International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Study Paper On: Ontology-Based Privacy Data Chain Disclosure Discovery Method for Big Data

Sonali T. Benke[1], Devidas S.Thosar[2], Kishor N. Shedage[3]

[1] M.E. Student, Computer Engineering, SVIT, Nashik
[2]PG Co-ordinator, Computer Engineering, SVIT, Nashik
[3]HOD, Computer Engineering, SVIT, Nashik

Abstract - *As a new software paradigm, cloud computing provides services dynamically according to user requirements. However, it tends to disclose personal information due to collaborative computing and transparent interactions among SaaS services. We propose a private data disclosure checking method that can be applied to the collaboration interaction process. First, we describe the privacy requirement with ontology and description logic. Second, with dynamic description logic, we validate whether SaaS services are authorized to obtain a user's privacy attributes, to prevent unauthorized services from obtaining their private data. Third, we monitor authorized SaaS services to guarantee privacy requirements. Therefore, we can prevent users' private data from being used and propagated illegally. Finally, we propose privacy disclosure checking algorithms and demonstrate their correctness and feasibility by experiments[7].*

To meet user's functional requirements, cloud computing and big data have become the most commonly used computing and data resources. Based on analysis, conversion, extraction and refinement for the big data, a disease can be prevented and group behavior can be predicted. However, each user's private data is also an element in big data. Users must provide private data to the service providers to meet their functional requirements. To gain economic benefits, some SaaS service providers have not been authorized to collect and analyze the user's sensitive private data, as a result, the user's private data is disclosed. In this paper, we propose a private data chain disclosure discovery method, to prevent a user's sensitive privacy information from being illegally disclosed. Firstly, we measure the similarity degree and cost of the disclosure of the private data.

Key Words: *Ontology, Privacy Disclosure Detection, Privacy Data Chain, Similarity Metric etc.*

1. INTRODUCTION

Big data usually include data sets that have sizes that are beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time, with characteristics that consist of volume, variety and velocity.[18] According to statistics, an average of 2 million users per second use the Google search engine; within one second, Facebook users share information more than 4 billion times, and Twitter handles more than 3.4 hundred million tweets per day. The amount of data grows exponentially every year, and threequarters of data are produced by people, for example, a standard American worker contributes 1.8 million MB every year. A large amount of personal privacy data can be mined for commercial purposes by agents. For example, Acxiomac queries more than 5 million personal data of consumers all over the world through data processing and analyzes individual behaviors and psychological tendencies with technologies, some of which are known as data association and logical reasoning.

In 2014, Adam Sadilekat University of Rochester and John Krumm in the Microsoft lab predicted a person's likelihood to reach a location in the future by analyzing the information in the big data, with an accuracy as high as 80%. A mobile application does not protect the location of the big data; as a result, a user's home address and other sensitive information can be disclosed through the triangulation reasoning method. Research shows that user attributes can be found by analyzing group features in a social network. For example, by analyzing a user's Twitter messages, the user's political leanings, consumption habits and other personal preferences can be found. Therefore, how to protect personal privacy information has become a hot research topic with respect to bigdata.

1.1 Objective

1. We can check the private data disclosure chain and the key private data, which can effectively prevent service participants from maliciously disclosing users' private data, increase the service trustworthiness, and provide a basis for a privacy safety-oriented trustworthiness measurement. Detailed contributions are showed as follows. Firstly, we get the relationships among privacy data by the mapping with knowledge ontology, and build the ontology tree. We also measure the similarity degrees, containing property similarity, object similarity and hierarchical similarity.

2. Secondly, we measure the cost of the disclosure of the private data with sensitivity grades and privacy disclosure vector. According to the similarity degree and cost of disclosure, the disclosure chain and key private data are detected in the process of interaction between user and SaaS service.

3. Thirdly, we propose a discovery framework for the private data chain and demonstrate its feasibility and effectiveness by experiments. Which provide a reference to develop a system software assuring the safety for personal privacy data in big data.

1.2 Literature Survey

A number of researches have been proposed by researchers for privacy preserving in big data. A detailed survey has been carried out to identify the various research articles available in the literature in all the categories of privacy preserving in big data, and to do the analysis of the major contributions and its advantages. Following are the literatures applied for assessment of the state-of-art work on the privacy preserving in big data. Here, few works has been analyzed.

Research article related to privacy preserving in data mining:

Data mining is a strategy where huge measures of both sensitive and non-sensitive information are gathered and analyzed. While circulating such private information, security protecting turns into a critical issue. Different strategies and procedures have been presented in security saving information mining to attempt this issue. The techniques can be listed under three major classical PPDM techniques.

Privacy preserving using Anonymization approach:

Asmaa et al.[13] have explained the Protect Privacy of Medical Informatics using K-Anonymization Model. Here, they displayed a structure and model framework for de-distinguishing health data including both organized and unstructured information. They exactly examine a straightforward Bayesian classifier, a Bayesian classifier with an inspecting based strategy, and a contingent irregular field based classifier for removing distinguishing traits from unstructured information. They, convey a k anonymization based system for de-recognizing the removed information to save most extreme information utility. Moreover, Tiancheng Li et al. [14] have explained the Towards Optimal k-anonymization which was a more flexible scheme for privacy preserving. Here, they introduced enumerate the algorithm for pruning approach for finding optimal generalization. Likewise, V.Rajalakshmi and G.S.Anandha Mala [11] amazingly advocated the Anonymization based on nested clustering for privacy preservation in Data

Mining. In their document, the dimension of clusters was preserved optimal to cut down the data loss. They elaborately discussed the technique, performance and outcomes of the nested clustering. Similarly, Yan Zhu and Lin Peng [15] have explained the Study on K-anonymity Models of Sharing Medical Information.

Privacy preserving using Clustering algorithm:

In this section, we discussed the research article based on privacy preserving using clustering algorithm. In S. Patel et al. proficiently introduced a privacy preserving distributed K-Means clustering of horizontally partitioned data which significantly alleviated the safety concerns in the malicious ill-disposed model. The vital growth involved the use of the secret transferring mechanism battered to code based zero knowledge identification technique.

Moreover, Bipul Roy brilliantly brought to limelight an innovative tree-based perturbation approach which was easily employed for tackling the perturbing data reflecting the hidden conveyances. In their approach, they made use of a Kd-tree stratagem to recursively divide a dataset into a number of diminutive subsets in such a way that the data records within each subset became further harmonized with every partition. When the partitioning process was fully carried out the confidential data in every subset were perturbed by the effective exploitation of the micro aggregation method.

Additionally, Alper Bilge and Huseyin Polat admirably brought to light the scalable privacy-preserving recommendation technique by means of bisecting k-means clustering. In their innovative privacy-preserving collaborative filtering scheme dependent on bisecting k-means clustering they introduced two pre-processing approaches.

Similarly, Ali Inan et al. intelligently advocated the Privacy preserving clustering on horizontally partitioned data. In the document, they competently created the dissimilarity matrix of objects from varied sites in a privacy preserving fashion which was employed for the purpose of the privacy preserving clustering and also the database joins, record linkage and other function which necessitated couple wise appraisal and assessment of individual private data objects horizontally disseminated to several sites.

In Jinfei Liu et al. efficiently brought to limelight the Privacy Preserving Distributed DBSCAN Clustering. In their innovative technique, they effectively tackled the hassles of the two-party privacy preserving DBSCAN clustering. At the outset, they brilliantly brought in two protocols for privacy preserving DBSCAN clustering over horizontally and vertically segregated data correspondingly and later widened them to the randomly segregated data.

Also, Pan Yang et al. excellently explained the Privacy-Preserving Data Obfuscation Scheme Used in Data Statistics and Data Mining. In their technique, they apportioned dissimilar keys to the diverse users, who were given divergent permissions to access the data. In the ultimate phase, a fine-grained grouping method founded on similarity was discussed.

2. MOTIVATION

Thanks to the advent of the Internet, it is now possible to easily share vast amounts of electronic information and computer resources (which include hardware, computer services, etc.) in open distributed environments. These environments serve as a common platform for heterogeneous users (e.g., corporate, individuals etc.) by hosting customized user applications and systems, providing ubiquitous access to the shared resources and requiring less administrative efforts; as a result, they enable users and companies to increase their productivity.[9]

These approaches are usually enough to manage users' access rights in closed environments, such as static organizations, which involve a limited number of entities and resources, and for which a manual management of access rules is feasible. However, manually managing privacy rules and access constraints in open environment, such as OSNs or the cloud, is not practical due to the following reasons:

(i) A large number of entities need to be managed. For example, Google Drive has billions of users and each user manages several types of resources.

(ii) The heterogeneous entities involved in these scenarios would likely have diverse privacy requirements. For example, for a cloud provider offering storage space, an organization involving employees, departments and resources would define significantly different privacy requirements than casual end-users.

(iii) The dynamicity and openness of such scenarios make the privacy requirements to change rapidly with respect to the type of services and users.

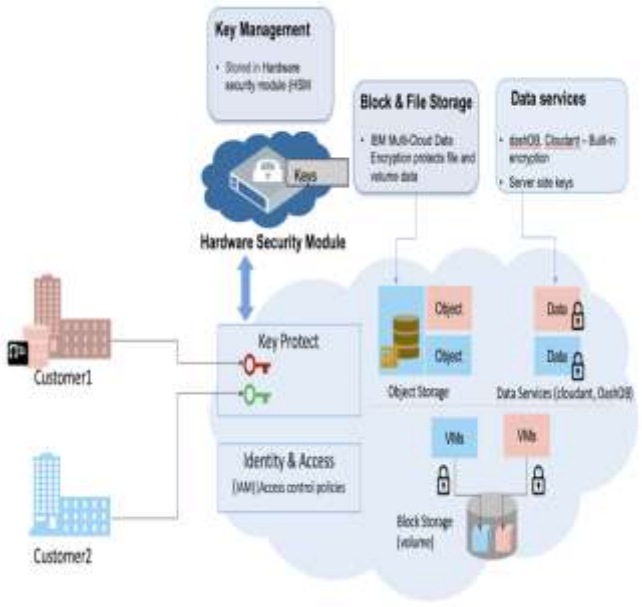

Fig -1: System diagram

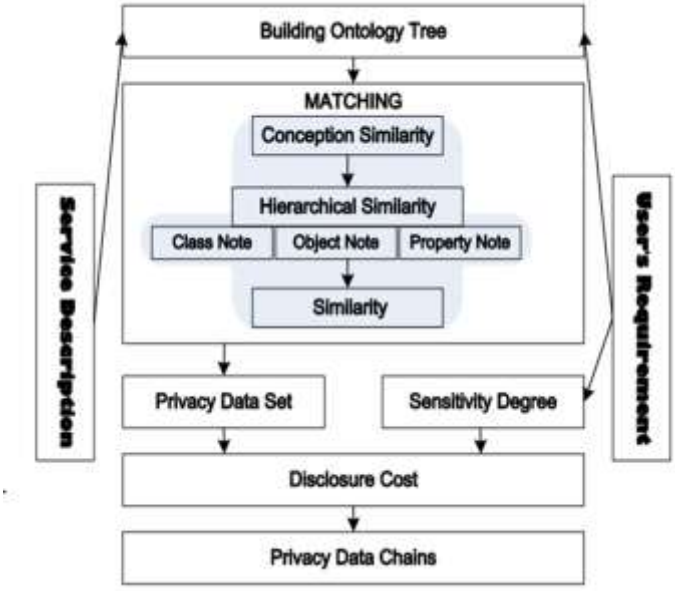

Fig -2: block diagram

3. CONCLUSION

According to the interaction characters among SaaS services, we propose a privacy disclosure checking method that satisfies users' requirements. We develop a prototype system, which describes the users' privacy requirements and extends BPEL and its execution engine to meet users' privacy requirements. We also design a case and run it on the prototype system to confirm the feasibility and correctness of our method. Our approach can check privacy disclosure behavior among SaaS services, which can effectively prevent service participants from maliciously disclosing users' privacy information, increase service credibility, and provide a basis for privacy protection-oriented credibility measurement. The next step is to detect the release of the data of users' privacy, analyze the data, and discretize the dataset that may be exposed to protect users' privacy before they are released.

REFERENCES

[1] Changbo Ke, Fu Xiao, Member, IEEE, Zhiqiu Huang, Yunfei Meng and Yan Cao ,Ontology-Based Privacy Data Chain Disclosure Discovery Method for Big Data", 2019 IEEE.

[2] Wang Y, Kung L A, Wang W Y C, et al. An integrated big data analyticsenabled transformation model: Application to health care [J]. Information & Management, 2018, 55(1): 64-79.

[3] He X, Ai Q, Qiu R C, et al. A big data architecture design for smart grids based on random matrix theory [J]. IEEE transactions on smart Grid, 2017, 8(2): 674-686.

[4] Peddinti S T, Ross K W, Cappos J. User Anonymity on Twitter[J]. IEEE Security & Privacy, 2017, 15(3): 84-87.

[5] Wan J, Tang S, Li D, et al. A manufacturing big data solution for active preventive maintenance [J]. IEEE Transactions on Industrial Informatics, 2017, 13(4): 2039-2047.

[6] Storey V C, Song I Y. Big data technologies and management: What conceptual modeling can do Data & Knowledge Engineering, 2017, 108: 50-67.

[7] Ke C, Huang Z, Cheng X. Privacy Disclosure Checking Method Applied on Collaboration Interactions Among SaaS Services[J].IEEE Access, 2017, 5: 15080-15092.

[8] Zeydan E, Bastug E, Bennis M, et al. Big data caching for networking: Moving from cloud to edge IEEE Communications Magazine, 2016, 54(9): 36-42.

[9] Imran-Daud, Malik, "Ontology-based Access Control in Open Scenarios: Applications to Social Networks and the Cloud", eprint arXiv: 1612.09527,Pub Date: December 2016

[10] Zhang X, Dou W, Pei J, et al. Proximity-aware local-recoding anonymization with mapreduce for scalable big data privacy preservation in cloud [J]. IEEE transactions on computers, 2015, 64(8): 2293-2307.

[11] Lv Y, Duan Y, Kang W, et al. Traffic flow prediction with big data: A deep learning approach [J]. IEEE Trans. Intelligent Transportation Systems, 2015, 16(2): 865-873.

[12] V Rajalaxmi,GSA MALA, Anonymization based on nested clustering for privacy preservation in data mining,IJCSE, 2013

[13] Sankita Patel, Viren Patel, Devesh Jinwala, "Privacy preserving distributed"

[14] Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science 7: 1–5

[15] Asmaa H. Rashid, A. F.Hegazy, "Protect privacy of medical informatics using K- anonymization madel", IEEE, April 2010

[16] Tiancheng Li, Ninghui Li, Towards optimal k-anonymization, Data & knowledge Engineering volume 65, issue 1, April 2008 pages 22-39

[17] Yan ZHU, Lin PENG, "Study on k-anonymity models of sharing medical information", IEEE, 2007[

[18] www.wikipedia.com