International Journal of Enhanced Research in Science Technology & Engineering, ISSN: 2319-7463
Vol. 3 Issue 6, June-2014, pp: (65-71), Impact Factor: 1.252, Available online at: www.erpublications.com

An Extended Frequent Pattern Mining for Privacy Preserving transactional data set using Probabilistic Models

Vijay kumar[1], Dr. T. Christopher[2]
[1]Research Scholar, Department of Computer Science, Bharathiyar University, Coimbatore, Tamil Nadu, India
[2]Head, Assistant professor of computer science, Government Arts College, Udumalpet, Tamil Nadu, India

Abstract: Privacy preservation of transactional data set has more influence in business solutions where the transactional data set of any organization can be used by many others for different purposes. Hiding sensitive information about the customers of any organization becomes the responsibility of the organization which implicates the trustworthy of organization on customers. We propose a new extended frequent pattern mining technique for the preservation of sensitive information about the customers. The general support count based frequent pattern mining technique is used to identify the sensitive transactions and based on that a probabilistic solution is used to generate the sanitized data set. The sanitized data set preserves the originality of the transactional data and the user could compute much information for their purpose. The probabilistic model computes the probability of sensitive attributes and used to generate the sanitized data set. We used random selection of attribute in the sensitive transaction to generate sanitized data set.

Keywords: Knowledge Hiding, Sanitized Data, Transactional Data, Frequent Pattern Mining.

Introduction:

Transactional data set is a collection of rows which represents the purchase pattern of various customers of any organization and generated in different ways like through online transaction, personal purchase, super market , mobile commerce and etc… In the growing trends of world , the people expects more from the product manufactures and their taste , interest changes in day to day life. In business point of view, marketing a product in customer location has many points, and to find an useful knowledge about the purchase habit of any customer they need transactional data set. By generating an useful knowledge from the transactional data set, it could be used to form a new business view so that the marketing strategy could be increased. Sometimes few product manufacturers would like to offer another product of different manufacturer but may oscillate in selection of a product. In that situation they need a complete transactional set. But the data owner will not be ready to expose the complete data where there are personal information about their customers. This is where sanitization becomes necessary, so that the privacy preservation has to be done on the transactional data set before handover the data set to other organization. Knowledge hiding is the process of hiding sensitive information from large data set without tampering the originality of the data set. While sanitizing the originality of the data set has to be retained and also should be useful for other to make some inference from the data set published. Consider, for instance, the example of a large retail company which sells thousands of different products, and has numerous daily purchase transactions. The large amount of transactional data may contain customer spending patterns and trends that are essential for marketing and planning purposes.

	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test	Viagra
Maran	×		×			×
Stephen	×		×			
Rathna		×		×	×	
Abi		×	×			
Prabu	×		×	×		

Fig 1: Original Purchase data set

From the figure 1, there are various records of different customer purchase, and the attributes Pregnancy test, Viagra are the private and sensitive information. Consider the transaction of Maran and Rathna who purchased Viagra and Pregnancy card

has to be preserved. If we simply expose these data the user confidentiality will be lost and will reduce the organizations trustworthy. So that the organizations have to release the data set without disclosing this information also with originality. To overcome this there must be a suitable method which should not disclose the privacy items and personal information, so that an attacker could not be able to identify or infer others purchase and so on. Also in case of business point of view the people's identity should not be visible; otherwise the product goal will not reach all the people.
To overcome the difficulties present in earlier methods, we propose a new frequent pattern mining technique based privacy preservation method using a probabilistic model.

Background

There are many implementations for privacy preservation; here we discuss few of them. In Extended Method for Privacy Preserving Association Rule Mining [1], is discussed which uses association rule mining techniques to preserve the privacy information. They presented a new approach that necessarily changes few transactions in the transaction database by decreasing support or confidence of sensitive rules without any side effect. Privacy-preserving Distributed Mining of Association Rules on Horizontally Partitioned Data sharing the data [2], addresses secure mining of association rules over horizontally partitioned data. The methods incorporate cryptographic techniques to minimize the information shared, while adding little overhead to the mining task A Hybrid Approach to Frequent Itemset Hiding [8], proposes a novel, exact border-based approach that provides an optimal solution for the hiding of sensitive frequent item sets. First it extends the original database by a synthetically generated database part - the database extension, second using binary integer programming it solves the constraint satisfaction problem then finally it provides approximate solution. Extending the original database for sensitive item set hiding is proved to provide optimal solutions to an extended set of hiding problems compared to previous approaches and to provide solutions of higher quality.

In [9] they proposed a random perturbation method to prevent re identification of records, by adding noise to the data. By adding noise to the data it loses the originality and it becomes incomplete one. In [10], it is shown that an attacker could filter the random noise, and hence, breach data privacy, unless the noise is correlated with the data. However, randomly perturbed data are not "truthful" [11] in the sense that it contains records which do not exist in the original data. Furthermore, random perturbation may expose privacy of outliers when an attacker has access to external knowledge.
Using quasi identifier to the records used in many methods, but in that case the original record can be matched with the few attributes and can be identified. To overcome this difficulty Samarthi in [12] introduced k-anonymity, a privacy-preserving paradigm which requires each record to be indistinguishable among at least k _ 1 other records with respect to the set of QID attributes. Records with identical QID values form an anonymized group. K-anonymity can be achieved through generalization, which maps detailed attribute values to value ranges, and suppression, which removes certain attribute values or records from the micro data.

LeFevre in [11], proposed optimal k-anonymity solutions for single-dimensional recoding, in that they used independent mapping for each attribute and in [4], he proposed a multidimensional recoding algorithm which maps the Cartesian product of multiple attributes. Mondrian outperforms optimal single-dimensional solutions, due to its increased flexibility in forming anonymized groups. Methods discussed so far perform global recoding, where a particular detailed value is always mapped to the same generalized value. Sensitive knowledge hiding application [13], present a single unified platform for the sanitization of the original data set. Where in [14] by integrating the advantages of both these techniques with the view of minimizing information loss and privacy loss. By making use of cryptographic techniques to store sensitive data and providing access to the stored data based on an individual's role, we ensure that the data is safe from privacy breaches. The trade-off between data utility and data safety of our proposed method will be assessed. MotifHider: A knowledge hiding approach to sequence masking [15], introduces a novel approach and a tool, called MotifHider, to sequence masking problem. MotifHider exploits sensitive knowledge hiding principles from database sharing. By hiding certain patterns, it provides successive motif discovery programs to avoid false discovery and rediscovery. At the same time, it avoids overly distortion of the input sequences so as to retain most of the authentic motifs. To overcome the problem of diversity and anonymity exists in all the previously discussed approaches we propose a new frequent pattern based preservation method using probabilistic model.

Proposed Method

The proposed method computes general support, count method to identify the frequent items and based on that frequent pattern is generated. For the generated frequent pattern support value is computed, whatever the pattern has support value greater than threshold is selected for sanitization. The probability is computed for the selected patterns and the probability is computed for the sensitive items.

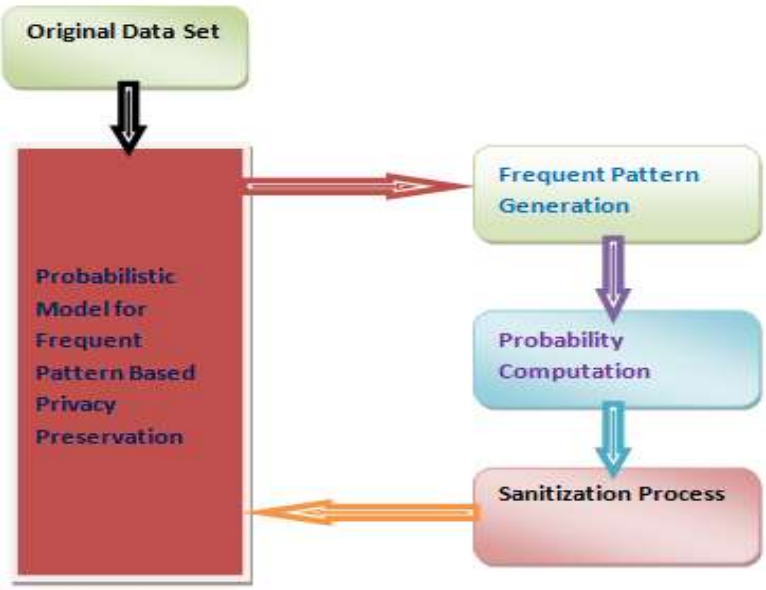


Fig 2: Proposed System Architecture.

A. Frequent Pattern Generation

Given a transactional data set T_s , the number of attributes N_a is identified and total number of transaction TN is computed. With the identified number of attributes, possible patterns set P_s is computed. Unlike generalized frequent pattern mining algorithm the support and count values are computed for the identified patterns in the pattern set P_s. Using computed support and count values the pattern P_i which has support value greater than support threshold S_t will be selected for sanitization process and other will be omitted.

Algorithm:

Step1: start
Step2: read data set T_s.
Step3: identify attribute set A_s.
Step4: compute number of transaction TN.
Step5: compute combination of possible patterns P_s = ((TN X §)/(K X(TN – K) X §)).
k- Specific pattern.
Step6: for each pattern P_i from P_s
Compute count = ø(P_i£(P_s)).
Ø- Number of pattern p_i contained in p_s.
Compute support = count/TN.
If support > ST then
Add P_i to Sanitization set S_s.
End.
End
Step7: return S_s.
Step8:End.
Computed sanitization pattern set will be returned for further processing and it contains the set of possible pattern set which can present in the publishing data set.

B. Probability Computation

From the original data set , sensitive items set SI_s , is identified at first stage. For each sensitive item SI_i , we compute the occurrence probability Ocp. The computed probability value will be used in publishing the sanitized data set. The sanitized data set will be merged with the computed probability value.

Algorithm:
Step1: start
Step2: read input data set T_s, initialize OCP.
Step3: identify sensitive items SI_s.
Step4: for each attribute SI_a from SI_s
OCP_a = (count of SI_a)/total number of records in T_s.
OCP(i) = OCP_a.
End
Step5: return OCP.
Step6: stop.

C. Sanitization Process

Generated frequent sanitization pattern set S_s is used for the sanitization process. From the original data set the records with the each pattern from S_s is selected for sanitization. The attributes present in the sensitive set SI_s is identified and merged to represent a single value. The sensitive attributes in the merged set M_s is replaced with the probability value OCP computed in the earlier phase. The final matrix or the merged set M_s is published for other usage.

Algorithm:
Step1: start
Step2: read original data set T_s.
Step3: read sanitization set S_s generated from frequent pattern process.
Step4: read sensitive item set SI.
Step5: for each pattern from S_s
Read record and add to M_s.
$M_{s(i)} = T_s(S_{s(i)})$.
End
Step6: for each attribute a from SI
Replace and merge values of M(i)(a).
Replace M(i)(a) = OCP(i).
End.
Step7: publish data set M.
Step8: end.

Results and Discussion

The proposed method generates good results compare to all other algorithms discussed earlier in this area. We used frequent pattern mining technique and probabilistic based privacy preservation method. The frequent pattern is identified in different way than general method. We compute support count values for set of combinations of patterns could be formed from the original data set. The proposed method retained the originality and also it preserves the private information. At this stage even if we declare the names with the sanitized data set the user cannot identify which record belongs to one and it is not possible to identify the purchase pattern of the user.

Names	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test
Siva	1	1	1	1	0
radha	0	1	1	1	1
Ram	1	1	0	1	0
Rajes	1	1	1	1	0
Saran	1	1	0	1	1

Names	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test
Kumar	1	1	1	1	0
Shela	1	1	0	0	0
Selva	1	1	0	0	0
Sivasankar	1	1	0	0	0
rohini	0	1	1	0	1

Table 1: shows the original data set

Pattern Type					Count/Support
Soap	Tooth Paste	Horlicks	Cream	Preg. Test	0/0.0
Soap	Tooth Paste				7/0.7
Soap	Tooth Paste	Horlicks			3/0.3
Soap	Tooth Paste	Horlicks	Cream		3/0.3
Soap	P.T	Horlicks	Cream		0/0.0
Soap	Tooth Paste	Horlicks	P.T		0/0.0
Soap	Tooth Paste	P.T	Cream		1/0.1
Tooth Paste	Horlicks	Cream	Preg. Test		1/0.1
Tooth Paste	Horlicks				5/0.5
Tooth Paste	Horlicks	Cream			5/0.4
Horlicks	Cream	Preg. Test			1/0.1
Horlicks	Cream				4/0.4
Cream	Preg. Test				2/0.4
Soap	Horlicks				3/0.3
Soap	Cream				5/0.5
Soap	P.T				1/0.1
Tooth Paste	Cream				6/0.6
Tooth Paste	P.T				3/0.3
Horlicks	Soap				3/0.3
Horlicks	Cream				4/0.4
Horlicks	P.T				2/0.4

Table 2 : shows the support count values for different pattern.

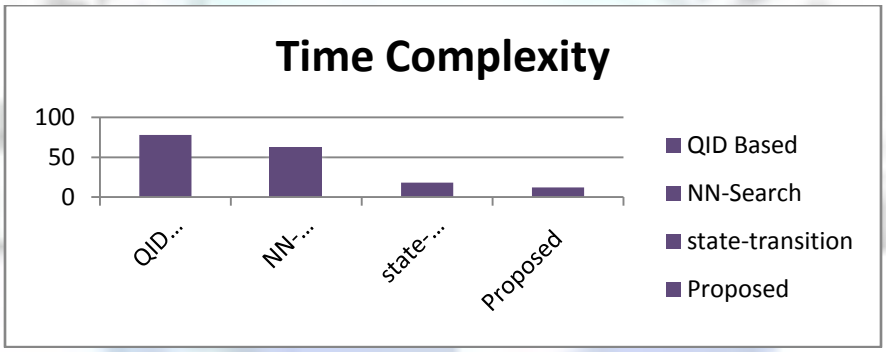
Pattern Type					Count/Support
Soap	Tooth Paste				7/0.7
Soap	Tooth Paste	Horlicks			3/0.3
Soap	Tooth Paste	Horlicks	Cream		3/0.3
Tooth Paste	Horlicks				5/0.5
Tooth Paste	Horlicks	Cream			5/0.4
Horlicks	Cream				4/0.4
Soap	Horlicks				3/0.3
Soap	Cream				5/0.5
Tooth Paste	Cream				6/0.6
Tooth Paste	P.T				3/0.3
Horlicks	Soap				3/0.3
Horlicks	Cream				4/0.4

Table 3: list of pattern set with count and support which is greater than threshold

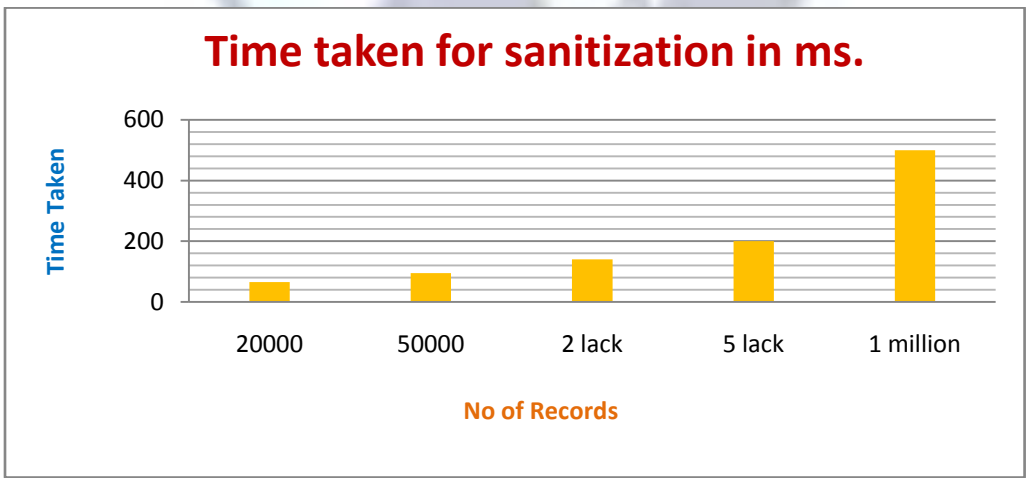
Pattern Type				Pregnancy Test	Count/Support
Soap	Tooth Paste			0.3	7/0.7
Soap	Tooth Paste	Horlicks			3/0.3
Soap	Tooth Paste	Horlicks	Cream		3/0.3
Tooth Paste	Horlicks				5/0.5
Tooth Paste	Horlicks	Cream			5/0.4
Horlicks	Cream				4/0.4
Soap	Horlicks				3/0.3
Soap	Cream				5/0.5
Tooth Paste	Cream				6/0.6
Tooth Paste	P.T				3/0.3
Horlicks	Soap				3/0.3
Horlicks	Cream				4/0.4

Table 4: shows the result of proposed system.

The table 5 shows the result of proposed system and it is clear that it shows the patterns and their count and support values, also it displays the value of pregnancy test.



Graph 1: shows the time complexity between other methods



Graph 2 : shows the overall time taken for sanitization process

Conclusion

We proposed a new frequent pattern mining based sanitization process for knowledge hiding and privacy preservation. The proposed method produces good results and it is efficient in time and space complexity. The data set published contains only the probability values not the original one, still the user can infer some knowledge but they cannot re generate the data set which is original.

References

[1]. Shikha Sharma, An Extended Method for Privacy Preserving Association Rule Mining, Volume International Journal of Advanced Research in Computer Science and Software Engineering, 2, Issue 10, October 2012.
[2]. M. Kantarcioglu and C. Clifton. Privacy-preserving distributed mining of association rules on horizontally partitioned data. In IEEE Transactions on Knowledge and Data Engineering Journal, volume 16(9), pages 1026–1037, Piscataway, NJ, USA, September 2004.
[3]. A. Gkoulalas-Divanis and V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," Proc. ACM Conf. Information and Knowledge Management (CIKM '06), pp. 748-757,
[4]. G.V. Moustakides and V.S. Verykios, personal comm., 2006.712 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009
[5]. Y.-H. Wu, C.-M. Chiang, and A.L.P. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 1, pp. 29-42, Jan. 2007.
[6]. X. Sun and P.S. Yu, "A Border-Based Approach for Hiding Sensitive Frequent Itemsets," Proc. Fifth IEEE Int'l Conf. Data Mining, pp. 426-433, 2005.
[7]. X. Sun and P.S. Yu, "Hiding Sensitive Frequent Itemsets by a Border-Based Approach," Computing Science and Eng., vol. 1, no. 1, pp. 74-94, 2007
[8]. Aris Gkoulalas-Divanis, Vassilios S. Verykios, "A Hybrid Approach to Frequent Itemset Hiding," ictai, vol. 1, pp.297-304, 19th IEEE International Conference on Tools with Artificial Intelligence - Vol.1 (ICTAI 2007), 2007
[9]. G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse, High-Dimensional Data," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.
[10]. A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-Anonymity," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2006.
[11]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," Proc. ACM SIGMOD, pp. 49-60, 2005.
[12]. K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," Proc. IEEE Int'l Conf. Data Eng.(ICDE), 2006.
[13]. Gokce, Sensitive knowledge hiding application , IEEE , Conference on Electrical, Electronics and Computer Engineering (ELECO), Page(s): 558 – 562, 2010.
[14]. Murugeswari.s, An Efficient Method for Knowledge Hiding Through Database Extension, IEEE conference on Recent Trends in Information, Telecommunication and Computing (ITC), Page(s): 342 – 344, 2010.
[15]. Abul O, MotifHider: A knowledge hiding approach to sequence masking , IEEE Conference on Computer and Information Sciences, pages 171-176 , 2009.