Improving Privacy in Web Mining by eliminating Noisy data & Sessionization

Rekha Garhwal
Computer Science Department
Om Institute of Technology & Management, Hisar, Haryana, India

Abstract: data mining automates the detection of relevant patterns in a database, using defined approaches. Data Mining are helpful in Web data analysis and and for sessionzation also. Sessionization is the determination of the number of visitors to a Web site. The user session identification is very important for the traffic characterization purpose. . You can use this document as both an instruction set and as a template into which you can type your own text. Data mining has been combined with privacy preserving algorithms for more number of years which works efficiently only for static data. Mainly the research study database is prepared by the organisations after implementing privacy and given for data mining

Keywords: Data Mining, Web Mining, Data Preprocessing, Web Structure Mining, KDD.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.

Web mining can be viewed as the extraction of structure from an unlabeled, semi-structured data set containing the characteristics of users/information respectively. A Web page contains many information blocks. Apart from the main content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements (for business purposes and for easy user access).Web mining uses many data mining techniques but it is not an application of traditional data mining due to heterogeneity and unstructured nature of the data on Web. The goal of Web Usage Mining is to capture, model and analyze the behavioral patterns and profiles of users interacting with a Web Site. Data preprocessing is one of the major part in Web data mining. It consists of data cleaning, page view identification, sessionization, data integration and data transformation.

In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

Data mining is actually the core step in Knowledge Discovery in Databases (KDD) process[1]. Though KDD is used synonymously to represent data mining, both these are actually different. Some preprocessing steps before data mining and post processing steps after data mining are to be completed to transform the raw data as useful knowledge [2]. Thus, data mining alone might not give you what you actually look for.

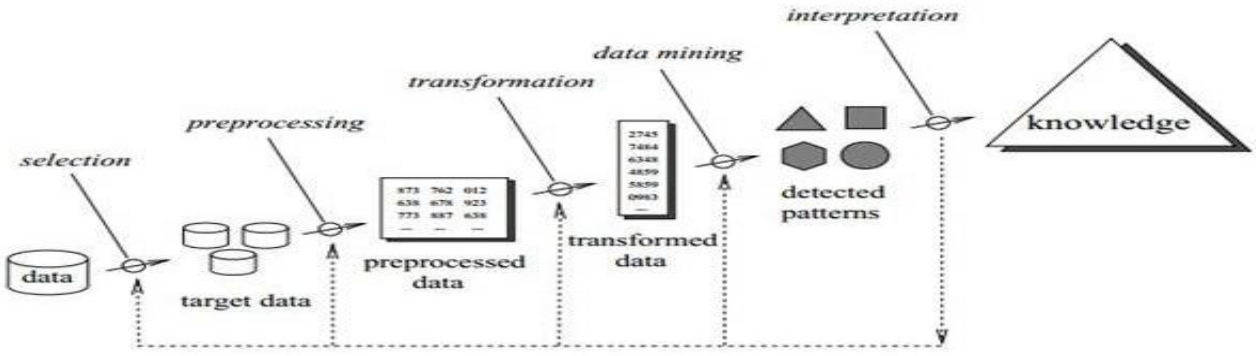

Figure 1 :- Knowledge discovery Process in Data Mining

Phases of Knowledge discovery process:-

Understanding: The first step understands your requirements. You need to have a clear understanding about the application domain and your objectives, whether it is to improve your sales, predict stock market etc. You should also know whether you are going to describe your data or predict information.

Selection of data set: Data mining is done on your current or past records. Thus, you should select a data set or subset of data, in other words data samples, on which you need to perform data analysis and get useful knowledge. You should have enough quantity of data to perform data mining.

Data cleaning: Data cleaning is the step where noise and irrelevant data are removed from the large data set. This is a very important preprocessing step because your outcome would be dependent on the quality of selected data. As part of data cleaning, you might have to remove duplicate records, enter logically correct values for missing records, remove unnecessary data fields, standardize data format, update data in a timely manner and so on.

Data transformation: With the help of dimensionality reduction or transformation methods, the number of effective variables is reduced and only useful features are selected to depict data more efficiently based on the goal of the task. In short, data is transformed into appropriate form making it ready for data mining step.

II. DATA MINING TOOLS

There are various tools can be used for the purpose of Web Data Mining are helpful in Web data analysis and and for sessionzation also. There are three main category of Data mining tools:-

A. Traditional Data Mining Tools

These tools are helpful in data patterns and trends by using a number of complex algorithms and techniques. These tool are used on both single user as well as multiuser operating system. Some of these tools are installed on the desktop to monitor the data and highlight trends and others capture information residing outside a database. The majority are available in both Windows and UNIX versions, although some specialize in one operating system only. In addition, while some may concentrate on one database type, most will be able to handle any data using online analytical processing or a similar technology.

B. Dashboards

Dashboard is mainly used to check or mintor the database. It refelects the change in data and updates in data to the user and this is shown with the help of Chart and table which enable the user to see the performance easily Historical data also can be referenced, enabling the user to see where things have changed (e.g., increase in sales from the same period last year). This functionality makes dashboards easy to use and particularly appealing to managers who wish to have an overview of the company's performance.

C. Text-mining Tools

This is third type and important data mining tool called text data mining tool. It is named as text data mining tool because it is capable of mining different type of data from different kind of text from Microsoft Word and Acrobat PDF documents to simple text files, for example. These tools scan content and convert the selected data into a format that is compatible with the tool's database, thus providing users with an easy and convenient way of accessing data without the need to open different applications.

When evaluating data mining strategies, companies may decide to acquire several tools for specific purposes, rather than purchasing one tool that meets all needs. Although acquiring several tools is not a mainstream approach, a company may choose to do so if, for example, it installs a dashboard to keep managers informed on business matters, a full data-mining suite to capture and build data for its marketing and sales arms, and an interrogation tool so auditors can identify fraud activity. Scanned content can be unstructured (i.e., information is scattered almost randomly across the document, including e-mails, Internet pages, audio and video data) or structured (i.e., the data's form and purpose is known, such as content found in a database). Capturing these inputs can provide organizations with a wealth of information that can be mined to discover trends, concepts, and attitudes.

III. SESSIONZATION TOOLS

The above define the category of the various Data Mining tool but there are various tools available for the Sessionization. Sessionization is the determination of the number of visitors to a Web site. The user session identification is very important for the traffic characterization purpose [3]. Web users transaction can be transformed into number of sessions by different strategies

A. WebSpy Vantage: - This is a very powerful tool that transforms raw log data into manageable information. Sessionization activity is efficiently handled by this software.It supports 200 log formats from many different vendors.

B. Relax :- It is a free specialized web server log analysis tool. Relax supports logs in RefererLog, Apache combined, NCSA extended/combined, TransferLog, and WebSTAR format.It is distributed under GNU General Public License(GPL).

C. Weblizer Xtended : - This is a very powerful tool for web analysis and produces many statistics related to traffic.It is also very good tool for sessionization purpose.

D. Web Lizer :- It is a fast, free web server log file analysis tool. It produces highly detailed, easily configurable usage reports in HTML format, for viewing with a standard web browser.The main thing is that it supports unlimited log file sizes and partial logs.

E. W3Perl :- This tool consists of set of Perl Script that can analyze log files for IIS,Apache,FTP,mail etc. upports sessions (length of time visitors spend on your site), RSS stats, referrers, keywords used on search engines, list of error pages invoked, classifcation of your visitors by countries, browser stats, screen sizes, real-time statistics, etc.

F. Analog Log File Analyzer: - This tool is very popular and used on number of web hosts to produce exhaustive reports of web data. It works in any operating system and freely available.

IV. WEB MINING AND EFFECT OF NOISY DATA ON WEB DATA

Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data. Although Web mining uses many data mining techniques, as mentioned above it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data [9]. Many new mining tasks and algorithms were invented in the past decade [10]. Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

Web structure mining: Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which, incidentally, is a key technology used in search engines. We can also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web content mining: Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc, for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer sentiments. These are not traditional data mining tasks.

Web usage mining: Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of clickstream data in usage logs in order to produce the right data for mining.

The Web mining process is similar to the data mining process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. Data mining on the Web thus becomes an important task for discovering useful knowledge or information from the Web. However, useful information on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation
bars, copyright notices, etc. Although such information items are functionally useful for human viewers and necessary for the Web site owners, they often hamper automated information gathering and Web data mining, e.g., Web page clustering, classification, information retrieval and information extraction. Web noises can be grouped into two categories according to their granularities:

A. Global noises: These are noises on the Web with large granularity, which are usually no smaller than individual pages. Global noises include mirror sites, legal/illegal duplicated Web pages, old versioned Web pages to be deleted, etc.

B. Local (intra-page) noises: These are noisy regions/items within a Web page. Local noises are usually incoherent with the main contents of the Web page. Such noises include banner advertisements, navigational guides, decoration pictures, etc.

The proposed cleaning technique is used for the analysis of both the layouts and the actual contents (i.e., texts, images, etc.) of the Web pages in a given Web site. Thus, our first task is to find a suitable data structure to represent both the presentation styles (or layouts) and the actual contents of the Web pages in the site. Document Object Model tree, which is commonly used for representing the structure of a single Web page, and showing that it is insufficient for our purpose. so we use the Style tree for evaluating the nodes in the style tree for noise detection.

V. SESSIONZATION PROBLEM AND ITS SOLUTION ON WEB MINING

There are mainly four problem occur in the phase of Sessionzation and the solution about these problem are also disscussed in this section

User may visit the site more than once so the server logs records multiple sessions for each user. The solution of above mentioned problem is given by Dynamic IP [6]. Dynamic IP addresses can create a problem during sessionization process.The solution of the problem is to establish cookie and URL encoding kind of mechanism to identify unique user from transaction.
HTTP protocol is stateless so it is impossible to determine when a user actually leaves the web site in order to determine session finish time. Heuristic based solution of above mentioned problem is given by many authors [5].

Problem due to caching which is performed either by proxy server or browser. Caching problem causes a single IP address to be associated with different user sessions so it is quite difficult to identify user based on IP address. This problem can be solved by two main perspective one is use of cookies [4] and second one is URL rewriting or by requiring the user to log in when entering the web site. Important information passed through POST method will not be available in server log so it is difficult to form a session. Packet Sniffing is an alternative method to collecting usage data through server logs.

So from the above disscussed solutions we can solve the problem related to sessionzation.

VI. PRIVACY CALCULATION IN DATA MINING

Data mining has been combined with privacy preserving algorithms for more number of years which works efficiently only for static data. Companies collect data about their customers to maximize their expected profit, scientists gather large repositories of observations to better understand nature and governments of many countries are collecting vast amounts of data to ensure the homeland security issue due to the globalization of conflicts and terrorism .When several different data repositories are combined, the data concerning even only a single person can be tremendously large and complex. Such data belongs to individuals who may be concerned with their privacy.The research study database is prepared by the organisations after implementing privacy and given for data mining [8].

The block diagram gives the general architecture of privacy preserving issue. The actual data being present, is analysed for implementation. The same data is modified by a privacy preserving algorithm and a new altered database is generated.

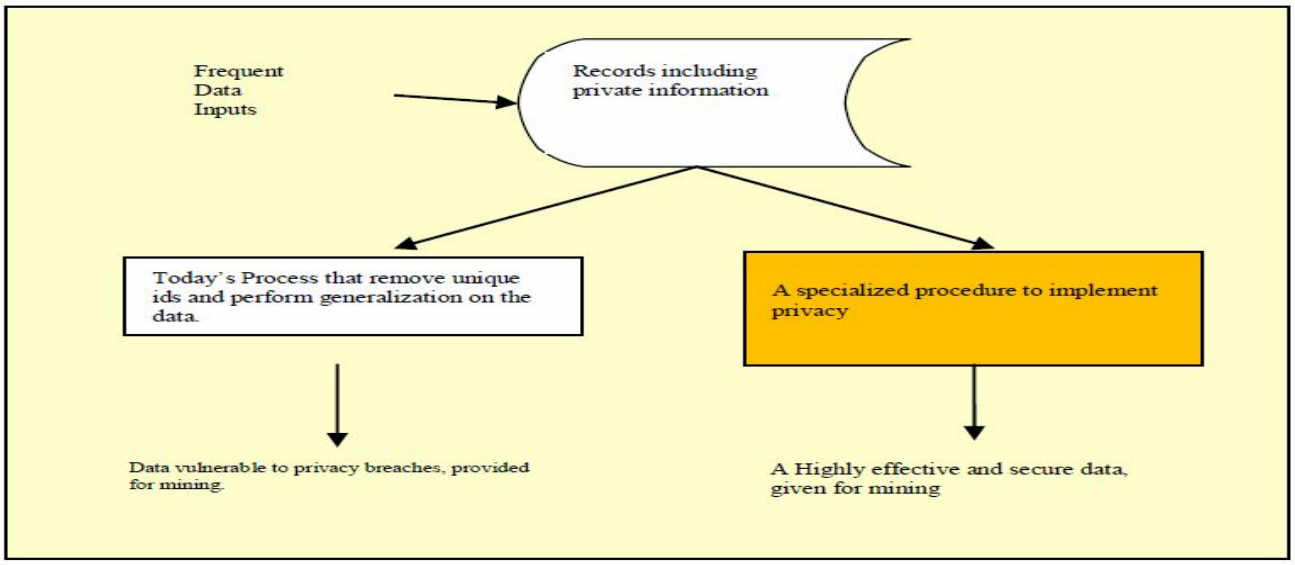

Figure 2. Architecture of a privacy preserving system

VII. METHOD FOR ANALYZING PRIVACY IN DATA MINING

CAHD: The methodology used is called as CAHD, Correlation- aware Anonymization of High-dimensional Data (CAHD). The procedure that is adopted is Anonymized group formation. It has transformed the data into a band matrix by performing permutations of rows and columns in the original table. It also does sorting with respect to Gray encoding. It specifies two methods to implement privacy[7].

The first category is based on approximate nearest neighbour (NN) search in high-dimensional spaces, which is efficiently performed through locality-sensitive hashing (LSH). the first method transforms the data into a band matrix by performing permutations of rows and columns in the original table. The outcome of the transformation is shown in Fig. This representation takes advantage of data sparseness and places nonzero entries near the main diagonal. The advantage is that neighbouring rows have high correlation, i.e., share a large number of common items.

$$\text{Number of distinct hash tables} = \frac{\log(1/\delta)}{-\log(1-p1\ w)}$$

where p1 is the probability of hashing two points to the same bucket in each individual hash table. To fulfill the privacy requirement, each sensitive transaction needs to be grouped either with nonsensitive transactions, or with sensitive ones with different sensitive items, such that the frequency of each sensitive item is reduced below 1/p in every group.

In the second category, they propose two data transformations that capture the correlation in the underlying data: 1) Reduction to a band matrix and 2) Gray encoding-based sorting. The second data transformation technique relies on sorting with respect to Gray encoding [8]: the QID items in each transaction t are interpreted as the Gray code of t. The transaction set is then sorted according to the rank in the Gray sequence. Fig. 1c shows the transformed data set.
The advantages of this method are it combines the advantages of both generalization and Permutation, the neighbouring rows have high correlation and it can be adopted for data stream. The major drawback is it addresses only binary databases. Data accuracy with clustering produces a decreased efficiency.

VIII. CONCLUSION

Data mining is extraction of relationships and useful patterns from various data sources, such as databases, texts, the web. Data Mining tools can be used to demostrate real world and business information. We can reduce the chances of fraud and improve audit reactions to potential business changes. We can ensure that risks are managed in a more timely and proactive fashion. In this paper we provided different strategies of sessionization. Depending upon the application area, one can select strategy. Along with solutions, this paper emphaisis on the problems that may arise in stages of sessionization and methods to resolve them. This paper also deals with different software tools available in market for any kind of web analytics. We also proposed an information based measure to detect noises by cleaning a page from a site using cleaning technique.

IX. REFERENCES

[1] G. Piatetsky-Shapiro, U. M. Fayyad, and P. Smyth. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[3] Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analyis. INFORMS Journal on Computing.

[4] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, 2000.

[5] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. Knowledge and Information Systems, 1(1):5–32, 1999.

[6] Myra Spiliopoulou, Bamshad Mobasher, Bettina Berendt, Miki Nakagawa, A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis.

[7] Gabriel Ghinita, Panos Kalnis, and Yufei Tao," Anonymous Publication of Sensitive Transactional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 2, FEBRUARY 2011.

[8] Z. Huang, W. Du, and B. Chen, "Deriving Private Information from Randomized Data," Proc. ACM SIGMOD, pp. 37-48, 2005.

[9] Chakrabarti, S. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan Kaufmann, 2002.

[10] Han, J. and Chang, K. C.-C. Data Mining for Web Intelligence, IEEE Computer, Nov. 2002.