International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Efficient Approach for Anonymizing Tree Structured Dataset using Improved Greedy Search Algorithm

Ruchira Warekar[1], Savitri Patil[2]

[1]P. G. Student, Department of Computer Engineering, GHRCEM, Wagholi, Pune, India

[2]Assistant Professor, Department of Computer Engineering, GHRCEM, Wagholi, Pune, India

Abstract: ***In companies and organizations collection of personal information is must. Hence this information collection is increasing day by day. This imposes the serious problem of maintaining the privacy of personal information. Data anonymization techniques presented recently in order to provide security to personal data of users. However such methods suffered from various limitations. Below four are recent research problems in this domain to achieve: a) in many practical cases there are strict utility requirements that cannot be met when more powerful guaranties are applied, b) there is often inability to characterize attributes as sensitive or non sensitive, c) the privacy protection law in most countries usually focuses on identity, and d) recent methods uses the greedy algorithm for achieving anonymization large scale tree structured dataset, but greedy algorithm are having limitations***.

Keywords: Privacy, anonymity , security , integrity.

1. Introduction

Increasing population and use of Smartphone's is resulting large scale collection of personal information at companies or organizations. Therefore the privacy related concerns are posing significant challenges to the data management community. Data anonymization techniques have been proposed in order to allow processing of personal data without compromising user's privacy. In this project we are addressing the problem of anonymizing tree structured data in the presence of structural knowledge. We are presenting k(m;n)-anonymity privacy guarantee which addresses background knowledge of both value and structure using improved and automatic greedy algorithm. The recent technique presented in which scalability is achieved by using basic greedy algorithm but due to the limitations of existing greedy algorithm we are proposing new automatic greedy algorithm to address this limitations.

1.1 Privacy

Privacy is the ability of an individual or group to seclude themselves, or information about themselves, and thereby express themselves selectively. The boundaries and content of what is considered private differ among cultures and individuals, but share common themes. When something is private to a *person*, it usually means that something is inherently special or sensitive to them. The domain of privacy partially overlaps security (confidentiality), which can include the concepts of appropriate use, as well as protection of information. Privacy may also take the form of bodily integrity.

1.2 Anonymity

"Anonymous" is used to describe situations where the acting person's name is unknown. It can be said as not using your own name, simply. Some writers have argued that namelessness, though technically correct, does not capture what is more centrally at stake in contexts of anonymity. The important idea here is that a person be non-identifiable, unreachable, or untraceable. Anonymity is seen as a technique, or a way of realizing, certain other values, such as privacy, or liberty.

2. Related Work

R. J. Bayardo and R. Agrawal, "*Data Privacy through Optimal k-Anonymization.*" Data de-identification reconciles the demand for release of data for research purposes and it demands individuals privacy. This paper proposes and evaluates an optimization algorithm for the powerful procedure of de-identification known as k-anonymization. A k-anonymized dataset has the property that each record is indistinguishable from at least other k - 1. More simple restrictions of optimized k-anonymity are NP-hard, leading to significant computational challenges. It present a new approach to exploring the space of possible anonymizations that tames the combinatorics of the problem, and it develop data-management strategies to reduce reliance on expensive operations like as sorting. Through experiments on real census data, the resulting algorithm can find optimal k-anonymizations under two illustrative cost measures and a wide range of k. The algorithm can produce good anonymizations in circumstances where the input data or input parameters restrict finding an optimal solution in reasonable time. This algorithm to explore the effects of various coding approaches and problem variations on anonymization quality and performance. This result signifying optimal k-anonymization of a non-trivial dataset under a general model of the problem.[1]

R. Chaytor and K. Wang, *"Small-domain randomization: Same privacy more utility"* Random perturbation is a promising technique for privacy preserving data mining. It

retains an original sensitive value with a certain probability and replaces it with a random value from the domain with the remaining probability. If the replacing value is chosen from a large domain, the retention probability must be small to protect privacy. For this reason, previous randomization based approaches have poor utility. In this paper, we propose an alternative way to randomize sensitive values, called small domain randomization. First, we partition the given table into sub-tables that have smaller domains of sensitive values. Then, we randomize the sensitive values within each sub-table independently. Since each sub-table has a smaller domain, a larger retention probability is permitted. We propose this approach as an alternative to classical partition-based approaches to privacy preserving data publishing. There are two key issues: ensure the published sub-tables do not disclose more private information than what is permitted on the original table, and partition the table so that utility is maximized. We present an effective solution.[2]

R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong , "*Publishing set-valued data via differential privacy*" This states set-valued data provides enormous opportunities for various data mining techniques. This mentioned the problem of preparing set-valued data for data mining tasks under the rigorous differential privacy model. All existing data producing methods for set-valued data are based on partition based privacy models, for example k-anonymity, which are unsafe to privacy attacks based on background knowledge. In contrast, differential privacy provides strong privacy guarantees individualistic of an adversary's background knowledge and computational power. Existing data publishing approaches for differential privacy, however, are not sufficient in terms of both utility and scalability in the context of set-valued data due to its high dimensionality. It indicate that set-valued data could be efficiently released under differential privacy with guaranteed beneficial with the help of context-free taxonomy trees. We propose a probabilistic top-down partitioning algorithm to produce a differentially private release, which scales linearly with the input data size. It also indicates the applicability of our idea to the context of relational data. We prove that our result is (ϱ, δ)-applicable for the class of counting queries, the foundation of many data mining tasks. We show that our approach maintains high informal for counting queries and frequent itemset mining and scales to large datasets through extensive demonstrate on real-life set-valued datasets.[3]

J. Cheng, A. W.-c. Fu, and J. Liu , " K-isomorphism : privacy preserving network publication against structural attacks." states Serious concerns on privacy protection in social networks have been increased in recent years; however, research in this area is still in its begining. The problem is demanding due to the diversity and complexity of graph data, on which an adversary can help many types of background knowledge to conduct an attack. Our investigations show that k-isomorphism, or anonymization by forming k pairwise isomorphic subgraphs, is both sufficient and necessary for the protection. The problem is shown to be NP-hard. We devise a number of techniques to enhance the anonymization efficiency while retaining the data utility. [4]

G. Cormode, "Personal privacy vs population privacy: learning to attack anonymization." states that Over the last decade great strides have been made in expanding techniques to compute functions privately. In particular, Differential Privacy gives strong promises about closure that can be drawn about an individual. In this paper, we consider the capability of an attacker to use data meeting privacy definitions to build an accurate classifier. Even under Differential Privacy, such classifiers can be used to deduce "private" attributes accurately in realistic data.[5]

3. Proposed Work

3.1 Solving approach

Increasing population and use of Smartphone's is resulting large scale collection of personal information at companies or organizations. Therefore the privacy related concerns are posing significant challenges to the data management community. Data anonymization techniques have been proposed in order to allow processing of personal data without compromising user's privacy. In this project I am addressing the problem of Anonymizing tree structured data in the presence of structural knowledge. I am presenting k(m;n)-anonymity privacy guarantee which addresses background knowledge of both value and structure using improved and automatic greedy algorithm. The recent technique presented in which scalability is achieved by using basic greedy algorithm but due to the limitations of existing greedy algorithm I am proposing new automatic greedy algorithm to address this limitations.

3.2 Anonymization

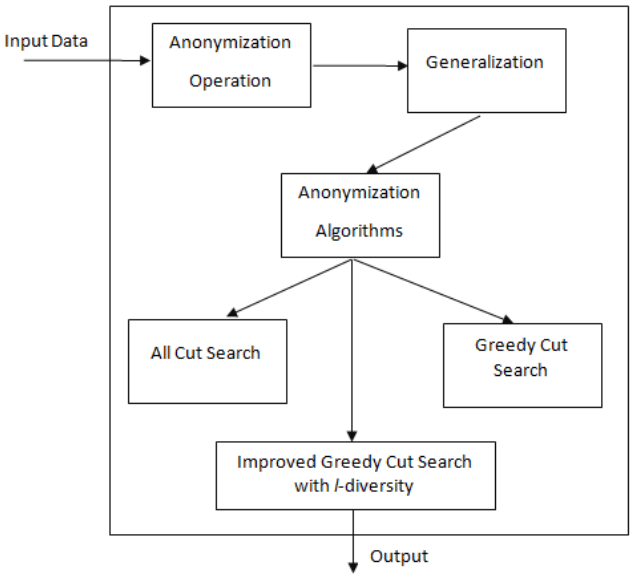

Figure 1: Architecture of the proposed anonymization system

Increasing progress in database, networking, and computing technologies, a large amount of personal data can be integrated and inspect digitally, leading to an increased use of data-mining tools to infer trends and patterns [4].This has raised universal concerns about maintaining the privacy of

individuals Combining data tables from multiple data sources allows us to draw conclusion which are not feasible from a single source. The traditional approach of releasing the data tables without breaking the privacy of individuals in the table is to de-identify records by removing the identifying fields such as name, address, and social security number.

3.3 k anonymity

While k-Anonymity forces one to derive an attribute value even if all but one of the records in a cluster have the identical value, the above clustering-based anonymization technique allows us to pick a cluster center whose value along this attribute dimension is the identical as the common value, thus enabling us to release more information without losing privacy.[1]

The concept of k-anonymity tries to express on the private table PT to be released, one of the main necessity that has been followed by the statistical community Agencies releasing the data, and according to which the released data should be equivalent related to no less than a certain number of respondents. The set of attributes involved in the private table, also externally obtainable and therefore exploitable for linking, is called quasi-identifier . The requirement just expressed is then translated in the k-anonymity requirement, which states that every tuple released cannot be related to fewer than k respondents

4. Scope of Work

- Some methods focusing to protect identity or other properties of single node which resulted into non-scalable method.
- Some methods provide privacy but data loss resulted.
- Greedy algorithm cannot work for some anonymization problems correctly.
- Greedy algorithm required manual working to achieve the correct results.

4 Outcomes

In this paper, we present efficient approach for anonymization tree structured data for improve datasets. We proposed new approach greedy algorithm to remove limitations of existing techniques. To present literature review of different techniques of personal information privacy preserving.

Additionally, we have presented anonymization techniques is able to scale large datasets. We demonstrate Comparative Analysis between existing and proposed system will done using performance metrics such as Execution time, Accuracy etc.

References

[1] R. J. Bayardo and R. Agrawal. Data Privacy through Optimal k-Anonymization. In ICDE, pages 217–228, 2005.
[2] R. Chaytor and K. Wang. Small-domain randomization: Same privacy more utility. In VLDB, 2010.
[3] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. PVLDB,4(11):1087–1098, 2011.
[4] J. Cheng, A.W.-c. Fu, and J. Liu. K-isomorphism: privacy preserving network publication against structural attacks. In SIGMOD, 2010.
[5] G. Cormode , Personal privacy vs population privacy: learning to attack anonymization.