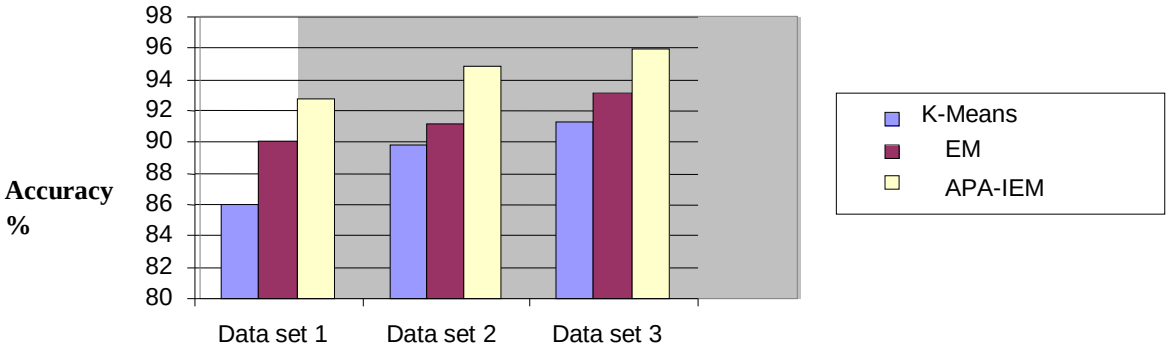

Figure 4.0 Performance analyses in terms of accuracy

From the above the table 1.0 and figures 4.0 shows Proposed APA-IEM performs better for all the three datasets with the highest accuracy of 92.74%, 94.81% and 95.96% than the K-MEANS and techniques.

VI. CONCLUSION

Mobile and Internet based services plays an important role in modern era. Several e-commerce Websites and the Business related services gathering Web user access behavior and their profiles to improve the services. Detection of the user profile based on location is the most important issue which is not handled completely. Adaptive Personalized access behavior from Web, social and mobile log improves the analysis accuracy. This thesis provides the methods and techniques used to find the user access behaviors from Wi-Fi connection log and Web log.

This research work consists of five-step process. The first step is data collection, the second step is SSID detection, the third step is SSID Semantic extraction, and the fourth step is data clustering using improved SVM with partial PCA utilization and the last step is customization. The main goal of this Proposed System is to study the users' preferences based on their previous access. The existing system doesn't provide the location based user access pattern. But the Proposed System gives the location based user access patterns and anomalies from Wi-Fi and Web log details. For this different type of algorithms and techniques are used. The first one is session based visiting patterns, to perform this user need to collect every session and performs SSID description extraction. With the help of Google adwords and semantic keywords, the location based SSID is extracted and grouped.

The hierarchical Dirichlet allocation process effectively finds the category and performs the group and anomaly. The results and output shows the Proposed System performs better than the existing one. This is the first work which handles Wi-Fi and Web log for user access pattern analysis.

REFERENCES

[1]. Shrivastava, Aditi, and NitinShukla. "Extracting Knowledge from User Access Logs." *International Journal of Scientific and Research Publications*2.4 (2012): 1.
[2]. Eirinaki, Magdalini, and MichalisVazirgiannis. "Web Mining for Web personalization." *ACM Transactions on Internet Technology (TOIT)* 3.1 (2003): 1-27.
[3]. Fu, Yongjian, KanwalpreetSandhu, and Ming-Yi Shih. "Clustering of Web users based on access patterns." *Proceedings of the 1999 KDD Workshop on Web Mining*. San Diego, CA. Springer-Verlag, 1999.
[4]. Cooley, Robert, BamshadMobasher, and JaideepSrivastava. "Web Mining: Information and pattern discovery on the world wide Web." *Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference on*. IEEE, 1997.
[5]. Mobasher, Bamshad, et al. "Using sequential and non-sequential patterns in predictive Web Usage Miningtasks." *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002.
[6]. Agosti, Maristella, Franco Crivellari, and Giorgio Maria Di Nunzio. "Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction." *Data Mining and Knowledge Discovery* 24.3 (2012): 663-696.
[7]. Tanasa, Doru, and Brigitte Trousse. "Advanced data preprocessing for intersites Web usage mining." *IEEE Intelligent Systems* 19.2 (2004): 59-65.
[8]. Thakare, Sanjay Bapu, and Sangram Z. Gawali. "An effective and complete preprocessing for Web Usage Mining." *International Journal on Computer Science and Engineering* 2.03 (2010): 848-851.