International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34251-34255

Min-max frequent pattern mining technique for privacy preservation in transactional data sets using support threshold

Vijaykumar
Research Scholar, Department of Computer Science,
Bharathiyar University, Coimbatore, Tamil Nadu, India.
Email: Vijayakumar123phd@gmail.com
Dr.T.Christopher
Head, Assistant professor of computer science,
Government Arts College, Udumalpet, Tamil Nadu, India.

Abstract- Privacy preserving the most dominant problem in knowledge sharing where the transactional information of organization is shared between organizations. Each organization has the responsibility to maintain the secrecy of user information or their transactional details while sharing information with other groups. There are many approaches has been discussed earlier for the preservation of user information, but struggles with maintaining the originality of data and data anonymity. We propose a navel approach called Max-Min frequent pattern mining, which identifies set of least frequent items using general pattern mining methods based on support threshold. Identified items are used as the base in identifying the other order frequent items and their patterns. The items with the least frequency and their appearance in the other item sets are computed for support and threshold values. Based on computed support threshold values, set of items are selected and identified as privacy items and are used for sanitization process. At the sanitization process the privacy items are sanitized using max-min values computed at the pattern mining process. The proposed sanitization approach has produces efficient result and the end user could obtain required information without loosing the originality.

Index Terms- Max-Min Values, Frequent Pattern, Privacy Preservation, Sanitization Process.

1. Introduction:

The business organization maintains much information about their customers which has personal details of customers also. Every organization has the responsibility to keep the secret information or personal information of their customer and cannot disclose them to others. Similarly the transactional data sets are the collection of records which contains information about the purchase pattern of their customers. The organizations use the purchase history or transactional log to generate business intelligence which could be used to make business decisions.

The organizations perform collaborative business at most situations and share the information about the customers of their own between them. For example, a car manufacturer may end up with a deal between a air conditioner manufacturing company. Before deciding or selecting the air conditioner manufacturer what the organization will do is, they perform a survey or generate a strategic report about how many car holders of that company has installed the air conditioner of that "X'. So the car manufacturer has their own data set and the air conditioner organization has their own transactional data set. Upon deciding a deal about reducing the price on giving an air conditioner as an offer with the car the AC organization will perform the business intelligence to identify the market of their product with the particular brand of car. Similarly both the organizations will perform the analysis to come to a deal of negotiation.

Now the organizations will share their data set between them to come to a conclusion between them, but they have the responsibility to protect the sensitive information about their customers. Similar to that in a market purchase data set, a customer may purchase many things using his credit card of a bank or anything. But the shopping agency has the responsibility to keep the sensitive purchase patterns and information without disclosing. This process is called privacy preservation and has many applications in various areas of real world.

Frequent patterns are the representation of user purchase method and the items purchased in a shopping cart. Every user has different purchase habits and whatever he purchases is generated as a pattern where each attribute of the pattern specifies whether he purchase the item or not. Each purchase sequence is called as a pattern and the same might be occurred for many times which we call frequency or frequent pattern. The frequent pattern can be used for many purposes in identifying the interested product of many user, so that the particular product can be focused well in market strategies.

Min-Max values specifies that there may be an item which has less support value in one item set but has many occurrences with other item sets which is greater than one. General apriori methods ignores this and other approaches also ignores this in identifying the privacy items. By computing min-max values , we can identify the privacy items which has low support in a single item set but has more frequency in other item sets.

2. RELATED WORKS:

There are many approaches has been discussed for preserving private information of users from transactional data set. We discuss few among them here for understanding purpose around the problem.

An Efficient Method for Knowledge Hiding Through Database Extension, propose a new solution 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.

A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, proposes a solution to this problem by managing unstructured data in to structured data using legacy system and distributed data partitioned method for gives distributed data for mining multi text documents. This frame work gives the testing of the similarities among text documents and privacy preserving meta data hiding technique, which are explored in text mining.

A Fuzzy Approach for Privacy Preserving in Data Mining addresses the problem of Privacy Preserving in Data Mining by transforming the attributes to fuzzy attributes. Due to fuzzification, exact value cannot be predicted thus maintaining individual privacy, and also better accuracy of mining results were achieved.

Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R. H. S. Items, propose two algorithms, ADSRRC (Advanced Decrease Support of R. H. S. items of Rule Cluster) and RRLR (Remove and Reinsert L. H. S. of Rule), for hiding sensitive association rules. Both algorithms are developed to overcome limitations of existing rule hiding algorithm DSRRC (Decrease Support of R. H. S. items of Rule Cluster). Algorithm ADSRRC overcomes limitation of multiple sorting in database as well as it selects transaction to be modified based on different criteria than DSRRC algorithm. Algorithm RRLR overcomes limitation of hiding rules having multiple R. H. S. items. Experimental results show that both proposed algorithms outperform DSRRC in terms of side effects generated and data quality in most cases.

Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining, present a novel hiding-missing-artificial utility (HMAU) algorithm is proposed to hide sensitive itemsets through transaction deletion. The transaction with the maximal ratio of sensitive to nonsensitive one is thus selected to be entirely deleted. Three side effects of hiding failures, missing itemsets, and artificial itemsets are considered to evaluate whether the transactions are required to be deleted for hiding sensitive itemsets. Three weights are also assigned as the importance to three factors, which can be set according to the requirement of users.

Hiding Sensitive Association Rules without Altering the Support of Sensitive Item, uses the data distortion technique where the position of the sensitive items is altered but its support is never changed. The size of the database remains the same. It uses the idea of representative rules to prune the rules first and then hides the sensitive rules. Advantage of this approach is that it hides maximum number of rules however, the existing approaches fail to hide all the desired rules, which are supposed to be hidden in minimum number of passes.

An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation, presents an index-based algorithm named SSAPP for exploring frequent sequential patterns in a distributed environment with privacy preservation. The SSAPP algorithm uses an equivalent form of a sequential pattern to reduce the number of cryptographic operations, such as decryption and encryption. In order to improve the efficiency of sequential pattern mining, the SSAPP algorithm keeps track of patterns in a tree data structure called SS-Tree. This tree is used to compress and represent sequences from a sequence database. Moreover, a SS-Tree allows one to obtain frequent sequential patterns without generation of candidate sequences.

A Novel Community Detection Algorithm for Privacy Preservation in Social Networks, presents a novel method for community detection with the assumption of privacy preservation is proposed. In the proposed approach is like hierarchical clustering, nodes are divided alliteratively based on learning automata (LA). A set of LA can find min-cut of a graph as two communities for each iteration. Simulation results on standard datasets of social network have revealed a relative improvement in comparison with alternative methods.

Identity-Based Privacy Preservation Framework over u-Healthcare System, proposes an identity-based privacy preservation framework over u-healthcare systems. Our framework is based on the concepts of identity-based cryptography and non-interactive key agreement scheme using bilinear pairing. The proposed framework achieves authentication, patient anonymity, un-traceability, patient data privacy and session key secrecy, and resistance against impersonation and replay attacks.

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 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.

All the above discussed method has the problem of identifying the privacy items from the large transactional data sets. To solve this problem, we propose a min-max frequent pattern approach for privacy preservation which will be explained in detail in the next section.

3. Proposed Method:

We proposed a Min-Max pattern mining approach for privacy preservation which has four stages namely Single Item Set Generation, Frequent Pattern Generation, Min Max Computation and Sanitization process. We discuss each stage in this section in detail.

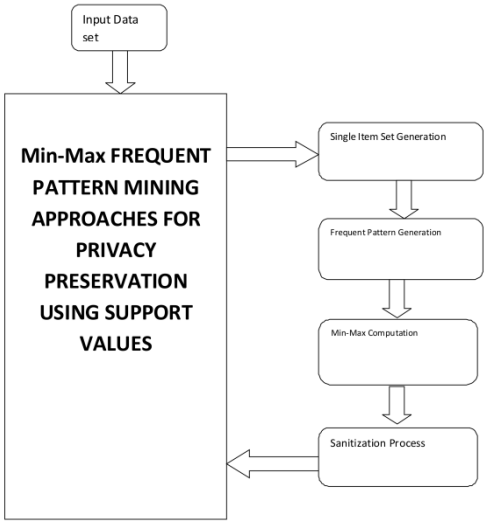


Fig 1: Proposed System Architecture.

3.1 Single Item Set Generation:

The single item set is generated for each item present in the transactional data set. For each single item we compute the support and count values which shows the frequency of purchase which will be used to compute min max values. From computed support count values, the item's which has less support than the support threshold are identified and selected for identifying the privacy item.

Algorithm:

Input: Transactional Data set Ts.

Output: Single Items Sets IS.

Step1: Identify distinct items in the transactional data sets.

DI= $\int \sum Item(Ts) \nexists NOI$

Step2: Compute number of distinct items

Noi =size(DI)

Step3: for each item I from DI

Compute Count = $\int_{i=1}^{size(Ts)} \sum Ti(I) == 1$

Compute support = Count/size(Ts).

If support<STh //support threshold

IS = $\sum Ti(IS) + I$

End

Step4: End

3.2 Pattern Generation:

The frequent pattern is generated on a given transactional data set T_s , where N specifies the number of attributes and TN specifies the total number of transactions available. Initially the number of attributes which forms the whole transaction set is identified and we generate combinatory of patterns set Ps. The combinatory of pattern set is computed according to the possible combinations which can be formed.

Algorithm:

Input: Data set Ts.

Output: Pattern set OPs.

step1: Compute Total number of transactions Tn =$\sum Ts$

Step2: Identify set of attributes Ats = $\int_1^{Tn} \sum Attr \nexists Ats$

Step3: Compute possible patterns Ps = $\int_1^N \forall (Tsi) \nexists Ps$

Step: stop

3.3 Min-Max Matrix Computation:

From generated single item set and pattern set , we generate the Min Max matrix using which the privacy items can be identified. For each single item I from SI , we compute the support and count values of each pattern where the item is present. For each items support and count values are stored in the min max matrix. Once the support values of each item has been computed then it will be used for privacy preservation. If an item present in the single item set but does not clear the multi item set threshold then it is considered as the private item. Identified privacy items are stored and used for generating the sanitized data set.

Algorithm:

Input: Single Item set SI, Pattern set Ops.

Output: Privacy items Pi.

Step1: initialize PI to null.

Step2: for each item I from SI

Identify the patterns where I is present.

PS = $\sum OPS(i) \in I$

For each pattern Pi from PS

Compute count of Pi = $\sum TS(i) == Pi$

End

Compute support = Count/size(Ts).

Add to MinMax matrix MM = $\sum MM(i) + Support$

If count>($\frac{1}{8} \times size(Ts)$) then

Add I to privacy item set PI = $\sum PI(i) + I$.

End.

End.

Step3: stop.

3.4 Sanitization Process:

The computed min max matrix and privacy item set is used to generate sanitized data set. From available min max values are used as the attribute value for the specific privacy item. Once the item which is sensitive is identified then the value at the sensitive item is represented with a min-max matrix values and used for data publishing. This shows the data set as original and the end user can infer required knowledge from the sanitized data set.

Algorithm:

Input: Transaction data set Ts, Min-Max minmax, Privacy Item PI

Output: Sanitized data set STs.

step1: for each attribute A_i from Ts

if $PI \in Ai\ Then$

Ts = $\int Ts(AI) = MinMax(Ai)$

end

end.

Step2: stop.

4. Results and Discussion:

The proposed method has produced efficient results in sanitization and produced good results. The proposed method generates single item set pattern using which it identifies the items which has less support value than the allowed threshold. 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.

Table1: shows the original data set

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
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

The table 2, shows the single item set pattern and their support values. If the threshold is 30% then the only item get selected in the single item set will be "Pregnancy Test". This will be add to the single item set and will be used to compute the sanitized data set.

Table2: shows the single item set and support value

Item Name	Support Value
Soap	0.85
Tooth Paste	1.0
Horlicks	0.42
Cream	0.42
Pregnancy Test	0.28

From table 3, if we count the total count value it becomes 11 and the support value becomes 15.85 which is less than the threshold of 30 so that the item is selected as the privacy sensitive item.

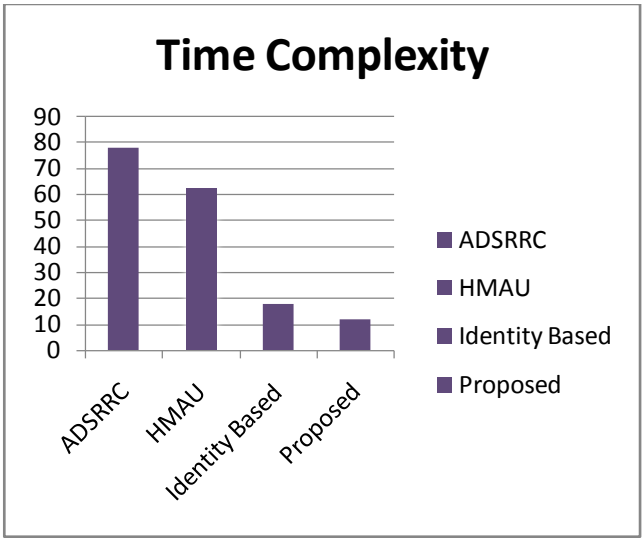
Table3: shows the support count values for different pattern with single item.

Pattern Type				Count/Support
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
Horlicks	Cream	Preg. Test		1/0.1
Cream	Preg. Test			2/0.4
Soap	P.T			1/0.1
Tooth Paste	P.T			3/0.3
Horlicks	P.T			2/0.4

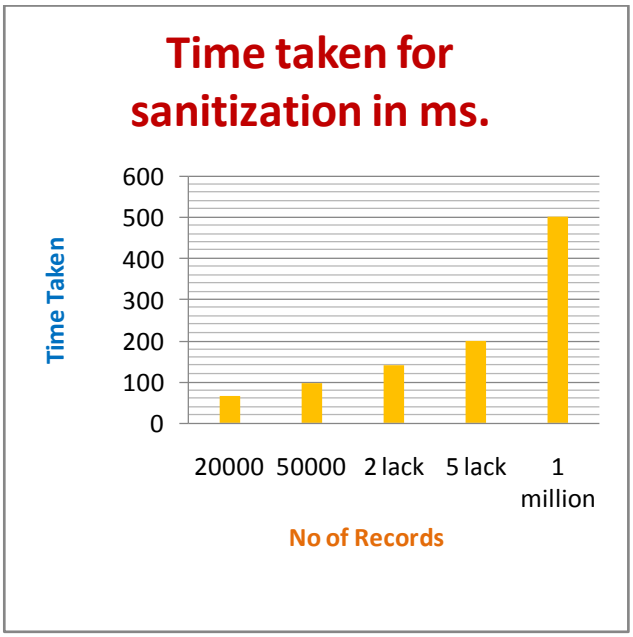
The table 4, shows the result produced by the proposed method and the pregnancy test has been hidden with the values of min max matrix and been published.

Table 4: Result of proposed system

Names	Soap	Tooth Paste	Horlicks	Cream	Pregnancy Test
Siva	1	1	1	1	11/15.0
radha	0	1	1	1	
Ram	1	1	0	1	
Rajes	1	1	1	1	
Saran	1	1	0	1	
Kumar	1	1	1	1	
Shela	1	1	0	0	
Selva	1	1	0	0	
Sivasankar	1	1	0	0	
rohini	0	1	1	0	



Graph1: shows the time complexity between other methods.



Graph2 : shows the overall time taken for sanitization process .

5. Conclusion

We proposed a new sanitization approach for data publishing with privacy preservation based on Min Max pattern generation technique. The proposed method computes single item set and for each item from the single item set we compute the support and count value. The items with less support threshold is added to the single item set and the other item set patterns are generated. For the other item set patterns , we compute the support threshold where the pattern has the single item. With the computed values a set of item is identified which has less support values. The identified item is specified with the values present in the min max matrix. The sanitized data set maintains the originality of the data and the end user can infer any information from that. The proposed method has produced efficient results with less time complexity.

References:

[1]. 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.

[2]. V. Thavavel, A generalized Framework of Privacy Preservation in Distributed Data mining for Unstructured Data Environment, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, January 2012

[3]. M Sridhar and Raveendra B Babu. A Fuzzy Approach for Privacy Preserving in Data Mining. International Journal of Computer Applications 57(18):1-5, November 2012.

[4]. Komal Shah, Amit Thakkar and Amit Ganatra. Article: Association Rule Hiding by Heuristic Approach to Reduce Side Effects and Hide Multiple R.H.S. Items. International Journal of Computer Applications 45(1):1-7, May 2012.

[5]. Chun Wei Lin, Reducing Side Effects of Hiding Sensitive Itemsets in Privacy Preserving Data Mining, The Scientific World Journal Volume 2014 (2014).

[6]. Dhyanendra Jain , Hiding Sensitive Association Rules without Altering the Support of Sensitive Item, International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[7]. Marcin Gorawski, An Efficient Algorithm for Sequential Pattern Mining with Privacy Preservation, Advances in Systems Science Advances in Intelligent Systems and Computing Volume 240, 2014, pp 151-161.

[8]. Fatemeh Amiri, A Novel Community Detection Algorithm for Privacy Preservation in Social Networks, Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, 2013, pp 443-450.

[9]. Kambombo Mtonga, Identity-Based Privacy Preservation Framework over u-Healthcare System, Multimedia and Ubiquitous Engineering Lecture Notes in Electrical Engineering Volume 240, 2013, pp 203-210.

[10] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, “Mondrian Multidimensional k-Anonymity,” Proc. IEEE Int’l Conf. Data Eng.(ICDE), 2006.

[11] Fatemeh Amiri, A Novel Community Detection Algorithm for Privacy Preservation in Social Networks, Intelligent Informatics Advances in Intelligent Systems and Computing Volume 182, 2013, pp 443-450.

[12] Kambombo Mtonga, Identity-Based Privacy Preservation Framework over u-Healthcare System, Multimedia and Ubiquitous Engineering Lecture Notes in Electrical Engineering Volume 240, 2013, pp 203-210.

[13] Abul O, MotifHider: A knowledge hiding approach to sequence masking , IEEE Conference on Computer and Information Sciences, pages 171-176 ,2009.

[14] 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