

# A Survey on User Data Classification and Profile Identification Using Wi-Fi Data

MS D. PRIYADARSHINI<sup>1</sup>, SREEDHANYA.C<sup>2</sup>

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India<sup>1</sup>

M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India<sup>2</sup>

**Abstract:** Mining terrorist data and analyzing user's profile category can be a key for different types of applications. In specific, many applications under anti-terrorism, cyber security and web security need complete user behavior and their Wi-Fi access history to improve the security by detecting abnormal web behavior. In data mining there are numerous studies with the above intentions are made a successful outcome. Classifying users profile and finding their profile category is a major task. This can be done using several data mining algorithm because the access log size is huge in size and contains more auxiliary information. Every user's access behavior differs from one location to another location. So finding the users based on the Wi-Fi access history is important. This survey brings the overall summary of data mining techniques and tools used to analyze and classify user profiles. This helps to know the user type like abnormal, normal, abnormal-terrorist and genuine user. And this survey gives cons of existing work and this may give a new idea to empower the user profile classification and analysis framework to detect terrorist or other kinds of users.

**Keywords:** Data Mining, Weblog Mining, Wi-Fi Log Mining, Normal Users (NU), Abnormal Genuine Users (AGU), Abnormal Terrorist User (ATU), Access Point (AP)

## I. INTRODUCTION

The increased growth of data mining techniques empowers different real time applications such as e-commerce, education, social network, cyber security analysis and anti-terrorism activities. Terrorist society having higher skill and source in internet to gather some illegal activities and make plans for blast or other threatens .Web source are easy to fetch information required for terrorists purpose and also for new cells to join the act of terror. The basic task is to detect and report any unusual suspected behavior that can work as a tool for development of defensive strategies against terrorism risk.

With more than two billion pages being created by millions of Web page author's and associations, the World Wide Web is growing tremendously. As like web, mobile usage is also growing tremendously and it creates a rich knowledge base. Even from the unique characteristics of web like hyperlink structure and its content and languages the knowledge can be extracted.

As like content, the location based user preferences and usages are also identified using Wi-Fi logs. Wireless Fidelity commonly known as Wi-Fi are broadly used internet access point nowadays and are used by majority of internet users throughout the world .the idea is centered on internet access through public Wi-Fi for web surfing and searching and gathering information through it. Wi-Fi are being provided by various Internet Service Providers(ISPs) as well as government these days at various public places like shopping malls ,railway stations, airports etc. So, analyzing weblog with Wi-Fi access history is often a better solution to classify the user profile.

Analysis of these characteristics often exposes interesting patterns and new knowledge. The extracted 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 main demand in extracting knowledge from web mining is the size of the Web and its unstructured and unstable content with its multilingual nature. To add further, the Web generates a huge amount of data in other formats that contain worthy information. To consider as an example, a Web server logs' information about user access samples can be used for information personalization or improving web content and it also helps to customize the web pages. In data mining, ML (Machine learning) techniques signify one possible approach to address the problem. In data mining, machine learning techniques have been employed in various important applications, which is related to the web access based user profile



analysis. This survey explains the number of techniques and tools available for the user profile analysis and few involved with Wi-Fi log to find the access point, and location details of the user.

The intention of Web mining is to detect useful information or knowledge from the Web page content, hyperlink structure along with user's web usage data. Various data mining techniques used by Web mining includes supervised learning [2] unsupervised learning [3] association rule mining [4] and sequential pattern mining [5]. Various exclusive characteristics of the Web make mining information that is useful and knowledgeable, as a fascinating and challenging task.

#### A. General issues in Mining Web data:

In the current scenario, web mining tasks are suffers from several issues, such as the quantity of data/information on the Web is vast and still increasing. Data of all types like unstructured texts, semi-structured Web pages, structured tables and multimedia files (images, audios, and videos) are present on the Web. Information on the Web is diverse. Information on the Web is related with others. There is some noise in the information on the Web. As the Web changes constantly, the information on it is lively. It is a virtual society which allows organizations, people and automated systems to network thought the web.

In web usage mining to detect user profile, the web data can be gathered and used in the framework of user profile classification, which can be gathered from mobile and personal computers. Such data are grouped in four categories based on:

**Content data** are offered to the end-user with appropriately structured. They can be simple text and images or structured data, like information got back from databases.

**Structure data** signify the way the content is organized. The content can be data entities used within a Web page, like HTML (Hyper Text Mark-up Language) or XML (Extensible Markup Language) tags, or they can be used to set a Website together, like hyperlinks linking one page to another.

**Usage data** signify a Web site's usage, such as a visitor's IP address, referrers' address, time and date of access, complete path (files or directories) accessed and other attributes that can be incorporated in a Web access log.

**User profile data** offer information about the users of a Website. A user profile holds information about the demographics for each user of a Web site, along with the information about their preferences and interesting areas. Registration forms or questionnaires make such information obtainable or they can be concluded by analysing Web usage logs [6].

#### B. Process of web log analysis

Web log analysis is used to get clear picture about evaluating the efficiency of a particular Web site, customer behaviour and help to figure out the success of a marketing campaign. Web mining task can be decomposed into the subtasks, namely [7]:

**a) Resource Finding:** The work of retrieving intended Web documents. By resource finding it denotes the method of retrieving the data that is either online or offline from the text sources existing on the web such as electronic newswire, electronic newsletters, the text contents of HTML documents attained by removing HTML tags and also the manual selection of Web resources.

**b) Information Selection and Pre-Processing:** Selecting and pre-processing particular information is retrieved from Web resources automatically. It is a kind of transformation methods of the original data that are retrieved in the IR process. These methods may be a kind of pre-processing as stated above such as stemming, stop words, etc. or a pre-processing aimed at getting the preferred representation such as transforming the representation to relational form, finding phrases in the training corpus or first order logic form, etc.

**c) Generalization:** It finds out common patterns at particular Web sites as well as across multiple sites automatically. Usage of machine learning or data mining techniques are in the process of generalization is done naturally. As the Web is an interactive medium, it allows Humans to play a vital role in the information or knowledge discovery process on the Web.

**d) Analysis:** This section involves Validating and/or interpretation of the patterns that are mined.

## II. LITERATURE REVIEW

Based on the Wi-Fi log, the terrorist exploiting the web for their communication, propagation of their agenda or preparing new active hosts for their plans, there is a growing need to develop methods that control these activities over the web. Over time there have been many surveys and studies on various methods for controlling this sort of traffic over the web. In this portion, we are going to discuss about the various studies and then explain our approach and its



unusualness. The profile classification contain different types of process such as data collection, pre-processing, profile identification etc., Web log preprocessing consists of Web pages, such as IP addresses, page references, and the date and time of accesses and user personal profiles such as user name, gender, age etc., [8]. And these data's are typically extracted from an Extended Common Log Format (ECLF) Server log [9] for web users and Wi-Fi log view for mobile users. Content preprocessing consists of converting the text, images, scripts, and multimedia data into forms that are useful for the web usage mining and user profile analysis process. Habitually this consists of performing content mining such as classification, association or clustering. In the context of web usage mining, the content of Web sites can be used to filter the input to the pattern discovery algorithms [8]. Web structure mining analyses the link structure of the web in order to identify relevant documents [8]. The structure of a site is created by the hypertext links. If the structure information includes the arrangement of various HTML or XML tags within a given page, then it is called as Intra page analysis. If the structure information is represented with hyper links (href) connecting one page to another is known as inter page. Various pre-processing techniques are available in data mining for the above three type of preprocessing. Author Castellano *et al.*[10] proposed a new approach for data preprocessing of web logs. Here the author proves the pre-processing is the first stage of web usage mining. In particular, the author presents LODAP (LOg DAta Preprocessor), which is a software tool which they designed and implemented in order to perform preprocessing of log data. This tool allows creating reports containing the results obtained in each step and information summaries mined from the analysis of the considered log files. But the main limitation of this paper is, this only applicable for a particular website rather than the whole analysis.

Author Ankit *et al* [11] has introduces many data pre-processing techniques helps tp identify unique users and user data session; this can be used to improve the performance features in web usage mining. The analysis of log files which can offer valuable insight into web site usage, those logs should be gathered properly. The given preprocessing techniques help to generate a useful log format from raw web log.

Reddy, K. Sudheer, M. Kantha Reddy, and V. Sitaramulu[12] analyzed and summarized about various details about data preprocessing activities that are necessary to perform Web Usage Mining (WUM). The author shows the importance of data preprocessing in web mining applications. The experiment of the authors helps to estimate data preprocessing importance and finally they provided a new preprocessing approach. The approach increases the quality of the data available and reduces the size. But the authors not concentrated on the pattern discovery process from the preprocessed data.

There are several feature extraction techniques available for we log mining process. The techniques such as association rule mining algorithms like Apriori, FPgrowth and EClat. Kumar, B. Santhosh, and K. V. Rukmani [13] offer an improved algorithm based on the original Apriori algorithm. The new algorithm designed to reduce the basic disadvantages of the apriori algorithm. The Apriori algorithm needs multiple scanning process of database. While applying this into the web log gives the huge amount of overhead. This also increases the delay of pattern extraction. The author recovered these issues by providing an additional property. It decides the candidate set generation iteration and reduces the total number of scans. Wang, Hengshan, Cheng Yang, and Hua Zeng., [14] introduced two prevalent data mining algorithms - FPgrowth and PrefixSpan into web usage mining. Authors enhance the FP growth into Maximum Forward Path (MFP), which is also used to find the patterns effectively. This eliminates the false pattern discovery problem. Sandeep Singh Rawat *et al.*, [15] proposed a custom-built Apriori Algorithm which is based on the existing Apriori algorithm, to find the effective patterns from the web log.

In the current scenario, there is a need to find the users and their profile is very important. In the literature, many works have been developed to handle the web users and their profiles. Currently, there are 2 dimensions to the understanding of the weblog, the first one is from mobile data and another one is web data such as social media data. This section reviews about the previous work for mining WI-FI logs.

Before mining the Wi-Fi log, the use of knowing Wi-Fi should be analyzed. The mobile devices have the capable of tracking users and the trajectories. Detection user profile is a common one. But detecting user profile based on the spatial attribute is a new one. To understand the users from such environment, one method has been proposed. This finds the user profile by user trajectories, which record users' location history of the user, since the trajectories involve in some extent users' interests and preferences. The authors from [16] developed a series of techniques for enabling user's profiles using GPS data. But the techniques suffer from several issues such as, getting location details from GPS is a tedious and high power consumption process. And there is a need of GPS activation to find the log of the user.

Finally the Yao-Chung Fan, Yu-Chi Chen, Kuan-Chieh Tung, Kuo-ChenWu, and Arbee L. P. Chen [17] proposed a new technique to find user preference by their Wi-Fi logs collected from their mobile devices. The authors utilizes the SSID (Service set identifier), which is collected from the Wi-Fi log. The authors just introduced a new effective method to handle the location of the user. But several works are not completed in the paper.



The concept of using Data-Mining for counter terrorism was previously proposed in [18] and the author described the idea of Data mining being a useful tool for detecting and preventing terrorism from the log analysis. It made the reader familiar with what actually data mining is and how we can use it against terrorist activities.

In the paper [19] author discusses various challenges that need to be overcome as well as how these threats can be categorized. Another paper was published in relevance to the prevention of terrorist related activities on the web, in which they developed a System called Advanced Terror Related System (ATDS). The Advanced Terror Detection System aimed at tracking down access to terror related sites by analyzing the content of information downloaded from the Web, which means the content classification.

Author in [20] developed a technique to find the types of pages visited by the users and classify their profiles accordingly. It operates in two modes: the training mode in which the interests of the pre-specified groups are derived by using algorithm like clustering and the detection mode which operates at real time and detects the text HTML pages which are related to terrorist websites. The ATDS focuses only on the textual content, analyzes them and if there are any sign of abnormality in the content it issues an alert. The inspiration of our proposed detection method is different from the above studies. Other systems try to detect terrorist websites by using the textual matter whereas our method uses only the search string to detect how prone the user is to be a terrorist.

Authors in [21] proposed decision tree based classification technique to find user interests and their profiles. Later authors in [20], proposed a novel way to classify passenger data using Naïve bayes algorithm against anti-terrorist. The algorithm classifies the users in three categories Normal User, Abnormal Genuine Users (AGU), Abnormal Terrorist User (ATU). And the classification is done on the basis of the probability of the appearance of the keyword from the synthetic terrorist database.

### **III. PROBLEM DEFINITION**

In Wi-Fi log analysis to perform user profile gathering based on location information is the aim of the proposed work. From the above literature and discussion, we find that the pre-processing is mandatory, because Wi-Fi Log file encloses noisy and un-semantic SSID which may affect results of the profile analysis process. Some of the web log file data are unnecessary for analytical process and could affect the performance of mining process.

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

- Attribute extraction
- Content pre-processing by cleansing
- User location and profile identification
- Gathering Session
- Terrorist and other class detection
- Training process
- Classification process

In this survey, we provide an overview of the Wi-Fi log and its feature. The main task is to "clean" the raw web and Wi-Fi log files and apply the classification 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.
- Train the application according to the classes.
- Clean the Wi-Fi logs and remove the redundant information and inappropriate service identification to find the type of service.
- Applying different data mining Techniques for finding the description of service id.
- Effective classifications to classify multi class of user profiles, these processes are need a complete and effective technique.

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.

### **IV. CONCLUSION**

Internet And Mobile Based Services Plays An Important Role In Modern Era. Several E-Commerce Websites and Business Related Services Gathering Web User Access Behaviour and Their Profiles but only few concentrated on the security related application such as cyber-crime and anti-terrorism. To Improve the Services, detection Of User Profile Based along with Location preference is considered as an important task.

Effective user profile classification to find and thwart terrorism from web, social and mobile log can improve the efficiency of the security related applications. This survey provided various techniques and methods used to classify user profiles. But the survey concludes that there is no sufficient research work has been done with user location to detect multi class of user profiles. So, analysing user profiles from social, Wi-Fi and web log will lead the research in an effective manner.

### REFERENCES

- [1]. Shrivastava, Aditi, and Nitin Shukla. "Extracting Knowledge from User Access Logs." *International Journal of Scientific and Research Publications* 2.4 (2012): 1.
- [2]. Eirinaki, Magdalini, and Michalis Vazirgiannis. "Web mining for web personalization." *ACM Transactions on Internet Technology (TOIT)* 3.1 (2003): 1-27.
- [3]. Fu, Yongjian, Kanwalpreet Sandhu, 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, Bamshad Mobasher, and Jaideep Srivastava. "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 mining tasks." *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on. IEEE*, 2002.
- [6]. Zaiane, Osmar R., and J. Luo. "Web usage mining for a better web-based learning environment." *Proceedings of conference on advanced technology for education*. 2001.
- [7]. 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.
- [8]. Tanasa, Doru, and Brigitte Trousse. "Advanced data preprocessing for intersites web usage mining." *IEEE Intelligent Systems* 19.2 (2004): 59-65.
- [9]. Thakare, Sanjay Bapu, and Sangram Z. Gawali. "A effective and complete preprocessing for Web Usage Mining." *International Journal on Computer Science and Engineering* 2.03 (2010): 848-851.
- [10]. Castellano, G., A. M. Fanelli, and M. A. Torsello. "Log data preparation for mining web usage patterns." *IADIS International Conference Applied Computing*. No. 10000. 2007.
- [11]. Ankit R. K., A. N. Chandni and K. D. Niyanta, 2013. A Complete Pre Processing Method for Web Usage Mining, *International Journal of Emerging Technology and Advanced Engineering* , 3(10):638-640.
- [12]. Reddy, K. Sudheer, M. Kantha Reddy, and V. Sitaramulu. "An effective data preprocessing method for Web Usage Mining." *Information Communication and Embedded Systems (ICICES)*, 2013 International Conference on. IEEE, 2013.
- [13]. Kumar, B. Santhosh, and K. V. Rukmani. "Implementation of web usage mining using APRIORI and FP growth algorithms." *Int. J. of Advanced Networking and Applications* 1.06 (2010): 400-404.
- [14]. Wang, Hengshan, Cheng Yang, and Hua Zeng. "Design and implementation of a web usage mining model based on upgrowth and prefixspan." *Communications of the IIMA* 6.2 (2015): 10.
- [15]. Sandeep, S. R. and R. Lakshmi, 2010, Discovering Potential User Browsing Behaviors Using Custom-Built Apriori Algorithm, *International journal of computer science and information Technology*, 2 (4): 28-37.
- [16]. Y. Zheng, X. Xie, and W.-Y. Ma, "Geolife: A collaborative social networking service among user, location and trajectory," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 32-39, Apr. 2010.
- [17]. Fan, Yao-Chung, et al. "A Framework for Enabling User Preference Profiling through Wi-Fi Logs." *IEEE Transactions on Knowledge and Data Engineering* 28.3 (2016): 592-603.
- [18]. Elovici, Y., Shapira, B., Last, M., kandell, A., & Zaafrany, O. (2004). Using data mining techniques for detecting terror related activities on the web. *Journal of Information warfare*, 3(1). 17-28.
- [19]. Elovici, Y., Shapira, B., Last, M., Zaafrany, O., Friedman, M., Schneider, M., et. Al. (2005). Content based detection of terrorists browsing an Advance Terror Detection System (ATDS).
- [20]. Singh, Saurabh, et al. "A novel way to classify passenger data using Naïve Bayes algorithm (A real time anti-terrorism approach)." *Next Generation Computing Technologies (NGCT)*, 2016 2nd International Conference on. IEEE, 2016.
- [21]. Suneetha, K. R., and R. Krishnamoorthi. "Classification of web log data to identify interested users using decision trees." *Proceedings of the International Conference on Computing Communications and Information Technology Applications*. 2010.