Information Security in Data Collector and Miner in Big Data Mining

Prof.Ms.NiketaV.Kadam[1], Prof.Anup G.Kadu[2], Prof.Ms.Preeti V.Dudhe[3], Prof.Ms.Maithili S.Deshmukh

Abstract- **The actual data mining task is the kind of semi-automatic or automatic analysis of large quantities of for the extra traction of previously unknown, interesting patterns for example, groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This can be performed using database techniques such as spatial indices. These patterns can be considered as a kind of the input data, and may be used in further analysis, in machine learning and predictive analytics. This is very useful in data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.**

Keywords- **Cluster, pattern mining etc.**

I. INTRODUCTION

In large data sets Data mining is useful for discovering patterns . It involves methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science with The main goal of data mining is to extract information (with intelligent method) from a data set and it is used to transform the information into a comprehensible structure which get used for further processing. Data mining is very much helpful for the the analysis of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data preprocessing, model and considerations, interestingness. To extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining) this method of data mining is used for the huge data sets. This can be done using database techniques such as spatial indices. [1][2]

This study paper clarifies about different discovery systems of HTTP Botnet recognition. On account of the unsafe impacts of botnets and the impressive enthusiasm among the exploration network in this field, we proposed overview of botnet look into which depict the botnet issue in worldwide terms and give distinctive discovery strategies. All identification procedures depend without anyone else life-cycle. This displays a fascinating property each phase of the life-cycle must be adequately completed if the botnet is to succeed. Consequently, interfering with the execution of only one phase in the botnet life-cycle renders the entire botnet futile. For location of HTTP botnet we can utilize signature based recognition method and conduct based discovery procedures We have assessed ebb and flow inquire about work in this field, and demonstrate that all safeguard endeavors are in truth centered around at least one of these stages. This audit is displayed here as a study of the most significant commitments in the field.

II. STAGES OF KNOWLEDGE DISCOVERY MODEL

There are four stages for knowledge discovery in databases (KDD) process as defined bellow:

1. Cleaning and Integration
2. Selection and Transformation
3. Data mining
4. Evaluation and presentation.

Data cleaning: noise and inconsistent data get removed at this first step of data cleaning.

Data integration: At this stage, the combination of multiple data sources performed

Data selection: Retrieval of data from the database which is relevant to the analysis task is performed.

Data transformation: where data are summarized or aggregations operations are used at this stage to transformed and consolidate data into forms appropriate for mining. Data mining: which is an most needed process where with the help of intelligent methods data patterns are extracted.

Pattern –evaluation: It is used to identify the truly interesting patterns representing knowledge based on interesting measures.

Knowledge presentation:To present mined knowledge to user's visualization and knowledge representation techniques are used. The basic form of the data in a table consists of following four types of attributes described bellow.

(i)Explicit Identifiers – it is a set of attributes containing information that identifies a record owner explicitly such as name, SS number etc.

(ii)Quasi Identifiers - is a set of attributes which is used to potentially identify a record owner when combined with publicly available data.

(iii)Sensitive Attributes - is a set of attributes which consists of sensitive person specific information such as disease, salary etc.

(iv)Non-Sensitive Attributes - is a set of attributes that creates no problem if revealed even to untrustworthy parties, for example, very general information of a person.

.

The identity or sensitive data about record owners data are to be hidden by using the approach known as Anonymization . It even assumes that sensitive data should be retained for analysis. It's obvious that explicit identifiers should be removed but still there is a danger of privacy intrusion when quasi identifiers are linked to publicly available data. Such attacks are called as linking attacks.[3].

Privacy Preserving publishing of Social Network Data

As face book and LinkedIn are Social networks sites which are responsible for significantly revolutionizing the way people interact. On the web this technique is very useful for connecting, interacting, communicating and sharing information .Most of the people are usually use social network sites for meeting old friends, making new friends, or for searching people who have the same interests or problems across various domains such as political, economic, and geographic borders. Social networking sites allow users to connect, post messages, send e-mails and instant messages to each other by creating personal information profiles that contain information such as photos, video and audio files

As every person is so much involved with the social networking sites hence the popularity of social network is enhanced and it has attracted many people. As a result, it became the routine of many people to publish anonym zed versions of the data collected from social network service users to third party consumers such as sociologists, (e.g., for studying social structure) , epidemiologists (e.g., to understand infectious disease dynamics) , businesses (e.g., to drive marketing campaigns and to enable better social targeting of advertisements) and criminologists (e.g., identifying insurgent networks and determining leaders and active cells) . Generally, the data collected by online social network operators is rich in content and relationships that are quite valuable to many third party consumers. As the social network data often comprises private and sensitive information about the social network users, it is imperative to ensure that any published social network data would not breach privacy of the social network users. And hence the social network operators release sanitized version of the social network data which is used by the third party consumers.[4]

2.1Privacy Breach Risks

When information deemed private and sensitive is disclosed to unauthorized individuals then the privacy breach is said to have taken place. Generally, online social network users have strong perception that the network operators keep their private information secures. The operators commonly anonymize the data before publishing it for use by the third party consumers, this all done by them to ensure privacy of the social network users. However is an increasingly important challenge for social network operators that is of maintaining online social networks users privacy in publishing social network data.

III. WORKING OF DATA MINER

Data miners are software applications that collect online information for a company's benefit. Many companies mine their own data to obtain a better grasp of how it's interrelated. They use this information for company presentations, which may eventually boost financial profits within their specific industries. More often than not, data miners are implemented without a computer user's knowledge. This is the case with the spyware enlisted by electronic storefronts to analyze consumer behavior.

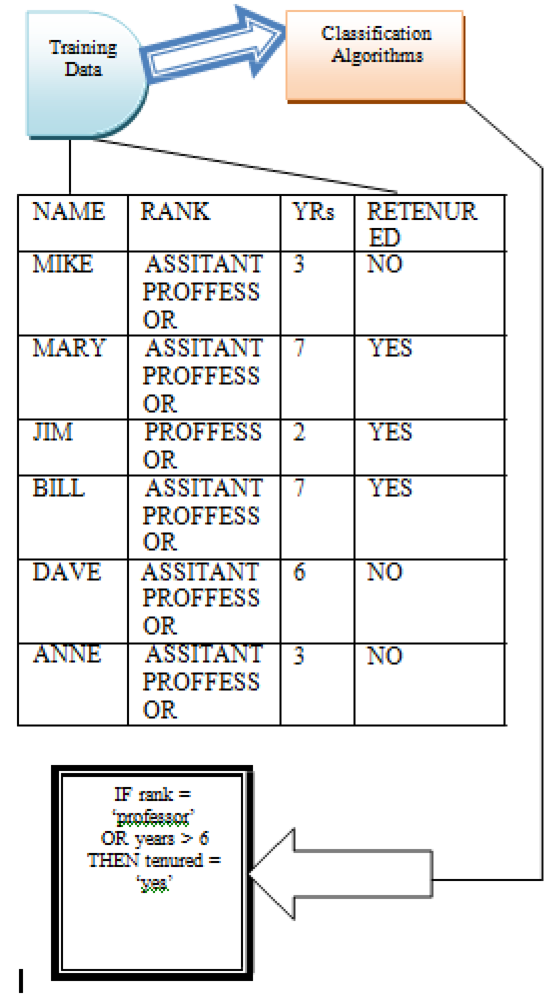
Spyware is uploaded onto users' computers without their consent after they click on something that triggers the automated spyware download. As they browse and search various online websites, online marketers and advertisers exploit this data by sending emails to clients containing product information based on the user's previous web searches and product interests. They may also link users to similar websites within a website they visit frequently, or remind online shoppers of their shopping history by suggesting new or similar products. Spyware can contain viruses, however, and for that reason and others, most online consumers are annoyed by these data mining techniques.

3.1 Concerns of Data Miner

The data miner applies data mining algorithms to the data acquired from data collector in order to discover useful knowledge which is desired by the decision-maker,. The protection issues accompanying the data mining operations are twofold. On one hand, security of the first information proprietor (i.e. the information supplier) will be traded off if individual data can be straight forwardly seen in the data and data break happens. Then again, furnishing with the numerous intense data mining systems, the data miner can find out different kinds of information underlying the data. sensitive data about the data owners Sometimes may not cover data mining results.

3.2 Approaches to Privacy Protection

Supervised learning (classification) is the most effective technique which is use for privacy protection.
Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations. Based on the training set new data get classified.Classification is a Two Step Process Model construction: It is used to describe a set of predetermined classes. Each tuple or sample is assumed to belong to a predefined class, as determined by the class label attribute. Training set is nothing but the set of tuples used for model construction. The model is represented as classification rules, decision trees, or mathematical formulae.
Model usage: It is used for classifying future or unknown objects by estimating accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set (otherwise over fitting). The model is used to classify new data, if the accuracy is acceptable.



NAME	RANK	YRs	RETENURED
MIKE	ASSITANT PROFFESSOR	3	NO
MARY	ASSITANT PROFFESSOR	7	YES
JIM	PROFFESSOR	2	YES
BILL	ASSITANT PROFFESSOR	7	YES
DAVE	ASSITANT PROFFESSOR	6	NO
ANNE	ASSITANT PROFFESSOR	3	NO

Figure 1: Model Construction

IV.ADVANTAGES AND LIMITATIONS

4.1 Advantages

1. Development of various data mining techniques possible only because of the PPDM approach.
2. Sharing of large amount sensitive data for analysis purposes is allowed.
3. It is able to track and collect large amounts of data with the use of current hardware technology.

4.2 Disadvantages

1. One of the major disadvantages of Privacy Preserving.
2. Data Mining is abundant availability of personal data.various technologies exist for supporting proper data handling, but much work remains and some barriers must be overcome order for them to be deployed.
3. All the above techniques are remarkably good but there is always extent for more enhancements.

V. CONCLUSION

Step by step instructions to protect sensitive data from the security dangers brought by data mining has turned into a hotly debated issue lately. In this report I reviewed the privacy issues identified with data mining by utilizing a user role based procedure. I separated two distinctive user roles that are generally required in data mining applications, i.e. data collector and data mine. Every client part has its own particular security concerns, subsequently the protection saving methodologies embraced by one client part are by and large unique in relation to those received by other.

REFRENCES

1. J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*, 2006.
2. L.Brankovic and V.Estivill Castro*,''Privacy issues in knowledge discovery and data mining*,'' in Proc. Austral. Inst. Comput. Ethics Conf., 1999, pp. 89–99.
3. R. Agrawal and R. Srikant, *''Privacy-preserving data mining,''*