High Performance Datafly based Anonymity Algorithm and Its L-Diversity

Zhi-ting Yu, School of Computer Engineering and Science, Shanghai University, Shanghai, China
Quan Qian, School of Computer Engineering and Science, Shanghai University, Shanghai, China
Chun-Yuan Lin, Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan, Taiwan
Che-Lun Hung, Department of Computer Science and Communication Engineering, Providence University, Taichung, Taiwan

ABSTRACT

Data anonymity, as an effective privacy protection method, has been widely used in real applications. High performance data anonymity algorithm is especially attractive for those massive data applications. In this paper, the authors propose a novel and efficient Datafly based data anonymity (Divide-Datafly) algorithm and the experimental results show that the proposed algorithm is not only more efficient than Datafly and Incognito, but also information loss less than KACA. Moreover, in order to improve the security of anonymous data, L-Divide-Datafly is presented that it combines Divide-Datafly and efficient distance based clustering. Experimental results show that L-Divide-Datafly achieves great performance both in execution time and Information loss.

Keywords: Divide-Datafly, High Performance Data Anonymity Algorithm, L-Divide-Datafly

1. INTRODUCTION

In this cloud computing era, vast number of information is collected in different ways every day. The cryptography becomes important role of Internet (Jasim et al., 2014). These data are commonly analyzed by companies and research institutions to find hidden values. Data mining promotes the development of data science and brings convenience to our daily life. For example, mining on medical data can improve diagnose accuracy and help doctors to find potential diseases. The companies can recommend clients what they are interested in based on their footprints.
in Internet. But these data may contain clients’ private information that is not allowed to be reached by others, especially medical records, home address, income and so on (Saranya et al., 2015). Sweeney pointed out that there were about 85% of Americans can be uniquely identified through the combination of their gender, birth and postcode (Sweeney, 2002a). In order to reduce the execution time of data anonymity, we proposed a novel algorithm Divide-Datafly based on Datafly that has a good performance in both execution time and information loss. Moreover, a L-diversity model, including L-diversity features of Divide-Datafly, has been presented to analyze this new kind of anonymity algorithm.

2. RELATED WORK

Data encryption, differential privacy, k-anonymity and many other technologies are proposed to protect the data privacy for users. The idea of k-anonymity is proposed by Samariti and L.Sweeney (Samariti and L.Sweeney, 1998). The key idea of kanonymity is to make individuals indistinguishable in a released table. A tuple representing an individual within the identifiable attributes has to be identical in at least (k-1) other tuples. This method has been widely used because of its simplicity.

The K-anonymity related algorithms could be divided into three types: global recoding, multidimensional recoding and local recoding. Global recoding algorithms, such as Datafly (Lefevre et al., 2005), Incognito (Sweeney, 2002b), TopDown (Fung et al., 2005) and BottomUp (Wang et al., 2004), require that all attributes of the tuples in dataset have the same generalization form. Although these algorithms have low computation complexity, they may cause over generalization. Multidimensional recoding, such as Mondrian, maps a set of values to another set of values, some of which are more general than the corresponding premapping values. But this model does not consider attribute hierarchical structures. Local recoding algorithms allow values of an attribute in different generalization domain. The information loss of these local recording anonymity algorithms is low, but the execution time of these algorithms is longer than that of global algorithm. Also, the model does not consider attribute hierarchical too. The typical local recoding algorithms are the KACA (Li et al., 2006), MDAV (Torra, 2004) and its L-diversity model (Jianmin et al., 2008). The optimal k-anonymity algorithm is considered as a NP-hard problem. Existing researches use heuristic strategies to gain an approximate optimal algorithm.

It is difficult to protect privacy just with k-anonymity model. Also, there are some attacks which k-anonymity is unable to resist, such as homogeneity attack, similarity attack and probability attack. Many algorithms are proposed to resist these attacks, such as p-sensitive k-anonymity (Truta & Vinay, 2006), (alpha,k)-anonymity (Wong et al., 2006), L-Diversity (Machanavajjhala et al., 2007), (a,d)-Diversity (Wang & Shi, 2009), t-closeness (Li et al., 2007), (ω; γ,k)-anonymity (Huang et al., 2014) and (l,t)-closeness anonymization (Yang et al., 2015).

In this paper, we analyzed the defect of global recoding and proposed a new algorithm Divide-Datafly. Through experiments, we compared the proposed algorithm with Datafly, Incognito and KACA. The experimental results on three different datasets show that, Divide-Datafly algorithm is suitable for dataset with numerical attribute. It improves the speed of anonymization and reduces the information loss. We also put forward an L-diversity model of the proposed algorithm based on clustering method and give experiments to analyze the execution time and information loss of it.

The rest of the paper is organized as follows. Section III introduces some prerequisite knowledge of data anonymity. Section IV describes the main process of Datafly and defects of global recoding. Section V presents the Divide-Datafly and the L-diversity model of Divide-Datafly in Section VI. Experiments are shown in Section VII and Section VIII makes a whole conclusion.
Distributed Storage Systems for Data Intensive Computing
[www.igi-global.com/chapter/distributed-storage-systems-data-intensive/62823?camid=4v1a](www.igi-global.com/chapter/distributed-storage-systems-data-intensive/62823?camid=4v1a)

A Simulator for Large-Scale Parallel Computer Architectures
[www.igi-global.com/article/simulator-large-scale-parallel-computer/42976?camid=4v1a](www.igi-global.com/article/simulator-large-scale-parallel-computer/42976?camid=4v1a)