PRIVACY PRESERVING ASSOCIATION RULE MINING USING ATTRIBUTE-IDENTITY MAPPING

IBRAHEEM JAFAR

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JANUARY 2017

Dedicated to my beloved parents, Alhaji Jafar Abubakar and Hajiya Rahmatu Ibraheem, for their wonderful support and encouragement.

ACKNOWLEDGEMENT

I first thank my Creator, Cherisher and Sustainer for His countless blessings on me. Among the blessings is enabling me to successfully complete this dissertation and the master programme in general.

My appreciation goes to my supervisor, Dr. Maheyzah Md. Siraj, for her enormous and vital support in various forms, from inception till completion of this dissertation. I thank the coordinator, master of computer science programme, Dr. Mohamad Shukor Bin Talib, for facilitating necessary coordination towards making the dissertation a success. Also, my appreciation goes to all academic and non-academic staff of faculty of computing, whose cumulative efforts transformed into completion of this dissertation and many more.

My profound gratitude goes to my lovely parents, Alhaji Jafar Abubakar and Hajiya Rahmatu Ibrahim for their support, prayer and guidance in all my endeavours. I also thank my beloved wife, Asiya Ahmad (Asibeb); my children, Abdulkadir Ibraheem (Amir) and Rahmatu Ibraheem (Amirah) for their endurance while I was away for this study. I remain thankful to all members of my family for their supports, prayers and understanding.

I especially express my appreciation to Kaduna State Government for supporting the study, through 2014/2015 Overseas Scholarship Programme. Similar appreciation goes to the Management of National Teachers Institute (NTI) Kaduna, for their support.

ABSTRACT

Association rule mining uncovers hidden yet important patterns in data. Discovery of the patterns helps data owners to make right decision to enhance efficiency, increase profit and reduce loss. However, there is privacy concern especially when the data owner is not the miner or when many parties are involved. This research studied privacy preserving association rule mining (PPARM) of horizontally partitioned and outsourced data. Existing research works in the area concentrated mainly on the privacy issue and paid very little attention to data quality issue. Meanwhile, the more the data quality, the more accurate and reliable will the association rules be. Consequently, this research proposed Attribute-Identity Mapping (AIM) as a PPARM technique to address the data quality issue. Given a dataset, AIM identifies set of attributes, attribute values for each attribute. It then assigns 'unique' identity for each of the attributes and their corresponding values. It then generates sanitized dataset by replacing each attribute and its values with their corresponding identities. For privacy preservation purpose, the sanitization process will be carried out by data owners. They then send the sanitized data, which is made up of only identities, to data miner. When any or all the data owners need(s) ARM result from the aggregate data, they send query to the data miner. The query constitutes attributes (in form of identities), *minSup* and *minConf* thresholds and then number of rules they are want. Results obtained show that the PPARM technique maintains 100% data quality without compromising privacy, using Census Income dataset.

ABSTRAK

Perlombongan peraturan persatuan mendedahkan corak tersembunyi yang penting dalam data. Penemuan corak membantu pemilik data untuk membuat keputusan yang tepat untuk meningkatkan kecekapan, meningkatkan keuntungan dan mengurangkan kerugian. Walau bagaimanapun, terdapat kebimbangan privasi terutama apabila pemilik data tidak melombong atau apabila banyak pihak yang terlibat. Kajian ini mengkaji privasi perlombongan menggunakan persatuan peraturan (PPARM) terhadap data mendatar dan sumber luar. Penyelidikan sedia ada tertumpu pada isu privasi dan memberi perhatian yang sangat sedikit untuk isu kualiti data. Sementara itu, lebih kualiti data, lebih tepat dan boleh dipercayai akan peraturan persatuan berkenaan. Oleh itu, kajian ini mencadangkan Sifat-Identity Pemetaan (AIM) sebagai teknik PPARM untuk menangani isu kualiti data. Diberi dataset, AIM mengenal pasti menetapkan sifat-sifat, nilai atribut bagi setiap atribut. Ia kemudian memberikan identiti 'unik' untuk setiap satu daripada sifat-sifat dan nilai yang berkaitan mereka. Ia kemudian menjana dataset dibersihkan dengan menggantikan setiap atribut dan nilai-nilainya dengan identiti yang sama mereka. Untuk privasi tujuan pemeliharaan, proses sanitasi akan dijalankan oleh pemilik data. Mereka kemudian menghantar data dibersihkan, yang terdiri daripada hanya identiti, untuk pelombong data. Apabila mana-mana atau semua pemilik data perlu keputusan ARM (s) dari data agregat, mereka menghantar pertanyaan kepada pelombong data. Pertanyaan ini merupakan sifat-sifat (dalam bentuk identiti), *minSup* dan *minConf* ambang dan kemudian beberapa peraturan yang mereka mahu. Keputusan yang diperolehi menunjukkan bahawa teknik PPARM mengekalkan 100% kualiti data tanpa menjejaskan privasidengan menggurakan data Census Income.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AIM	-	Attribute - Identity Mapping
ARM	-	Association Rule Mining
CQA	-	Certified Quality Auditor
IP	-	Internet Protocol
LHS	-	Left Hand Side
minConf	-	Minimum Confidence
minSup	-	Minimum Support
PPARM	-	Privacy Preserving Association Rule Mining
PPDM	-	Privacy Preserving Data Mining
RHS	-	Right Hand Side
US	-	United States
WEKA	-	Waikato Environment for Knowledge Analysis

LIST OF SYMBOLS

==>	Association
D	Unsanitized dataset
D'	Sanitized dataset

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Information and communication technology, high performance hardware components such as processor and storage capacity, rapid and efficient access to social networks, advanced database technologies, among other factors, have made it possible to generate high volume of data efficiently. Reduction in cost of hardware, computerization of many tasks and access to the technologies also contribute greatly to the efficient production of data.

Data, to be meaningful, need to be processed and analyzed to generate useful information out of it. The process of information generation from data is mostly applicable to small-scale data. Also, in the small-scale data, the kind of information to be generated is known already; for example, finding an average score from a set of scores. However, for huge amount of data, there are interesting patterns and knowledge that can be of high value for decision making. The activity of realizing such patterns and knowledge out of huge amount of data is known as data mining (Xu et al., 2014).

Data mining is a domain that involves many methods from many areas such as statistics, database management, information retrieval, data warehousing, machine learning and pattern recognition (Patel and Donga, 2015). There are three basic data mining techniques, which are classification clustering and association rule.

Classification, as the name implies, involves assigning class value to an instance of database/dataset. Before the actual classification stage, learning stage should take place. At the learning stage, sample dataset, known as training set, with class labels for each instance is provided. Based on class labels of the training set, classifiers or models are generated. Next stage is the actual classification. Given an instance of dataset, whose class value is unknown, a classifier is employed to estimate appropriate class value to the instance. Classification is categorized as supervised learning because class labels are predefined in the training set (Patel and Donga, 2015).

Clustering involves grouping of instances, based on similarities in their attribute values. Each group or cluster is given a description label. Unlike classification, clustering is categorized as unsupervised learning. This is because the clusters' labels are not predefined (Jain and Srivastava, 2013).

The third technique, association rule, generates rules that describe relationships among instances of dataset. It is also known as frequent patterns/itemset mining, because it tells how frequent patterns/itemsets occur together. The frequency is usually expressed in percentage (Jain and Srivastava, 2013).

As the technology for generating data and mining it improves, threat to data security and privacy also increases, which creates data privacy concern. In the case of data mining, privacy preserving data mining PPDM, emerged as a general research area and measure to preserve privacy in data mining. Specifically, privacy preserving association rule mining PPARM, emerged as PPDM wing for association rule, which is the focus of this research.

1.2 Problem Background

PPARM is a research discipline that seeks to facilitate privacy of both data and data mining result in association rule mining with minimum or no compromise to data and data mining result quality. There are two (2) major aspects in PPARM, which are data distribution and privacy preservation approach.

Data distribution is based on number of sources of data. This can be single source or multiple sources. When there is single source of data, the data distribution is called centralized; whereas when data come from multiple sources, the data distribution is called distributed. In both distributions, privacy issues arise especially when there is need to outsource data for mining purpose (Abdel Wahab et al., 2014).

Data outsourcing is a situation where data owners give their data to third party, for mining purpose. This happens majorly because of two (2) reasons. Firstly, the data owner does not have resources and/or expertise needed for mining purpose. This is the case of centralized data. Secondly, multiple data owners/parties need data mining result their data aggregate. This is the case of distributed data (Wadkar and Shelke, 2015). This research is based on distributed data with horizontal partitioning. That is, the various data owners have the same database schema, but different instances.

Recent research works in PPARM include that by Abdel Wahab et al. (2014). The work proposed DiffGen algorithm, which facilitates privacy preservation by generalization. While the research focused on scalability to large dataset, generalization leads to information loss and hence data quality is negatively affected.

Santhoshi and Rao (2015) proposed a cryptographic based PPARM approach which uses 1-1 substitution cipher and encryption of items in the transaction. While the work assured data quality, the dataset used does not cater for multi-party data ownership.

Similarly, Vemuri and Nutalapati (2015) proposed Rob Frugal algorithm, which is also based on 1-1 substitution cipher and addition of fake transactions. While 1-1 substitution cipher does not reduce data quality, addition of fake transactions does.

Wadkar and Shelke (2015) proposed another PPARM algorithm based on Rob Frugal and *k*-means clustering. Use of *k*-means algorithm is applicable only on numeric data, whereas real-life dataset contains both numeric and non-numeric data. Similarly, the *k*-means clustering serves the function of generalization, which leads to information loss and hence reduces data quality.

Similarly, Mohammed et al. (2016), proposed a PPARM technique called Stochastic Standard Map. While the work ensured data quality, the dataset on which the approach is tested is not a real-life dataset.

It is obvious therefore, most research works paid less attention to data quality. Meanwhile, the more the data quality, the more accurate and reliable will association rules be. Qi and Zong (2012) highlighted issues to consider when developing a privacy preservation model for data mining. Among the issues is data utility, which is another name for data quality. Also, Abdel Wahab et al. (2014) identified three (3) issues in developing a PPARM model. First is ensuring privacy of data, second is ensuring data quality of the generated rules and third is ensuring privacy of the association rules.

1.3 Statement of the Problem

Data quality is a key issue in PPARM. This is because privacy preservation techniques achieve their objectives by modifying dataset in order to sanitize it. This however, results to reduction in data quality.

This research proposed Attribute-Identity Mapping (AIM), as privacy preservation technique to address the problem data quality in PPARM.

1.4 Research Goal

The goal of this research is address problem of data quality in PPARM, using AIM privacy preservation technique.

1.5 Research Objectives

Based on the problem statement of the research, below are the objectives of the research:

i. To design and implement Attribute-Identity Mapping (AIM), as privacy preservation technique, to sanitize the original dataset

ii. To apply Association Rule Mining (ARM) technique on the sanitized dataset, to generate identity-based ARM results

iii. To make use of the identity-based ARM results, to facilitate recovery of actual ARM results

1.6 Research Scope

The research has under-listed as its scope:

i. Based on data distribution, the research accommodates horizontally portioned data, where data sources maintain the same schema but different instances

ii. The datasets used for implementation of this research is Census Income dataset

iii. The data mining software used is Waikato Environment for Knowledge Analysis (WEKA), version 3.8

iv. Performance evaluation is based on data quality metrics, which are accuracy, dissimilarity measure and completeness

1.7 Significance of the Research

The research contributes to knowledge in field of privacy preserving association rule mining as follows:

i. The research introduced Attribute-Identity Mapping (AIM), as privacy preservation technique in collaborative association rule mining, by introducing Attribute-Identity Mapping technique

ii. The technique ensures data quality in addition to privacy in PPARM

1.8 Organization of the Research

The research is organized into chapters, comprising of six (6) chapters. Chapter 1 gives an overview of what the research is all about. Topics highlighted in the chapter are research overview, problem background, problem statement, research goal, objectives and scope.

Chapter 2 is on literature review. Highlights of the chapter are PPDM conceptualization, Design of PPDM model, existing PPDM techniques, research focus, association rule mining, privacy preserving association rule mining (PPARM). Existing research works in PPARM were highlighted and research gap in the PPARM research area was identified.

Chapter 3 is research methodology. It discussed research framework which is divided in phases. Each phase corresponds to a research objective. Also, discussed in the chapter is research instrumentation. The chapter concludes with details about dataset used in the research.

Chapter 4 discussed design and implementation of the research objectives. These are design and implementation of AIM technique; association rule mining on both sanitized and unsanitized datasets; and recovery of actual association rule mining results from identity-based results.

Chapter 5 is on results and discussion. It highlighted about the experimental results obtained from Chapter 4. The results were discussed and evaluated based data quality metrics.

Chapter 6 highlighted research problem, steps taken to tackle the problem and research achievements. Finally, future works in the research area were highlighted.

REFERENCES

Abdel Wahab, O., Hachami, M. O., Zaffari, A., Vivas, M. & Dagher, G. G. (2014). *DARM: a privacy-preserving approach for distributed association rules mining on horizontally-partitioned data.* Proceedings of the 18th International Database Engineering & Applications Symposium. ACM, 1-8.

Aggarwal, C. C. & Philip, S. Y. (2008). *A general survey of privacy-preserving data mining models and algorithms,* Springer.

Agrawal, R., Imieliński, T. & Swami, A (1993). *Mining association rules between sets of items in large databases.* Acm sigmod record. ACM, 207-216.

Agrawal, R. & Srikant, R (1994). *Fast algorithms for mining association rules.* Proc. 20th int. conf. very large data bases, VLDB. 487-499.

Agrawal, R. & Srikant, R (2000). *Privacy-preserving data mining*. ACM Sigmod Record. ACM, 439-450.

Ahmed, A. B. E. D. & Elaraby, I. S. (2014). *Data Mining: A prediction for Student's Performance Using Classification Method*. World Journal of Computer Application and Technology, 2**,** 43-47.

Aldeen, Y., Salleh, M. & Razzaque, M. (2015). *A comprehensive review on privacy preserving data mining*. SpringerPlus, 4**,** 1-36.

Bertino, E., Fovino, I. N. & Provenza, L. P. (2005). *A framework for evaluating privacy preserving data mining algorithms*. Data Mining and Knowledge Discovery, 11**,** 121-154.

Bertino, E., Lin, D. & Jiang, W. (2008). *A survey of quantification of privacy preserving data mining algorithms*. Privacy-preserving data mining. Springer.

Bora, S. P (2011). *Data mining and ware housing*. Electronics Computer Technology (ICECT), 2011 3rd International Conference on, 8-10 April 2011 2011. 1-5.

Brickell, J. & Shmatikov, V. (2009). *Privacy-preserving classifier learning*. Financial Cryptography and Data Security. Springer.

Coronel, C. & Morris, S. (2016). *Database Systems: Design, Implementation, & Management*, Cengage Learning.

Dowd, J., Xu, S. & Zhang, W. (2006). Privacy-preserving decision tree mining based on random substitutions. *Emerging Trends in Information and Communication Security*. Springer.

Elrahman, A. & Idris, O. a. E. (2013). *Comparative study of k-anonymity algorithms for privacy preserving datamining.* Universiti Teknologi Malaysia, Faculty of Computing.

Gavison, R. (1980). Privacy and the Limits of Law. *The Yale Law Journal*, 89**,** 421-471.

Helm, L. & Hahsler, P. D. D. M. (2007). *Fuzzy Association Rules*. Vienna University of Economics and Business Administration.

Indumathi, J. (2012). *A Generic Scaffold Housing the Innovative Modus Operandi for Selection of the Superlative Anonymisation Technique for Optimized Privacy Preserving Data Mining*, INTECH Open Access Publisher.

Irvine, U. C. (1987). Congressional Voting Records Dataset University California Irvine.

Irvine, U. C. (1996). *Census Income Dataset*. University California Irvine.

Jain, D., Khatri, P., Soni, R. & Chaurasia, B. K. (2012). *Hiding sensitive association rules without altering the support of sensitive item(s*). Advances in Computer Science and Information Technology. Networks and Communications. Springer.

Jain, N. & Srivastava, V. (2013). *Data Mining techniques: A survey paper*. IJRET: International Journal of Research in Engineering and Technology, 2**,** 2319-1163.

Jain, Y. K., Yadav, V. K. & Panday, G. S. (2011). *An efficient association rule hiding algorithm for privacy preserving data mining*. International Journal on Computer Science and Engineering, 3**,** 2792-2798.

Jyoti, D. R. (2015). *A Review on Privacy Preserving Data Mining*. International Journal of Emerging Trends in Science and Technology, 2.

Kavitha, M. R. & Vanathi, D. (2014). *A Study Of Privacy Preserving Data Mining Techniques*. International Journal, 3.

Le, H. Q., Arch-Int, S., Nguyen, H. X. & Arch-Int, N. (2013). *Association rule hiding in risk management for retail supply chain collaboration*. Computers in Industry, 64**,** 776-784.

Lee, G., Chang, C.-Y. & Chen, A. L. (2004). *Hiding sensitive patterns in association rules mining*. Computer Software and Applications Conference. COMPSAC 2004. Proceedings of the 28th Annual International, 2004. IEEE, 424-429.

Liu, B. (2007). *Web data mining: exploring hyperlinks, contents, and usage data*, Springer Science & Business Media.

Llc, T. (2016). *ARFF (stable version)* [Online]. Available: https://weka.wikispaces.com/ARFF+(stable+version) [Accessed 24/05/2016].

Matwin, S. (2013). *Privacy-Preserving Data Mining Techniques: Survey and Challenges*. In*:* CUSTERS, B., CALDERS, T., SCHERMER, B. & ZARSKY, T. (eds.) Discrimination and Privacy in the Information Society. Springer Berlin Heidelberg.

Mohammed, R. S., Hussien, E. M. & Mutter, J. R (2016). *A novel technique of privacy preserving association rule mining*. 2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA), 9-10 May 2016 2016. 1-6.

Natarajan, R., Sugumar, D. R., Mahendran, M. & Anbazhagan, K. (2012). *A survey on Privacy Preserving Data Mining*. International Journal of Advanced Research in Computer and Communication Engineering, 1.

Nithi and Maheyzah. (2015). *K-Anonymity Data Obfuscation and Tree-Based Data Perturbation in Privacy Preserving Data Mining.* Master of Computer Science (Information Security), Universiti Teknologi Malaysia.

Oliveira, S. R. & Zaiane, O. R (2002). Privacy preserving frequent itemset mining. Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14. Australian Computer Society, Inc., 43-54.

Parmar, S., Gupta, M. P. & Sharma, M. P. (2015). *A Comparative Study and Literature Survey on Privacy Preserving Data Mining Techniques*.

Patel, K. & Donga, J. (2015). *Practical Approaches: A Survey on Data Mining Practical Tools*. Foundations, 2.

Prakash and Singaravel (2014). *A Review on Approaches, Techniques and Research Challenges in Privacy Preserving Data Mining*. Australian Journal of Basic and Applied Sciences, 8(10) July 2014, Pages: 251-259.

Qi, X. & Zong, M. (2012). *An Overview of Privacy Preserving Data Mining.* Procedia Environmental Sciences, 12**,** 1341-1347.

Reddy, O. H. & Singh, P. (2015). *Preserving Privacy in Data Mining by Data Perturbation Technique.* International Journal, 4.

Sakthi, R. & Umarani, V. (2013). *BSS Homomorphic Encryption: A Privacy model for large transactional database*.

Santhoshi, P. & Rao, M. E. (2015). *Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases*.

Saranya, K., Premalatha, K. & Rajasekar, S. S (2015). *A survey on privacy preserving data mining*. Electronics and Communication Systems (ICECS), 2015 2nd International Conference on, 26-27 Feb. 2015. 1740-1744.

Setiabudi, D. H., Budhi, G. S., Purnama, I. W. J. & Noertjahyana, A (2011). *Data mining market basket analysis' using hybrid-dimension association rules, case study in Minimarket X*. Uncertainty Reasoning and Knowledge Engineering (URKE), 2011 International Conference on, 2011. IEEE, 196-199.

Shorayha and Maheyzah. (2015). *Integration of PSO and Clustering Algorithms for Privacy Preserving Data Mining.* Master Master, Universiti Teknologi Malaysia.

Skarkala, M. E., Maragoudakis, M., Gritzalis, S. & Mitrou, L. (2011). *Privacy preserving tree augmented Naïve bayesian multi-party implementation on horizontally partitioned databases*. Trust, Privacy and Security in Digital Business. Springer.

Srikant, R. & Agrawal, R (1996). *Mining quantitative association rules in large relational tables*. Acm Sigmod Record. ACM, 1-12.

Srinivas, K., Rani, B. K. & Govrdhan, A. (2010). *Applications of data mining techniques in healthcare and prediction of heart attacks*. International Journal on Computer Science and Engineering (IJCSE), 2**,** 250-255.

Sukhdev, S. & Vasava, H. Privacy Preserving Data Mining With Classification And Encryption Methods.

University of California Irvine, U. (2016). *Congressional Voting Records Data Set* [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records [Accessed 10/04/2016].

Vaidya, J., Clifton, C. W. & Zhu, Y. M. (2006). *Privacy preserving data mining*, Springer Science & Business Media.

Vaidya, J., Kantarcıoğlu, M. & Clifton, C. (2008). *Privacy-preserving naive bayes classification.* The VLDB Journal—The International Journal on Very Large Data Bases, 17**,** 879-898.

Vaidya, J., Shafiq, B., Basu, A. & Hong, Y (2013). *Differentially private naive bayes classification*. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), IEEE/WIC/ACM International Joint Conferences on, 2013. IEEE, 571-576.

Vemuri, G. & Nutalapati, A. (2015). *Protected Association Rule Mining on Outsourced Transaction Databases.*

Wadkar, D. D. & Shelke, S. N. (2015). *A Novel Technique of Privacy Protection Mining of Association Rules from Outsource Transaction Databases*.

Waikato, T. U. O. (2016). *Weka 3: Data Mining Software in Java* [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 07/04/2016].

Wikipeadia. (2015). *Certified Quality Auditor (CQA)* [Online]. Wikipeadia. Available: https://en.wikipedia.org/wiki/Certified_Quality_Auditor [Accessed 09/ 10/ 2016].

Xu, L., Jiang, C., Wang, J., Yuan, J. & Ren, Y. (2014). *Information Security in Big Data: Privacy and Data Mining*. Access, IEEE, 2**,** 1149-1176.

Yi, X. & Zhang, Y. (2013). *Equally contributory privacy-preserving k-means clustering over vertically partitioned data*. Information Systems, 38**,** 97-107.