Paper on Privacy Preserving Data in Web Log Mining

Brijal Kharad[1] Ruchika P Dungarani[2]

[1,2]P.G Student

[1,2]Grow More Institute, Gujarat, India

Abstract— Web log mining is a great source of information and knowledge, where a numerous of users would search their interest. The data present is in form of structured and text data. So, different kinds of data model can be implement with web data for design discovery[3] Researches on protecting private data in the application of web data mining possess practical value. Introducing fundamental concepts of web log mining and private data and then it puts first privacy preserving mining model based on evolutionary algorithm of cloud model, joining with evolutionary algorithm and cloud model theory. Web Log mining is the result of web usage mining which contains information of web access of various users. Study of log files provides the full information of the access patterns of the users, example biography of the user's behavior, operating system used, particular time period of usage in the way of successful/unsuccessful transactions etc.[3]

Key words: web log mining, private data preservation

I. Introduction

Privacy-preserving data mining on the Web is one of the new trends in privacy and security research[4]. It is driven by one of the major policy issues of the information that the right to privacy. this research field is very new we have already seen great interests in it the recent proliferation in PPDM techniques is evident, the interest from academia and industry has grown quickly, and. In this we discuss the problems in defining privacy and how privacy can be violated in data mining.[4]

Then the definition of privacy preservation in data mining, we analyze the implications of the Organization for Economic Cooperation and Development (OECD). data privacy principles in knowledge discovery. some policies for PPDM based on the OECD privacy guidelines. We also introduce a taxonomy of the existing PPDM techniques and a discussion on how these techniques are applicable on Web data. we suggest some privacy requirements that are related to industrial initiatives, and point to some technical challenges as future research trendsin PPDM on the Web.

II. Motivation of work

Web log file analysis began with the purpose to offer to Web site administrators a way to ensure adequate bandwidth

A. *What Is Privacy Preserving Data In Web Log Mining?:*

Web log mining is a vital research realm in applying web data mining. Web log mining together with web content mining, web constitution mining consists the content ofweb mining. Its mining process is made up of data preprocessing, mode discovering and mode analyzing.Data preprocessing is the base for the whole web log mining process, including data preprocessing, user identification, secession identification, transaction identification, path completion and formatting. Web server access log and applied service log are the chief data sources of web log mining. Semi-structured data in the communication process of user and net site, which is consisted of using record data, content data, structured data and user data, is settled as the main mining target. By using data mining technology, web log mining realizes the mining of web server log document so as to obtain user's browsing pattern and to provide personalized service recommendation for user. Simultaneously, it evokes privacy preserving problem which becomes a bottleneck and conditions further development of web log mining, accompanying with the popularization of information-based net.

World Wide Web is a global area and rich source of information. Day by day number of web sites and its users are increasing. Information extracted from may sometimes do not turn up to desired expectations of the user. A refined approach, referred as Web Mining, which is an area of Data Mining dealing with the extraction of interesting knowledge from the World Wide Web, can provide better result. While surfing the web sites, users' interactions with web sites are recorded in web log file. These Web Logs are abundant source of information. Such logs when mined properly can provide useful information for decision making.

Web Mining on WWW is classified into three areas, (1)Web usage mining, (2)Web structure mining and (3) Web content mining. Web usage mining tries to find user's behavior in Web accesses and is mostly approached from access log analysis. Web structure mining finds linkage structure of the complex Web structure, which often means the hypertext structure. The linkage analysis may find some community on Web. Web content mining aims and server capacity to their organization[2]. This field of analysis made great advances with the passing of time, and now e-companies seek ways to use Web log files to obtain information about visitor profiles and buyers activities The analysis of Web log may offer advices about a better way to improve the offer, information about problems occurred to the users,in web log mining and even about problems for the security of the site. Traces about hacker attacks or heavy use in particular intervals of time may be really useful to configure the server and adjust the Web site. The concept of initial findings on a specific aspect that is highly relevant for personalization

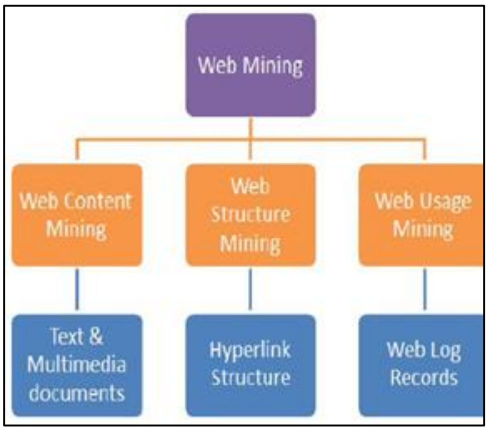
III. Problem statement

Scenario 1Suppose we have a server and many clients in which each client has a set of sold items The clients want the server to gather statistical information about associations among items in order to provide recommendations to the clients. However, the clients do not want the server to know some strategic patterns also called restrictive association rules). In this context, the clients represent companies and the server is a recommendation system for an e-commerce application, for example ,fruit of the clients collaboration. In the absence of rating, which is used in collaborative filtering for automatic recommendation building, association rules can be effectively used to build models for on-line recommendation. When a client sends its frequent itemsets or association rules to the server, it must protect the restrictive itemsets acording to some specific policies. The server then gathers statistical information from the non-restrictive itemsets and recovers from them the actual associations.[4]

Scenario 2Two organizations, an Internet marketing company and an on-line retail company, have datasets with diff erent attributes for a common set of individuals. These organizations decide to share their data for clustering to find the optimal customer targets so as to maximize return on investments. How can these organizations learn about their clusters using each other's data without learning anything about the attribute values of each other? Note that the above scenarios describe diff erent privacy preservation problems. Each scenario poses a set of challenges. For instance, scenario 1 is a typical example of collective privacy preservation, while scenario 2 refers to individual's privacy preservation. How can we characterize scenarios in PPDM? One alternative is to describe them in terms of general parameters. [4]

Outcome -Refers to the desired data mining results. For instance, someone may look for association rules identifying relationships among attributes, or relationships among customers 'buying behaviors as in sce nario 1, or may even want to cluster data as in scenario maximize return on investments. Privacy PreservationWhat are the privacy preservation requirements? If the concern is solely that values associated with an individual entity not be released (e.g. personal information), techniques must focus on protecting such information. In other cases, the notion of what constitutes "sensitive knowledge" may not be known in advance. This would lead to human evaluation of the intermediate results before making the data available for mining. [4] automatic extraction of desirable information from Web. It studies to use the Web as a big dictionary or databases. It is mainly approached from the text analysis of the HTML documents through WWW.

IV. Web Mining Categories



\

1) Web Content Mining It refers to knowledge discovery in which the main objects are the traditional collections of multimedia documents such as text, images, videos, audios, which are embedded in or linked to the Web pages There are two types of approaches in Web Content Mining, Database Approach and Agent Based Approach. Agent based Approach includes intelligent search agent, information filtering and categorization, and personalizedweb agent.Database approach includes Multilevel Databases and Web Query System Various techniques to extract data from Web mining exist.[5]
2) Web Structure Mining It is the process of inferring knowledge from the organization and links on the Web. It works on the hyperlink structure of the web. The graph structure can provide information about ranking or authoritativeness and enhance search results of a page through filtering.[5]
3) Web Usage Mining has been defined as the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of Web- based applications. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis Structure of information should be good which will allow extracting knowledge from log files. Web Usage Mining may be applied to data such as contained in logs files. A log file contains information related to the user queries on a website. Web usage mining may be used to improve the website structure or giving recommendations to visitors . Log Files contains fields like client IP address, access time, date, HTTP request method, document size, path of the resource on the Web server, protocol used for the transmission., status code of the server and number of bytes transmitted in the transaction In Web Mining, data can be collected at the server-side, client-side, proxy servers, or obtained from an organization's database. Each type of data collection differs not only in terms of the location of the data source, but also the kinds of data available, the segment of population from which the data was collected, and its method of implementation [Web usage mining is achieved first by reporting visitors traffic information based on Web server log files and other source of traffic data. For discovery and analysis of usage patterns from the available data, it is necessary to perform three steps: Preprocessing, Pattern Discovery, Pattern Analysis The information on the web is scattered and unstructured and does not match user's requirements always. One of the solutions may be the development of ontology or updating an existing ontology. The other solution is identifying web user's profiles and analysis of their visit patterns using web usage mining methods. Therefore, first , we illustrate about the development of an ontology with an example and second.[5].

V. Appilication of Privacy Preserving Data in Web Log Mining

Web mining is the application of data mining technologies on the Internet to extract interesting, useful patterns and implicit information from activities related to WWW. Web mining technique is introduced to e-commerce recommendation system, by means of data mining technology to be used for automatically, quickly discovering the visitor's browsing patterns from web log data.[4] Basing on visitor's browsing patterns, the site can efficiently and automaticly dynamic adjust web pages' content, and recommend right items for each customer to provide personalized recommendation service. With the personalized goods recommendation service e- commerce system attracts more visitors.

Personalized service recommendation is a hot research spot of web log mining in the application realm.. In the E-commerce era, web log mining has become the main instrument and method for enterprises to get involved in competition and to aid decision making. It, on one side, brings about convenience for users; on the other side, threatens user privacy.

VI. Conclusion

According to our observations, the performances of the algorithms are strongly depends on the support levels and the features of the data sets. There are Various work has been done using the web log mining patterns Like to usage patterns from web data in order to understand better serve the web based application. application of the Web Mining Algorithm for web log analysis which is applied to identify the context associated with the web design of an e commerce web portal that demands security. Its use In the E-commerce era, web log mining has become the main instrument and method for enterprises to get involved in competition and to aid decision making. On the basis of web log data's features, this paper brings forward a privacy preserving mining model based on cloud model's evolutionary algorithm

References

[1] Jiangchang-bin,chen Li, school of management Wuhan Univercity of technology WHUT,Wuhan,P.R China,43070,privacy preserving data in web log mining.

[2] Maristella Agosti and Giorgi MariaDi Nunzi Department Of information Engineering – Univercity of Padua Via Gradegnigo 6/a,35131 padova,Italy,web log mining:a Study of User Sessions

[3] Amit Vishwakarma,M.tech scholar,TIT science,Bhopal KedarNatha singh Asst.Proffesor TIT science Bhopal A survey on Web Log mining Pattern,Discovery.

[4] Stanley R.M.Oliveira 1,2 Embrapa Informatica Agropecu Andr´eTosello,209-Bara˜oGeraldo 13083-886,Campinas sp Brasil Osmar R Za¨ıane2 Department of Computing Science University of Alberta Edmonton, AB Canada, T6G 1K7, Privacy-Preserving Data Mining on the Web: Foundations and Techniques

[5] Sanjay Kumar Malik, sdmalik@hotmail.com Nupur Prakash Malik@rediffmail.com S.A.M. Rizvi, School of Comp.Sc. Jamia Millia Islamia, University School of I.T. GGS Indraprastha University New Delhi samsam_rizvi@yahoo.com , Ontology and Web Usage Mining towards an Intelligent Web focusing web logs