PEZZOTTAITE JOURNALS

Volume 7, Number 4, October – December' 2018
ISSN (Print): 2319-9016, (Online): 2319-9024
SJIF (2016): 7.368, SJIF (2017): 7.418

THE RESEARCH OF BIG DATA ANALYTICS OF PRIVACY PRESERVATION IN UBIQUITOUS ENVIRONMENT

Suneetha V.[3] Dr. Y. S. Kumara Swamy[4] Ramakrishnan M. R.[5]

ABSTRACT

Big data analytics has created opportunities for researchers to process huge amount of data but created a big threat to privacy of individual. Data processed by big data analytics platforms may have personal information, which needs to be taken care of when deriving some useful results for research. Existing privacy preserving techniques like, anonymization requires having dataset divided in the set of attributes like, sensitive attributes, quasi identifiers, and non - sensitive attributes. With the organized data, it may possible to have such a distribution but in unstructured data, it is very difficult to identify sensitive attributes and quasi identifiers. The development of the Data Science, Sets off the third waves of the world information industry. The data mining technology plays a vital role in the development and promotion of Data Science and ubiquitous environment, but it causes leakage problem of privacy information at the same time. In the light of the data mining association rules and randomized response method. We propose a new method, suppressible randomized response method (SRRM), and introduce the data mining algorithm of privacy protection based on SRR. Finally, this paper evaluates the privacy of the method.

KEYWORDS

Big Data Analytics, Data Science, Data Mining, Association Rules, Privacy Preserving, Ubiquitous Environment, Randomized Response etc.

INTRODUCTION

Gaining access to high-quality data is a vital necessity in knowledge-based decision-making. However, data in its raw form often contains sensitive information about individuals. Providing solutions to this problem, the methods and tools of privacy-preserving data publishing enable the publication of useful information while protecting data privacy.

Privacy preservation is crucial in ubiquitous computing environment. Without this, users feel uneasy to use and live in the UC environment. The implementation of privacy safeguard or privacy enhancing technologies is going to be a long road. Understanding the challenges & issues of privacy protected in ubiquitous computing, is helpful to design and implement privacy aware system. In this paper, we try to describe privacy in ubiquitous computing briefly.

Big data can be defined as, "The data sets so large so large or complex that are difficult to process using traditional data processing applications". Size of big data may be in zeta bytes (increasing proportionally with time). Big data has characteristics of 3Vs; Volume (large amount of data), Velocity (speed of data generation and processing), and Variety (structured, unstructured, or semi-structured data).

Big data analytics is very helpful in various fields like, medical science, national security, semantic web, social media, etc. On the other hand, it creates threat to an individual as it has capacity to store and process large amount of data very quickly and accurately, due to advancement in technologies like, NoSQL data models, Hadoop, Map- reduce, etc. Therefore, instead of seeing the picture as big data analytics vs. privacy, we need to have individual privacy preservation with almost all advantages of big data analytics. A privacy preserving technique is required,

[3] HoD-MCA, Dayananda Sagar College of Arts Science & Commerce, Karnataka, India, mcabu@dayanandasagar.edu
[4] Professor, Department of CSE, Nagarjuna College of Engineering & Technology, Karnataka, India, ysklds wamy2@yahoo.com
[5]Chief Architect, Crimson Infotech, Karnataka, India, maggekris@gmail.com

which maintains a trade-off between privacy and utility of individual's data. To understand the importance of privacy in big data analytics, privacy issues with big data analytics have been discussed. The Internet of Things is a new system that can communicate with the real world. It is also a virtual network including the ubiquitous data perception, the information transmission mainly by wireless and the intelligent information processing through the sharing information platform. In today's world each individual wish that his private information is not revealed in some or the other way. Privacy preservation plays a vital role in preventing individual private data preserved from the praying eyes. Anonymization techniques enable publication of information, which permit analysis and guarantee privacy of sensitive information in data against variety of attacks. It sanitizes the information. It can also keep the person anonymous using encryption technique. There are various anonymization techniques and algorithms available, which are discussed in this paper.

The wide application of Internet of Things in the production and living must be with more knowledge discovery. In this process, the data mining technology plays a positive role. The data of the sharing platform on the Internet of Things most comes from the data being closely related to people's life, such as position, environment and habits of work and living. These data are sensitive information for most users, and some malicious users cannot access them. Therefore, we have to consider the privacy leakage problem in the sharing platform on the Internet of Things.

Everything is available on mobile nowadays. People are sharing lot of information on mobile phones. Often, mobile sends data to the service provider without user's knowledge. Identifying the person using his mobile data and the details provided by the service provider is very easy. Therefore, privacy in mobile data is very important. Text message analysis is an example of unstructured big data analytics in mobile. Mobile application like WhatsApp is using text message analysis in their mobile number verification method. In such method, a verification number sent to the registered mobile number and if it is same in which app is installed, app will automatically write the number in verification box as soon as it arrives via SMS.

PRIVACY ISSUES IN HEALTH CARE DATA

Big data analytics and genome research having real time access to patient record helps doctors to take decisions. Electronic Health Records (EHR) helped a lot to digitize the health care system and EHR incentive program motivates hospitals to create an accurate and complete EHR. On the other hand, EHR having personal information of patient may lead to privacy breach. Therefore, privacy-preserving analysis is required and data need to be anonymized or encrypted before data analysis.

Privacy Issues in Social Media Data

Social media is one of the biggest revolutions in past decade. People on social media are sharing Lot of information. Sometimes, people close to you shares some information about you, which you do not want disclose on social media. This may lead to privacy violation of an individual. For e.g.; you have taken a sick leave from office to watch a cricket match and one of your friend checked-in, you on Facebook so people comes to know that you were not sick but enjoying match. Though, privacy settings are there in Facebook to approve tag so if someone tagged you, it need to be approved before it is posted on your wall but it is going to appear on your friend's wall as soon as he is posts with link to your profile.

Privacy Issues in Web Usage Data

Intel want to make its internal website dynamic (appearance of the website changes as per the access pattern of users, viz. links visited by most number of users should be on the first page to save click time and improve productivity) based on web usage data of all the users of the website. With browser information and IP address from web usage data any user can be identified and whatever activities he is performing on line may be detected. Therefore, user privacy is violated by such system. Sedayao et al., have suggested a model in which symmetric key encryption is used to anonymize the sensitive data identified based on predefined tags like, IP address and user ID from semi structured web usage data.

RELATED CONCEPTS AND WORK

Data mining is to extract knowledge people are interested in from a large number of data. **The association rule is one of them, the definition is as follows:**

The following is a formal statement: Let $I = \{i_1, i_2,.....i_k\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. We say that a transaction T contains N, a set of some items in I, if $N \subseteq T$. An association rule is an implication of the form $N \rightarrow Y$, where $Y \subseteq I$, and $N \cap Y = \phi$.

The one of the data mining methods in privacy preserving of randomized response method firstly put forward by Warner. The randomized response technology means when we use a random turning needle to handle and involve sensitive problems. if a sensitive problem has two options, Y and N, the needle is only seen by respondents and it respectively points to Y and N with fixed probability. Finally, according to the probability distribution of needle pointing, investigators get the maximum likelihood estimator of option Y and N taking shares respectively in investigation. Based on it, this proposes a new randomized response method -- suppressible randomized response method (SRRM), which makes the original data be hidden before mining.

SUPPRESSIBLE RANDOMIZED RESPONSE METHOD (SRRM)

In order to express more intuitively, we assume that the data is the data set of market basket. Each commodity is an item with an identified number, customer' each shopping is a transaction, which is expressed by sequence of {0,1}, and the length is the total number of items.

The suppressible randomized response method is firstly exchanging and hiding the original data before the mining of data without the limit of alternatives of randomized parameters. In addition, the specified methods are as follows:

Firstly, giving randomized parameter p_a, a = 1,2,3, $0 \leq p_a \leq 1$, and $\sum p_a = 1$, and then set $f_1=n$, $f_2=1$, $f_3=0$, in the item n, $n \in \{0,1\}$, and the pi probability selected value of randomized function f(n) is f_j j=1,2,3.

To set the items of total number is x, the transaction $N=(n_1,n_2,....n_x)$ which is expressed by a sequence of {0,1} and the distributed transaction $R=(r_1,r_2,...r_x)$ can be calculated through function R = F(N), and $r_i = f(n_i)$. That is to say, the value of r_i is n_i with the probability of p_1 similarly p_2 equals the probability of n_i and p_3 equals the probability of 0.

THE MINING TECHNOLOGY BASED ON SRRM

By the data transformation and data hiding of SRRM, transaction sets D can get a forged transaction sets D'. In the progress of generating the frequent item sets, the most crucial point is to figure out the support of item sets. We will introduce how to compute the support of k-itemsets in the following discussion, and then give the mining algorithm based on Apriori algorithm.

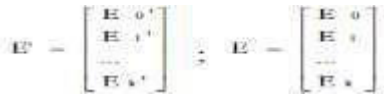
Computing the Support of k-Itemsets

Let $A= \{i_1,i_2,....,i_k\}$ is a k-itemsets. In the case of every item in A, which will be handled by the same randomization parameter, we can make use of some optimization strategy to reduce the computing complexity of the item sets support. When every item in A uses the same randomization parameter, transaction in D including Ai will have the same rate with transaction in D' including Ai which has been handled by SRRM. That is the reason.

$$l_{ij} = \sum_{t=\max(0,i+j-k)}^{m(i,j)} E_j^t * (P_1 + P_2)^t \, P_3^{j-t} * E_{k-j}^{i-t} * P_2^{i-t} * (P_1 + P_3)^{k-i-j+t} \quad (1)$$

Regarding to k-itemsets A, and transaction T in D, there are k+1 possible value in $\{T \cap A\}$. We take the serial sequence E_0, E_1,....E_k as the ratio of every transaction in D. For example, as a 3-item sets, all the transactions in D

will be divided into {000}, {011,101,110},{111}, and E2 is the ratio of two items in A. Similarly, for the transaction T' in D', there are k+1 possible values in |T' ∩ A| as well. Similarly, for the transaction T' in D', there are k+1 possible values in |T' ∩ A| as well. We also take the serial sequence E_0',E_1'....,E_k' as the ratio of every transaction in D'.



Then we have E' = LE, and; L = $[I_{ij}]$ is a (k+1)x (k+1) matrix. L_{ij} exactly represents that D including Aj, changes into a ratio in D' including Ai after it is handled by SRRM.

If L is reversible, let L-1 = $[a_{ij}]$, E= L-1E', yet E_k is just the support of k -itemsets which we are computing.

$E_k = a_{k,0}E_0' + ak,1E_1' + ...+a_{k,k}E_k'$ (2)

Firstly, according to D', we can get E_j' and solve $a_{k,j}$ using L, then the support of k-itemsets A comes out. The time complexity and space complexity of this algorithm is O(k).

In addition, we need to notice E0'+E1'+.....+Ek' =|D'| = N, so, there is one item among all the E_j' can be obtained without any computing. Generally, the value of E_0' is bigger than anyone else. Because of this reason, we can get the value of E_0' by way of E_0'

$= N - (E_0' + E_1' + ... + E_k')$.

The Complete Mining Algorithm

Using the computation formula mentioned above, we can figure out the association rules in which we are interested with the help of various frequent item sets generation algorithm available. In this paper, using Apriori algorithm, here is the specific frequent item sets generative algorithm, which has been handled by SRRM.

Input: D': The transactional databases handled by SRRM:
min_sup: The minimum support count threshold.
output: M : frequent itemsets in D'
scan D', for each item i ϵ I count i.count;
Mi = {i ϵ I |, ((i.count /N) - p2) / p1 ≥min_sup};
for (k=2;$M_{k-1} \neq \phi$; k++)
E_k = apriori_gen (M_{k-1}); // Generate candidate k-itemsets E_k
for each transaction t ϵ D' // scan D for counts
for (j-1; j ≤k; j++)
$E_{i,j}$ = partial_subset (E_k,t,j); // transaction t just
contains candidates k-itemsets of item j
for each candidate eϵE_{ij}
e.count++;
for each candidate e ϵ E_k
e.count = a_{ko}.e.$count_0$ + $a_{k,1}$.e.$count_1$+....+a_k,k.e.$count_k$;
Ek = {eϵ E_k | e.count ≥min_sup;
return M =U$_k$M$_k$;

PRIVACY ASSESSMENTS

The original intention and ultimate goal of research of privacy protection data mining methods is to do data mining and knowledge discovery, and search the potential, valuable patterns and rules based on the premise that protects privacy data properly, therefore, the level of privacy has become the primary factor when evaluating a kind of method.

According to the calculation formula of privacy damage coefficient B[4]:

$B = P_{\text{ratio of real data}} X P_{\text{probability of real data recognized}} + P_{\text{ratio of non real data}} X P_{\text{probability of non real data recognized}} X P_{\text{probability of non-real data reverted}}$.

Assuming that the ratios of real metadata in all method are the same, computing the privacy damage factor:

Randomization parameter p_1=p. We take the value $p_2 = p_3=(1-p_1)/2$. In this way, the probability of being 0 and 1 of non-real data will be exactly same, and cannot be reverted. Otherwise, non-real data will be possible to be recognized and then reverted. E.G., if we have $p_2 =1-p_1$, $p_3=0$, then all the data having the value of 0 that has been handled is real data, thereby the protection degree will be reduced greatly. This method, which takes the average value of 0 and 1 in practice, is not only convenient but also in favor of privacy preserving.

In this case, $B = p1 \frac{p1}{2} = \frac{2p}{p+1}$,

when $0 < p < \sqrt[1/]{2}$, it is a relevant ideal selection range of randomization parameter, and better in privacy.

CONCLUSION

In this paper, we proposed a new method of randomized response -SRRM. Then for the data which handled after SRRM, by giving a simple and highly efficient algorithm to create frequent item sets, finally to realize a new mining methods of updated associated rule of privacy preserving. We have also analyzed the SRRM way to choose the randomized parameter to strengthen the data's privacy. In brief, the privacy preserving of sharing on ubiquitous environment will be one of the hotspots and focal points of Internet of Things industry's development and study, but whether in one of the theoretical level or in the technical level both of them have many problems, which need further investigation and discussion. We hope that more and more effective privacy - preserving data mining algorithms will be proposed and they will play an important role in the application of the IOT and seamless connectivity in the future.

REFERENCES

Lai, T., Li, W., Liang H, and Zhou, X. 2008. ***FRASCS: A Framework supporting context sharing Young Computer Scientists***. In: The 9th International Conferences for ICYCS 2008, November 18-21, p. 919-924.

Langheinrich, M. 2002. ***A privacy awareness system for Ubiquitous computing Environments.*** In: Borriello, G., and Holmquist, L.E. (eds). UbiComp 2002. LCNS, vol. 2498, p. 237-245. Springer, Heildelberf (2002).

Warner, S. L. 1965. Randomized response: A Survey technique for eliminating evasive answer bias. ***Journal of the American Statistical Association***. 60, 63-69.

Dong, A. 2007. ***Privacy-preserving Associatio Rules Mining***. DaLian: Dalian Jiaotong University.

Agrawal, R., and Srikant, R. 2000. ***Privacy preserving data mining***. In: Weidong, C., and Jeffrey F. (eds) Proc of the ACM SIGMOD Conf. on Management of Data, p. 439 - 450. Dallas: ACM Press.

"***Big data definition***" [online] Available: http://en.wikipedia.org/wiki/Big data

P. J. Sadalage, and M. Flower. 2012. ***NoSQL Distilled: A Brief guide to the Emerging world of Polyglot Persistence***, 1st ed. Addison-Wesley Professional.

L. Garber. 2013. Security privacy, policy, and dependability roundup. ***IEEE Security and Privacy***, vol. 11, No. 2, p. 6-7, March.

"***Ehr***" [Online] Available: http://www.himss.org/library/ ehr/

J. Sedayao, R. Bharadwaj, and N. Gorade. (June, 2014). "***Making big data, privacy, and anonymization work together in the enterprise: Experiences and issues***" In 2014 IEEE International Congress on Big Data. p. 601-607.

B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. (June 2010). "***Privacy preserving data publishing: A survey of recent developments***". ACM computing surveys, Vol. 42, No. 4, p. 14:1-14:53.

Retrieved from https://link.springer.com/content/pdf/10.1007/978-3-642-32427-7_67.pdf

Retrieved from https://link.springer.com/content/pdf/10.1007%2F978-3-642-32427-7_67.pdf

Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-32427-7_67

Retrieved from https://www.ijernd.com/manuscripts/v1i1/V1I1-1174.pdf

Retrieved from https://www.csee.umbc.edu/~kunliu1/research/privacy_review.html

Retrieved from https://epdf.tips/intermediate-statistics-and-econometrics-a-comparative-approach968f89dc0...

Retrieved from https://scinapse.io/authors/2765451417

Retrieved from https://www.crcpress.com/Introduction-to-Privacy-Preserving-Data-Publishing-Concepts-and-T...

Retrieved from https://www.ijcseonline.org/full_paper_view.php?paper_id=3224

Retrieved from https://www.ijcseonline.org/pub_paper/106-IJCSE-05332-12.pdf

Retrieved from https://www.researchgate.net/publication/234060246_Finding_the_True_Frequent_Itemsets

Retrieved from https://www.researchgate.net/publication/304108753_Governing_Internet_of_Things_Issues_app...

Retrieved from https://en.wikipedia.org/wiki/Probability

Retrieved from https://www.csee.umbc.edu/~kunliu1/research/

Retrieved from https://searchdatamanagement.techtarget.com/definition/big-data

Retrieved from http://shodhganga.inflibnet.ac.in/bitstream/10603/44185/11/11_chapter%203.pdf

Retrieved from https://www.researchgate.net/publication/224197797_Context_protecting_privacy_preservation...

Retrieved from http://www.ijsrd.com/articles/IJSRDV2I2230.pdf

Retrieved from https://blogs.oracle.com/oraclemagazine/having-sums%2c-averages%2c-and-other-grouped-data

Retrieved from https://en.wikipedia.org/wiki/Big_data

Retrieved from https://en.wikipedia.org/wiki/Social_media

Retrieved from https://www.sas.com/en_us/insights/articles/big-data/big-data-in-healthcare.html

Retrieved from https://blog.hootsuite.com/social-media-data/

Retrieved from https://www.freshersworld.com/jobs/companies

Retrieved from https://sproutsocial.com/insights/social-media-data/

Retrieved from https://www.researchgate.net/publication/322752340_Data_mining_using_Association_rule_base...

Retrieved from https://www.researchgate.net/publication/322752340_Data_mining_using_Association_rule_base...

Retrieved from https://en.wikipedia.org/wiki/Web_mining

Retrieved from https://www.researchgate.net/publication/4324149_Connections_between_Mining_Frequent_Items...

Retrieved from https://www.datapine.com/blog/big-data-examples-in-healthcare/

Retrieved from https://www.oktopost.com/blog/social-media-data/

Retrieved from https://www.datasciencecentral.com/profiles/blogs/what-is-big-data-and-why-should-you-care

Retrieved from https://www.rogerebert.com/reviews/a-simple-favor-2018

Retrieved from https://www.sdxcentral.com/sdn/network-virtualization/definitions/what-is-a-virtual-networ...

Retrieved from https://www.pdpjournals.com/docs/88016.pdf

Retrieved from https://intellipaat.com/blog/7-big-

Retrieved from https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life/

*****