International Journal of Early Childhood Special Education (INT-JECSE)
DOI:10.9756/INTJECSE/V14I6.337 ISSN: 1308-5581 Vol 14, Issue 06 2022

DESIGN AND ANALYSIS ON MECHANISMS FOR PRIVACY PRESERVING BIG DATA MINING

[1]P Latha, [2]Dr.C D Kumawat
Research Scholar Mewar University,Chittorgarh,Rajasthan
Research Supervisor Mewar University,Chittorgarh,Rajasthan

Abstract
Incredible amounts of data is being generated by various organizations like hospitals, banks, e-commerce, retail and supply chain, etc. by virtue of digital technology. Not only humans but machines also contribute to data in the form of closed circuit television streaming, web site logs, etc. Tons of data is generated every minute by social media and smart phones. The voluminous data generated from the various sources can be processed and analyzed to support decision making. However data analytics is prone to privacy violations. One of the applications of data analytics is recommendation systems which is widely used by ecommerce sites like Amazon, Flip kart for suggesting products to customers based on their buying habits leading to inference attacks. Although data analytics is useful in decision making, it will lead to serious privacy concerns. Hence privacy preserving data analytics became very important. This paper examines various privacy threats, privacy preservation techniques and models with their limitations, also proposes a data lake based modernistic privacy preservation technique to handle privacy preservation in unstructured data.

1. Introduction

Big data refers to data with characteristics like Volume, Velocity and Variety. As big data became an important research area, increase in the size of data leads to more chances of breaches of privacy of individuals. As big data needs to have huge amount of computing resources, it is done using MapReduce kind of programming in cloud computing environment. Public clouds providing distributed programming frameworks are simultaneously used by thousands of users across the globe. As multiple parties are involved in big data processing, there is ever growing risk of losing privacy. There are many researchers contributed towards privacy preserving data mining and privacy preserving data publishing. However, big data is different as it has got characteristics like volume, velocity and variety. Big data deals with different kinds of data, huge amount of data and streaming data as well. Therefore it is essential to have mechanisms specific to big data for privacy preservation while giving data for mining.

As cloud became a reality and cloud computing is used by people to outsource their storage and computational needs, it is essential to have mechanisms for preserving privacy of big data. The big data needs different approach in preserving privacy. The rationale behind this is that the data is in different formats such as structure, uunstructued and semi-structured. The big data life cycle includes multi-source big data generation, big data storage, and big data processing. When value is added to enterprise that mine big data, it is very useful proposition. On the other hand, if big data is not considered, it is like having partial facts in hand and taking accurate decisions is not possible. There are many conditions in which privacy of data may be lost.

1. When personal information available is combined with other external datasets it is possible to infer sensitive data from big data. It causes disclosure of sensitive data and this kind of attack made by adversaries is known as inference attack
2. Sometimes, personal information is collected and used in any business. It infract adds value to business. However, the individual's habits of shopping and other activities may reveal sensitive information.
3. During storage and processing of data, there is possibility of data leakage. This could lead to disclosure of sensitive information to have privacy breaches.

Many privacy preserving approaches came into existence. They include privacy preserving data publishing (PPDP), privacy preserving data classification, privacy preserving clustering, privacy preserving association rule mining, and data anonymization techniques. There are integrity verification techniques as well. They are Provable Data Possession (PDP), Proofs of Retrievability (POR), and public auditing. The existing research on privacy preserving big data mining reveals the need for protecting big data from privacy attacks. Privacy preserving big data mining is still an open problem to be addressed. The aim of the research is to investigate the present state of the art of privacy preserving big data mining, propose a comprehensive privacy framework for big data mining and implement the same.

2. Literature Survey

This section provides review of literature on protection of privacy of big data. Gheid and Challal [1] proposed a privacy preserving K-Means algorithm which protects privacy of data. They used cryptography-free multi-party additive scheme for horizontally portioned data. As cryptographic solutions degrade performance of a system, it became essential to have alternative solution. As the data mining in distributed environment causes privacy

issues, their proposed solution performs clustering on big data with privacy preserved. Big data refers to data that is streamed as well. Krempl*et al*. [2] investigate on the problem of protecting privacy of big data which is streaming. They found challenging issues in mining such streaming data with respect to privacy protection of big data. The challenges include streamed pre-processing, handling incomplete information, dealing with skewed distributions, handling delayed information, and selection of information dynamically. With respect to active learning of data streams the challenges identified are uncertainty related to convergence, need for perpetual validation, temporal budg*et al*location, and performance bounds. Other challenges related to challenges of aggregation and challenges of learning. With respect to mining there are two challenges related to mining. They include incompleteness of information and finding properties of privacy preserving in existing datasets.

Privacy is nothing but non disclosure of sensitive information. It needs to be protected when data is given for publishing or mining. Li *et al*. [3] proposed a light weight framework known as L-EncDB which enables privacy preserving queries in cloud copying. As cloud computing helped people to move their databases to cloud, it became an important research to ensure privacy to such data outsourced to cloud. These researchers provided light weight encryption scheme to be applied to the data to be outsourced. The traditional encryption techniques are not suitable for outsourced data as they are not light weight. While performing data mining on outsourced data it is essential to preserve privacy of data. Privacy preserving database encryption is employed in order to have privacy preserving database queries. The encryption technique used here has different approaches such as format-preserving encryption, fuzzy query encryption and order preserving encryption.

Xu *et al*. [4] investigated privacy and data mining in big data in the context of information security. Their work is related Privacy Preserving Data Mining (PPDM). Many of the researchers focused on privacy risk due to data mining operations. However, they did not focus on unwanted disclosure of sensitive information. Xu *et al*. viewed this problem with wide perspective in order to protect sensitive information from disclosure. Their approach is related personalized privacy that is adapted to each user of cloud. They believed that different users have different responsibilities with respect to privacy of data. Pitre and Kolekar [5] reviewed various aspects of data mining with big data. They opined that data mining with big data can provide comprehensive business intelligence. Such intelligence is used to make well informed decisions. Many latent patterns can be obtained from mining big data that considers structured, unstructured and streaming data. However, they found that privacy is a big challenging to be considered and addressed carefully.

Differential privacy is one of the techniques used to protect privacy of data. Lin *et al*. [6] proposed a methodology for privacy protection of big data. Sensitive big data is protected using the different privacy scheme. Especially the data collected from healthcare units using wearable devices is very sensitive. Such data needs to be protected from misuse. The bulk of data collected from sensor network needs to be utilized and processed without causing privacy issues. Abawajy*et al*. [7] explored different privacy models of big data. Two attack models are identified. First model is to make an attack on published data to infer sensitive information while the second one is to have probabilistic attack on the data. The identifiers in a data set are classified into 4 types. They are known as explicit identifiers, quasi identifiers, sensitive attributes and non-sensitive attributes. Explicit identifiers can help identify individuals uniquely. Quasi identifiers are the identifiers that do not directly reveal identity of people. However, adversaries can use them to have inference attacks. Sensitive attributes are the attributes that contain information specific to a person. Non-sensitive attributes are the attributes which are not sensitive and disclosing such data is not at all harmful.

Shokri and Shmatikov [8] focused on the concept of deep learning while preserving privacy of big data. They used artificial neural network for the purpose of deep learning. Massive data collection is needed to have deep learning. In such cases privacy of data is an important concern. Privacy in machine learning can be achieved using secure multi-party computation (SMC). The SMC can help protect intermediate computations and thus provide privacy to data in collaborative environment. While extracting complex features from high dimensional data, usage of multi-layer neural network is one of the widely used techniques to achieve deep learning. The privacy violating scenario such as multi-party communication is considered for the study.

3. Research Methodology

The research problem identified is privacy of big data in the context of big data mining which became frequent in the recent past. Different privacy attacks such as inference attack are to be addressed in the context of big data where the data has got attributes like volume, velocity and variety. Due to the complexity of data, heterogeneity in data types and continues growth of data, it is not easy to consider all these and achieve a comprehensive privacy preserving data mining solution. Nevertheless, this research focuses on the providing such framework which can cater to the needs of people who involve in big data mining. The privacy of individuals in the data is protected by the framework and this can be reused by individuals and organizations across the globe once its concept is proved. The proposed framework is as shown in Figure 1.

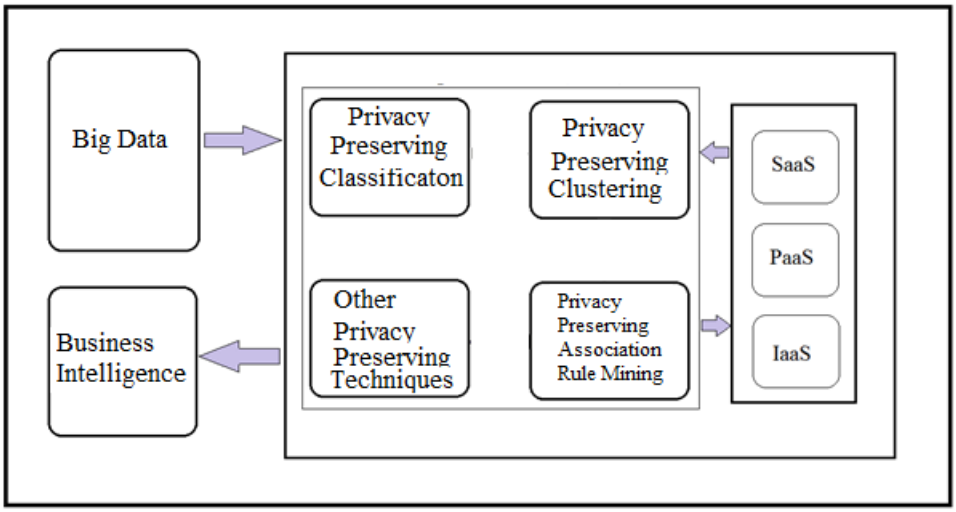

Figure 1: Shows conceptual framework that caters to privacy preserving big data mining

As shown in Figure 1, different privacy preserving data mining techniques are proposed and implemented. The privacy preserving clustering, classification, association rule mining is built explicitly that work for various datasets. Then a placeholder for other privacy preserving techniques is provided to support future work on the framework with respect to building new privacy preserving techniques. Different techniques for preserving privacy such as k-Anonymity, l-Diversity, t-Closeness and differential privacy are explored to have a comprehensive framework that considers certain measures to know misuseability of data and take appropriate decisions on privacy preserving data mining techniques. Anonymization is the process of converting sensitive data into different formats so as to avoid inference attacks on data. A specific attack model is built that helps in evaluating the proposed framework. The attack model includes privacy attacks such as known background information attack and data heterogeneity attack. It also includes any other inference attack which discloses identity of individuals by matching known data with external unknown data.

Algorithms are proposed and implemented to realise the privacy preserving data mining operations. Datasets are collected from Internet sources that are related to social networking or sensor networks or any data that has general characteristics of big data. Then a prototype application is built to demonstrate the proof of the concept. The techniques employed as part of the framework are evaluated using different evaluation procedures. The research results are evaluated against research hypothesis conceived after secondary research. The experimental results are presented with useful research insights.

CONCLUSION

No concrete solution for unstructured data has been developed yet. Conventional data mining algorithms can be applied for classification and clustering problems but cannot be used in privacy preservation especially when dealing with person specific information. Machine learning and soft computing techniques can be used to develop new and more appropriate solution to privacy problems which include identity disclosure that can lead to personal embarrassment and abuse.

There is a strong need for law enforcement by governments of all countries to ensure individual privacy. European Union is making an attempt to enforce privacy preservation law. Apart from technological solutions, there is a strong need to create awareness among the people regarding privacy hazards to safeguard themselves form privacy breaches. One of the serious privacy threats is smart phone.

References

1. Zakaria Gheid and YacineChallal. (2016). Ecient and Privacy-Preserving k-Means Clustering for Big Data Mining. *IEEE*. p1-9.
2. Georg Krempl,IndreZliobaite,DariuszBrzezinski,Eyke H¨ ullermeier,MarkLast,VincentLemaire,TinoNoack,AmmarShaker,SonjaSievi,MyraSpiliopoulou and Jerzy Stefanowski. (2014). Open Challenges for Data Stream Mining Research. *SIGKDD*. p1-64.
3. JinLi ,Zheli Liu , Xiaofeng Chen , FatosXhafa , Xiao Tan and Duncan S. Wong . (2014). L-EncDB: A lightweight framework for privacy-preserving data queries in cloud computing. *Computer Systems*. p1-10.
4. LEI XU, CHUNXIAO JIANG, JIAN WANG, JIAN YUAN AND YONG REN. (2014). Information Security in Big Data: Privacy and Data Mining. *IEEE*. P1-28.
5. Rohit Pitre AND Vijay Kolekar. (2014). A Survey Paper on Data Mining With Big Data. *IJIRAE*. 1 p1-3.
6. Chi Lin ,Pengyu Wang , Houbin Song , Yanhong Zhou , Qing Liu and GuoweiWu. (2013). A differential privacy protection scheme for sensitive big data in body sensor networks. *IEEE*. p1-19.

7. Nancy Victor ,Daphne Lopez and Jemal H. Abawajy. (2016). Privacy models for big data: a survey. *IEEE*. p1-16.
8. Reza Shokri and Vitaly Shmatikov. (2015). Privacy-Preserving Deep Learning. *ACM*. p1-12.
9. ABID MEHMOOD, IYNKARAN NATGUNANATHAN, YONG XIANG, GUANG HUA AND SONG GUO. (2016). Protection of Big Data Privacy. *IEEE*. 4 , p1-14.
10. Zakaria Gheid and YacineChallal. (2015). An Efficient and Privacy-preserving Similarity Evaluation For Big Data Analytics. *IEEE*. p1-9.