A Dissertation Proposal

On

Preserving Privacy in Big Data Mining

Submitted To

Lovely Professional University

In partial fulfillment of the requirement for the award of degree of

MASTER OF PHILOSPHY (M.Phil)

In

COMPUTER SCIENCE

Submitted by: **Supervised by:**

Shabnum Rehman **Dr. Anil Sharma**

Reg. No. 11512234

SCHOOL OF COMPUTER APPLICATION

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA (PUNJAB)

Certificate of the Supervisor

This is to declare that the work Preserving Privacy in Big Data Mining is a section of work completed by Shabnum Rehman under my guidance and supervision for the degree of Master of Philosophy (M.Phil) in Computer Science of Lovely Professional University, Phagwara, Punjab, India. To the best of my knowledge, the present work is the result of her original analysis and study and all other sources of information used have been acknowledged. No part of the project report has been submitted for any other degree or diploma to any other university.

Supervisor: Dr Anil Sharma

Signature: ……………………

Date: 09.12.2016

Declaration

I hereby admit that this thesis entitled "***Preserving Privacy in Big Data Mining***" is an authenticate record of my own original work carried out for the award of degree of M.Phil. (Computer Science) and all ideas and references have been duly acknowledged. The matter presented in the dissertation has not been submitted in part or fully to any other university or institute for the award of any degree.

Date: 09.12.2016 **Signature of Candidate:**

Shabnum Rehman

Reg.no: 11512234

Acknowledgement

I would like to sincerely thank Allah and all those who contributed in one way or another to this study. Words can inadequately express my deep gratitude to my guide, **Dr Anil Sharma, Assistant Professor**, Lovely Professional University, for his constant and constructive guidance. He supported me in every possible way since the beginning of my research work. Furthermore, it has been a memorable and enjoyable experience for me to work with him.

I feel very short to express my heartiest gratitude and sincere thanks to my parents and my sisters for their support and all they have comprised for me during the tenure of my research work. My appreciation also goes to my friend Jatinder Singh Bains for his moral support throughout my difficult time.

Lastly and most importantly, I wish to thank all those who were directly and indirectly involved in the completion of this work thereby making it a total success.

Thank you

Table of Contents

List of Tables

Chapter 1

Introduction

1. Introduction

1.1. Big Data

We live in a world where huge amount of data is collected, stored and these data are in the different fields. It might be in any format a document, a graphical format, video, audio or any other format. The massive data in the organizations till now was just ordinary data maintained in the databases. All of sudden this gigantic data termed as big data got popular [1]. Big data refers to incredible huge and complex data that becomes complicated to process using conventional applications [2]. Data may come from different sources such as transactions, social media, images, audios and videos. We can say that big data is generated by the machines consisting of both structured and un-structured data [3]. It’s not possible for traditional systems like SQL and RDBMS to deal with big data due to its scalability and complexity. Big data is characterized by: volume, velocity, variety and veracity [4]. Volume is defined as the amount of data that is to be mined so that valuable information is collected as it’s not possible for traditional systems to manage the huge amount of data. The range of data is very large such as terabytes, petabytes, Zettabytes [5], [6], [7], [8]. Every day twitter and facebook handles more than 40 billion photos from its users. So analysts need to construct the structural design to handle such kind of data [9]. Velocity refers to the speed at which data is generated and analyzed within a perfect time because a nanosecond of delay may cause a huge amount of loss in some companies. In some real time situations it becomes very essential to process data as soon as possible [10] specifically in online purchases and credit card payments but when we consider traditional systems they are not able to manage the velocity of big data as its one of the issue in big data to deal with the overflow of information within required time [11]. In simple words velocity means how fast data is produced and how quickly it must be processed [12]. The third characteristic of big data is variety which is considered as one of the most important property of big data. The data that has been stored in the database is without having any structure i.e., it may consist of structured, semi-structured and un-structured data types including text, audio, video, images [13], [4]. So here we can say that no structure of big data poses a great challenge for the traditional technologies to deal with such data that is completely un-structured [14], [15]. Next is veracity that considers inconsistencies of data flow. When we deal with high volume, speed and variety of data to get accurate data without any inconsistency is not possible [16]. The quality of data that is collected differs greatly. To implement privacy and security mechanism on uncertain data is often challengeable task [17]. From the huge companies like Google, Face book, and twitter we came to know about the need of term big data. Such organizations contain vast amount of data about individuals that needs to be analyzed in order to find some valuable information. Nowadays the term big data is used universally in every field like finance, banking, and marketing where every day or we can say every second data flows through workstations which are well managed and stored in order to find out the consumer behavior or the market trends to gain profitability [5],[6]. Another period of exploration began where existing information mining methods are considered for security saving. As the improvement and utilization of web expands the danger against the security and it make difficult issue. The business organizations that hold vast amount of data about the ongoing activities of their clients. To make utilization of the data, the data holders make use of data mining techniques to extract knowledge and compromise the privacy of their clients. It’s very challengeable for the businesses to mine new patterns of knowledge while securing one’s privacy. When the data owners reveal the output/result of mined data, they need to disclose just that sort of information which is not uncovering any individual's personal data.

1.2. Data Mining

Data mining is a promising and the fastest developing fields in computer science and engineering. In the middle 1990’s, data mining appeared as a strong tool [18] with the aim of examining the data stored in the databases from different perspectives by performing various operations so that interesting patterns and hidden knowledge is uncovered from the large data sets that is further used to predict the future benefits and trends with the help of different data mining techniques. Sometimes data mining is referred as knowledge hub or knowledge discovery in databases although data mining is a part of knowledge discovery process [19], [20]. The main reason that attracted a lot of consideration in information technology the innovation of important information from huge databases towards the area of data mining is because the impression of “we are information rich however data poor”. There are huge volumes of data but however, we are not able to turn them into valuable information and knowledge for making decision in business. To make valuable information we may require huge collection of data. In order to take absolute advantage of data, different tools and techniques are required for extraction of data, summarization of data, and the detection of patterns from raw data. Data mining is one of the powerful technologies that assist business organizations to focus on most valuable information stored in their data warehouse. Data mining helps in predicting future patterns and also businesses can make out convenient knowledge driven decision [21]. Data mining tools can respond queries that traditionally were extremely tedious, making it impossible to determine. Different tools organize databases for uncovering unknown knowledge; discovering analytical information that specialist may neglect since it lies outside their prospects. Beside the effective utilization of data mining tools and techniques, there are numerous threats to privacy. By using different data mining techniques one can easily reveal others private data or knowledge. So, before discharging database, private data of an individual or an organization must be hidden from any un-authorization access. Privacy is mainly essential because it may have harm full consequences on someone's life. There are several data mining techniques used in different data mining projects like for example: association rule, clustering, classification, prediction, sequential patterns and decision tree. We will briefly examine one of the technique i.e. association rule mining.

1.3. Association Rule Mining

Association is one of the most widely used data mining technique in which interesting patterns are obtained on the basis of relationship between items in the same transactions [22]. Association technique is sometimes called as relation technique. Association is usually used in market basket analysis in order to find out the products that customers purchase together frequently. The unearthed relationships can be represented in the form of association rules or set of frequent items. In [23] presented association rule for identifying consistency between items in large transaction data verified by point of scale systems in supermarkets. For example, on the basis of historical data stored in the databases retailers can identify the customers buying habits like if customer is buying diapers then it's obvious the customer will also purchase beer along with the diapers therefore; they put diapers and beer together. Hence the rule suggests that there exists a strong relationship between diapers and beer, using these rule retailers also try to identify the new opportunities for cross selling their products to the customers [24, [25]. Other than market basket data, association analysis is pertinent to other application spaces such as bioinformatics, medical diagnosis, web mining and scientific data analysis. In the analysis of earth science data, for instance the association examples may uncover interesting correlation among the sea, land and the atmospheric processes. Association rules are if/then statements used to identify relationship among data in the data repositories. In association rule mining there are two main steps involved i.e. support and confidence to recognize the relationships and also some rules are generated by analyzing data for frequent if/then pattern. Association rules generally need to convince a user specified minimum support and a user specified minimum confidence at the same time. Support specifies how frequently items appear in the database and confidence indicates the number of times the if/then statement have been found to be true. In association rule, information pattern should be analyzed and also the data structure of database so that we may be able to find out the preferred plans to keep the stability between the accuracy of database and the privacy or private information

[26]. There are algorithms like Apriori and Eclate used for generating frequent item sets so that association rules are mined proficiently

1.4. Apriori Algorithm

Apriori algorithm is one of the classical and the most popular association rule mining algorithm used to identify all the frequent individual items in the database. It continues by recognizing the individual items that occur frequently in the database and expand them to larger and larger item set as long as those item sets become visible adequately frequently in the database. Frequent item sets recognized through Apriori can be used to decide association rules which emphasize common patterns in the database for example, market basket analysis. Apriori algorithm employs the breadth first search and data structure of Apriori algorithm is very simple, clear and easy to understand [27]. Apriori algorithm follows Apriori property which states that if an item set occurs frequently, then all of its subsets must also be frequent. If an item set is not frequent, then any of its superset is never frequent [28], [29]. Apriori algorithm follows two steps:

- Generate Phase: In this phase candidate (K+1) item set is generated using K-item set; this phase is responsible for generating CK candidate set.
- Prune Phase: Here the candidate set is pruned and large frequent item set are generated using “minimum support” so, that these large frequent items can be used as the pruning parameter.

1.5. Eclate Algorithm

Eclate is another frequent pattern mining algorithm which carries out a breadth first search on the dataset and recognizes the support of item sets by performing intersection transaction lists [30]. Apriori is a basic frequent pattern mining algorithm in which database is scanned again and again which is very time consuming process. To overcome the drawbacks of Apriori, we have Eclate algorithm, uses the vertical database design which clusters all the relevant information in an item set. Each processor calculates the frequent items set from one equivalence class before proceeding to the next level. Thus the database is scanned only once [31]. The main steps involved in Eclate are: Firstly, to obtain all the frequent 1-item sets database is scanned, then candidate 2-item sets are generated from 1-item sets, after that get all frequent 2- item sets by extracting non-frequent candidate item sets; then candidate 3-item set is generated from frequent 2-item set and get all the frequent 3-item sets by extracting non-frequent item sets; repeat the process until no candidate item can be generated. In Eclate support is counted where as confidence is not counted at all. This algorithm is best suitable for small datasets and therefore, require less time for frequent pattern generation than Apriori [32].

Chapter 2

Review of Literature

In [33] some real issues related to big data storage and management are highlighted. The various challenges are also discussed that might be faced within next few years due to the epidemic growth of data. In addition to velocity, volume, variety data complexity is another characteristic of big data that define the importance of big data with respect to its complexity. Handling of big data is the real issues highlighted by the author.

In [34] some insights about big data issues, challenges and tools are discussed. Also provides basic concept of big data and some basic fundamental properties of big data like velocity, volume, heterogeneity, are talked about different source from which data is generated and examined. Big data has a huge significance in different activities like social media, sensor information, and log storage and risk analysis.

In [35] some issues related to privacy are mentioned and author suggests we can make use of system verification for big data using Map Reduce, Data processing and privacy preserving for global recording anonymization. Integration of Map Reduce, if used for analyzing data may provide better privacy.

In [36] proposed a big data processing model consisting of 3 tier architecture, I tier is focusing on accessing data and arithmetic computing, and the II tier is concentrating on the user privacy issues and the III tier is all about the challenges faced while mining the complex and dynamic data. Using this model we may require high performance computing platforms.

In [37] a multi-level security with masking algorithm is proposed to identify the sensitive columns in the data. Data is maintained in two ways scrambled data and cleared data. Cleared data is send to a database where strict rules are applied and only scrambled data is send to data miner by data collector. Before extracting knowledge data miner connects the present data to the sensitive data by applying decryption if data matches, then its provided to decision maker if not then its returned back to the data miner for further improvement. The problem with this is every time data is matched to sensitive data stored in the database which is very time consuming.

In [38] proposed a cryptographic algorithm in order to protect the data by converting plan text into cipher text using encryption schemes. This approach is suitable when there are more than two parties involved and need to perform computations based on their private data and never expose their output to any other party such problem is known as secured multiparty computation problem. The only problem with this is that less sensitive data that can be useful in big data analytics is also encrypted and is not accessible.

In [39] a technique is introduced namely K-anonymity. In this every record is alike to at least another k-1 other records on the possibly recognized variables. K anonymity can be achieved using generalization (replaces original value with less specific consistent value) and suppression (replaces original value by some special character like *). The only issue with this technique is that it doesn’t give attention to the links between the sensitive attributes so there is still outflow of sensitive data.

In [40] authors proposed a generalization algorithm generally called as bottom up approach in order to deal with the scalability of data. It replaces original value with a less specific consistent value. The structure for generalization can be obtained by making a tree of user’s original data set and various operations can be performed on specific ranges. This way of anonymzing data may be considered as efficient because generalization compresses the user’s data as data increases. Identifying the best generalization is the key to climb up the hierarchy at each iteration.

In [41] an encryption technique is developed called as Homomorphic encryption. It’s basically a form of encryption that allows performing some specific computations on cipher text and encrypted results are obtained. The decrypted results are then matched to the results of operations that are performed on plain text. This approach is useful to deal with un-trusted party because neither the input is unveiled nor the internal state of the encrypted data.

In [42] top down specialization approach is introduced that provides security and preserves sensitive data of the user by partitioning the large data sets into two phases; in first phase data is anonymized and intermediate results are created. While in second phase those results generated in first phase are integrated to get the final result. The only problem with this approach is that if the data set is too large it becomes difficult to apply anonymization to the data and there remains fair of privacy losses while portioning the data.

In [43] a method for securing two party high multi-dimensional private data is introduced called data mash up technique. This technique generally mash up the data on users end before its send to the third party. Only the ordinary data is exposed to third party, and the sensitive data is hidden by performing encryption before it is revealed to the other party. The issue with this technique is mashing up large data sets will require lot of time.

In [44] mentioned a technique called differential privacy. It's a method that doesn't allow clients to have access to the database. It's totally opposed to anonymization, there is no need to modify the data but an interface exits that calculates results and adds distortion to the results after this results are displayed. The only aim of this technique is to shrink the possibilities of individual recognition while querying the data. One problem with this method is that an analyst should know the query before using it.

In [45] a proxy re-encryption technique is discussed. This technique involves only sharing of cipher text securely over multiple times. Neither the message and sender's identity nor the receiver's identity is disclosed. Basically it follows an encryption scheme that allows converting the cipher text of particular key into an encryption of the same message by using another separate key.

In [46] some of the detailed technologies have been discussed like generalization, bucketization, and multiset-based generalization, one attribute per column, slicing and slicing with suppression. By using these techniques a different level of privacy can be achieved. When we consider generalization technique, it becomes difficult to apply on high dimensional data. Bucketization fails to maintain the membership disclosure, so they have mentioned slicing technique that can be used to overcome the above problems.

In [47] introduced a new technique called as slicing technique, which is basically an anonymzing technique in which data is partitioned vertically as well as horizontally. In vertical partitioning, attributes that highly correlate to each other are cluster into column. In horizontal partitioning column values are sorted randomly so that no column values can be linked. The vital idea behind slicing is to break the relationship cross columns, but to preserve the relationship within each column. To deal with the high dimensional data slicing technique is a best approach.

In [48] proposed a hybrid technique by combining randomization and generalization technique. First, data randomization is performed and after that generalization method is applied to the randomized data, this methods result in high information loss.

In [49] proposed a genetic algorithm for hiding sensitive rules, which investigates how sensitive rules must be guarded against malicious data miner. In this algorithm, a fitness function is computed, transactions are selected on the basis of this value, and the sensitive items in these transactions are transformed by crossover and some mutation operations without facing any loss of data. Using this method, sensitive rules are hidden, fake rules can’t be generated and non sensitive rules are not changed.

In [50] some limitations of existing Apriori algorithm like wasting time in scanning the entire database identifying on the frequent item sets. An improved Apriori is presented that helps in reducing the time depending on scanning only some transactions. The time consumed by improved Apriori in each and every one value of minimum support is lesser than the original Apriori.

In [51] an algorithm decrease support rule cluster is used to maintain privacy for sensitive association rules in database. Sensitive association rules are clustered on the basis of certain criteria by changing some transactions and hide many rules at a time. Besides it offer privacy for sensitive rules at certain level while ensuring database quality. The performance of this algorithm is better than other existing approaches.

In [52] describes about Apriori algorithm and how it’s used for discovering nearby frequent patterns. It’s one of the most suitable algorithm used for transactional databases. Apriori can be applied to different applications like telecommunication, network analysis, banking, market analysis and many more.

There are numerous methods for taking care of issue of protection of data mining. Although the main aim is to bestow privacy to the data, by performing review based on privacy here are some research findings that can help researchers to overcome the problems faced by existing techniques.

2.1. Research Findings

After performing review we find out different privacy preserving techniques that are used to maintain privacy whenever data mining is applied. We find out some of the challenges faced by these privacy techniques shown in Table 1.

Techniques	***Challenges***
Cryptographic technique	- Not good for large databases - Difficult to scale when more than a few parties are involved - Non sensitive data is also encrypted that can be useful for analytics
K-anonymity	- Doesn't not give attention to the links between sensitive data - It cannot protect against attacks based on background knowledge - Does not applied to high-dimensional data
Homomorphic encryption	- Computational overhead is very high, because computations are performed on encrypted data - Not applicable for large datasets
Top down specialization approach	- Loss of privacy - Leads to its inadequacy in handling large-scale data sets
Differential privacy	- The only issue with this technique is computation complexity
Data mash up technique	- Mashing large scale of data requires lot of time - Mashing of data may cause loss of accuracy
Anonymization through Generalization	- Causes loss of information

	- Doesn’t preserve attribute correlations - Each attribute is generalized separately - Identifying the best generalization is the key to climb up the hierarchy at each iteration
Bucketization	- Doesn't preserve membership disclosure - Publishes QI values in their original form - Needs clear separation between quasi identifiers and sensitive attributes
Slicing	- Mostly attributes are grouped randomly which is not efficient - It’s not clear how attribute disclosure is preserved - Utility of data is lost because of fake tuples

Table 1: Privacy techniques and challenges

Techniques	**Parameters**			
	Linkage Property	***Information Loss***	***Type of Data***	***Privacy Preserved***
Cryptographic Technique	Low	Low	Micro Data	High
Homomorphic encryption	Low	Low	Micro Data	High
Proxy re-encryption	Low	Low	Micro Data	High
Data Mash up Technique	Low	High	High Dimensional	High
Differential Privacy	Low	Low	Micro Data	High
K-Anonmization	High	Low	Micro Data	Low

Top Down Specialization Technique	Low	High	Micro Data	Low
Generalization	High	Very High	Micro Data	Low
Bucketization	High	Low	Micro Data	Low
Slicing Technique	Very Low	Low	High Dimensional	High
Hybrid Approach	Low	High	High Dimensional	High

Table 2: Comparison of Various Privacy Preserving Techniques

Above table 2: provides a comparative analysis of different privacy techniques based on different parameters like linkage property, information loss, type of data, and privacy preserved. The analysis shows that no single method is reliable in all areas. Each method performs in a different way depend on the size of data and the type of application.

2.2. Scope of the Study

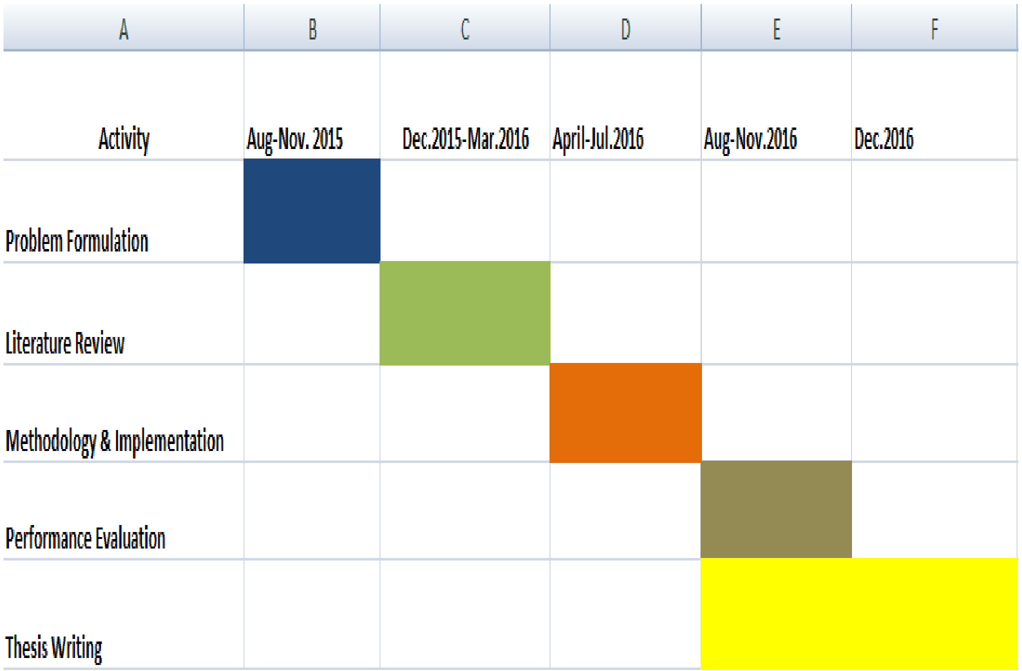
To survive in a rapidly expanding increasingly competitive environment business organizations need solutions to store huge amount of data and to find unseen insights from the large store of data. We are using association rule with Apriori for the market basket dataset. It performs analysis and finds out the products that customers purchase together. After performing analysis we are generating the item sets that are in encrypted format. Here we are applying privacy by making the readable data into unreadable format so that privacy of data is well maintained. While generating item sets we analyzed that maximum number of items get processed from the data set which helps in reducing the number of iterations. We are comparing two algorithms, in first of the algorithm we are applying privacy to the data set by performing encryption and in second one no privacy of data is ensured. While comparing these two algorithms we observed that the data sets that are encrypted provide less number of iterations than that of algorithm in which there is no privacy of data ensured. We also found that the time taken by these iterations in encrypted data set is also reduced.

2.3. *Objective*

The main objective of this work is:

- To introduce a privacy preserving technique that will not reveal the sensitive information of an individual.
- Testing of technique based on some specific dataset.

2.4. *Timelines*

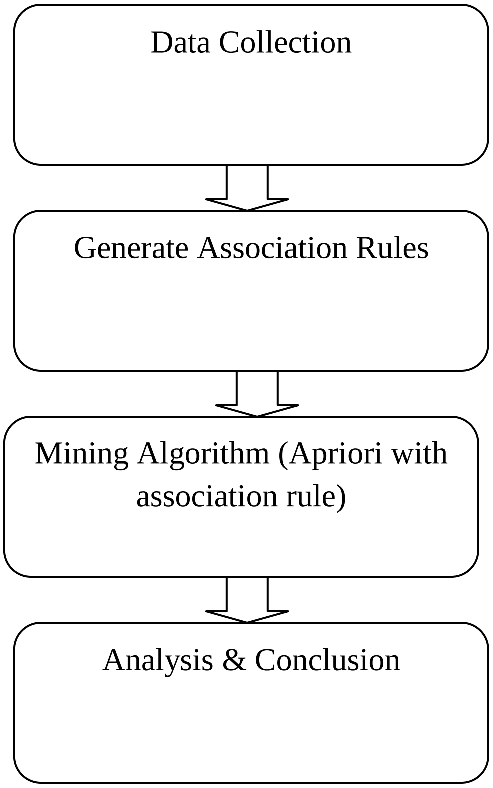


A	B	C	D	E	F
Activity	Aug-Nov. 2015	Dec.2015-Mar.2016	April-Jul.2016	Aug-Nov.2016	Dec.2016
Problem Formulation	🟦				
Literature Review		🟩			
Methodology & Implementation			🟧		
Performance Evaluation				🟫	
Thesis Writing				🟨	🟨

Chapter 3

Methodology

3.1. Methodology


Data Collection
Generate Association Rules
Mining Algorithm (Apriori with association rule)
Analysis & Conclusion

We have considered some market basket data with a set of items and transactions. Apriori algorithm is used to find out the most frequent item sets based on minimum support threshold. In our work we are generating item sets in encrypted format. It’s also observed that when item sets are generated maximum number of items get processed from the data set which reduces the number of iterations. As for the tool is concerned we are using MATLAB 2013 in which we are comparing two algorithms, in first algorithm there is no privacy ensured and in second algorithm privacy is ensured. In the next step we are comparing the number of iterations in algorithm where privacy is ensured with the algorithm in which there is no privacy employed. We are also comparing the time taken by the iteration in both of the algorithms.

Chapter 4

Result & Discussion

4.1. Result

4.1. Data Set: To achieve the final goal collected market basket data which is one of the most common and useful types of data analysis for marketing and retailing. Market basket data identifies the items sold in a set of baskets or transactions. Association rules use the Apriori algorithm to generate association rules that describe how items tend to be purchased in groups. By using Apriori algorithm we generated association rules in encrypted format. After generating these rules we compared the number of iteration in encrypted data with the number of iteration in which no encryption is done. Terminology used in Market Basket:

- Items: items are the objects in which we are identifying the associations
- Transactions: these are the group of instances of items that co-occur together
- Rules: rules are generally if/then statements; for example, if a customer buys milk then he/she will also purchase bread together.

Screen Shots:

4.2. Total Number of Iterations for Each Item Set: In this section we will show some existing iterations on item set in which no privacy is ensured and some new iteration in which privacy is ensured.

4.2. *Time taken by each iteration in an item set:* this section will provide the information about the time taken by the iterations in existing data (without ensured privacy) and new data (encrypted data).

4.2. Result Table:

For determining privacy, we compared the number of frequent item sets generated from the original dataset and the encrypted data. Our experimentation and analysis showed that the number of iterations generated in encrypted data is less than that which is generated in original dataset. So, if any one extracts the encrypted dataset, will not be able to get suitable results in terms of frequent item sets and strong association rules. So, in this way privacy of association rules along with database quality is well maintained.

Levels	Existing Iterations	New Iterations
C1	2.025%	1.568%
L1	1.640%	1.186%
C2	3.486%	3.029%
L2	3.798%	4.255%
C3	4.222%	2.851%
L3	4.070%	2.243%

Table 3: Comparison of iterations in existing data and new data

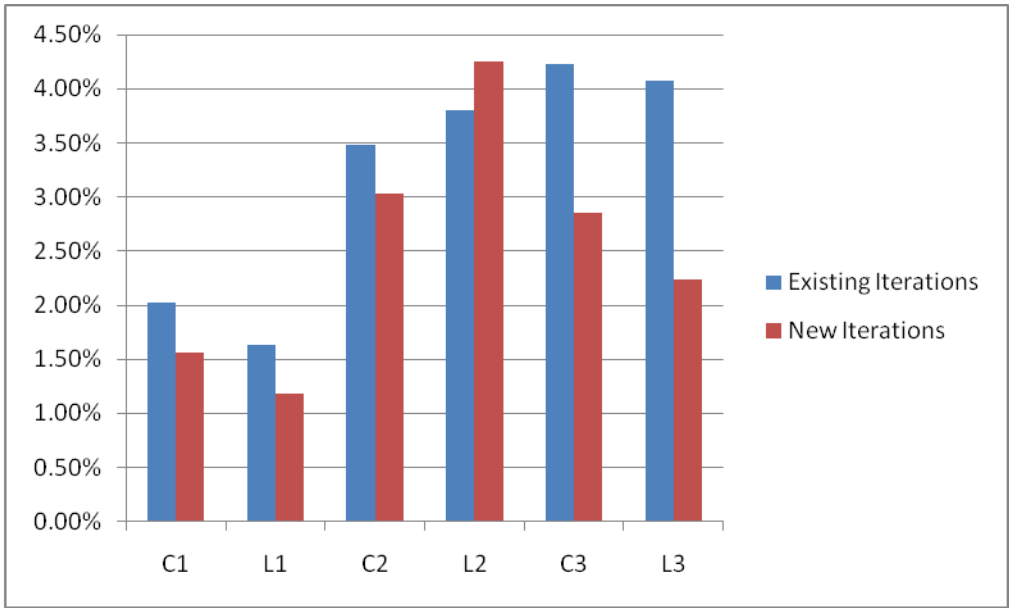


Figure 1: Comparison of iterations in existing data and new data

As shown in Figure 1, our experiment shows the comparison between the number of iteration in new item set and the existing item set. We observed that the number of iterations in new item set is less than the no of iteration in existing dataset. While considering privacy, less the number of iteration will lead to the reduction of privacy breaches.

Levels	Existed Time Calculated	New Time Calculated
C1	3.395 sec	2.938 sec
L1	4.381 sec	3.924 sec
C2	4.857 sec	4.400 sec
L2	5.169 sec	4.712 sec

C3	6.506 sec	6.049 sec
L3	4.070 sec	3.614 sec

Table 4: Comparison of time taken by iterations in original data and sanitized data

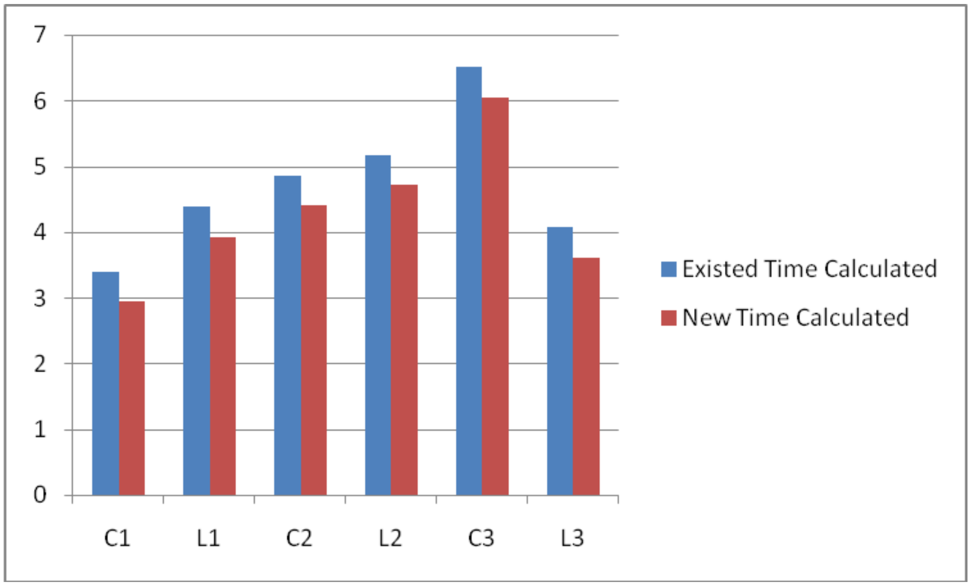


Figure 2: Comparison of time taken by iterations in existing data and new data

In Figure 2, comparison of time taken by completing iterations in new data and existing data is shown. It’s clearly shown that the total times taken by the iterations in the new data are lesser than the time taken by the iterations in existing data. Thus helps in reducing the time complexity of an algorithm.

Chapter 5

Conclusion & Future Scope

5. Conclusion & Future Scope

Preserving Privacy in data mining is a new body of research focusing on the implications originating from the application of data mining algorithms to large public databases. In our research work we did comparison of two algorithms; first algorithm is employed with encrypted item sets and next is without encrypting item sets. Comparison of these two algorithms is done on the basis of iterations per item set and the total amount of time taken in each iteration. While comparing the new item set (encrypted item sets) with the original item set (not encrypted item sets) it's observed that the total number of iterations in new item set is less than the old item set. The time consumed by iteration in encrypted item set is also reduced to that of original item sets. So, in this way privacy of association rules along with data quality is well maintained.

In the future work, the field of data mining requires some powerful techniques that will maintain the privacy and quality of the data.

References

[1] V.K. Vishakha, S. Alokkumar, "A Security and Privacy Preserving in Big Data", IORD Journal of Science and Technology, Vol.2 Issue 3, pp.32—37, 2015.

[2] C. Jinchuan, C. Yueguo, D. Xiaoyong, L. Cuiping, L. Jiaheng, Z. Suyun, Z. Xuan, “ Big data challenge: a data management perspective”, Research Article, Springer, DOI: 10.1007, Vol.7 No.2, pp. 157—164,2013.

[3] G. S. Poonam, B. L. Desai, “Big Data Mining: Challenges and Opportunities to Forecast Future Scenario”, International Journal of Innovative Research in Computer & Communication Engineering, Vol.3 Issue 6, pp.5228—5232,2015.

[4] A. T. H. Ibrahim, Y. Ibrar, B.A. Nor, M. Salimah, G. Abdullah, U.K. Samee, “The Rise of Big Data on Cloud Computing: Review and Open Research Issue”, Elsevier, DOI: 10.1016, pp.98—115, 2014.

[5] Ishwarappa, J. Anuradha, “A brief introduction on big data 5Vs characteristics and Hadoop technology”, International Conference on Intelligent Computing, Communication & Convergence Elsevier, DOI: 10.10.16, pp. 319 — 324, 2015.

[6] A.T. Alexandru, “Big Data Challenge, Database Systems Journal”, Vol. IV No.3, pp.31—40, 2013.

[7] S. Seref, S. Duygu, “ Big Data: A Review”, IEEE, pp.42—47, 2013

[8] H. Koichiro, Yokohama, “Social Issues of Big Data and Cloud”, International Conference on Availability, Reliability and Security IEEE, DOI: 10.1109, pp.506—511, 2013.

[9] Shilpa, K. Manjit, “Challenges and Issues during Visualization of Big Data”, International Journal for Technological Research in Engineering, Vol.1 Issue 4, pp.174—176, 2013.

[10] P. Kamakashi, “Survey on Big Data and Related Privacy Issues”, International Journal of Research in Engineering and Technology, Vol.3 Issue 12, pp.68—70, 2014.

[11] T. Raghav, G.D. Kanishka, N. Asoke, “Big Data Security Issues and Challenges”, International Journal of Innovative Research in Advanced Engineering, Vol.2, pp.15—20, 2015.

[12] C. Yosepu, P. Srinivasulu, S. Bathala, “A Study on Security and Privacy in Big Data Processing”, International Journal of Innovative Research in Computer and Communication Engineering, DOI: 10.15680, Vol.3 Issue 12, pp.12292—12296, 2015.

[13] Z. Kudakwashe, M. Mainford, G. Trust, “A Survey of the Security Use Cases in Big Data”, International Journal of Innovative Research in Computer and Communication Engineering, Vol.2 Issue 5, pp.4259—4266, 2014.

[14] B. Elisa, “Big Data- Security and Privacy”, IEEE International Congress on Big data, DOI: 10.1109, pp.757—761, 2015.

[15] S. Jaskaran, S. Varun, “Big Data: Tools and Technologies in Big Data”, International Journal of Computer Applications, Vol.112 No.15, 2015.

[16] T. Bharti, M. Manish, “Data Mining for Big Data: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol.4 Issue 5, pp. 469—473, 2014.

[17] A. Pradeep, S.D. Srikari, Z. Xiaowen, “Hadoop Eco System for Big Data Security and Privacy”, IEEE, 2015.

[18] M.H. Tekieh, B. Raaheni, “Importance of Data Mining in Healthcare: A Survey”, International Conference on advances in Social Networks Analysis and Mining IEEE, 2015.

[19] D. Himel, S. Tanmoy, B. Madhusudan, E.A. Mohammad, “An Approach to Protect the Privacy of Cloud Data from Data Mining Based Attacks”, IEEE, DOI 10.1109, pp.1106—1115, 2013.

[20] K.S. Dileep, S. Vishnu, “Data Security and Privacy in Data Mining: Research Issues & Preparation”, International Journal of Computer Trends & Technology, Vol.4 Issue 2, pp.194—200, 2013.

[21] N. Padhy, P. Mishra, R. Panigrahi, “The Survey of Data Mining Applications and Future Scope”, International Journal of Computer Science Engineering and Information technology, Vol.2, No.2, DOI: 10.5121, pp. 43—58, 2012.

[22] N.H. Domadiya, U.P. Rao, “Hiding Sensitive Association Rules to Maintain Privacy and Data Quality in Database”, 3rd IEEE International Advance Computing Conference, 2013.

[23] R. Agarwal, T. Imielinski, A. swami, “Mining Association Rules Between sets of Items in Large Databases”, International Conference on Management of Data – SIGMOD, pp. 207—216, 1993.

[24] C.N. Modi, U.D. Rao, D.R. Patel, “Maintaining Privacy and Data Quality in Privacy Preserving association Rule Mining ”, 2nd International Conference on Computing Communication and Networking Technologies IEEE, 2010.

[25] J. Zheng, L. Yan, “Research on the Improvement of Apriori Algorithm and its Application in Intrusion Detection System”, IEEE, pp. 105—108, 2015.

[26] S. Wu, H. Wang, “Research on Privacy Preserving Algorithm of Association Rule Mining in Centralized Database”, International Symposium on Information Processing IEEE, DOI: 10.1109, pp.131—134, 2008.

[27] K. Zhang, J. Liu, Y. Chai, J. Zhou, Y. Li, “A Method to Optimize Apriori Algorithm for Frequent Items Mining”, 7th International Symposium on Computational Intelligence and Design IEEE, DOI: 10.1109, pp.71—75, 2014.

[28] M.G. Ingle, N.Y. Suryavanshi, “Association Rule Mining using Improved Apriori Algorithm”, International Journal of Computer Applications, Vol.112, No.4, pp. 37—42, 2015.

[29] J. Yabing, “Research of an Improved Apriori Algorithm in Data Mining Association Rules”, International Journal of Computer and communication Engineering, Vol.2, No.1, pp. 25—27, 2013.

[30] Z. Ma, J. Yang, T. Zhang, F. Liu, “An Improved Eclate Algorithm For Mining Association Rules Based on Increased Search Strategy”, International Journal of Database Theory and Application, Vol.9, No.5, DOI: 10.14257, pp.251—266, 2016.

[31] S. Solanki, N. Soni, “A Survey on Frequent Pattern Mining Methods Apriori, Eclate, FP Growth”, International Journal of Computer Techniques, pp.86—89.

[32] K. Vani, “Comparative Analysis of Association Rule Mining Algorithms Based on Performance Survey”, International Journal of Computer Science and Information Technologies, Vol.6, No.4, pp.3980—3985, 2015.

[33] S. Kaisler, F. Armour, J.A. Espinosa, W. Money, “Big Data: Issues and Challenges Moving Forward”, 46th Hawaii International Conference on System Science IEEE, DOI: 10.1109, pp.995—1004, 2013.

[34] K. Avital, M. Wazid, R.H. Goudar, “Big Data: Issues, Challenges, Tools and Good Practices”, IEEE, pp.404—409, 2013.

[35] S. Venilla, J. Priyadarshini, “Scalable Privacy Preservation in Big Data: A Survey”, Elsevier, DOI: 10.1016, pp.369—373, 2015.

[36] W. Xindong, Z. Xingquan, Q.W. Gong, D. Wei, “Data Mining with Big Data”, IEEE Transaction on Knowledge & Data Engineering, DOI: 10.1109, Vol.26 No.1, pp.97—107, 2014.

[37] R.C. Muni, “Data mining and Security in Big data”, International Journal of Advanced Research in Computer Engineering and Technology,vol.4 Issue 3, pp.1065—1069, 2015.

[38] I.H. Nasrin, C. Bharadwaj, R. Sandip, “A Novel Method for Preserving Privacy in Big-Data Mining”, International Journal of Computer Applications, Vol.103 No 16, pp.21—25, 2014.

[39] S. Salini, V.K. Sreetha, R. Neevan, “Survey on Data Privacy in Big Data with K- Anonymity”, International Journal of Innovative Research in Computer and Communication Engineering, DOI: 10.15680, Vol.3 Issue 5, pp.3765—3771, 2015.

[40] B. Manimaran, S. Muthusundari, “Data Anonymization through Generalization Using Map Reduce on Cloud”, IEEE International Conference on Computer Communication and Systems, pp.39—42, 2014.

[41] M. Sangeetha, P. Anishprabhu, S. Shanmathi, “Homomorphism Encryption Schema for Privacy Preserving Mining of Association Rules”, International Journal of Innovative Research in Science & Engineering, 2013.

[42] C.M.F. Benjamin, W. Ke, S.Y. Philip, “T op-Down Specialization for Information and Privacy Preservation”, 2Ist International Conference on Data Engineering IEEE, 2005.

[43] S. Indhu, J. Perm, “Secure Two Party High Dimensional Private Data Using Data Mash Up”, International Journal of Computer Science and Information Technology, Vol.5 (1), pp. 644—645, 2014.

[44] G. Anjana, C. Nikita, “Privacy Preservation in big Data”, International Journal of Computer Applications, Vol.100, No.17, pp.44—47, 2014.

[45] M. Nijitha, Y. Kalpana, “Privacy-Preserving Cipher Text Multi Sharing Control for Big data Storage”, International Journal of Latest Trends in Engineering and Technology, Vol.6 Issue 3, pp.136—142, 2016.

[46] C.K. Preet, G. Tushar, M. Vanita, “Analysis of Data Security by Using Anonymization Technique”, 6th International Conference-Cloud System and Big Data Engineering IEEE, pp.287—293, 2016.

[47] R. Kavita, G. Parmeet, “A Review on Anonymization Techniques for Privacy Preserving Data Publish”, International Journal of Research in Engineering and Technology, Vol.4 Issue 11, pp.228—231, 2015.

[48] L. Savita, L. Lata, “Privacy Preserving In Data Mining Using Hybrid Approach”, Fourth International Conference on Computational Intelligence and Communication Networks IEEE, DOI:10.1109, 2012.

[49] S. Narmadha, S. Vijayarani, “Protecting Sensitive Association Rules in Privacy Preserving Data Mining Using Genetic Algorithms”, International Journal of Computer Applications, Vol.33, No.7, pp.37—43, 2011.

[50] M. A. Maolegi, B. Arkok, “An Improved Apriori Algorithm For Association Rules ”, International Journal on Natural Language Computing, DOI: 10.5121, Vol.3, No.1, pp.21—29, 2014.

[51] Y. J. Aniket, R. D. Virendra, S. B. Sagar, P. K. Hardik, “Privacy Preserving Association Rule Mining In Retail Industries”, International Journal of Advanced Research in Computer and Communication Engineering, DOI: 10.17148, Vol.4 Issue3, 2015.