Preservation of Privacy in Data Bases

Manish Tiwari and Rakesh Rathi

***Abstract*—In the modern world data privacy is very important during the Transaction process. Whenever user give his/her information he is not sure whether the information will be secure or not. Privacy Preservation mainly deals with the hiding of the user data with his/her choice. In our paper we basically dealing with the privacy issue that is how it can be secure during the time it travels to databases/data warehouse. For this purpose we implemented the algo based on bayes Estimation. This algo basically works on bayes probability that is we basically applied on the email system. Whenever the user types any mail his whole mail will be process through our algo which uses the word list from the stored data base and if the mail is fine and not contains any threat material than it will get stored in the data base. Our algo basically counts the conditional probability of the words from the number which is given in the data base. First of all words are assigned the threat numbers according to the previous usage in the transaction process in internet than their total count is done which finally uses in the probability. This probably if greater than assigned probability i.e. in our case .5 than it will get stored in the database otherwise it gets discarded.**

***Index Terms*—Bayes, database, probablity**

I. Introduction

With the advance of the information age, data collection and data analysis have exploded both in size and complexity. The attempt to extract important patterns and trends from the vast data sets has led to a challenging field called Data Mining. When a complete data set is available, various statistical, machine learning and modeling techniques can be applied to analyze the data.

Privacy-Preserving Data Mining (PPDM) has emerged to address this issue. The research of PPDM is aimed at bridging the gap between collaborative data mining and data confidentiality. It involve s many areas such as statistics, computer sciences, and social sciences. It is of fundamental importance to homeland security, modern science, and to our society in general.

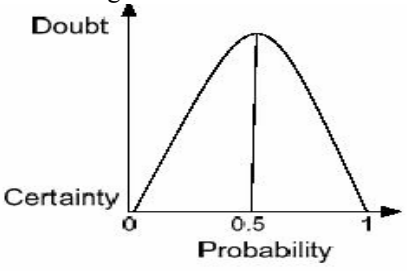
II. Related Work

Sabah Al-Fedaghi first proposed using randomization to solve PPDM problems [2]. In their randomization scheme, a random number is added to the value of a sensitive attribute. For example, if xi is the value of a sensitive attribute, xi + r, rather than xi, will appear in the database, where r is a random value drawn from some distribution. It is shown that given the distribution of random noises, recovering the distribution of the original data is possible. The randomization techniques have been used for a variety of privacy preserving data mining work Non Zhang, Shengqaom, Wei Zaho challenged the randomization schemes, and they pointed out that randomization might not be secure [4]. They proposed a random matrix-based Spectral Filtering (SF) technique to recover the original data from the perturbed data. Their results have shown that the recovered data can be reasonably close to the original data. The results indicate that for certain types of data; randomization might not preserve privacy as much as we have believed.

A. Bayes Theorem

Probability is defined as a quantitative measure of uncertainty state of information or event. It has an index which ranges from 0 to 1. It is also approximated through proportion of number of events over the total experiment. If the probability of a state is 0 (zero), we are certain the state will not happen. However if the probability is 1, the event will surely happen. A probability of 0.5 means we have maximum doubt about the state that will happen. The following section describes some of the basic probability formulas that will be used:

Conditional probability: The probability of an event may depend on the occurrence or non-occurrence of another event. This dependency is written in terms of conditional probability: P (A|B).“The probability that A will happen given that B already has” or “the probability to select A among B” Notice that B is given first, and we find the Proportion of A among B:



$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B|A) = P(A \cap B) / P(A)$$

$$P(A|B) = P(A \cap B) / P(B)$$

$$P(A \cap B) = P(B|A)\,P(A) = P(A|B)\,P(B)$$

Fig. 1. Bayes probability

III. Experiment Results

Privacy is an important mechanism for securing data, work or development strategically. Its success in hiding the mentioned fields has been a motivation and also a guide to

Manuscript received March 11, 2012; revised June 13, 2012.

The authors are with the Department of Information Technology and Computer Science Government Engineering College Ajmer (e-mail: manish.j.tiwari@gmail.com)

DOI: 10.7763/IJFCC.2012.V1.100

bring it to purpose elsewhere. Data warehouse having extremely high amount of data tempts a mechanism that could be of help we took the case study of hospital management in which we made the proper web interface. The main form which we made from an algo prospective was the online appointment form, feedback form and the patient desk. In online appointment form patient can fix the appointment with the specified department by filling his/her details. In this we hide his/her first name so that if anybody wants to access his information would not be able to do so. In patient desk the information can be retrieved only when the correct email id is given by the patient. The hidden name will again be displayed in correct form from the data base. Secondly in feedback form the feedback is checked using the baye's algo and if that doesn't contain any threat words than it will get stored in the database.

After implementation and testing phase we find some loopholes in our designing and implementation so we revalidated and corrected on errors like slow speed. We have designed the system from the point of simulation which can be used for commercial purpose after certain modification. The privacy of data as well as content has been proposed at the dimension level. The case study depicts the proper scenario for the requirement and need of the proposed bayes estimation strategy.

Tabular Description Of calculated value:

TABLE I : RESULT SUMMARY

Mail Type	Total Count	P(mail Type)
Spam	508042	.209
Non Spam	1925892	.79
	243934	1

Here values like .20 represents 20% and .79 represents 79 %Comparison of our developed software with previously developed are:

IV. CONCLUSION

In this paper, we studied about privacy and how to preserve data from the threat. The algorithm which we applied successfully for this purpose is Baye's probability. One can use any platform for the implementation purpose as this algo is platform independent. The algorithm can be applied to other techniques also for privacy preservation.

Evaluation Matrix	**Application developed**	**Commercial Software**
Compliance with existing database	Yes	No
Compliance with large database	Yes	May or may not
Handling High Quality of data	Yes	May or may not
Flexible and Expandable	Yes	Yes
Compatibility With All Software	Yes	Yes
Shows the before and after value changes	Yes	Yes

V. ACKNOWLEDGEMENT

The authors thank Government Engineering College Ajmer for the resources without which the research would not have been possible.

REFERENCES

[1] L. Compagna, P. E. Khoury, F. Massacci, and R. Thomas, "How to capture, model, and verify the knowledge of legal, security, and privacy experts: a pattern-based approach," *ICAIL'07*, pp. 4-8, Palo Alto, CA USA, 2007.

[2] S. S. A. Fedaghi, "Beyond Purpose-Based Privacy Access Control," *18th Australasian Database Conference* (ADC 2007)

[3] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "Implementing P3P Using Database Technology," in *Proceedings of the 19th International Conference on Data Engineering* (ICDE'03)

[4] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu. "Hippocratic Databases," in *Proceedings of the 28th VLDB Conference*, Hong Kong, China, 2002

[5] K. Karaca and A. Levi, "Towards a framework for security analysis of multiple password schemes," October 2001 EC '01: in P*roceedings of the 3rd ACM conference on Electronic Commerce*, 2001.

[6] S. A. Fedaghi, "How sensitive is your personal information," March 2007 SAC '07: in *Proceedings of the 2007 ACM symposium on applied*, 2007.

[7] J. J. Yan, "A note on proactive password checking," September 2001 NSPW '01: in *Proceedings of the 2001 workshop on New security paradigms Publisher*: ACM, 2001.

[8] N. Zhang and S. W. Zaho, "A new scheme on privacy preserving data classification," KDD'05, pp. 21-24, 2005, Chicago, Illinois, USA, 2005.