International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 9 Issue 1 – JULY 2014.

Data Mining Analysis Using Query Formulation In Aggregated Distributed Database

M.Deepika[#1] and V.Vijayadeepa[*2]

[1]PG Scholar, Muthayammal College of Arts&science

[2] Associate Professor, Muthayammal College of Arts&science

Abstract- Recommender systems are becoming increasingly important to individual users and businesses for providing personalized Recommendations analyze data efficiently, Data mining systems are widely using datasets with columns in horizontal tabular layout. Preparing a data set is more complex task in a data mining project, requires many QLA queries, joining tables and aggregating columns. Conventional RDBMS usually manage tables with vertical form. Aggregated columns in a horizontal tabular layout returns set of numbers, instead Relational databases are acceptable repository for structured data; integrating , Query Formulation Algorithm with a relational DBMS is an essential research issue for database programmers.

Keywords: Aggregation, Data Mining, Query Formulation Algorithm.

I. INTRODUCTION

Network security consists of the requirements and policies adopted by a network administrator for preventing and monitor unauthorized access, misuse, alteration, or rejection of a computer network and network-accessible resources. Network security involves the authoritative access to enter into the network which is controlled by the network administrator. Users can enter into the network by using user ID and password for authenticate information that allows them access to information and programs within their authority. Network security covers a range of computer networks, In that public and private that are used in everyday jobs conducting transactions and interactions among businesses, government agency and individuals.

Networks can be confidential, such as within a company, and others would be open to public access. Network security is implicated in organizations, enterprise, and other types of institutions. It secures network, as well as protecting and managing operations for various network processes. The common and simple way of protecting a network resource is to give the unique name and a unique password for each authentication. There are two of the primary reasons for the increased attention on information leak.

There are two of the primary reasons for the increased attention on information leak. The first has to do with the state of information security when it comes to protecting against the external threats. Viruses and worms spreading on the internal network infect systems, corrupting data, and sap away network bandwidths are high profile issues that demand immediate attention. They have an importance based on the potential damage they can do, as well as the visibility they have to both users and management.

Data mining [11] is the process of data analyzing from different perspectives and summarize those data into useful information. It can be used to increase revenue and costs,. Data mining allows users to analyzing the data from different dimensions or angles, categorizing and summarizing the relationships to be identified.

In data mining several data preprocessing techniques are available such as data cleaning, data integration ,data transformation and data reduction. In the today's world the data is dirty, incomplete, noisy errors, and inconsistent. Discrimination is the ability or the power to make fine distinctions. The treatments are the consideration of finding the particular class or category instead of finding the individual list in partiality or prejudice.

We have taken Adult data set from the UCI repository. Adult data set is extracted from the census bureau database information. The way to prevent the private information in public networks we have taken the adult data set. The private information like race, sex etc. in the adult data set to be prevented using the data mining techniques.

Rapid miner is the data mining tool that used for classification and data transformation. In classification Naïve bayes classification we used and for data transformation normalize, map, and discretize to transform those data. The data which are sensitive and the information which should not show to public those types of information are meant to be private. Several data preprocessing techniques are used to clean the data and the personal information in public network to be prevented and the information should not show to any third party enter into the network.

II. Literature review

Data mining is the extraction of useful information from the large collections of data. Negative social perceptions about data mining in which [10] privacy discrimination prevention in data mining. Data collection from the large set of data to be taken. Classification rule mining have done the way of making the automated collection of data. They are using the relationship between the direct and indirect discrimination prevention in data mining. Nowadays online social networking websites allow users to publish about their information [9] They are connecting their friends through those networks. Some of the information in the network to be private.

In this paper they are exploring the relationship between the friends and how they are launching inference attacks on those networks. Social networking websites to predict private information using the friendship links and their details to predict those private information in the network. Taking those individual user information from the network and how to prevent those information in the face book. Detecting attack by evasion technique is a challenge for intrusion detection and intrusion prevention systems. Here they are using five common evasion technique to evade the system. Dos attack disable [14] the system to find the resources. Packet splitting chop the data into packets. so the system may not completely reassemble the packets for signature matching. Payload mutation and shell code mutation helps the attackers to evade the system.

Text password is the popular user authentication in networks for security. Users often select weak passwords and they are reusing the same passwords across different websites. Reusing passwords causes a domino effect typing passwords into untrusted computers suffers password. [6].The password stealing attacks to snatch passwords such as phishing, key loggers and malware. The one time password that helps to authenticate those networks.

False positives and False negative occur in intrusion detection system/intrusion prevention system. The mechanism for finding the false positive/negative with many IDSs/IPSs .Collecting the FP and FN cases from real-world traffic [3] and analysis for finding the attacks in the network security. Missing attack signature in the design is the main cause of FN cases. An intrusion detection or prevention that monitors the activities of a given location and decides whether these malicious activities or normal based on system integrity, confidentiality and the availability of information resources. False positives and False negatives cause several problems.

The problem of rapid anomaly detection in computer network traffic. The problem statistically analyse the sequential change point detection, They proposed a new anomaly detection method. The method is based on the multi-cyclic (repeated) Shiryaev–Roberts detection procedure where the likelihood ratio is replaced with the linear-quadratic score. This can be done using [1] in real-world network security applications both pre-attack and post-attack distributions to be made for different from hypothesized distributions such as Gaussian or Poisson. Many change point detection schemes, our method is also of practically no computational complexity and easy to implement.

Data to be collected automatically for the use of mining the intrusion and crime detection. Large corporations like banks, insurance companies are increasing the facilities for mining the data about the customers and their employees in the way of detecting potential intrusion, fraud or crime. Mining algorithms [11] from the trained datasets regarding gender, race, and religion those information’s are to be very sensitive and those information’s to be prevented from the public networks.

Differential privacy has the computations be insensitive for changing in any particular single person record to be private. There are restricting data leaks through the various results. The privacy preservation [4] ensures unconditionally safe access for the data and does not require from the data miner in any expertise in way of privacy. A naïve utilization for the interface to construct the privacy preservation in data mining.

In the prediction the important one we want to find is the input and the goal for finding those observation. By using the dyadic prediction the input consists of pair of items and the goal to predict the particular value. This prediction involves collaborative filtering and to predict the ratings for movie and link prediction. The main goal [8] is to find the presence or absence of an edge between the nodes in the graph structure.

Many social networking websites allow the users to protect their personal information like their profiles from the public. They are showing how an adversary can exploit an online social network with a combination of both private and public user profiles to predict the private attributes of users. The problem of finding the [15] relational classification and the friendship links for the single user profile and group.

Analyzing the large data set from the social networks. This data set provides a view of analyzing the nature of attacks possible in the social networking sites. The work of finding [12] the information for large dataset and mining those information from those various data set. The significant correlations between the attackers and the group of customers using those data set. Using those data set finding the better understand vulnerability life cycle.

III. Proposed Work

In our proposed work the personal information to be prevented using the data mining techniques like Naïve bayes classifier and data transformation techniques to be applied. In the Fig 1 shows the system architecture for securing personal information using data mining in public networks (SPI-DM).

Random Key Generation:

A Random key is used to encrypt and decrypt the data being encrypted /decrypted. Symmetric-key algorithms have been used for a single secret key. It requires the key is to keep secret. Public-key algorithms use a public key and a private key. The set of one-time password is established by hash function through multiple hashing. N one time password can be created by producing N hashes on input c. N be the Length of the hash function. c be the secret key sharing between the user and the server.

$$\delta_0 = \mathcal{H}^N(c) \qquad (1)$$

δ_0 the number of times the key is to be performed for opening our network. $\mathcal{H}$ be the hash function for finding the number of hash function to be performed.

3.2 Naïve Bayes Classifier

Naïve bayes classifier is a probabilistic classifier based on applying bayes theorem with naïve independence assumptions.A more explanatory term for the underlying probability model would be independent feature model for appliying the classifier. For some types of probability models, naive Bayes classifiers is to be trained very efficiently in a supervised learning method. In many practical applications, parameter evaluation for naive Bayes models to be used for maximum likelihood in other words one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Despite their naive design and it appears that oversimplified assumption, naive Bayes classifiers have worked quite well in many complex real-world situations. In an analysis of the Bayesian classification problem showed that there are sound theoretical reasons for the apparently implausible efficiency of naive Bayes classifiers. Bayesian a comprehensive comparison with other classification algorithms showed that Bayes classification is outperformed by other approaches.

Bayes theorem provides a way of calculating probability, P(xi),from P(ci), P(xi),and P(xi|ci). Naive Bayes classifier assumes that the effect of the value of a predictor P(xi)on a given class P(ci) is independent of the values of other predictors in the naïve bayes classifier.This assumption is called class conditional independence.

$$P\left(\frac{c_i}{x_i}\right) = \frac{P(x_i/c_i)P(c_i)}{P(x_i)} \qquad (2)$$

Data Transformation

A data transformation converts a set of data values from the data format is illustrated and the source data can be converted into the data format of a destination data system. Data maps data elements from the source data system to the destination data system and captures any alteration that must occur code generation that creates the actual transformation program.

Normalization

Normalization is the data preprocessing technique for finding the rescale attribute value to fit in a specific range of values. Normalization technique is to normalize the attribute values for the attributes we are selected to find the values in the specific range. Here we have taken the attribute to rescale the preprocessing techniques to find the values. Collecting data is important while dealing with attributes of different units and scales for those normalization. Normalization technique should have the same scale of values for the comparision between the attributes.Here we have taken the attributes like age,workclass, education, marital status and sex. Normalization is a technique used to level the playing when looking at attributes that widely vary in size as a result of the units selected for representation .Label is the race to normalize the values.

Discretize

Discretizing is the data preprocessing technique that converting the selected numerical attribute into the nominal attribute by discretizing the numerical attribute into a specified number of bins. Automatically equal number of bins is generated but the range of those bins values may vary in these operation.The thresholds values of all the bins is selected in the way of finding bins that contains the same number of numerical values. The parameters are used to specify the number of bins in the values. The Discretize By Frequency operator creates bins in such a way that the number of unique values in all bins are equal. By Binning operator creates bins in such a way that the range of all bins is also equal. Here we are the nominal into the numerical attribute for finding the values for those sensitive attribute like race, school and their personal information.

Map

This Map operator can be applied on both numerical and nominal attributes. This operator can be used to replace nominal values (e.g. replace the value 'black' by the value 'others') as well as numerical values. A single mapping can be specified using the parameters replace what and replace by as in Replace operator. Multiple mappings can be specified through the value mappings parameter. Additionally, the operator allows denying a default mapping. This operator allows you to select attributes to make mappings in. This operator allows you to specify a regular expression. Attribute values of selected attributes that match this regular expression are mapped by the specified value mapping.

Table 1. A Sample of the most conservative detail values

Detail name	Detail value	Likelihood
Group member	George	45
Favorite movies	End of the spear	14
Activites	College republicans	58
Favorite books	Redeeming	63
Interests	Vegetarianism	11
Favorite music	deerhoof	22

IV. EXPERIMENTAL ANALYSIS

In this chapter we are discussing about our experimental analysis and design. We have used Rapidminer tool for result analysis. Rapidminer is the open source data mining tool. This will give the accuracy for classification and data transformation. Adult data set consists of 14 attributes in our experiments we have take the attributes like age ,work class, education, race, marital status and sex. For classification and the predicted values we have taken the work class and race for predicting the values. The selected attribute will be the predicted value for that naïve bayes classifier. Data transformation is used for transform the data in the data set and the information which would be replace can be made.

Table 2. A Sample Values from Adult Data Set

Attributes	Naïve bayes	Data transformation
Age	25	45
Workclass	45	37
Education	77	47
Marital status	20	56
Race	10	24
Sex	58	54

The details of the most conservative from the face book data set collected as shown in Table 1.that describes the details of a particular user information in the face book data set. Table 2 describes the Naïve bayes classification and transformation in the adult data set and shows the performance analysis of Naïve bayes classification and transformation from the Adult data set.

In our proposed work we have taken Adult data set from the UCI repository. Data transformation is to transfer the data in the network. In data transformation we use mapping, discretization and normalization for securing the information like sex and race information in the network.

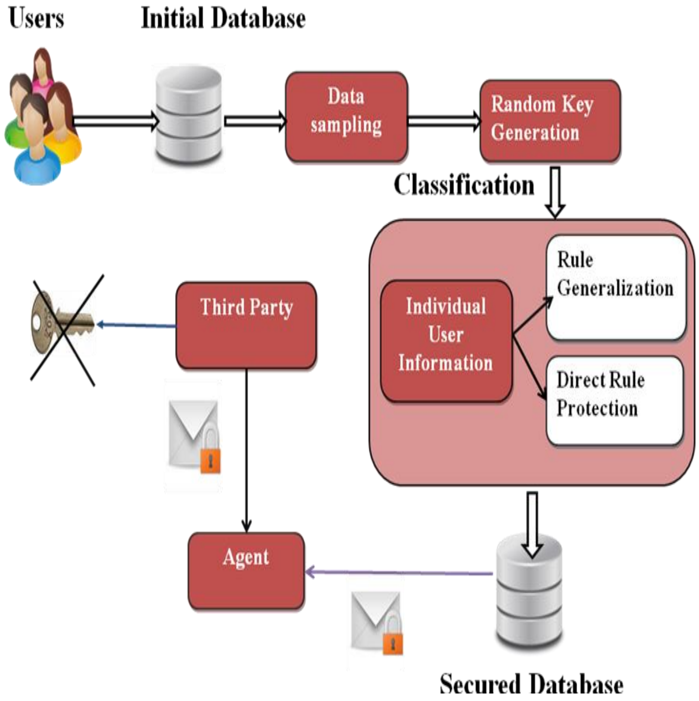
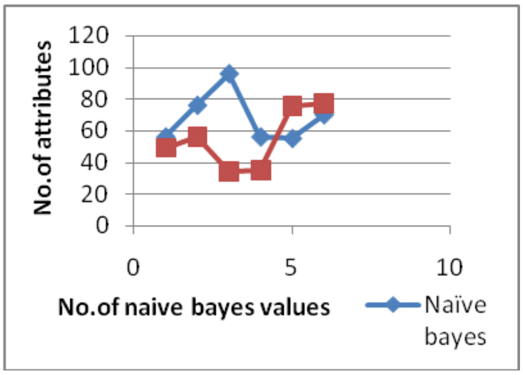


Fig 1: System Architecture for SPI-DM



V. CONCLUSION AND FUTURE WORK

This work proposed the securing personal information using data mining in networks. In this work initially Random key generation is used to authenticate the user is the trust person to enter into the network. Informations are collected from the adult data set in UCI repository. Using the naïve bayes classifier the attributes information are classified. The data transformation is applied to transform the data in the network. This classification and data transformation helps to identify the sensitive attributes. Once if the attribute is sensitive that information is to prevented from public networks. In future a method for preventing the personal information in social network and Detecting attacks possible in social network.

VI. REFERENCES

[1] G. Alexander Tartakovsky, S.Aleksey Polunchenko, and Grigory Sokolov ,"Efficient Computer Network Anomaly Detection by Changepoint Detection Methods", IEEE Transactions on signal processing, Vol. 7, No. 1, February 2013.

[2] Z.Adam and J.Marek Druzdzel," Knowledge Engineering for Bayesian Networks :How Common Are Noisy-MAX Distributions in Practice," IEEE Transactions on Systems, Man, And Cybernetics: Systems, Vol. 43, No. 1, January 2013.

[3] H.Cheng-Yuan, L.Yuan-Cheng, I-Wei Chen, Fu-Yu Wang, and T.Wei-Hsuan, "Statistical Analysis of False Positives and False Negatives from Real Traffic with Intrusion Detection/Prevention Systems," Topics In Network Testing, 2012.

[4] G.Alexander Tartakovsky, S.Aleksey Polunchenko, and Grigory Sokolov ,"Efficient Computer Network Anomaly Detection by Changepoint Detection Methods", IEEE Transactions on signal processing, Vol. 7, No. 1, February 2013.

[5] Z.Adam and J.Marek Druzdzel," Knowledge Engineering for Bayesian Networks :How Common Are Noisy-MAX Distributions in Practice," IEEE Transactions on Systems, Man, And Cybernetics: Systems, Vol. 43, No. 1, January 2013

[6] H.Cheng-Yuan, L.Yuan-Cheng, I-Wei Chen, Fu-Yu Wang, and T.Wei-Hsuan, "Statistical Analysis of False Positives and False Negatives from Real Traffic with Intrusion Detection/Prevention Systems," Topics In Network Testing, 2012.

[7] A. Friedman and A. Schuster, "Data Mining with Differential Privacy," Proceedings .,16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 493-502, 2010.

[8] P. Fogla and W. Lee, "Evading Network Anomaly Detection Systems: Formal Reasoning and Practical Techniques," In Proceedings of ACM Conference on Computer and Communications Security (CCS), Oct.–Nov. 2006.

[9] S.Hung-Min, C.Yao-Hsin, and Yue-Hsun," oPass: A User Authentication Protocol Resistant to Password Stealing and Password Reuse Attacks, "IEEE Transactions on Information Forensics and Security, Vol. 7, No. 2, April 2012.

[10] R. Kohavi and B. Becker, "UCI Repository of Machine Learning,"http://archive.ics.uci.edu/ml/datasets/Adult, 1996.

[11] A. Menon and C. Elkan, "Predicting Labels for Dyadic Data," Data Mining and Knowledge Discovery, Vol. 21, pp. 327- 343,2010.

[12] H.Raymond ,K. Murat, and T. Bhavani ,"Preventing The Private Information Inference on Social Networks, "IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 8, August 2013.

[13] H.Sara , F.Josep Domingo, "A methodology for Direct and Indirect Discrimination Prevention in Data mining, "IEEE Transactions on Knowledge and Data Engineering Vol. 25, No. 7, July 2013.

[14] H.Sara,F.JosepDomingo,andA.Martı´nezBalleste´,"Discrimination Prevention in Data Mining for Intrusion and CrimeDetection," Proceedings IEEE Symposium on Computational Intelligence in Cyber Security (CICS '11), pp. 47-54, 2011.

[15] S.Sathya Chandran , B. Sandeep , R. Marc Eisenbarth," Examining Intrusion Prevention System Events from Worldwide Networks," BADGERS'12, October 15, 2012,

Author's Details:

V.Vijayadeepa received her B.Sc degree from university of Madras and M.Sc degree from Periyar University. She has completed her M.Phil at Bharathidasan University.She is having 11 years of experience in collegiate teaching and She is a Head of the department of computer applications in Muthayammal college of Arts and Science affiliated by Periyar University. Her main research interests include personalized Web search, Web information retrieval, data mining, and information systems.

M.Deepika received her B.com.(CA), degree in Muthayammal College of arts and science from Periyar University, Salem (2007-2010) Tamil Nadu (India). Then finished MCA, degree in Veltech multitech Dr.Rangarajan Dr.Sakunthala Engineering college from anna university, Chennai (2010-2013) Tamil Nadu (India). She is the M.Phil Research Scholar of Muthayammal College of Arts and Science ,Rasipuram. Periyar university, salem. Her area of interest is Data mining.