氏　　名	MOHAMMAD RASOOL SARRAFI AGHDAM
学位(専攻分野)	博士(情報学)
学 位 記 番 号	総研大甲第 1928 号
学位授与の日付	平成29年3月24日
学位授与の要件	複合科学研究科　情報学専攻 学位規則第6条第1項該当
学 位 論 文 題 目	Achieving High Data Utility K-Anonymization Using Similarity-Based Clustering Model
論 文 審 査 委 員	主　査　教授　曽根原　登 教授　越前　功 教授　合田　憲人 特任教授　山田　茂樹　国立情報学研究所 教授　小舘　亮之　津田塾大学

(別紙様式 2)
(Separate Form 2)

論文の要旨
Summary (Abstract) of doctoral thesis contents

Nowadays, in the communication, smart devices, social networks and Big Data era, privacy has become one of the main concerns of all socially active individuals. Mainly because of advanced information processing, storage capacity, data mining technologies and the necessity of information sharing in social life. There are lots of government agencies, organizations and service providers that collect and store huge amount of information containing personal information of individuals as their common procedure. The collected micro-data which is a combination of categorical and numerical attributes, contains identifying attributes (e.g., Date of Birth, Sex and Zip code) and private sensitive attributes (e.g., Salary, Credit Card Records and Disease). Sharing the collected micro-data for research and education purposes would be very helpful for researchers and data miners to investigate the correlation between different attributes and get some useful outcomes. However, individual's privacy is one of the main concerns in data publishing especially when releasing datasets involving human subjects contain private sensitive information. Even though information such as name and social security number are discarded in the shared dataset re-identification of individuals is still very much possible due to the existence of combination other identifying attributes. Therefore to protect the privacy of individuals, a model that is widely used for privacy preservation in publishing micro-data, is k-anonymity proposed by Samarati and Sweeney. It suggests "For every record in a released dataset there should be at least k-1 other records identical to it along the quasi-identifier attributes". K-anonymity from clustering aspect is defined as "k-anonymity is clustering with constrain of minimum k tuples in each group". K-anonymity protects the privacy of individuals by modifying the values of identifying attributes through generalization and suppression. Through this modification some information loss occurs. Information loss in k-anonymity model is an unfortunate and inevitable consequence. This information loss reduces the utility of anonymized-data and makes the anonymized-data to be less accurate and accordingly less useful for further analysis.

In addition, there is a trade-off relationship between the privacy and data utility. Due to this trade-off, performing anonymization with maximum privacy and attaining maximum data utility is not possible. Moreover, the problem of optimal k-anonymization and computational complexity of finding an optimal solution for the k-anonymity problem has been proven to be NP-hard. Furthermore, real world and census datasets contain both numerical and categorical type data. As a matter of fact most of the QID attributes in micro-data are assume to be categorical. The combination of numerical and categorical attributes makes anonymization process

rather complicated and very often results in an inefficient anonymization with very high information loss. Most of the previous approaches and techniques to achieve k-anonymity suffer from huge information loss and very low data utility. Also most of the approaches are mainly designed for continuous numerical attributes and in case of considering categorical attributes, they depend on hierarchical taxonomies or require some additional information, which more often than not, are not defined or available in real life applications.

Therefore, in order to maximize the utility of anonymized-data in real life applications in this work a new approach is proposed. The proposed model is called Similarity-Based Clustering (SBC) Anonymization. It is based on clustering and local recoding anonymization method. SBC model concentrates on clustering the original dataset containing both numerical and categorical attributes efficiently based on given k value so after anonymization the information loss kept as minimum as possible.

SBC model suggests a new similarity measurement and distance calculation based on the measured similarity for categorical attributes so the total distance between tuples can be calculated for clustering. This approach does not depend on hierarchical taxonomies regarding categorical attributes. Based on the proposed model a bottom-up greedy algorithm for k-anonymization is proposed and evaluated on two different real datasets. Our extensive study on information loss and data utility show that the proposed algorithm based on SBC model in comparison with existing well-known algorithms offers data utility above 80% and reduces the information loss to less than 20% within the wide range of various k values.

Keywords: Privacy Preserving, Data Mining, Anonymization, Algorithm and Security

(別紙様式 3)
(Separate Form 3)

博士論文審査結果の要旨

Summary of the results of the doctoral thesis screening

本学位論文は、一般化を用いるk-匿名化処理に伴う情報損失の削減手法に関する。プライバシー保護とデータ活用の両立が問題となる実データセットは、数値と分類属性情報の組み合わせ情報からなり、ロングテール型の分布特性を示す。従来のk-匿名化処理では階層的分類による一般化処理が行われるが、実データセットでは情報損失が大きいという課題がある。本研究では、このような問題を解決するため、データセット内の分類属性情報の生起確率とデータレコード間の距離による類似性評価に基づいたクラスタリングを行い、生成された各クラスタをローカル・リコーディングによってデータ匿名化するSimilarity-Based Clustering（SBC）手法が考案されている。本提案手法を実社会のシステムで収集される大規模データセットに適用してデータ有用性を検証した結果、全てのデータセットで広範囲なプライバシー保護レベルk(2～100人)で、情報損失量は20%以下に軽減されている。また、従来のMondrian, Datafly, Incognito匿名化手法と比較すると、情報損失量が平均で35%改善されている。

本学位論文は、5章から構成される。第1章では、研究の背景や目的について述べる。第2章では、データ・マイニングやパブリッシングにおけるデータ・プライバシーの課題について論じている。個人データのデータベース連結によって生じるプライバシー侵害の事例を取り上げ、プライバシーの保護と活用の両立問題について述べている。次に、同じ属性データの組み合わせを持つレコードが、少なくともk個存在し、属性データからの個人識別がk人未満に絞り込めないというk-匿名化処理の技術的課題について述べている。k-匿名化処理では、一般化、削除、雑音付加、データ入れ替えなどによって、再識別性リスクを1/kに減少させ、プライバシー保護を行う。このk-匿名化処理の課題として、データが数値と分類属性情報を有し、しかもそれらがロングテール分布を示す実データセットである場合、データ匿名化処理後の情報損失が大きく、有用性が減少する問題がある。

第3章にて、クラスタリングとローカル・リコーディングに基づくSimilarity-Based Clustering（SBC）手法を提案している。提案手法は、個人を間接的に特定できる可能性がある準識別情報を、生起確率とデータレコード間距離による類似性評価に基づいてクラスタリングし、生成された各クラスタをローカル・リコーディングによって匿名化する方法である。本手法は、数値と分類属性情報の組み合わせからなるデータセットを同一の評価基準でクラスタリングできることに特徴がある。

第4章では、本提案手法を2つのデータセットに適用し、データ有用性の定量的な評価結果について述べている。SBC手法と代表的なMondrian, Datafly, Incognitoの3つの手法を情報損失の観点から比較している。その結果、プライバシー保護レベルk=2を設定した場合、既存手法と比較してデータ有用性は平均40.6%、最大で62%改善し、k=100に設定した場合でも、平均32.3%、最大で40%改善できている。このように広範なk（2～100）に対し、情報損失量が20%以下に抑えられている。最後に、第5章で研究の結論、今後の進展等について述べている。このように、提案したSBC手法は、広範囲のプライ

(別紙様式 3)
(Separate Form 3)

バシー保護レベルに対してデータ有用性の高い有効な匿名化処理手法であり、実社会のデータセットに適用可能となっている。

なお、研究成果として、出願者は主著で査読付きジャーナル論文 2 篇(IEICE Trans. on Information and Systems, (full paper)), IJCSDF International Journal of Cyber-Security and Digital Forensics, (full paper))、査読付き国際会議論文 (CyberSec2013, (full paper)) 1 篇を発表した。また、共著論文で査読付きジャーナル論文 1 篇(JIP Journal of Information Processing Society of Japan, (full paper))を発表した。したがって、学位論文に充分なレベルであることを判断した。