Journal of Engineering Sciences

Vol 10, Issue 9, Sept /2019
ISSN NO: 0377-9254

PRIVACY-PRESERVING FREQUENT ITEM-SET MINING WITH DIFFERENTIAL PRIVACY

[1]V Leena Parimala, [2]Md Fayaz, [3]Bushra Tahseen
[1]Assistant Professor , Assistant Professor ,Associate Professor
DEPARTMENT OF CSE
Dr. K V SUBBA REDDY INSTITUTE OF TECHNOLOGY, KURNOOL

Abstract: In recent years, people have an interest in coming up with differentially non-public data processing algorithms. several researchers square measure functioning on style of information mining algorithms which supplies differential privacy. during this paper, to explore the chance of coming up with a differentially nonpublic FIM , cannot simply accomplish high info utility and a high level of protection, in addition offers time effectiveness. to the present finish, the differential non-public FIM based on the FP-growth algorithmic program, that is talk about to as PFP-growth. The Privacy based mostly algorithm consists of a pre-processing half and a mining half. among the preprocessing half, to boost the utility and privacy exchange, a very distinctive smart good ripping technique is anticipated to transform the information. Privacy is the most essential in today's world for online as well as offline data. Frequent Item sets Mining (FIM) it is a typical data processing task and has gained abundant attention. Due to the consideration of individual privacy, various studies have been focusing on privacy-preserving FIM problems. Differential privacy has emerged as a promising theme for shielding individual privacy in data processing against adversaries with impulsive information. In this work we propose an efficient, privacy preservation based frequent item sets mining (FRM) algorithm on large as well as high dimensional data called Frequent Item set Mining Privacy Preservation (FIM_PP). In light of the thoughts of examining and exchange truncation utilizing length limitations, our calculation lessens the calculation force, decreases mining affectability, and subsequently enhances information utility given a settled protection spending plan. Partial experimental analysis show the proposed system evaluation show the how proposed system provides best results than existing systems.

Keywords: Wearable sensors, healthcare, big data, cloud computing, authentication, security

I. INTRODUCTION

Frequent item set mining is a well-recognized data mining problem. The discovery of frequent item sets can serve valuable economic and research purposes, e.g., mining association rules [5], predicting user behavior [3], and finding correlations [11]. Publishing frequent item sets, however, may reveal the information of individual transactions, compromising the privacy of them. In this paper, we study the problem of how to perform frequent item set mining (FIM) on transaction databases while satisfying differential privacy. Differential privacy [17] is an appealing privacy notion which provides worst-case privacy guarantees. In recent years, it has become the de facto standard notion of privacy for research in private data analysis. The key challenge in private FIM is that the dimensionality of transactional datasets is very high. While effective techniques for differentially private data publishing have been developed for low-dimensional datasets (e.g., [23, 33]), these techniques do not apply to high-dimensional data. In fact, even for the weaker privacy notion of k-anonymity, the curse of high dimensionality effect is well known [4].

Our work is inspired by Bhaskar et al.'s KDD10 paper [8], in which they propose an approach to privately publish top k frequent item sets and their frequencies. Their approach first selects k item sets from the set of all item sets that include at most m items, and then adds noise to the frequencies of these selected item sets. This approach works reasonably well for small k values; however for larger values of k, the accuracy is poor. The main reason is that for larger k values, one has to set the size limit m to be larger (e.g., 3, 4, or higher). This results in a very

large candidate set from which the algorithm must select the top k, making the selection inaccurate.

In this paper we propose a novel approach that avoids the selection of top k item sets from a very large candidate set. More specifically, we introduce the notion of basis sets. A θ-basis set B = {B1, B2, . . . , Bw}, where each Bi is a set of items, has the property that any item set with frequency higher than θ is a subset of some basis Bi. A good basis set is one where w is small and the lengths of all Bi's are also small. Given a good basis set B, one can reconstruct the frequencies of all subsets of Bi's with good accuracy. One can then select the most frequent item sets from these. We also introduce techniques to construct good basis sets while satisfying differential privacy. Finally, we have conducted extensive experiments, and the results show that our approach greatly outperforms the existing approach.

We call our approach PrivBasis. It meets the challenge of high dimensionality by projecting the input dataset onto a small number of selected dimensions that one cares about. In fact, PrivBasis often uses several sets of dimensions for such projections, to avoid any one set containing too many dimensions. Each basis in B corresponds to one such set of dimensions for projection. Our techniques enable one to select which sets of dimensions are most helpful for the purpose of finding the k most frequent item sets.

II. LITERATURE SURVEY

Pan, Zhaopeng et.al [1] FP-growth algorithm produces all the frequent item sets without producing a large number of candidate items. However, when the item set is too large, the branch of the spanning tree will be long, occupying the space too large, and the mining efficiency will be reduced. In this paper, the method of using dynamic insert node FP - tree structure, and all the back pointer, to generate a new type of FP - tree. This article also proposes Max-IFP maximum frequent patterns mining algorithm, using the new generation of FP - tree dug up all the maximum frequent item sets. The experimental results show that the new FP-tree occupies a smaller space, and the algorithm proposed in this paper is shorter and more effective than other algorithms when mining the maximum frequent item sets.

Fahrudin, TresnaMaulana et.al [2] No proof of malady (NED) is carcinoma patient condition standing that it indicates that they will life, no realize the cancer by tested, and with none symptoms of cancer in period of times, after they received primary treatment. Patient condition status which it indicates that they NED is a critical status, because it involves the treatment type and patient cancer condition factors. The examines about breast cancer problem in data mining technical side, especially to discover the patterns of NED-breast cancer patient using cancer registry data from Oncology Hospital. Its patterns are discovered through the relationship of among features begin from 1dimensional, 2-dimensional, 3-dimensional, and n-dimensional. They applied association rules mining using Apriori and FP-Growth algorithm, which both have the advantage and drawback. Apriori algorithm involves all generation of candidate item sets and multiple database scans, but it makes high consuming iteration. While FP-Growth algorithm extracts the frequent item sets directly from FP-Tree, it make the advantage of FP-Growth that is faster process needs only scan the database once. This paper experiment shown that the association result of Apriori and FP-Growth is almost similar, 10-highest confidence value represented 100% confidence of association rule on breast cancer dataset with support value up to 50%.

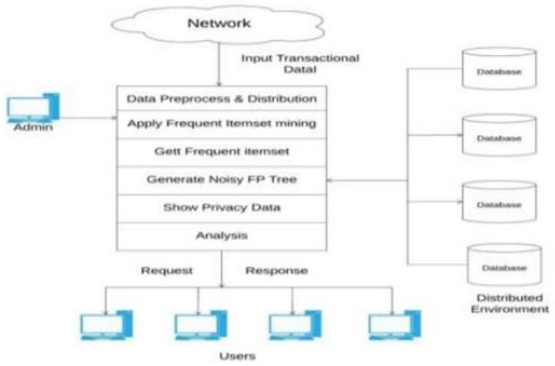
According to Xun, Yaling, Jifu Zhanget.al [4] Existing parallel mining algorithms for frequent itemsets lack a mechanism that allows automatic parallelization, load effort, data distribution, and fault tolerance on large clusters. As a solution to the present disadvantage, we tend to tend to vogue a parallel frequent item sets mining rule called FiDoop victimization the MapReduce programming model. to realize compressed storage and avoid building conditional pattern bases, FiDoop incorporates the frequent things ultrametric tree, instead of typical FP trees. In Fi Doop , Map Reduce jobs unit enforced to finish the mining task. at intervals the crucial third Map Reduce job, the mappers severally decompose itemsets, the reducers perform combination operations by constructing very little ultrametric trees, and jointly the actual mining of those trees individually. we have a tendency to tend to implement FiDoop on our in-house Hadoop cluster.

we have a tendency to tend to indicate that FiDoop on the cluster is sensitive to knowledge distribution and dimensions, as a results of itemsets with entirely completely different lengths have different decomposition and construction costs. to enhance FiDoop's performance, we tend to tend to develop a employment balance metric to measure load balance across the cluster's computing nodes. we tend to develop FiDoop-HD, Associate in Nursing extension of FiDoop, to hurry up the mining performance for high-dimensional data analysis. in depth experiments victimization real-world celestial spectral data demonstrate that our projected resolution is economical and ascendable.

Zhang, and Hou Ying [8]. With the coming of big data time, the huge number of IDS log makes the traditional computing technology and systems cannot cope and deal with the needs of the analysis of security log, so large-scale computing power has become a prerequisite for the effective implementation of data mining technology. Based on the Hadoop framework, the applies the parallel frequent item sets mining algorithm to the Snort Intrusion Detection System, which solves the problem that Snort-IDS cannot judge the security event itself. At the same time, it also solves the problem of decreasing the processing speed due to the enormous increase of data. So that the system has the ability of detecting new intrusions, enrich and improve the SnortIDS functional system and the performance

III. PROPOSED SYSTEM

In the proposed research work to design and implement a system for FIM using privacy preservation approach. This work also carried out an efficient, differential private frequent item sets mining algorithm over large scale data. Based on the ideas of sampling and transaction truncation using length constraints, our algorithm reduces the computation intensity, reduces mining sensitivity, and thus improves data utility given a fixed privacy budget



Mining has turned into a blasting subject of research in Computer Science. This fast growing phenomenal is driven by numerous reasons. First, information mining and information warehousing are to a great degree fertile with inquire about issues but then present an outrageous helpful instrument to oversee huge measure of information. Besides, information develops at an exponential rate and Internet technology has made it simple to suspect that information from everywhere throughout the world so companies and associations these days find themselves immersed with information, and anxious to extricate useful information from them to benefit their business

IV. IMPLEMENTATION AND RESULTS

Admin

In this module, the Cloud has to login by using valid user name and password. After login successful he can do some operations such as List all users and authorize, View all company users and authorize Add all company name and view, View all company details with rank and reviews,View all companies by Frequent Item sets Mining using FP-Tree format and give link on company name view its details, View all user search transaction by keyword, Show search ratio by keyword, Find top k Frequent item sets by ranks View all companies rank by chart, View all search ratio by keyword in chart

View and Authorize Users

In this module, the admin can view the list of users who all registered. In this, the admin can view the

user's details such as, user name, email, address and admin authorizes the users.

Production Company

In this module, there are n numbers of Owners are present. Owner should register before doing any operations. Once registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful Owner will do some operations like View your profile, Add company data set, View your company details with reviews and rank, View user search transactions on your company, View other related companies by Frequent Itemsets Mining using FP-Tree format and give link on company name view its details

Users

In this module, there are n numbers of users are present. User should register before doing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user will do some operations like View your profile, Search companies by keyword and show all related companies by FP-Tree format and give link on company name view its details, view its details with image(increment rank),review , show other review also, find search ratio, View your search transactions by keyword

Algorithm 1 gives the BasisFreq algorithm for computing the noisy counts of all itemsets in C(B). In the algorithm we compute noisy counts, which can be translated into frequencies easily. The key ideas of the algorithm are as follows. Each basis Bi divides all possible transactions into 2|Bi| mutually disjoint bins, one corresponding to each subset of Bi. For each X ⊆ Bi, the bin corresponding to X consists of all transactions that contain all items in X, but no item in Bi \ X

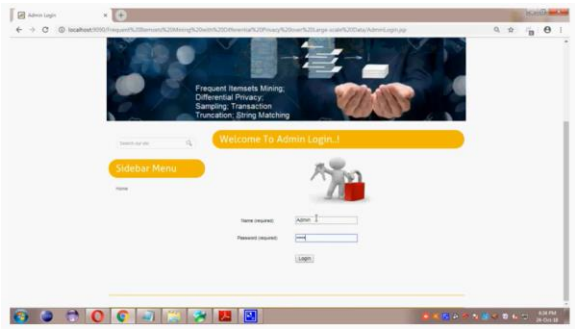
Algorithm 1 BasisFreq: Privately Releasing Frequent Itemsets using Basis Sets

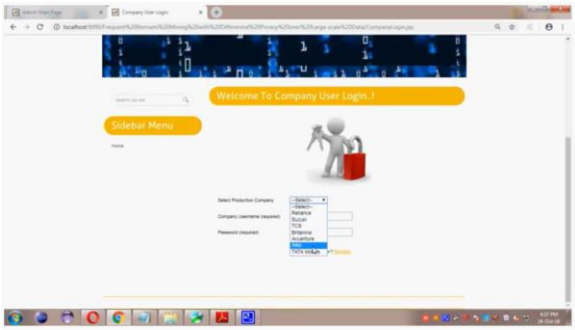
Input: Transactional dataset D, $\mathbf{B} = \{B_1, \cdots, B_w\}$, k, differential privacy budget ϵ.

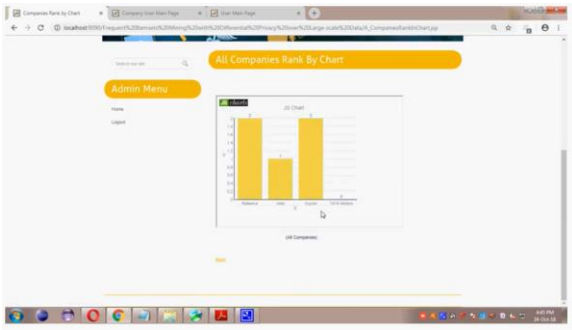
Output: Top k frequent itemsets in $\mathcal{C}$ and their frequencies.

function BasisFreq($D, \mathbf{B} = \{B_1, \cdots, B_w\}, k, \epsilon$)
  for $i = 1 \rightarrow w$ **do**
    for $j = 0 \rightarrow 2^{|B_i|} - 1$ **do**
      $b[i][j] \leftarrow \mathsf{Lap}\left(\frac{w}{\epsilon}\right)$
    end for
  end for
  for all $t \in D$ **do**
    for $i = 1 \rightarrow w$ **do**
      $b[i][t \cap B_i] \leftarrow b[i][t \cap B_i] + 1$
    end for
  end for
  $C \leftarrow \emptyset$
  for $i = 1 \rightarrow w$ **do**
    for all $X \subseteq B_i$ **do**
      $nc \leftarrow \sum_{Y \subseteq B_i | X \subseteq Y} b[i][Y]$
      $nv \leftarrow 2^{|B_i| - |X|}$
      if $C(X)$ is undefined **then**
        $C(X).nc \leftarrow nc$
        $C(X).v \leftarrow nv$
      else
        $v \leftarrow C(X).v$
        $C(X).nc \leftarrow \frac{nv}{v+nv} C(X).nc + \frac{v}{v+nv} nc$
        $C(X).v \leftarrow \frac{v \cdot nv}{v+nv}$
      end if
    end for
  end for
  $R \leftarrow$ the k elements in X's with highest $C(X).nc$
  return R
end function

RESULTS







V. CONCLUSION

In this paper we examine the problem of design a private DP-RElim with differential privacy ,which consist of preprocessing phase and mining phase. In first phase to better enhance utility exchange off, utilizing keen part strategy. In mining stage, a run time estimation technique is proposed to counterbalance the data misfortune brought about by exchange part. By using dynamic reduction method to dynamically decrease the amount of noise added to guarantee privacy during the mining process. The DP-RElim algorithm is time efficient and can achieve both utility and good privacy. The dynamic reduction and run-time estimation methods are used in phase to enhance the quality of the results. Recursive depends on a stage by step end of things from the exchange database together with a recursive preparing of exchange subsets. This calculation works without entangled information structures furthermore, permits us to discover visit itemset effectively.

REFERENCES

[1] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li," Differentially Private Frequent Itemset Mining via Transaction Splitting" IEEE Transaction on knowledge and Data Engineering, vol. 27, No. 7, July 2015.

[2] C. Zeng, J. F. Naughton, and J.-Y. Cai, On differentially private frequent itemset mining," International Conference on Very Large Data Bases, Vol. 6, August 2012.

[3] N. Li, W. Qardaji, D. Su, and J. Cao, Privbasis: Frequent itemset mining with differential privacy", International Conference on Very Large Data Bases, Vol. 5, August 2012.

[4] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, in SIGMOD, 2000.

[5] J. Vaidya and C. Clifton, Privacy preserving association rule mining in vertically partitioned data, in KDD,2002.

[6] Cynthia Dwork. "Differential Privacy" ICALP, Springer, 2006.

[7] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, in Proc. 20th Int. Conf. Very Large Data Bases, 1994, pp. 487499.

[8] J. Han, J. Pei, and Y. Yin, Mining frequent patterns without candidate generation, in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000

[9]. Christian Borgelt,"Simple Algorithms for Frequent Item Set Mining" Advances in Machine Learning II,Springer,2010.

[10]. Luna, José María, et al. "Apriori versions based on mapreduce for mining frequent patterns on big data." IEEE transactions on cybernetics (2017).