I J C T A, 9(6), 2016, pp. 2737-2743 ISSN: 0974-5572
© International Science Press

An Empirical Study of Security in Text Mining For Large Datasets

Kumaran U.* and Neelu Khare**

ABSTRACT

Text mining is the process to extract relevant information from large volume of database. Nowadays, security with text mining is contributing very important role to extract relevant information in secure and effective way on websites and social networking. In existing, there are many research work are done in terms of privacy. Here, some research article provides intrusion detection techniques, Anonymization Cryptographic, Perturbation and k-anonymity, Sanitization, Blocking-Based, space transformation, Noise Addition etc. However, these techniques have key complexities issues; accuracy problem, data classification issues and time consumption which are required to be addressed. After study of many research articles, this empirical study work is noticed that there is a need to develop some technologies to extract the relevant information from large volume of datasets in online environment with privacy and without compromising accuracy and data retrieval time in text mining. This work also plans to work efficient data visualization with filtering facilities. Finally, this survey work elaborates existing approach details along with limitation and represents the technical gap between current requirement and available technology

Keywords: Text mining, privacy persevering, data extraction, secure multi-party computation, Anonymization, encryption.

1. INTRODUCTION

Nowadays, security in text mining is become challenging task due to wide utilization of data from various resources in diverse form. Generally, text mining process permits an organization to utilize large volume of data to build up correlations and relationships among data to enhance the business. To extract relevant information from websites or social networking portal with privacy is not easy because of malicious attack, un-trusted user etc. For data protection in text mining, there are several research works are done. However, these existing techniques have key complexities issues, accuracy problem, data classification issues and also time consumption is high. There is still a huge gap between current requirements and existing approaches. Privacy is the major and important requirement in data mining because organization or industry either are using centralized server or cloud server. In order to analyze a distributed data with user data protection, security is a compulsory. This research work is studied the current requirement in text timing along with available resources.

In literature, there are many techniques are implemented in privacy preserving data mining to protect the data namely as an intrusion detection techniques, randomization, Anonymization, cryptographic, perturbation, sequential pattern mining, encryption, secure multi-party computation and k-anonymity etc.[13], [5]. This empirical study is gone through every existing approach individually and it is noticed that these existing approaches have key complexities issues, accuracy problem, data classification issues and data retrieval is more. In this scenario, it can be easily realized that how much efficient privacy is compulsory in text mining. In India, government is providing Voter ID card, PAN card, Aadhar card and with specific details of candidates.

* Research Scholar, School of Information Technology and Engineering, VIT University, Vellore, India, *Email: kumaran.u@vit.ac.in*

** Research Supervisor, School of Information Technology and Engineering, VIT University, Vellore, India *Email: neelu.khare@vit.ac.in*

After collecting the candidate information, it generates unique identification number for candidates. All collected data is officially maintained by government server but data operations are operated with specific text mining algorithms. Here, Data volume is high, if these data is compromised with malicious attack or external personal then how data operation can be continued in normal way.

To alleviate these issues, this empirical study is planning to develop an efficient and secure mechanism to extract the relevant information with privacy without affecting the text mining features. This approach should be highly effective in the terms data classification and minimal extraction time. Here, this mechanism should support a multi-keyword extraction, fuzzy supports with multiple groups of classifications. This work also plans to work efficient data visualization with filtering facilities. It minimizes data retrieval time and saves physical data storages. Mechanism should be represented the data in multiple point of data view with multiple group association which reduce the data analyst time to visualize or study the data in quick view. Meanwhile, it is also applicable for organization head to view the large volume of data in less time.

The rest of paper follows as: Section 2 expressed the literature work which is close to proposed agenda. Section 3 introduces the various existing approaches along with their limitation. In section 4, author concludes the overall work with future work.

2. LITERATURE WORK

In this section, this elaborates previous works which are related to proposed work. Clifton et al. [1] discussed about Privacy preserving of data mining to getting valid results without learning the underlying data values. Agrawal et al. [2] developed models for aggregated data to utilize these models without access to precise information of individual data in records. Mahajan et al. [3] elaborated about privacy preserving association rule sequential mining approach to provide privacy preservation for data mining. Divya et al. [4] developed differential privacy approach which offers the strong individual protection, interactive and non-interactive method for their applications to prevent the data from the unauthorized user or customer.

Panackal et al. [5] represent trade off between privacy and information loss creates a bottleneck while developing generic solutions. This article also explores many PPDM techniques based on classification hierarchy. Bonath et al. [6] expressed the fundamental concept of privacy preserving in data mining to develop novel standards and privacy algorithms in data mining. Uniyal et al. [7] surveyed various methods, techniques used for privacy-handling in data mining to process of patterns discovery for large data sets. Ilavarasi et al. [8] reviewed many Anonymization techniques to display generalization and bucketization of data with privacy.

Dhivakar et al. [9] elaborated recent approaches which are involved in privacy preservation like a randomization, Anonymization, perturbation and distributed privacy preservation methods. It also explained about computational and theoretical limits with privacy preservation for high dimensional data sets. Janbandhu et al. [10] expressed privacy preserving in data mining of many techniques along with their advantages and disadvantages. It also discussed about present limitations and scope for future research in privacy preserving data mining. Patel et al. [11] introduced a certain transformation approach to deal with the privacy during mining. This approach main objective is to provide more accuracy of specific data and preserving privacy of original data. It focused on Geometric data perturbation approach to analyze the large volume of data sets. Thuraisingham et al. [12] expressed privacy problem in the form of inference problem and introduce the notion of privacy constraints. It considered the developments on privacy-preserving data mining to maintain that privacy during extraction of useful information from datasets.

Hamza et al. [13] introduced preserving for both data mining (PPDM) and data publishing (PPDP) approach to share sensitive data for analysis purposes. It focused on k-anonymity model which in turn led to other models such as confidence bounding, l-diversity, t-closeness, (á, k)-anonymity, etc. This approach minimizes the information loss and such an attempt provides a loophole for malicious attacks. Liu et al. [14] focused on utilization of additive and matrix multiplicative data perturbation techniques in privacy preserving data mining (PPDM). Vaghashia et al. [15] explained privacy preserving data mining (PPDM) algorithms to consider a mining result with accuracy and as well privacy for textual datasets. Yale et al. [16] described a privacy of data mining along with cloud data to provide secure information. This approach considered to provide privacy of cloud based large volume of data.

Nivetha et al. [17] explored Sequential pattern mining to find relevant pattern in the data set. Predicting the sequence datasets leads to violate the privacy and disclose sensitive patterns related to medical records, business secrets etc. It also explained many approaches like anonymity, randomization, secure multiparty computation, sequential pattern hiding etc. Guajardo et al. [18] developed homomorphism Paillier cryptosystem to build a secure protocol for secure modulo reduction. This protocols permits for efficient multiparty computation of statistics like mean, variance and median for medical data or very sensitive information. Bhatti et al. [19] presents a framework to reduce false positive rate, ambiguity and provide accurate information for detection engine. This technique cleans network data or incomplete data and it is highly configurable. Kantarcioglu et al. [20], analyzed trade-offs between performance and security in privacy-preserving distributed data mining approaches. To enhance the privacy performance, these approaches are divvied in two protocols namely semi-honest model with zero knowledge proofs and malicious models.

Blanton et al. [21] expressed prior privacy-preserving data mining approach to tackle collusion attack problems, extract the data in effective and secure data. Nadiammai et al. [22] integrated data mining extraction process with Intrusion detection approach to identify the relevant, hidden data of interest for the user effectively with less execution time. They presented EDADT approach, Hybrid IDS model and Semi-Supervised Approach to data classification, labelled data leakage, distributed denial service (DDoS) attack problems. Lindell et al. [23] surveyed general paradigms and notions of secure multiparty computation in privacy-preserving data mining domains to identify constructions for secure multiparty computation.

Hoff et al. [24] designed new method for selecting parameter in soft s sets for NTRUEncrypt public key cryptosystem. This approach protects against new hybrid meet in the middle and lattice reduction of attack. Hoff et al. [25] enhanced the NTRUSign signature scheme to analysis its security, along with several parameter preferences. NTRUSign operations provided a good privacy model but not intended to be the last word on parameter generation for NTRUSign. Natrajan et al. [26] studied about Privacy Preserving Data Mining (PPDM) with the privacy driven from personally identifiable information. This paper discussed about various approach like a data partition, data modification, data restriction technique in PPDM to avoid the data access from unauthorized users. Kumari et al. [27] explored about many methods to privately preserve the data holder for medical data. It utilized following methods namely as an “Anonymization”, “Suppression”, “Generalisation” and “Data Hiding” on different area of dataset. Khazali et al. [28] developed an approach to preserve the privacy of the published data by modifying the graph by adding the smallest number of edges. This approach produces a quantitative value of missing information due to the generalization of the labels.

3. CURRENT APPROACH

This section represents the available approaches in privacy preserving data mining area. Here, it explains all the approaches along with their features and limitations in table 1 which helps to identify the gap between requirement and available mechanisms.

3.1. Randomization

Randomization approach is allow-cost and effective approach to data mining with privacy. In order to promise performance evaluation of individual security of data mining, randomization approach can be utilized. This approach protects user data of records randomly based on some true information of user data. This approach is adding or multiplying random values to numerical records or by deletion true data by addition of fake value to set of attributes.

3.2. Anonymization

In this phase, Anonymization is introduced to protect individual identification and hide sensitive information. In order to apply privacy in data mining, k-anonymity approach was presented to generalization and suppression of data. However, it is quite be difficult for an imposter to recognize identity of individual dataset collections. For every data release, every combination of values indistinctly at least should be matched to k-1 respondents. Here, generalization assists to replace the value with semantically reliable value. Suppression reduces the correctness of applications and does not disclose any true information of datasets. This approach minimizes the risk to detect the exact information.

3.3. Secure multi-party computation

In this phase, secure multi-party computation approach (SMC) represented a secure protocol which states that security of each data section can be disclosed with one more parties. Here, everyone knows some of the secure data participate in a protocol which produces text mining results. However, this mechanism is unable to control privacy when data is shared with third party.

3.4. Sequential Pattern Hiding

In this phase, Sequential pattern hiding introduced to conceal the credential patterns that can from published data without affecting the data and general interesting patterns. It has more composite semantics sets compare than item sets. It provides effective solution with high utility. However, there is no guarantee for data privacy.

3.5. Encryption

Here, encryption techniques applied to resolve the privacy issues. By utilization of cryptographic techniques Encryption technique, mutual trust between un-trusted parties problem can be resolved easily. An encryption technique is effective for tight security in text mining. However, this approach is not flexible in the terms of efficiency and key complexities problem.

3.6. Perturbation Techniques

A perturbation technique protects the data through perturbing the data via mechanisms. It provides privacy for data mining through pair of conflicting requirements systems. This approach is effective to maintain data privacy by selecting specific information during the data perturbation process. This system is capable for multiple data transformation to achieve high level data utility. However, this approach is not flexible for good data transformation with privacy.

3.7. Data Modification Techniques

Data modification techniques modify the original attributes of a database which requires to be contributed. So, privacy preservation is ensured. The transformed database is made available for mining and must fulfil the privacy requirement of data. The main objective of this approach is to find an appreciate balance between privacy preserving and knowledge disclosure. This approach is divided in following way:

3.7.1. Noise Addition Techniques

This approach adds some noise (e.g., information not present in a particular tuple or transaction) in original data to avoid the identification of confidential information. In other words, noise is added to confidential attributes by randomly shuffling to prevent the discovery of patterns which are not supposed to be extracted.

3.7.2. Space Transformation Techniques

This method is designed to protect the underlying data values which is subjected to cluster without jeopardize the similarity among objects. This method does not only meet privacy requirements but also assure the valid clustering results.

3.8. Data Restriction Techniques

Data restriction methods worked on to provide limitation to cluster the result through either generalization or suppression data. It restricts to some patterns which should not disclosed. These approaches are divided in following group namely as Blocking-based technique and Sanitization-based technique.

3.8.1. Blocking-Based Techniques

This approach objective to protect some sensitive information when data are distributed for mining. The private information contains sensitive association rules and classification rules that must remain private. Before preceding the data for mining, data holder have to confirm how much data can be inferred or evaluated from large databases. It also works to minimize the data leakage.

3.8.2. Sanitization-Based Techniques

The sanitization-based technique is used to protect the credential information through strategically suppressing some attributes in transactional databases. This method generalizes the information to preserve the privacy with classification. This method is used to remove or hide the group of sensitive association rules which contain sensitive information.

3.9. Data Ownership Techniques

These techniques are developed a method to ensure that who is data owner. During the data transmission, this method is used to provide read/write authentication and ensure that right receiver. This approach is applicable for Web mining and on-line business-to-business (B2B) interactions.

Table 1
Features and drawback of current approach for security in data mining and privacy for parameter reductions.

Approaches	*Features*	*Drawbacks*
Randomization	• This approach is simple & effective. • It can be easily implement during data collection.	This approach is not capable for multiple attribute datasets.
Anonymization	• This approach is used to protect credential information while giving true information. • This approach is support generalization and suppression methods.	This approach has limitation to detect the malicious attacks and potential can be misused.
Secure Multi-Party Computation	• It provides secure protocol to disclose information with one more parties.	This mechanism is unable to control privacy when data is shared with third party.
Sequential Pattern Hiding	• It concerned about credential patterns. • It has more composite semantics sets compare than item sets	There is no guarantee for data privacy.
Encryption	• It supports strong privacy in data mining. • This approach is capable to establish mutual trust among un-trusted parties.	This approach has low efficiency due to key complexity problem.

(*contd...*)

(*Table 1 contd...*)

Approaches	*Features*	*Drawbacks*
Perturbation Techniques	• This approach is capable for multiple data transformation to achieve high level data utility	This approach is not flexible for good data transformation with privacy.
Noise Addition Techniques	• This method adds some noise in original data to avoid the identification of confidential information. • This system also maintain the privacy of discovery patterns	This technique does not maintain the classification accuracy and data retrieval time.
Space Transformation Techniques	• This system protects the underlying data values which is subjected to cluster without jeopardize the similarity among objects • It maintains the privacy with the valid clustering results.	However, this approach does not provide multi visualization feature during data filtering.
Blocking-Based Techniques	• This system protects some sensitive information when data are distributed for mining • It also works to minimize the data leakage	This systems consume more time to extract the relevant information from web.
Sanitization-Based Techniques	• It protects the credential information through strategically suppressing some attributes in transactional databases. • This method is used to remove or hide the group of sensitive association rules	This system unable to minimize data extraction time and enhance the data classification accuracy.

4. CONCLUSION

In this paper, this empirical study work studied many research articles and survey paper to fulfil the gap between current requirement and available technology. In details, this paper studied about intrusion detection techniques, Anonymization, Cryptographic, Perturbation, k-anonymity, Sanitization, Blocking-Based, space based transformation techniques, Noise Addition etc but it notices that these techniques have key complexities issues, data classification accuracy issues, data retrieval time is high. In behalf of privacy in data mining, authors feels that technology should be required to extract the relevant information from large volume of datasets with privacy without affecting mining result, classification, and data retrieval time in text mining. This research work can also help to implement some novel technologies to represent the data with multiple types of visualization in data mining. It minimizes the data analyst time to generate report. Meanwhile, it is also applicable for organization head to view the large volume of data in less time. Finally this survey work presents many existing approach with limitation like intrusion detection techniques, Anonymization, Cryptographic, Perturbation and k-anonymity Sanitization, Blocking-Based, space transformation, Noise Addition which are worked and security enhancements in text mining.

In future, author plan to develop and secure and efficient framework for human recruitment system to extract the query from large volume of dataset with minimal retrieval time.

REFERENCES

[1] Clifton C., Kantarcioglu M., & Vaidya J., “Defining privacy for data mining”, In National Science Foundation Workshop on Next Generation Data Mining, Vol. 1, No. 26, 2002, pp. 126-133.

[2] Agrawal R., & Srikant R., “Privacy-preserving data mining”, In ACM Sigmod Record, Vol. 29, No. 2, 2000, pp. 439-450.

[3] Mahajan S. M., & Reshamwala A. K., “Data Mining Ethics in Privacy Preservation-A Survey”, International Journal of Computer Theory and Engineering, Vol. 3, No. 4, 2011, pp. 529-533.

[4] Divya S., Kumar B. S., & Karthik S., “A Survey on Privacy Preserving Data Mining Techniques using Differential Privacy”, International Journal of Engineering Research and Technology, Vol. 3, No. 11, 2014, pp. 496-498.

[5] Panackal J. J., & Pillai A. S., “Privacy Preserving Data Mining: An Extensive Survey” In conference proceeding of

ACEEE International Conference on Multimedia Processing, communication and Information Technology, 2013, pp. 297-304.

[6] Bonath R., Devaki K., Meghavath D., & Vijaya G., "A Report of the Privacy in Data Mining: Speakers Survey", International Journal of Innovative Science and Modern Engineering, Vol. 2, Issue 4, 2014, pp. 4-6.

[7] Uniyal V., Panchpuri S., & Kamboj G., "A Survey of Privacy-Handling Techniques and Algorithms for Data Mining", International Journal of Technology Innovations and Research, Volume 8, 2014, pp. 1-17.

[8] Ilavarasi A. K., Sathiyabhama B., & Poorani S., "A Survey on Privacy Preserving Data Mining Techniques", International Journal of Computer Science and Business Informatics, Vol. 7, 2013, pp. 1-12.

[9] Dhivakar K., Mohana S., "A Survey on Privacy Preservation Recent Approaches and Techniques", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 11, 2014, pp. 6559-6566.

[10] Janbandhu S., Chaware S.M., "Survey on Data Mining with Privacy Preservation", International Journal of Computer Science and Information Technologies, Vol. 5, No. 4, 2014, pp. 5279-5283.

[11] Patel J. D., Patel S., "A Survey on Data Perturbation Techniques for Privacy Preserving in Data Mining", International Journal for Scientific Research & Development, Vol. 3, Issue 01, pp. 52-54, 2015.

[12] Thuraisingham B., "Privacy-Preserving Data Mining: Developments and Directions Journal of Database Management, Vol. 16, No. 1, pp., 2005.

[13] Hamza, N., & Hefny, H. A., "Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing", Journal of Information Security, Vol. 4, No. 2, 2013, pp. 101-112.

[14] Liu K., Giannella C., & Kargupta, H., "A survey of attack techniques on privacy-preserving data perturbation methods", In Privacy-Preserving Data Mining Springer US, 2008, pp. 359-381.

[15] Vaghashia H., & Ganatra A., "A Survey: Privacy Preservation Techniques in Data Mining", International Journal of Computer Applications, Vol. 119, No. 4, 2015, pp. 20-26.

[16] Yale A. R., & Borkar, P., "Survey Paper on Data Mining in Cloud Computing", International Journal of Science and Research, Vol. 4, Issue 3, 2015, pp. 2126-2128.

[17] Nivetha P.R., & Thamarai S. K., "A Survey on Privacy Preserving Data Mining Techniques", International Journal of Computer Science and Mobile Computing, Vol. 2, Issues 10, 2013, pp. 166-170.

[18] Guajardo J., Mennink B., & Schoenmakers B., "Modulo reduction for paillier encryptions and application to secure statistical analysis", In Financial Cryptography and Data Security, Springer Berlin Heidelberg, 2010, pp. 375-382, 2010.

[19] Bhatti D. G., & Virparia P. V., "Data Pre-processing for Reducing False Positive Rate in Intrusion Detection", International Journal of Computer Applications, Vol. 57, No. 5, 2012, pp. 15-19.

[20] Kantarcioglu M., & Kardes O., "Privacy-preserving data mining in the malicious model", International Journal of Information and Computer Security, Vol. 2, No. 4, 2008, pp. 353-375.

[21] Blanton M., "Achieving full security in privacy-preserving data mining", In proceeding of 2011 IEEE Third International Conference on Social Computing (Social Com), Privacy, Security, Risk and Trust (PASSAT), 2011, pp. 925-934.

[22] Nadiammai G. V., & Hemalatha M., "Effective approach toward Intrusion Detection System using data mining techniques", Egyptian Informatics Journal, Vol. 15, No. 1, 2014, pp. 37-50.

[23] Lindell Y., & Pinkas, B., "Secure multiparty computation for privacy-preserving data mining", Journal of Privacy and Confidentiality, Vol. 1, No. 1, 2009, pp. 55-98.

[24] Hoff stein J., How grave-Graham N., Pipher J., Silverman, J. H., & Whyte, W., "Hybrid lattice reduction and meet in the middle resistant parameter selection for NTRU-Encrypt," IEEE Submission/contribution, 2007 pp. 1363-1381.

[25] Hoff stein J., How grave-Graham N., Pipher J., Silverman, J. H., & Whyte W., "Performance Improvements and a Baseline Parameter Generation Algorithm for NTRUSign" IACR Cryptology ePrint Archive, 2005, pp. 274-292.

[26] Natarajan R., Sugumar R., Mahendran M., and Anbazhaga K., "A survey on Privacy Preserving Data Mining", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 1, Issue 1, March, pp. 102-112, 2012.

[27] Kumari D. A., Vineela Y., Krishna T. M., & Kumar B. S., "Analyzing and Performing Privacy Preserving Data Mining on Medical Databases", Indian Journal of Science and Technology, 9(17), pp. 1-9, 2016.

[28] Khazali M. J., Sargolzaei E., & Keikha F., "Privacy Preserving Approach of Published Social Networks Data with Vertex and Edge Modification Algorithm", Indian Journal of Science and Technology, 9(12), pp., 2016.