# Preserving Privacy in Web Log Mining in Cloud

Ila Chandrakar, Dr. Vishwanath R H, Dr. Venugopal K R
*Research Scholar, Professor, Principal*
*REVA University, REVA University, UVCE*
*Bangalore, India, Bangalore, India, Bangalore, India*

**Abstract-** Web log data consists of information of usage of web pages of any website which is stored in web server. If these web servers use cloud for storing log data, then this information may be disclosed to curious cloud. In this data, some fields are sensitive for websites. However, if cloud is dishonest, then it may disclose this sensitive data to other competitive web sites owner for their business profits. To the best of our knowledge, privacy preserving in data mining is restricted to data owner database but there is no literature existing for providing privacy in web log data in cloud environment. Hence we propose the need of securing web usage log data along with other data before storing it to cloud. The original web usage data is first hidden and then stored in cloud which results in better security for web log data hence cloud won't be able to extract original web log data.

**Keywords**—Privacy Preserving, Web log data, Web log mining

## 1. INTRODUCTION

Data mining is a technique of extraction of interesting pattern from the large database. This information can be used by the data owner for taking different types of decisions for business and marketing. The interesting patterns can be used for market basket analysis, data analytics, predictive analysis etc. There are different types of data mining techniques like clustering, classification, association rule mining, outlier detection etc. It serves to extract different types of information for different purpose. If these patterns of data owner's databases is shared among each other, it is good for their business collaboration etc. but for every data owner, some data are sensitive, disclosure of which can harm their business profits. So it is necessary to protect sensitive data of the data owners from others.

Web log file contains the information about the usage of any web server by any user. These log files stores data about User Name number of Bytes Transferred, IP Address, Access Request, Time Stamp, Result Status, URL referred and User Agent. The username contains the information about the particular user who has used the website. Generally it stores IP address of the system used for surfing the website, time stamp gives the information that for how much time the user has used the website i.e. the start and end time. It also stores the information about the number of bytes of data transferred during the visit of the website. The user agent is the information about the web browser used by the user to get the information from the web server.

Association rule mining is the data mining technique which is used to extract the correlation between the items in the database. It is basically used to extract information from transaction type of databases which contains different items and their presence or absence in particular transaction. For example in how many transactions, shampoo and conditioner are purchased together. These information can be extracted using two values support and confidence. Let *D* be a database which contains rows as different transactions and columns as different items for those transactions. . Let *IT= {IT_ 1, IT_2, IT_3, IT_n}* be a set of items. Consider T denotes a transaction in database *D, TID* is a unique identifier given to each transaction. An association rule of the form $P \rightarrow Q$ can be generated if $P \subseteq IT$, $Q \subseteq IT$ and $P \cup Q = \varphi$. Support of any item tells what is the frequency of occurrence of the item in database. Confidence is used to find the correlation of occurrence of two items in the database. Support of *rule* $P \rightarrow Q$ and can be calculated by using the following formula (1).

$$\text{Support } (P \rightarrow Q) = |P \cup Q|/N \qquad (1)$$

where$P \cup Q$ is the number of transactions in which both *P and Q* are present and n is the total number of transactions in the database. Confidence of rule $P \rightarrow Q$ can be calculated using formula (2).

$$Confidence\ (P \rightarrow Q) = |P \cup Q|/|P| \qquad (2)$$

Where*P* denotes number of transactions in which item P is present.

Association Rule Mining can also be used in mining correlation between different types of information of web log data for extracting information about the behavior of the usage of website by users. Clustering is an unsupervised learning data mining technique which is used to create clusters of similar kind of data items. In this technique, the given population can be divided into different groups such the data items in a group should be more similar to data items in another group. Many techniques like k means clustering and hierarchal clustering are used for clustering. Classification is a supervised learning data mining technique which assigns different categories to the data based on their features. Different algorithms like K nearest neighbor and decision tree algorithms are generally used for classification. These techniques can be used in web log data like finding the age group who are buying a particular product.

The paper is organized as follows: The different methods for different types of web log mining used by researchers in their paper are explained in II section. Section III contains the proposed work in which web log data mining in cloud is explained and the idea for securing web log data of data owners from cloud is discusses. Section IV is conclusion of the paper.

## 2. WEB LOG DATA MINING

The extraction of interesting pattern from web log data is called web log data mining. This extracted pattern is very useful for deciding different business strategies. For example, if we want to find the pattern information about the purchase of items of customers for any e-commerce websites, we can use some data mining techniques like association rule mining to find the correlation of the sale of two items together from the website. These types of information is very useful for websites to decide strategies. to increase the sale of those items. The mining of web log data can be done in different ways for extracting different types of patterns.

Web usage mining shown in figure 1 consists of three stages: Data Preprocessing, Pattern Discovery and Pattern Analysis. In data preprocessing, data is changed into some format which can be easily understood and processed. This step consists of cleaning of data and identification of user and session. In pattern discovery stage, different data mining techniques like association rule mining, classification, clustering are applied to preprocessed data to extract interesting information. In pattern analysis stage, the extracted pattern are analyzed using different techniques like OLAP(Online Analytical Processing) and query mechanism to find the useful information which can be used for business strategies.

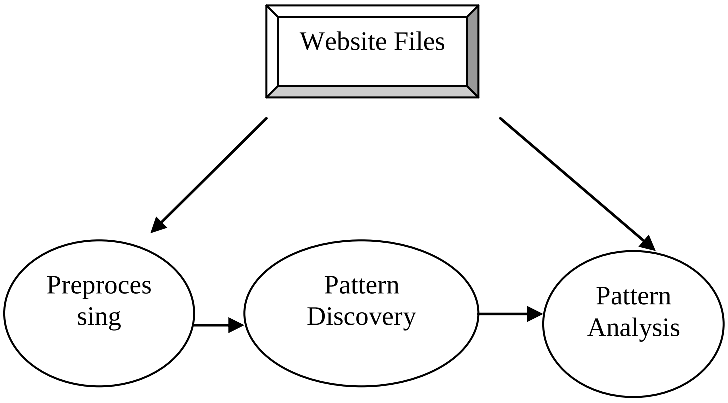

Figure 1. Web Log Mining Process

### A. Web Log using Clustering for Recommendation:

Deptii D. Chaudhari[1] has proposed a technique Web Usage Mining (WUM) which uses logs containing the information about the web pages accessed by the users to evaluate the logs using data mining techniques. Then the data is ranked based on its number of views. So this technique recommends the user for different web pages based on their previous web page visits and the rank of pages based on other user usage. In this paper, clustering technique is used for grouping of users. The user’s choices is extracted based on their profiles and then clustering is applied in to this data to categorize users based on their interests. Then this system recommends the user for the web sites by using the information about interest and rank of pages.

### B. Web Log Mining using Classification:

Suharjito[2] has used k-nearest neighbor classification technique for web usage mining which can be used to extract interesting patterns. In this work, the aim is to solve the problem, the users are facing in using the website. So first the users website usage are observed and extract the information about the web pages which are taking more time to load using k-nearest neighbor classification technique. This information is used by the website owner like bank to improve the performance of the website so that customer should not face problems because it may affect their business.

### C. Web Log Mining using Association Rule Mining market basket analysis:

Association rule mining can be used to extract information from web log to find correlation between different items in web log data. It is very useful for e-commerce website so that they can find sale of different products together and decide the marketing strategies. Zhang[3] has used matrix Apriori algorithm in web log data Sogou search log. In this work, the order details web log data for customers are processed using association rule mining. So the mining results give the details of the products which are purchased together by customers frequently. If user search for some page then this system recommends another page based on other users choice of visiting pages. The recommendation of pages depends on the support and confidence values of visiting count of these pages. If support and confidence is more for two pages then the customer who visits first page is recommended for another page. Association rule mining is more effective in finding relation between user behavior and transaction.

### D. Anomaly Detection from Log Files Using Data Mining Techniques:

Jakub[15] has used data mining techniques for dynamic rule creation to find anomalies in web log data. In this work, first association rules are generated from the web log data. Then these rules are applied to observe anomalies among transactions. The flow of detection procedure is: First file is loaded from training data set, then it is checked whether that log file belongs to particular transaction or not. If it belongs to that transaction then store that record but if doesn't belong to that transaction then check if that satisfies the particular anomaly profile, if it is, then transform that transaction to remove the anomaly and then rules are generated.

**E. Web Log Data Mining in Cloud:**

In previous section, we explained different methods of using data mining techniques in web log data of any private website for extracting interesting patterns. But now data is not just data, it is converted into Big Data and many websites keep their data in cloud because their resources are not enough to store such a huge data. Along with their data set, the web site owners also store their web log data information in cloud.

**F. Preserving Privacy in Data Mining:**

Privacy preserving data mining is the technique by which the data mining results can be preserved from unauthorized access. It is very important when data owners store their data in cloud. Data owners trust cloud that it doesn't disclose data owner's data to others but it may be curious so it is necessary to secure sensitive data of data owners before storing it to cloud. Many researchers proposed many techniques for privacy preserving using different techniques[6-13] which ensures that sensitive data will be safe from cloud and data owners.

## 3. PROPOSED WORK

### Preserving Privacy in Web Log Mining in Cloud:

As we know, privacy preserving techniques are used to secure sensitive data from data owner's database so that it should not be disclosed to other data owners and cloud. Privacy preserving web log mining is a technique to secure web log data like IP address, time used, and web page visited etc. We consider a system model in which two or more e-commerce websites store their all types of data including web log data in cloud. Let us consider these websites are Amazon, Flipkart, Myntra, Voonik etc. These websites don't have enough storage resources so they store their product details, purchase details, IP address, location details who used websites, for how much time they used, which product they were surfing, which products are in wish list, which products are in cart. This information is very sensitive for each website owners because if this information is disclosed then it can be used by other competitive websites for increasing their sale. If other similar websites know about the search history of products and location of customer of any website, then they can use that information for sending advertisement for their websites to those customers.

The web log information of any website are stored in cloud. So cloud can be curious and it can mine the interesting patterns about the usage of web sites. So some method should be proposed to secure this sensitive web log data so that cloud should not be able to mine the interesting patterns from this data. The proposed system is shown in figure 2.

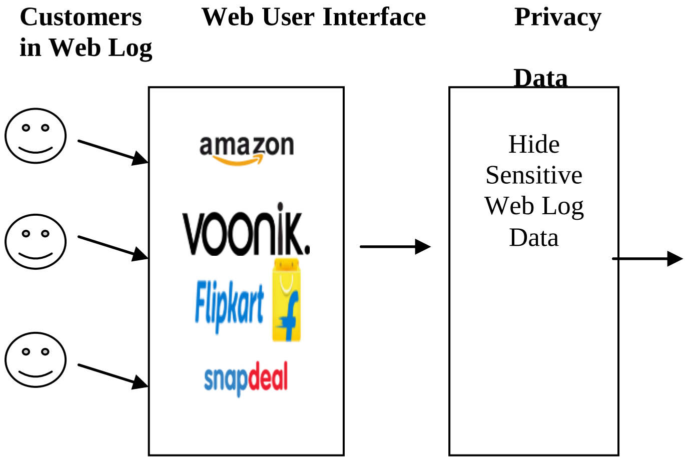

**Figure 2.** Privacy Preserving Web Log Mining

In proposed work, the sensitive part of web log data generated by the e-commerce website is first hidden and then stored in cloud. If web log data consists of the information given on table 1. This data gives information about the IP address of a customer and from which location is the web pages accessed for given amount of time. If it is some e commerce website like Amazon, it gives the information about the product search information using the web page usage information.

**Table 1**. Web Log data

| Ip address | Username | Age | Location | Website | Duration (spending time) |
|---|---|---|---|---|---|
| 192.168.1.1 | Arjun | 25 | Chennai | Flipkart | 25 |
| 192.168.1.2 | Abimanyu | 27 | Chennai | Flipkart | 20 |
| 192.188.1.3 | Uma | 30 | Madurai | Snapdeal | 17 |
| 192.168.1.4 | Jaya | 21 | Madurai | Snapdeal | 10 |
| 192.168.1.5 | Ravi | 27 | Namakkal | Amazon | 39 |

Hence, if this data is disclosed to cloud and cloud becomes dishonest and disclose data to other similar websites like Flipkart. Then using this information, these websites can send advertisement of those products with competitive price and offers to those IP address to increase their sale. So these sensitive fields should be hidden from cloud and other data owners. To achieve this, data owners should first hide the sensitive part of web log data before storing it to cloud. Table 2 shows the web log data with hidden information:

**Table 2.** Hidden Sensitive Data

| Ip address | Username | Age | Location | Website | Duration (spending time) |
|---|---|---|---|---|---|
| ****** | *** | 25 | ****** | ****** | 25 |
| ****** | *** | 27 | ****** | ****** | 20 |
| ****** | *** | 30 | ****** | ****** | 17 |
| ****** | *** | 21 | ****** | ****** | 10 |
| ****** | *** | 27 | ****** | ****** | 39 |

After hiding the data, only non-sensitive data is visible which does not disclose important information. So our work identifies that not just database but log data of web usage can also be sensitive to be disclosed to others hence privacy should be applied to this data before storing it to cloud.

## 4. CONCLUSION

Many researchers have worked in providing privacy to the database of the data owners but we tried to identify the privacy issues in web log data when data owner websites store their log data in cloud. So our work propose that website owners should apply privacy preserving methods to web log data along with other data before storing it to cloud. It ensures non-disclosure of sensitive log data with regard to cloud and other competitive websites.

## REFERENCES

[1]Deptii D. Chaudhari, “A Choice Based Recommendation System using WUM and Clustering” IEEE International Conference on Data Mining and Advanced Computing (SAPIENCE) 2016.

[2]Suharjito, Diana, Herianto,” Implementation of Classification Technique in Web Usage Mining of Banking Company” IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA) 2016.

[3]Hanxiao Zhang, Wei Song, Lizhen Liu, Hanshi Wang, “The Application of Matrix Apriori Algorithm in Web Log Mining”, IEEE 2nd International Conference on Big Data Analysis 2017.

[4]Bhupendra Kumar Malviya, Jitendra Agrawal, “A Study on Web Usage Mining: Theory and Applications”, IEEE Fifth International Conference on Communication Systems and Network Technologies 2015.

[5] J. Breier and J. Branišová, "Anomaly detection from log files using data mining techniques". In Information Science and Applications, pages 449–457. Springer, 2015. doi: 10.1007/978- 3-662-46578-3_53.

[6]J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in Proc. SIGKDD, 2002, pp. 639–644.

[7]J. Zhan, S. Matwin, and L. Chang, "Privacy-preserving collaborative association rule mining," in Proc. DBSEC, 2005, pp. 153–165.

[8]S. Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets," Inf. Sci., vol. 177, no. 2, pp. 490–503, 2007.

[9]F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy- preserving mining of association rules from outsourced transaction databases," IEEE Syst. J., vol. 7, no. 3, pp. 385–395, Sep. 2013.

[10]F. Giannotti, L. V. S. Lakshmanan, A. Monreale, D. Pedreschi, and W. Wang, "Privacy- preserving data mining from outsourced databases," in Proc. CPDP, 2011, pp. 411–426.

[11] X. Ge, L. Yan, J. Zhu, and W. Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," in Proc. SEDM, Jun. 2010, pp. 345–350.

[12.J.-L. Lin and J. Y.-C. Liu, "Privacy preserving itemset mining through fake transactions," in Proc. ACM Symp. Appl. Comput.(SAC),Seoul,SouthKorea,Mar. 2007, pp. 375–379.[Online].Available: http://doi.acm.org/10.1145/1244002.1244092

[13]F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-preserving data mining from outsourced databases,"in Proc. SPCC2010 Conjunction with CPDP, 2010, pp. 411–426.

[14] Kato, Hisayoshi, Hironori Hiraishi, and Fumio Mizoguchi. "Log summarizing agent for web access data using data mining techniques." IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th. IEEE, 2001.

[15] Breier, Jakub, and Jana Branišová. "Anomaly Detection from Log Files Using Data Mining Techniques." Information Science and Applications. Springer Berlin Heidelberg, 2015. 449-457.