International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
Volume 30, Issue 04, Quarter 04 (Oct-Nov-Dec 2018)
ISSN (Online): 2347-601X
www.ijemhs.com

S-CPM: A Cell Generalization Principle

Sudharani Mahapatra
Aryan Institute of Engineering & Technology, Bhubaneswar

Abstract: Timely data analysis on a wide variety and a large volume of dataunveil valuable information or new insights. The analysis results could be used to innovate new avenues in health care service, business and e-service, etc. However, releasing, storing and reusing sensitive data to third parties results in breaching the data privacy of the individual. To combat privacy breach invasion, privacy-preserving techniques such as suppression,generalization and encryption-based privacy models have been proposed in the literature. The widely used privacy preservation model k-anonymity model prevents record-linkage invasions but fails to satisfy monotonicity property. It has more data distortion and fails to defend semantic-similarity,closeness, nearest-neighborhood data privacy breaches. Moreover, existing approaches are not scalable for the large-scale data set. The paper proposes a semantic similarity two-phase cluster based privacy preservation model. The proposed model considers both numerical and categorical attribute values for data anonymization. Two-phase clustering contains two phases. Inthe first phase, the t-centroid clustering algorithm is designed and used to partition a set of transaction records of data set D into a set of t-centroids based on the Euclidean distance between transaction records. In the second phase, the neighborhood-aware hierarchical clustering algorithm is designed. It is used to split a set of transaction records within clusters based on neighborhood aware attribute values. Two-phase clustering operations are carried out in parallel and scalable for Big Data sets. The proposed privacy model relies on cell generalization to combat records linkage and semantic-similarity, closeness, nearest-neighborhood privacy breach invasion. All experiments are carried out on two different datasets: Income-Census (KDD) and Bank Credit Card dataset. The experimental results demonstrate that the proposed privacy model can combat privacy breach invasion with cell generalization principles. The proposed privacy model is scalable and time efficient for large-scale data sets.

Keywords: Privacy Preservation Model, Cell Generalization, Transaction Records, Clusters, Quasi-Identifiers and Sensitive Attributes

Introduction

The rapid technological development in information, communication and proliferation of mobile devices enabled millions of users to use social networks, sensors surveillance systems, IoT-enabled healthcare applications, e-Learning and e-Commerce applications for various purposes (Lv *et al*., 2017; Ang *et al*., 2020 and Zheng *et al*., 2020). All these applications are a source of data deluge in different formats (i.e., text, audio, video, image, etc.) (Tsui *et al*., 2019 and D'Alconzo *et al*., 2019). The data with different formats are generated at a higher speed and it is referred to as Big Data. Big Data is characterized by volume, velocity, variety and traditional methods are not appropriate to handle data that explode at an exponential rate (Manyika *et al*., 2011). Moreover, the data generated from different sources could be unstructured, semi-structured, or structured and make it difficult to process, store and maintain privacy (L'heureux *et al*., 2017). The systematic and time-bound analysis of Big Data gives actionable and profitable insights; these insights could be more useful in enhancing business, defining new strategies, take profitable management decisions (Liang *et al*., 2018). The research challenges and issues in the systemic analysis of Big Data have attracted research and the scientific community. In recent years, Big Data analytics in the cloud environment privacy preservation has been a hot research topic.

Despite the difficulty in storing and processing Big Data, Big Data could be effectively utilized to understand the trends of users on social networks, trends in business, proliferate new research solutions to these complex problems. With the great potential of Big Data, it is easy to gather store user personal information. However, commercial social network platforms have started sharing user personal information with the purpose of profit. The reuse/misuse of User personal information by social network platforms is a violation of personal data privacy and a breach of data integrity. For example, most common social network platforms such as Amazon, Flipkart, e-bay perform analytics on user data to extract user shopping frequency, pattern, priorities, likes and dislikes. Social media like Facebook do extensive analytics on user habits, social status, social relationships, list out family members, friends, colleagues and store user personal data. YouTube suggests the videos to the user based on the user search track on the browser. Under various circumstances, social network platforms breach the user's privacy (Liu Zhang, 2020; Mehmood *et al*., 2016; Zhou *et al*., 2019; and Bhaskar and Shylaja, 2021):

- To find the user preferences over product or services, business companies retrieve user personal information from social networks platforms
- Secretive personal information is stored in a public database and new inference from the public database may reveal confidential information of the user to others
- Storing and processing of user personal information in an unprofessionally and unsecured manner may result in the disclosure of the user's data privacy

To preserve the privacy of user data, extensive research work has been going in recent years. A well-known and widely accepted approach has been presented to protect Big Data privacy or hide private personal information, while some agglomerated data are open for data analysis purposes. The existing privacy models are either not scalable or inefficient due to velocity, volume and variety of data (Li *et al*., 2009; Aggarwal *et al*., 2010; Fung *et al*., 2010). Moreover, privacy models introduce noise and falsify data to protect the privacy of data. The existing privacy models cannot withstand record-linkage, sensitive attribute attacks, data distortion and are unable to maintain monotonicity properties. Therefore, designing a privacy model that preserves privacy with low distortion and combat sensitivity attribute and linkage attribute attack is a challenging and open research problem for largescale datasets. This study proposes a privacy protection model that combat privacy breach attacks, data distortion and satisfying monotonicity property. The paper proposes a two phase cluster-based privacy model that minimizes both data distortion and privacy breach attacks; the paper considers both numerical and categorical attribute values for data anonymization. Two-phase clustering contains two phases; in the first phase, *t-centroid clustering algorithm* is designed and used to partition a set of transaction records of data set D into a set of t-centroids based on Euclidean distance (i.e., the similarity between quasi-identifiers of transaction records) Between two transactions records. In the second phase, neighborhood aware hierarchical clustering algorithm is designed and used to split a set of transaction records within clusters based on neighborhood-aware attributes values (i.e., the similarity between categorical and sensitive attribute values). Two-phase clustering operations are executed in parallel and are scalable for Big data set.

Conclusion

This study proposes a Semantic-aware Cluster-based Privacy Model (S-CPM) that adopts cell generalization for anonymization and to thwart data privacy breach invasion with cell generalization, the proposed privacy model can combat privacy breach invasion, scalable and time-efficient. The proposed model includes multiple numerical, categorical sensitive attributes. This study proposes a scalable two-phase cluster based privacy model to protect privacy breach invasion with cell generalization. The two-phase clustering combines the benefits of the point-assignment and hierarchical clustering approach. In the first phase, this study leverages the point assignments technique to split the dataset and nearest neighbor, or closest records are grouped to form a cluster. In the second phase, quasi-identifiers attribute similarity and semantic-similarity of sensitive values of transaction records are considered to merge clusters. A Series of experiments are conducted to investigate the efficiency and scalability of the proposed approach.

The data set size is large enough to assess the effectiveness of the proposed model. Approximately the size of the cluster is 1000 for different sizes of the dataset. The values of k-anonymity parameter (i.e., k = 10), weight of semantic similarity (i.e., w_C = 0.5, w_N = 0.5), stopping condition (i.e., θ = 5, τ = 0.001) and ten computation nodes make a model to combat privacy breach invasion with cell generalization principles. The proposed privacy model is scalable and time efficient for large-scale data set. Future research explores the adoption of proposed research work for data anonymization through a bottom-up approach. Future plans to investigate scalable and robust data anonymity privacy solutions against adversaries' privacy breach attacks.

References

Aggarwal, G., Panigrahy, R., Feder, T., Thomas, D., Kenthapadi, K., Khuller, S., & Zhu, A. (2010). Achieving anonymity via clustering. ACM Transactions on Algorithms (TALG), 6(3), 1-19. doi.org/10.1145/1798596.1798602

Ang, K. L. M., Ge, F. L., & Seng, K. P. (2020). Big educational data & analytics: Survey, architecture and challenges. IEEE access, 8, 116392-116414. doi.org/10.1109/ACCESS.2020.2994561

Bhagyashri, S., & Gurav, Y. B. (2014). Privacy-preserving public auditing for secure cloud storage. IOSR Journal of Computer Engineering (IOSR-JCE), 16(4), 33-38. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.933.623&rep=rep1&type=pdf

Bhaskar, R., & Shylaja, B. S. (2021). Dynamic Virtual Machine Provisioning in Cloud Computing Using Knowledge-Based Reduction Method. In Next Generation Information Processing System (pp.193-202). Springer, Singapore. doi.org/10.1016/j.suscom.2018.01.002

D'Alconzo, A., Drago, I., Morichetta, A., Mellia, M., & Casas, P. (2019). A survey on Big Data for network traffic monitoring and analysis. IEEE Transactions on Network and Service Management, 16(3), 800-813. doi.org/10.1109/TNSM.2019.2933358

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In 2015 IEEE Symposium Series on Computational Intelligence (pp. 159-166). IEEE. doi.org/10.1109/SSCI.2015.33

Fung, B. C., Wang Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. ACM Computing Surveys (Csur), 42(4), 1-53. doi.org/10.1145/1749603.1749605

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of associationrules on horizontally partitioned data. IEEE transactions on knowledge and data engineering, 16(9), 1026-1037 doi.org/10.1109/TKDE.2004.45

Kanwal, T., Anjum, A., & Khan, A. (2021). Privacy preservation in e-health cloud: Taxonomy, privacy requirements, feasibility analysis and opportunities. Cluster Computing, 24(1), 293-317. doi.org/10.1007/s10586-020-03106-1

L'heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. (2017). Machine learning with Big Data: Challenges and approaches. Ieee Access, 5, 7776-7797. doi.org/ 10.1109/ACCESS.2017.2696365

Li, J., Tao, Y., & Xiao, X. (2008, June). Preservation of proximity privacy in publishing numerical sensitive data. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 473-486). doi.org/10.1145/1376616.1376666

Li, N., Li, T., & Venkatasubramanian, S. (2009). Closeness: A new privacy measure for data publishing. IEEE Transactions on Knowledge and Data Engineering, 22(7), 943-956. doi.org/0.1109/TKDE.2009.139

Liang, F., Yu, W., An, D., Yang, Q., Fu, X., & Zhao, W. (2018). A survey on Big Data market: Pricing, trading and protection. Ieee Access, 6, 15132-15154 doi.org/10.1109/ACCESS.2018.2806881

Liu, Z., & Zhang, A. (2020). Sampling for Big Data profiling: A survey. IEEE Access, 8, 72713-72726. doi.org/10.1109/ACCESS.2020.2988120

Lv, Z., Song, H., Basanta-Val, P., Steed, A., & Jo, M. (2017). Next-generation Big Data analytics: State of the art, challenges and future research topics. IEEE Transactions on Industrial Informatics, 13(4), 1891-1899. doi.org/10.1109/TII.2017.2650204

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R.,Roxburgh, C., & Hung Byers, A. (2011). Big data:The next frontier for innovation, competition andproductivity. McKinsey Global Institute. https://catalog.lib.kyushu-u.ac.jp/opac_detail_md/?lang=0&amode=MD824 & bibid=3144682

Mehmood, A., Natgunanathan, I., Xiang, Y., Hua, G., &Guo, S. (2016). Protection of big data privacy. IEEEaccess, 4, 1821-1834. doi.org/10.1109/ACCESS.2016.2558446