Frequent Itemsets Mining with Differential Privacy: A Survey

P. Suganthi Malarvizhi[1]**, D.Poorani**[2]

[1]AP (Senior Grade), Department of Computer Science & Engineering, VCET, TN, India. Email: suganthivenu97@gmail.com.
[2]Department of Computer Science & Engineering, VCET, TN, India. Email: paripoorani2@gmail.com

Abstract - Ecommerce-oriented Data mining is a very promising area. It can automatically predict trends in customer spending, market trends which guide company to build personalized business intelligence web site, bring huge business profits. Association rule learning is one of popular and well researched method for discovering interesting relations between variables in large databases in data mining. A systolic tree algorithm is several times faster than the implementation of the FP-growth algorithm and UP-growth algorithm to implement in large dataset. We propose a system, designed to perform weighted rule mining for transaction dataset. In this system, automatic weight estimation scheme is used. Each item is assigned a weight value with reference to the request count and sequence.

Keywords - Frequent itemset, Systolic tree, Association rule mining, FP-tree, Pattern mining.

I. INTRODUCTION

A. Introduction to Data Mining

Data Mining is the process of discovering hidden patterns in the large data set and establishes relationships to solve problems. Practically, the large pre-existing databases are examined to generate new information, for which it is referred as “Knowledge Discovery in Databases”. Data mining comprises three intertwined scientific disciplines: Statistics, Machine learning and Artificial Intelligence. Statistics is the numerical study of data relationships. Artificial intelligence is the human-like intelligence displayed by software and/or machines. Machine learning has algorithms that can learn from data to make predictions. Over the last decades, we have advanced in processing power and speed; thus, enabling us to move beyond manual, tedious and time consuming works to quick, easy and automated data analysis. More the data is collected; more the work needed to uncover relevant insights. So, Data mining is used to analyze those big data’s and predict meaningful relationship.

The first in a data mining is describing the data that summarizes its statistical attributes, visually reviewing by using charts and graphs, and look for potentially meaningful links among variables. The Collection, exploration and selection of the right data are critically important. But data description alone can never provide an action plan. To overcome, a predictive model based on patterns, which are determined from known results is built, and then the model is tested outside the original sample result. A good model must not be confused with reality just as a road map is not an actual representation of the actual road, but it can be a useful guide to understanding the business.

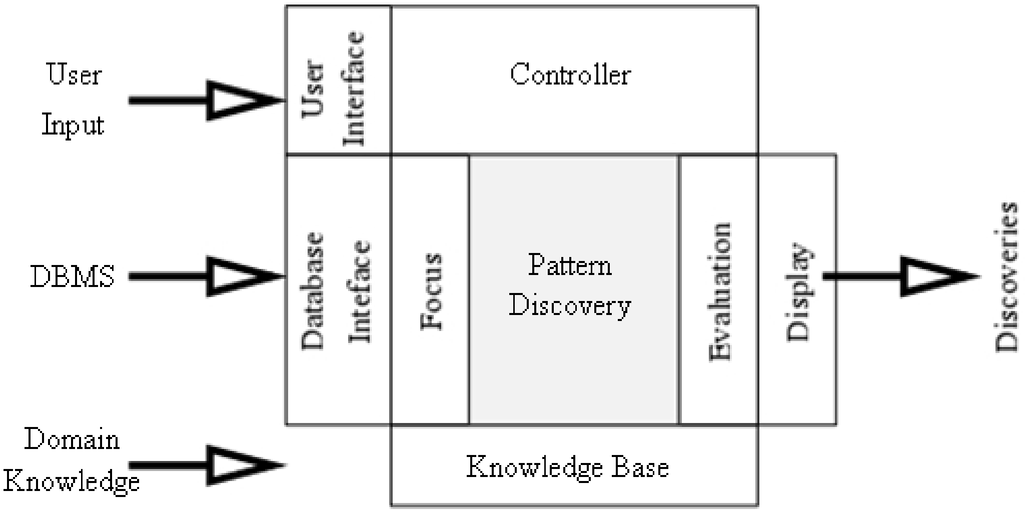

Fig. 1 Mining data from a database

The last step is to empirically verify the model. For example, from the database of customers that already responded to a particular offer, a model is built predicting which prospects are likeliest to respond to the same offer. What can be completely relied on this prediction? Send a mail requesting portion of the new list and see what results you get.

B. Data Mining & Warehousing

The data to be mined is first extracted from a data warehouse into a data mining database or data mart (Figure 1.2). If the data is already in a data warehouse then it is a real benefit. As it is seen later, there is a similarity in the problem of cleansing data for a data warehouse and for data mining. If the data has already been cleansed for a data warehouse, then it will not need further cleaning in order to be mined. Furthermore, many problems of data consolidation are already addressed and will be placed in maintenance procedures.

The data mining database will be logical rather than the physical subset of the data warehouse, provided that it can support the additional resource demands of data mining. If it cannot, then it will be better to do with a separate data mining database.

A warehouse is not a must for data mining. A large data warehouse can be set to consolidate data from multiple sources, resolves data integrity problems, and then loaded into a query database which is a tremendous task, and sometimes takes years and costing millions of dollars. However, it is possible that data can be mined from one or more transactional databases by simply extracting it into a read-only database (Figure 1.3). This new database then functions as a type of data mart.

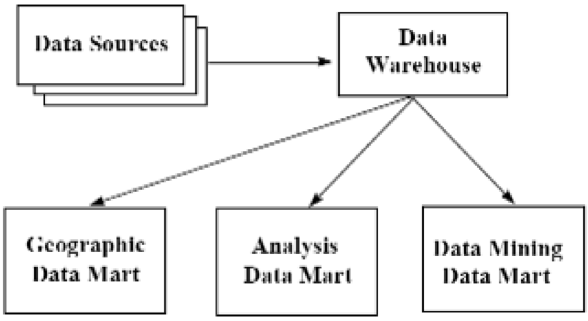

Fig. 2 Data mining data mart extracted from a data warehouse

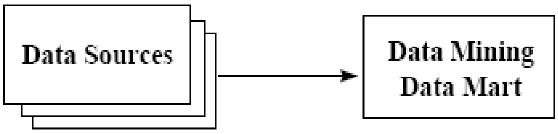

Fig. 3 Data mining data mart extracted from operational database

C. Data Mining Applications

Data mining is popular because of the unordinary contribution it can effect. It can be used for controlling costs and contributing for revenue to increase. Most organizations use data mining to help administer every phase of customer life cycle, also brining in new customers, revenues to be increased from existing customers, and accommodating valuable customers. By determining different aspects of good customers, a company will target possibility with similar aspects. By marking customers that have brought a distinct product can focus attention on similar customers who did not brought the same product. By marking customers who have exited, a company must act to keep the customers that are at risk for exiting, because it is less expensive to prolong a customer than to bring new customers. Data mining offers profit across an expansive spectrum of industries. Telecommunication companies and credit card companies are two leaders for using data mining to detect fraudulent use of their services. Insurance companies and stock exchange companies are also attentive in applying same technology to reduce fraudulent activities. A medical application is another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Companies active in the financial markets make use of data mining to resolve market and industry aspect as well as to anticipate individual company and stock performance. Retailers achieve more use of data mining to determine which product to be stocked in distinct stores, and also to be assessed for the development of promotions and coupons. Pharmaceutical firm mines large database of chemical compounds and genetic material to recognize substances that might be applicant for developing as agents for the treatments of disease.

D. Data Mining Techniques

Data mining techniques are association, clustering, classification, prediction, sequential patterns and decision tree. Association technique is called relation technique because it discovers patterns based on a relationship between items in the same transaction. Classification is based on machine learning and is used to classify items in a set of data into a predefined set of classes or groups. Unlike classification, clustering technique defines the classes and puts objects in each class.

Prediction, as the name implies, discovers the relationship between independent variables and relationship between dependent and independent variables. Sequential patterns analysis mainly seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period. Decision tree technique is most commonly used data mining technique because it makes the users to understand easily.

Discovering useful patterns hidden in databases plays a critical role in many data mining tasks, such as frequent pattern mining, weighted frequent pattern mining, and high utility pattern mining. Association rule mining is a procedure used to find frequent patterns, correlations, associations or casual structures from data sets can be found in various kinds of databases like relational databases, transactional databases, and other forms of data repositories.

Association rule mining is a rule based machine learning for discovering interesting relations between variables in large databases. It intends to identify strong rules discovered in databases using some measure of interestingness. This rule-based method also generates new rules as it analyses more data.

In order to select interesting rules from set of all possible rules, it uses constraints on various measures of significance and interest. The best known constraints are support and confidence. Support is an indication constraint of how frequently the itemset appears in the dataset. Confidence constraint is an indication of how often the rules have been found to be true. Algorithms are proposed for generating association rules. Few well known algorithms used are Apriori algorithm, Eclat and FP-Growth algorithm.

II. LITERATURE SURVEY

A. Robust Semi-Supervised Subspace Clustering via Non-Negative Low-Rank Representation

In LRR-based semi-supervised subspace clustering methods, the label information is not used to guide the affinity matrix construction so that the affinity matrix cannot deliver strong discriminate information. However, these methods cannot guarantee an overall optimum since the affinity matrix construction and subspace clustering are often independent steps. So, a robust semi-supervised subspace clustering method based on non-negative LRR (NNLRR) to address these problems. By bringing together the LRR framework and the Gaussian fields and harmonic functions method in a single optimization problem, the supervision information is especially incorporated to guide the affinity matrix construction and subspace clustering are practiced in one step to guarantee the overall optimum. The affinity matrix is obtained by exploring a non-negative low-rank matrix that represents each sample as a linear combination of others. The sparse constraint on the affinity matrix is imposed explicitly such that the affinity matrix obtained by NNLRR is non-negative low-rank and sparse.

B. Rule-based system to detect energy efficiency anomalies in smart buildings

The energy use of rapidly growing world already has concerns over the exhaustion of energy resources and heavy environmental impacts. As a result, a trend of green and smart cities has been increasing. To respond to this increasing trend of smart cities with buildings every time more complex, a new method is proposed to solve energy inefficiencies detection problem in smart buildings. This solution is based on the rule-based system developed through data mining techniques and applying the knowledge of energy efficiency experts. A set of useful energy efficiency indicators to detect anomalies is also proposed. The data mining system is developed over the knowledge extracted by a full set of building sensors. So, the results of this process provide a set of rules that are used as a part of a decision support system for the optimization of energy consumption and the detection of anomalies in smart buildings.

C. Robust echo state networks based on correntropy induced loss function

A method correntropy induced loss function (CLF) is proposed. CLF is robust to outliers over the process of correntropy which is extensively applied in information theoretic learning. The planned method improves the anti-noise capacity of echo state network and overcomes the problem of being sensitive outliers which are common in real-world tasks. The echo state network with CLF achieves the basic plan of echo state network, but changes the commonly used mean square error (MSE) criterion with CLF. The stochastic gradient descent method is used to better the objective function. The proposed method is next verified in nonlinear system identification and chaotic time-series prediction. The experimental result demonstrates that this method is robust to outliers and surpasses the echo state networks with Bayesian regression and Huber loss function.

D. Item recommendation using tag emotion in social cataloging services

A tag-based recommendation method considering users emotions in tags is proposed. The tag weight is based on the rating and the emotion value of a tag. The emotion value of the tag is got by using SenticNet. A High-Order Singular Value Decomposition is applied. The evaluation results that used emotion play an important role in recommendation. Due to the overburden of contents, the user experiences from difficulty in selecting items. The social cataloging services grant users to preoccupy items and share their opinions, which controls in not only thyself but other users to choose new items. The recommendation system diminishes the problem of the choice by advocating the items seeing the behavior of the people and the aspects of the items. A tag-based recommendation method considering the emotions reflected in the user’s tags is proposed. Since the item estimation by the user is made after exhaust the item, the feelings of the user obtained during exhaustion is directly reflected in ratings and tags. The rating has long-term valence on the item, and the tag characterizes the detailed feelings. It is assumed that the basic emotion of the tag attached to the item is the user rating for an item, and the unique emotion value of the tag adjusts the emotion of tag. The relationships between users, items, and tags as a three-order tensor is represented and tensor factorization is applied. The result of the experiment is that the proposed method achieves better recommendation performance than baselines.

E. Embedding cryptographic features in compressive sensing

Compressive sensing (CS) has been widely considered and applied in many fields. Recently, the way to perform secure compressive sensing (SCS) has become a topic of a growing interest. The existing works on SCS usually take the sensing matrix as a key and only considered as preliminary explorations on SCS. Firstly some possible encryption models for CS are proposed. These models provide a new point of view and stimulate further research in both CS and cryptography. It is demonstrated that random permutation is an acceptable permutation with overwhelming probability, which can effectively relax the Restricted Isometry Constant for parallel compressive sensing. Moreover, random permutation is applied to design a secure parallel compressive sensing scheme. Security analysis indicates that the proposed scheme can achieve the asymptotic spherical secrecy. The realization of chaos validates the feasibility of one of the proposed encryption models for CS. Results verify that the embedding random permutation based encryption enhances the compression performance and the scheme possesses high transmission robustness against additive white Gaussian noise and cropping attack.

F. Top-k frequent itemsets via differentially private fp-trees

Frequent itemset mining is a core data mining task and has been studied broadly. By nature, frequent itemsets are aggregates over many individuals and would not seem to pose a privacy threat; an attacker with strong background information can learn private individual information from frequent itemsets. This led to differentially private frequent itemset mining, which protects privacy by giving inexact answers. An approach that first identifies top-k frequent itemsets, and then uses it to construct a compact, differentially private FP-tree is introduced. Once the noisy FP-tree is built, the support of all frequent itemsets can be derived from it without accessing the original data. The result of experiment shows that the designed algorithm gives substantially higher levels of privacy than prior approaches.

G. Differentially private frequent itemset mining via transaction splitting

Designing algorithms for differentially private data mining has been a growing interest for a decade. One of the fundamental problems in data mining is Frequent Itemset Mining (FIM). The possibility of proposing a differentially private FIM algorithm which does not only offers high time efficiency, but alongside achieves high data utility and a high degree of privacy is explored. A differentially private FIM algorithm is proposed based on the FP-growth algorithm, which is referred to as PFP-growth. The PFP-growth algorithm comprises of a preprocessing phase and a mining phase. In the preprocessing phase, a novel smart splitting method is proposed to transform the database to improve the utility and privacy tradeoff. The preprocessing phase needs to be performed only once for the given database. To offset the information loss caused by transaction splitting in the mining phase and a run-time estimation method to estimate the actual support of itemsets in the original database is constructed. It leverages the downward closure property and also a dynamic reduction method to effectively reduce the amount of noise added to guarantee the privacy during mining process is used. It is shown that PFP-growth algorithm is ϵ-differentially private in formal privacy analysis. Extensive experiments on real datasets illustrate that the PFP-growth algorithm substantially outperforms the state-of-the-art techniques.

H. A two-phase algorithm for mining sequential patterns with differential privacy

Frequent sequential pattern mining is an important task in fields such as biology, finance and more. Moreover, release of these patterns is raising increasing concerns on individual privacy. The differential privacy framework which provides formal and provable guarantees of privacy is the sequential pattern mining problem which is explored. Due to the nature of the differential privacy process which agitates the frequency results with noise and the high dimensionality of the pattern space is challenging to mine. For mining both prefixes and substring patterns, a novel two-phase algorithm is proposed. This approach took advantage of the total properties of the data in the first phase to build a model-based prefix tree which helps to mine prefixes and a candidate set of substring patterns. The frequency of the substring patterns is farther clarified in the successive phase where a novel transformation of the original data to reduce the perturbation noise is engaged. The results of extensive experiment using real datasets shows this approach is effective for mining both substring and prefix patterns in comparison to the state-of-the art solutions.

I. Differentially private frequent sequence mining via sampling-based candidate pruning

Exploration on the problem of mining frequent sequences under the rigorous differential privacy model is done. The possibility of proposing a differentially private frequent sequence mining (FSM) algorithm that can achieve both high degree of privacy and a high data utility is explored. It is founded that in differentially private FSM, the amount of required noise is proportionate to the number of candidate sequences. If those unpromising candidate sequences are effectively pruned, the utility and privacy tradeoff can be seriously improved. By leveraging a sampling-based candidate pruning technique, a novel differentially private FSM algorithm, PFS2 is proposed. It supports the general gap-constrained FSM in the context of differential privacy and is the first algorithm to support. The gap constraints in FSM can be used to restrict the mining results to a dominated set of frequent sequences. The core is to utilize sample databases to prune the candidate sequences generated based on the downward closure property is done in PFS2 algorithm. In particular, the noisy local support of candidate sequences in the sample databases to estimate which candidate sequences are potentially frequent is used. To increase the accuracy of such private estimations, a gap-aware sequence shrinking method is proposed to carry out the length constraint on the sample databases. To calibrate the amount of noise required by differential privacy, a gap-aware sensitivity computation method is designed to obtain the sensitivity of the local support computations with different gap constraints. A threshold relaxation method is designed to relax the user-specified threshold for the sample databases, to decrease the probability of misestimating frequent sequences as infrequent.

Extensive experiments on real datasets illustrate that the PFS2 algorithm can privately find frequent sequences with high accuracy.

J. Mining frequent graph patterns with differential privacy

Discovering frequent graph patterns in a graph database offers valuable information for variety of applications. If the graph dataset contains sensitive data of individuals such as mobile phone call graphs and web-click graphs, releasing discovered frequent patterns may present a threat to the privacy of individuals. Differential privacy has recently emerged as the standard for private data analysis due to its provable privacy guarantee. The first differentially private algorithm for mining frequent graph patterns is proposed. It first shows that previous techniques on differentially private discovery of frequent itemsets cannot apply in mining frequent graph patterns due to the inherent complexity of handling structural information in graphs. This challenge is addressed by proposing a Markov Chain Monte Carlo (MCMC) sampling based algorithm. The techniques do not rely on the output of a non-private mining algorithm than previous methods. Instead, it is observed that both frequent graph pattern mining and the guarantee of differential privacy can be unified into an MCMC sampling framework. In addition, the privacy is established and utility guarantee of the algorithm and proposes an efficient neighboring pattern counting technique as well. The result of the experiment shows that the designed algorithm is able to output frequent patterns with good precision.

K. Differentially private frequent sub graph mining

Mining frequent sub graphs from a collection of input graphs is an important task in data mining research. Moreover, if the input graphs contain sensitive information, releasing frequent subgraphs may pose considerable threats to individual's privacy. It is studied that the problem of frequent subgraph mining (FGM) under the rigorous differential privacy model. A novel differentially private FGM algorithm is introduced which is referred to as DFG. First privately identify frequent subgraphs from input graphs, and then compute the noisy support of each identified frequent subgraph. In particular, to privately identify frequent subgraphs, frequent subgraph which identification approach which can improve the utility of frequent subgraph identifications through candidates pruning is recommended. To compute the noisy support of each identified frequent subgraph, a lattice-based noisy support derivation approach is devised, where a series of methods has been proposed to improve the accuracy of the noisy supports. Through formal privacy analysis, it is proved that the DFG algorithm satisfies ε-differential privacy. Results of extensive experiment on real datasets show that the DFG algorithm can privately find frequent subgraphs with high data utility.

III. CONCLUSION

When sensitive data – data that risks the privacy of the people it describes – needs to be modeled, decision trees lend themselves well to the strict requirements of differential privacy. In this research a novel differentially private algorithm for frequent item sets mining. The algorithm features better data utility and better computation efficiency. On the basis of our analysis we proposed efficient algorithm was then proposed. The experimental results show that the execution result is reduced much due to the considerable reduction in number of records on high F-Score and low relative error. There are many approaches to improve the efficiency of the algorithm. The efficiency would improve to a greater extent if the number of unwanted item sets is reduced. The efficiency of the proposed algorithm reduces if the number of items and its support increases with the number of transaction being constant.

REFERENCES

[1] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, “Robust semi-supervised subspace clustering via non-negative low-rank representation”, IEEE Trans. Cybern., vol. 46, no. 8, pp. 1828-1838, Aug. 2016.

[2] Pena, F. Biscarri, J. I. Guerrero, I. Monedero, and C. León, “Rule-based system to detect energy efficiency anomalies in smart buildings, a data mining approach”, Expert Syst. Appl., vol. 56, pp. 242-255, Sep. 2016.

[3] Y. Guo, F. Wang, B. Chen, and J. Xin, “Robust echo state networks based on correntropy induced loss function”, Neurocomputing, vol. 267, pp. 295-303, Dec. 2017.

[4] H. Lim and H.-J. Kim, “Item recommendation using tag emotion in social cataloging services”, Expert Syst. Appl., vol. 89, pp. 179-187, Dec. 2017.

[5] Y. Zhang et al., “Embedding cryptographic features in compressive sensing,” Neurocomputing, vol. 205, pp. 472-480, Sep. 2016.

[6] J. Lee and C.W. Clifton, “Top-k frequent itemsets via differentially private fp-trees”, in Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 931-940, 2014.

[7] S. Su, S. Xu, X. Cheng, Z. Li, and F. Yang, “Differentially private frequent itemset mining via transaction splitting”, IEEE Trans. Knowl. Data Eng., Vol. 27, No. 7, pp. 1875-1891, Jul. 2015.

[8] Bonomi and L. Xiong, “A two-phase algorithm for mining sequential patterns with differential privacy”, in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage., pp. 269-278, 2013.

[9] S. Xu, S. Su, X. Cheng, Z. Li, and L. Xiong, “Differentially private frequent sequence mining via sampling-based candidate pruning”, in Proc. IEEE 31st Int. Conf. Data Eng. (ICDE), pp. 1035-1046, Apr. 2015.

[10] Shen and T. Yu, ``Mining frequent graph patterns with differential privacy," in Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 545-553, 2013.

[11] S. Xu, S. Su, L. Xiong, X. Cheng, and K. Xiao, “Differentially private frequent subgraph mining”, in Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE), pp. 229-240, May 2016.