INVESTIGATION AND NEED OF PRIVACY PRESERVING IN MOBILE SENSING SYSTEM

G. DIVYA
Research Scholar,
PG & Research Department of Computer Science,
Tiruppur Kumaran College for Women,
Tirupur, India.
E-mail : maragathamani109@gmail.com

Ms. P.BANUMATHI M.Sc(CT)., M.Phil.,
Assistant Professor,
Department of Computer Applications,
Tiruppur Kumaran College for Women,
Tirupur, India.
E-mail : banutech@gmail.com

ABSTRACT:

Today, we are able to discover and store increasing amounts of detailed data on human activities: automated payment systems record our purchases; search engines record logs of our queries on the Internet; mobile devices record the trajectories of our movements; and so on. Such information is at the heart of the idea of a 'knowledge society', in which our understanding of social phenomena is sustained by data mining. It is evident that maintaining control of personal data is increasingly difficult and cannot simply be accomplished by de-identification.

Our idea is to inscribe privacy protection into the knowledge discovery technology by design, so that the analysis incorporates the relevant privacy requirements from the start. In recent years, privacy-preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. A number of algorithmic techniques have been designed for privacy-preserving data mining. In this paper, we provide a review of the state-of-the-art methods for privacy. We discuss methods for randomization, *k*-anonymization, and distributed privacy-preserving data mining.

Keywords: Privacy, Preserving, Data Mining, Mobile, Sensing, Knowledge Discovery, Techniques and Algorithms.

I. INTRODUCTION

Data mining is a technique that deals with the extraction of hidden predictive information from large database. It uses sophisticated algorithms for the process of sorting through large amounts of data sets and picking out relevant information. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. With the amount of data doubling each year, more data is gathered and data mining is becoming an increasingly important tool to transform this data into information. Long process of research and product development evolved data mining.

This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

The most commonly used techniques in data mining are:

***Artificial neural networks*:** Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

Genetic algorithms: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of evolution.

***Nearest neighbor method*:** A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbor technique.

***Rule induction*:** The extraction of useful if-then rules from data based on statistical significance.

There is a rapidly growing body of successful applications in a wide range of areas as diverse as: analysis of organic compounds, automatic abstracting, credit card fraud detection, financial forecasting, medical diagnosis etc. Some examples of applications (potential or actual) are:

- A supermarket chain mines its customer transactions data to optimize targeting of high value customers
- A credit card company can use its data warehouse of customer transactions for fraud detection
- A major hotel chain can use survey databases to identify attributes of a 'high-value 'prospect.

Applications can be divided into four main types:

- Classification
- Numerical prediction
- Association
- Clustering.

Data mining using labeled data (specially designated attribute) is called supervised learning.

Classification and numerical prediction applications falls in supervised learning. Data mining which uses unlabeled data is termed as unsupervised learning and association and clustering falls in this category.

II. DATA MINING & PRIVACY

Data mining deals with large database which can contain sensitive information. It requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. Advancement of efficient data mining technique has increased the disclosure risks of sensitive data. A common way for this to occur is through data aggregation.

Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.

What data mining causes is social and ethical problem by revealing the data which should require privacy? Providing security to sensitive data against unauthorized access has been a long term goal for the database security research community and for the government statistical agencies. Hence, the security issue has become, recently, a much more important area of research in data mining. Therefore, in recent years, privacy-preserving data mining has been studied extensively.

PPDM can be classified according to different categories. These are

Data Distribution- The PPDM algorithms can be first divided into two major categories, centralized and distributed data, based on the distribution of data. In a centralized database environment, data are all stored in a single database; while, in a distributed database environment, data are stored in different databases. Distributed data scenarios can be further classified into horizontal and vertical data distributions. Horizontal distributions refer to the cases where different records of the same data attributes are resided in different places. While in a vertical data distribution, different attributes of the same record of data are resided in different places. Earlier research has been predominately focused on dealing with privacy preservation in a centralized database. The difficulties of applying PPDM algorithms to a distributed database can be attributed to: first, the data owners have privacy concerns so they may not willing to release their own data for others; second, even if they are willing to share data, the communication cost between the sites is too expensive.

Hiding Purposes - The PPDM algorithms can be further classified into two types, data hiding and rule hiding, according to the purposes of hiding. Data hiding refers to the cases where the sensitive data from original database like identity, name, and address that can be linked, directly or indirectly, to an individual person are hided. In contrast, in rule hiding, the sensitive knowledge (rule) derived from original database after applying data mining algorithms is removed. Majority of the PPDM algorithms used data hiding techniques. Most PPDM algorithms hide sensitive patterns by modifying data.

Privacy Preservation Techniques - PPDM algorithms can further be divided according to privacy preservation techniques used. Four techniques – sanitation, blocking, distort, and generalization -- have been used to hide data items for a centralized data distribution. The idea behind data sanitation is to remove or modify items in a database to reduce the support of some frequently used item sets such that sensitive patterns cannot be mined.

The blocking approach replaces certain attributes of the data with a question mark. In this regard, the minimum support and confidence level will be altered into a minimum interval. As long as the support and/or the confidence of a sensitive rule lie below the middle in these two ranges, the confidentiality of data is expected to be protected. Also known as data perturbation or data randomization, data distort protects privacy for individual data records through modification of its original data, in which the original distribution of the data is reconstructed from the randomized data. These techniques aim to design distortion methods after which the true value of any individual record is difficult to ascertain, but “global” properties of the data remain largely unchanged. Generalization transforms and replaces each record value with a corresponding generalized value.

III. PRIVACY PRESERVING IN MOBILE SENSING SYSTEM

Mobile devices such as smart phones are gaining an ever increasing popularity. Most smart phones are equipped with a rich set of embedded sensors such as camera, microphone, GPS, accelerometer, ambient light sensor, gyroscope, etc. The data generated by these sensors provides opportunities to make sophisticated inferences about not only people (e.g., human activity, health,

location, social event) but also their surrounding (e.g., pollution, noise, weather, oxygen level), and thus can help improve people's health as well as life. This enables various *mobile sensing* applications such as environmental monitoring [1], traffic monitoring [2], healthcare [3], etc.

In many scenarios, aggregation statistics need to be periodically computed from a stream of data contributed by mobile users [4], in order to identify some phenomena or track some important patterns. Although aggregation statistics computed from time-series data is very useful, in many scenarios, the data from individual user may be privacy-sensitive, and users do not trust any single third-party aggregator to see their data in clear text.

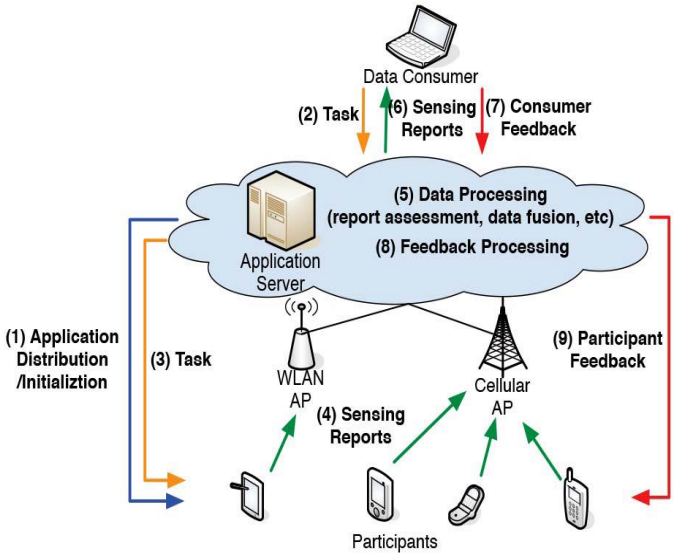

Figure 1: - Mobile Sensing System

In general, privacy preservation occurs in two major dimensions: users' personal information and information concerning their collective activity. We refer to the former as individual privacy preservation and the latter as collective privacy preservation, which is related to corporate privacy in (Clifton et al., 2002).

- ***Individual privacy preservation*:** The primary goal of data privacy is the protection of personally identifiable information. In general, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Thus, when personal data are subjected to mining, the attribute values associated with individuals are private and must be protected from disclosure. Miners are then able to learn from global models rather than from the characteristics of a particular individual.
- ***Collective privacy preservation*:** Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. We refer to the protection of sensitive knowledge as collective privacy preservation. The goal here is quite similar to that one for statistical databases, in which security control mechanisms provide aggregate information about groups (population) and, at the same time, prevent disclosure of confidential information about individuals. However, unlike as is the case for statistical databases, another objective of collective privacy preservation is to protect sensitive knowledge that can provide competitive advantage in the business world.

In the case of collective privacy preservation, organizations have to cope with some interesting conflicts. For instance, when personal information undergoes analysis processes that produce new facts about users' shopping patterns, hobbies, or preferences, these facts could be used in recommender systems to predict or affect their future shopping patterns. In general, this scenario is beneficial to both users and organizations. However, when organizations share data in a collaborative project, the goal is not only to protect personally identifiable information but also sensitive knowledge represented by some strategic patterns.

IV. PERFORMANCE EVALUATION METRICS

The new system aims at simultaneously achieving satisfactory performance on three objectives: privacy, accuracy and efficiency. The specific metrics used to evaluate Privacy Preserving performance w.r.t. privacy; accuracy and efficiency are given below.

Privacy Metrics: - Privacy can be computed at various levels, namely: Basic Privacy (BP), Re interrogated Privacy (RP) and Strict Privacy (SP).

Accuracy Metrics: - The association rule mining errors can be quantified in terms of *Support Error* _ and *Identity Error*.

Efficiency Metric: - This metric determines the runtime overheads resulting from mining the distorted database as compared to the time taken to mine the original database. This is simply measured as the inverse ratio of the running times between Apriori on the original database and executing the same code augmented with new system on the distorted database.

V. CONCLUSION

This paper gives data mining that protects privacy and technologies involved in mobile sensing system fields. The idea of privacy to protect data mining is to extract information from active data. Applications use some data mining classification, suite, computation, and association rules and so on. This paper proposes a privacy-preserving identification mechanism for mobile sensing systems, which can not only protect participants' privacy, but also recognize participants' identities to ensure data trustworthy. The mechanism is based on trusted computing.

VI. REFERENCES

[1] Xie, Q.; Wang, L. Efficient privacy-preserving processing scheme for location-based queries in mobile cloud. In Proceedings of the IEEE International Conference on Data Science in Cyberspace, Changsha, China, 13–16 June 2016.

[2] Li, X.Y.; Jung, T. Search me if you can: Privacy-preserving location query service. In Proceedings of the IEEE INFOCOM, Turin, Italy, 14–19 April 2013; pp. 2760–2768.

[3] Nishide, T.; Sakurai, K. Distributed paillier cryptosystem without trusted dealer. In Information Security Applications; Springer: New York, NY, USA, 2010; pp. 44–60.

[4] Dou, Y.; Zeng, K.C.; Yang, Y. Poster: Privacy-Preserving Server-Driven Dynamic Spectrum Access System. In Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, Paris, France, 7–11 September 2015; pp. 218–220.

[5] Duckham, M.; Kulik, L. A formal model of obfuscation and negotiation for location privacy. In Pervasive Computing; Springer: New York, NY, USA, 2005; pp. 152–170.

[6] Shankar, P.; Ganapathy, V.; Iftode, L. Privately querying location-based services with SybilQuery. In Proceedings of the 11th International Conference on Ubiquitous Computing, Orlando, FL, USA, 30 September–3 October 2009; pp. 31–40.

[7] A. V. Ev¯mievski, J. Gehrke, and R. Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proc. of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 211{222, San Diego, CA, USA, June 2003.

[8] A. V. Ev¯mievski, R. Srikant, R. Agrawal, and J. Gehrke. Privacy preserving mining of association rules. In *Proc. of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 217{228, 2002.

[9] C. Farkas and S. Jajodia. The inference problem: A survey. *ACM SIGKDD Explo-rations Newsletter*, 4(2):6{11, December 2002.

[10] S. E. Fienberg. Privacy and confidentiality in an e-commerce world: Data mining, data warehousing, matching and disclosure limitation. *Statistical Science*, 21:143{154, 2006.

[11] S. E. Fienberg and J. McIntyre. Data swapping: Variations on a theme by Dalenius and Reiss. *Journal of O±cial Statistics*, 21:309{323, 2005.

[12] American Association for Artificial Intelligence. Fraud detection and prevention. avail-able from http://www.aaai.org/aitopics/html/fraud.html. visited on 07.07.06.

[13] Murat Kantarcio_glu, Jiashun Jin, and Chris Clifton. When do data mining results violate privacy? In Proc. of ACM SIGKDD, pages 599{604, New York, NY, USA, 2004. ACM Press.

[14] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. On the privacy preserving properties of random data perturbation techniques. In Proc. of ICDM, page 99, Washington, DC, USA, 2003. IEEE Computer Society.

[15] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? In FOCS, pages 531{540, 2008.

[16] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Incognito: E_cient full domain k-anonymity. In Proc. of SIGMOD, pages 49{60, New York, NY, USA, 2005. ACM Press.

[17] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Mondrian multi-dimensional k-anonymity. In Proc. of ICDE, April 2006.

[18] Kristen LeFevre, David J. DeWitt, and Raghu Ramakrishnan. Workload-aware anonymization. In KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 277{286, New York, NY, USA, 2006. ACM.

[19] "Privacy Preserving Data Mining - IBM Research: Almaden: San Jose

[20] D.Aruna Kumari, Dr.K.Rajasekhara rao, M.suman "Privacy Preserving Clustering in DDM using Cryptography"in TJ-RJCSE-IJ-06.

[21] Maloji Suman,Habibulla Khan,M. Madhavi Latha,D. Aruna Kumari "Speech Enhancement and Recognition of Compressed Speech Signal in Noisy Reverberant Conditions " Springer -Advances in Intelligent and Soft Computing (AISC) Volume 132, 2012, pp 379-386.