A Framework for Mining Frequent Patterns and Protecting Data from Privacy Leakage

[1]Raghavender K V , [2]Vasavi Prasanna,[3]T.Sai Krishna Reddy,[4]M.Pavani,[5]T.S. Vishnu

1Associate professor , Department of CSE, in Malla Reddy Engineering College (Autonomous), Hyderabad, TS, India

[2]Student, Department of CSE, in Malla Reddy Engineering College (Autonomous), Hyderabad, TS, India

[3]Student, Department of CSE, in Malla Reddy Engineering College (Autonomous), Hyderabad, TS, India

[4]Student, Department of CSE, in Malla Reddy Engineering College (Autonomous), Hyderabad, TS, India

[5]Student, Department of CSE, in Malla Reddy Engineering College (Autonomous), Hyderabad, TS, India

kvraghu2011phd@gmail.com,vasaviprasanna99@gmail.com,skricky69@gmail.com,
madellapavani234@gmail.com, vishnusachin26@gmail.com

Abstract- Mining frequent item sets is an important requirement in data mining domain. It is used in various rea time applications to discover trends or patterns. The trends or patterns provide business intelligence or customer behavior that can make help in making well informed decisions. There are many approaches available in the literature to mine frequent item sets. However, efficiency and privacy preserving approach is to be given high importance. In order to achieve this, an algorithm proposed. It is named as Fast and Privacy Preserving Item set Mining (FPPIM). The algorithm takes privacy budget and dataset besides support and confidence as input. It makes use of a tree structure known as POC tree to hold data. Then mines item sets that frequent in a faster way and with privacy preserving. The two measures such as support and confidence are used to prune the results and improve quality. A prototype application is built to evaluate the proposed algorithm. Two benchmark datasets from UCI machine learning repository and a synthetic dataset are used for the empirical study. The results revealed the usefulness of the proposed algorithm. It showed better performance with existing ones.

Keywords- Frequent Item Sets Mining; Differential Privacy; Sampling; Transaction Truncation; String Matching

1. INTRODUCTION

Frequent item sets mining with differential privacy refers to the problem of mining all frequent item sets whose supports are above a given threshold in a given transactional dataset, with the constraint that the mined results should not break the privacy of any single transaction. Current solutions for this problem cannot well balance efficiency, privacy and data utility over large scaled 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. In recent years, with the explosive growth of data and the rapid development of information technology, various industries have accumulated large amounts of data through various channels.

To discover useful knowledge from large amounts of data for upper-layer applications (e.g. business decisions, potential customer analysis, etc.), data mining has been developed rapidly. It has produced a positive impact in many areas such as business and medical care. Along with the great benefits of these advances, the large amount of data also contains privacy sensitive information, which may be leaked if not well managed. From the literature [3], [5], [8] and [10], it is understood that there have been efforts to improve the state of the art in frequent itemset mining. However, there is need for faster and privacy preserving algorithm. Our contributions in this paper are as follows.

1.A faster and privacy preserving algorithm known as Fast and Privacy Preserving Itemset Mining (FPPIM).

2.A prototype application is built to demonstrate proof of the concept.

3.The algorithm is evaluated with different benchmark datasets and a synthetic dataset besides comparing results with the state of the art.

The remainder of the paper is structured as follows. Section 2 provides review of literature. Section 3 presents the proposed system in detail. Section 4 presents experimental results while section 5 concludes the paper.

2. RELATED WORK

Different approaches in data mining are used to extract business intelligence from historical data. However, frequent itemset mining is widely used phenomenon. Knowledge discovery from databases is explored in [1] and [2]. There are many applications for frequent itemset mining. It can help in identifying interesting and hidden information or trends or customer behavior. It is used in different domains including education as studied in [3] and [4]. In [5] different concepts related to datamining are explored. Supervised learning methods is investigated in [6] while a data mining approach to solve problems in power distribution systems is elaborated in [7].

Fast mining of frequent itemsets with an underlying data structure is explored in [8] while similar kind of approach is followed to generate association rules in [9]. Frequent item set mining with more speed is defined in [10] and [11]. In [12] associations among features is exported while the [13 focuses on association rules with ranking concept. In [14] various datasets required by association rule mining are provided.

The research carried out in [15], [16] and [17] is linked to frequent itemset mining and association rule generation that is very useful to extract actionable knowledge from databases. From the literature, it is understood that it is essential to have faster item set mining and also preserve privacy. Towards this end, an algorithm is proposed and a porotype is built in this paper to demonstrate proof of the concept.

3. PROPOSED SOLUTION

The proposed solution is a web based application where the proposed algorithm runs. It has different users involved as we can see in the real world. It has different production companies, users and administrator. These roles are used to control access to different users. The application generates synthetic data on which frequent itemset mining is performed with the proposed algorithm. It also has provision to work with external datasets that can be used to mine frequent itemsets with privacy preserved.

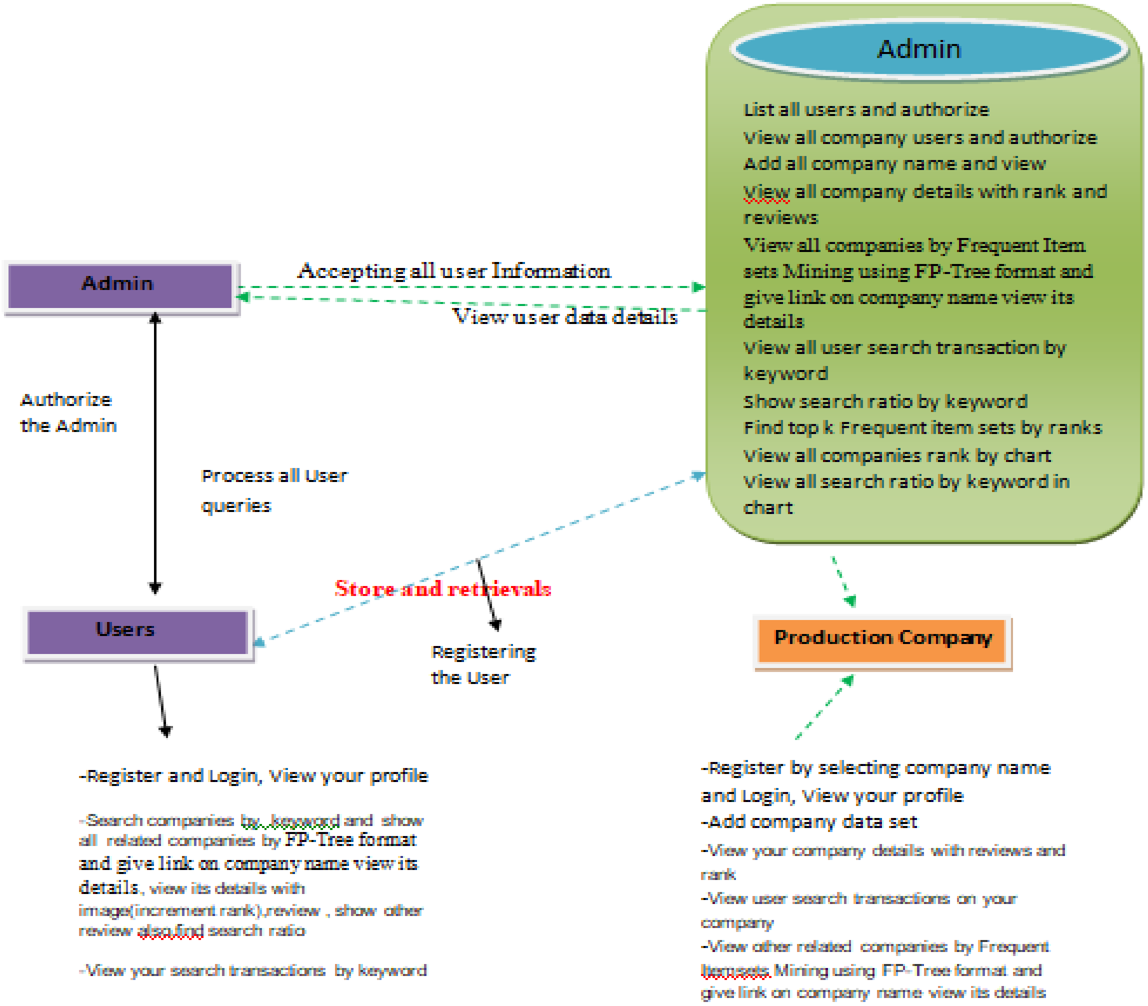
Admin
List all users and authorize
View all company users and authorize
Add all company name and view
View all company details with rank and reviews
View all companies by Frequent Item sets Mining using FP-Tree format and give link on company name view its details
View all user search transaction by 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
Admin
Accepting all user Information
View user data details
Authorize the Admin
Process all User queries
Store and retrievals
Users
Registering the User
Production Company
-Register and Login, 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
-Register by selecting company name and Login, 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

Figure 1: System architecture

As shown in Figure 1, there is system architectur showing different components like admin, user, production company and data mining activities. It enables the users to interact with the system wth appropriate functions. It generates companies dataset on which frequent itemset mininig is done with privacy preservation. 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 Item sets Mining using FP-Tree format and give link on company name view its details.

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.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.

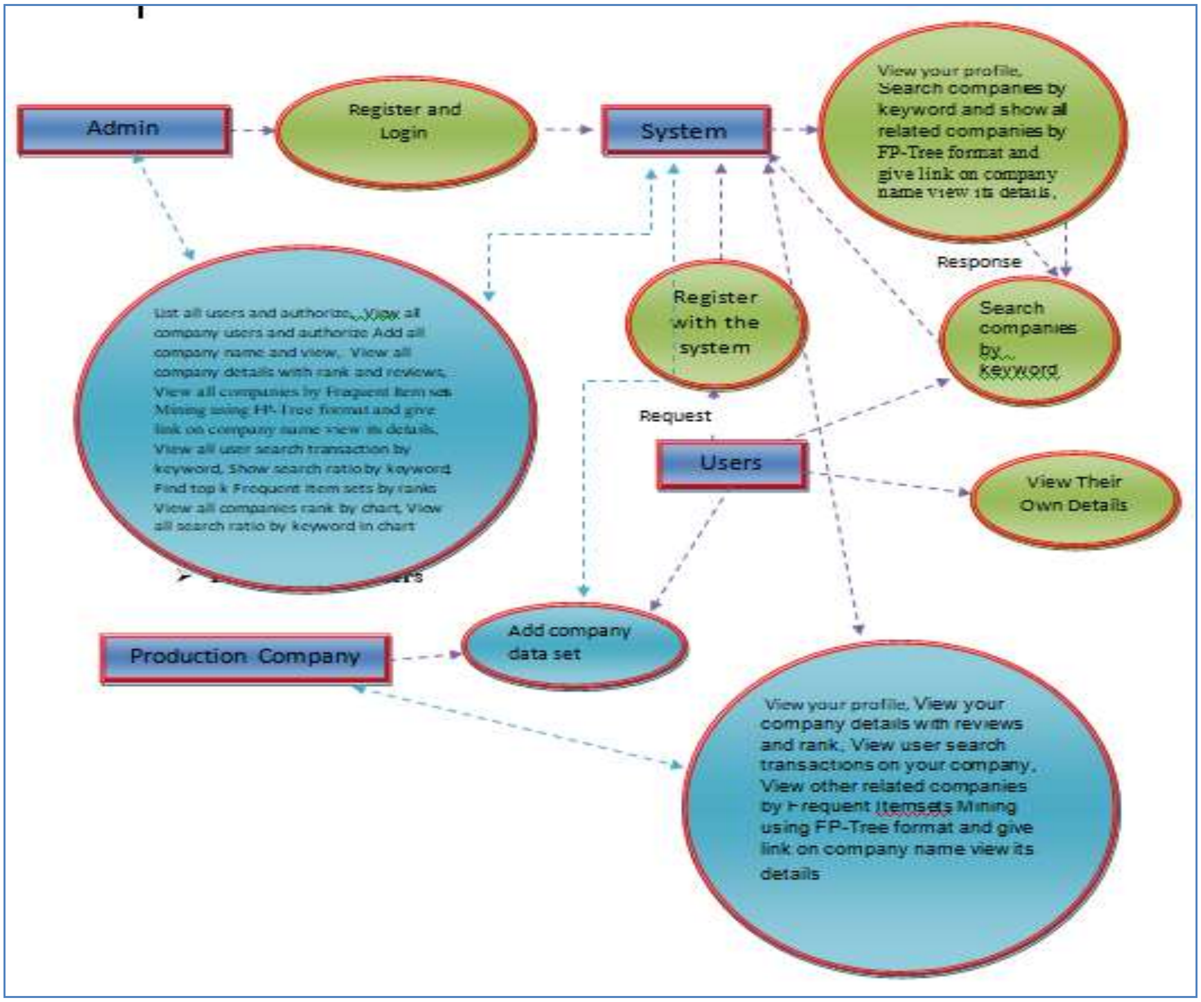
Admin
Register and Login
System
View your profile, Search companies by keyword and show all related companies by FP-Tree format and give link on company name view its details,
Response
List all users and authorize, View all company users and authorize Add all company name and view, View all company details 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
Register with the system
Search companies by keyword
Request
Users
View Their Own Details
Production Company
Add company data set
View your profile, 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

Figure 2: Data flow diagram

As shown in Figure 2, it is evident that there are many processes involved in the system. These processes are associated with different files in the system like user role and admin role. There are production companies involved as well. An algorithm based on POC tree is defined. We propose a novel differential private frequent itemsets mining algorithm for big data by merging the ideas, which has better performance due to the new sampling and better truncation techniques. We build our algorithm on POC-Tree for frequent itemsets mining. In order to solve the problem of building POC-Tree with large-scale data, we first use the sampling idea to obtain representative data to mine potential closed frequent itemsets, which are later used to find the final frequent items in the large-scale data.

ID	Items	Ordered Frequent Items
1	a, c, g, f	c, f, a
2	e, a, c, b	b, c, e, a
3	e, c, b, i	b, c, e

4	b, f, h	b, f
5	b, f, e, c, d	b, c, e, f

Table 1: Sample transaction database

The POC-tree for the data present in Table 1 is as shown in Figure 3. The tree is constructed for further processing while discovering frequent item sets.

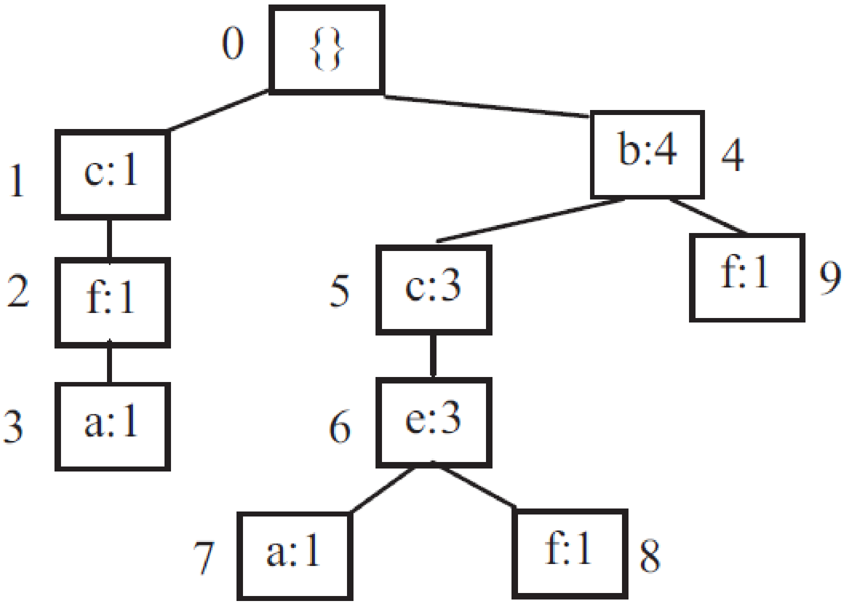

Figure 3: POC-tree

As shown in Figure 3, the POC tree is shown for the data presented in Table 1. This tree is more efficient and can help in improving performance of item set mining. The proposed algorithm is as follows.

```
Algorithm: Fast and Privacy Preserving Itemset Mining (FPPIM)
Inputs: Dataset D, support sup, confidence conf, privacy budget p
Output: Frequent Item Sets F' with Privacy
    1.  Start
    2.  Initialize vector POC to hold POC tree
    3.  Initialize vector AR to hold association rules
    4.  Initialize F to hold frequent item sets
    5.  Construct POC from D
    6.  Find frequent 1-itemsets
```

```
Scan POC tree for finding frequent 2-itemsets
F = Mine all frequent (>2) item sets that are compatible with sup and conf
For each frequent item f in F
    f'=ApplyDifferentialPrivacy(f, p)
    add f' to F'
End For
Return F'
End
```

Algorithm 1: PP-FIM algorithm

As shown in Algorithm 1, PP-FIM takes dataset, support, confidence and privacy budget as inputs. It generates a POC tree based on the given dataset. It is the tree which is light weight and support faster navigation. It generates frequent item sets from the POC faster. Then, they are pruned based on the support and confidence. Afterwards the differential privacy is applied to frequent itemsets based on the given privacy budget. Thus the itemsets are slightly anonymized to preserve privacy and also ensure that data utility is not lost.

4. EXPERIMENTAL RESULTS

Experiments are made with different datasets and privacy budget. Observations are made in terms of F-score and relative error (RE). This section presents the results and compare with the existing algorithm.

Privacy Budget	**F-Score**					
	Mushroom Dataset With Existing	**Mushroom Dataset With Proposed**	**Retail Dataset With Existing**	**Retail Dataset With Proposed**	**Companies Dataset With Existing**	**Companies Dataset With Proposed**
0.1	0.85	0.89	0.58	0.63	0.59	0.64
0.25	0.95	0.97	0.68	0.74	0.69	0.75
0.5	0.95	0.98	0.73	0.8	0.74	0.9
0.75	0.96	0.99	0.75	0.82	0.76	0.83
1	0.98	0.99	0.77	0.85	0.78	0.86

Table 2: Shows experimental results

As shown in Table 2, it has F-score values for existing and proposed systems on multiple datasets. The F-score is captured against a given privacy value.

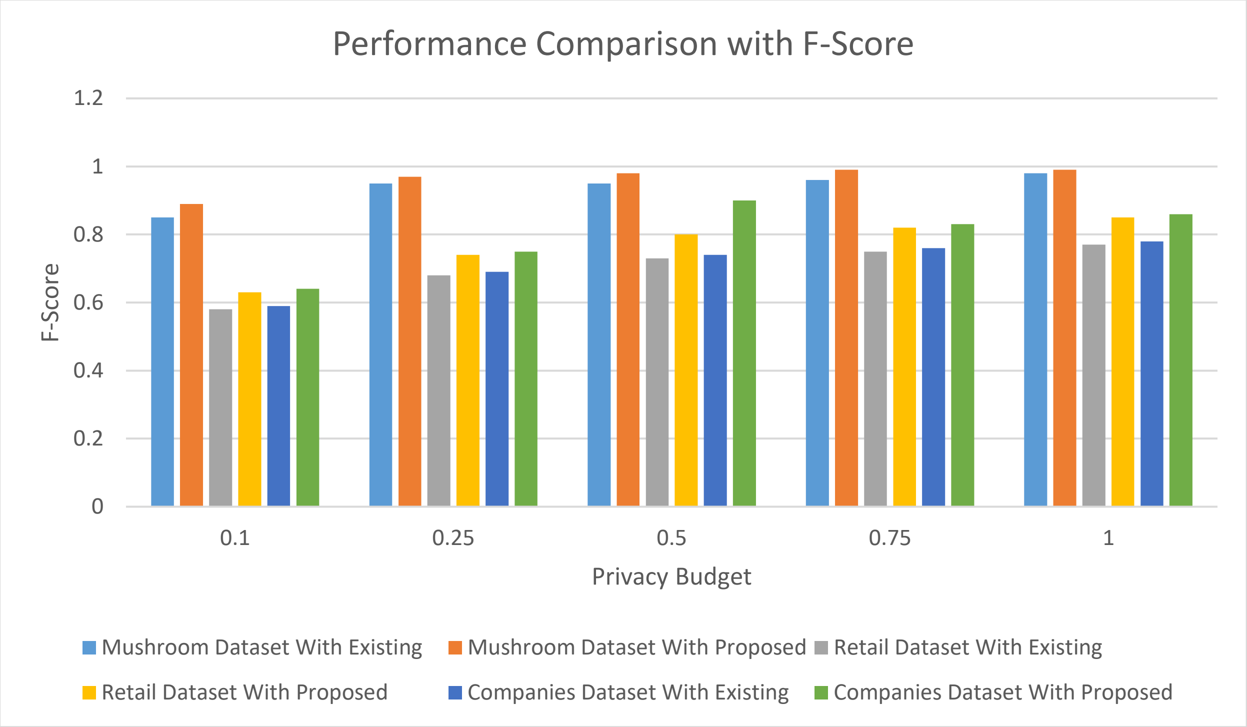

Figure 4: Shows experimental results

As shown in Figure 4, horizontal axis shows privacy budget. Vertical axis shows performance of the algorithms in terms of F-Score (a measure used to know accuracy of frequent item set mining). The privacy budget has its influence on the performance. The proposed system showed better performance over existing on different datasets.

Privacy Budget	**RE**					
	Mushroom Dataset With Proposed	**Mushroom Dataset With Existing**	**Retail Dataset With Proposed**	**Retail Dataset With Existing**	**Companies Dataset With Proposed**	**Companies Dataset With Existing**
0.1	0.045	0.052	0.17	0.24	0.18	0.25
0.25	0.04	0.09	0.15	0.19	0.16	0.2
0.5	0.03	0.08	0.19	0.28	0.2	0.29
0.75	0.02	0.07	0.17	0.22	0.18	0.23
1	0.02	0.01	0.13	0.18	0.14	0.19

Table 3 Shows experimental results with RE

As shown in Table 3, it has RE values for existing and proposed systems on multiple datasets. The RE is captured against a given privacy value.

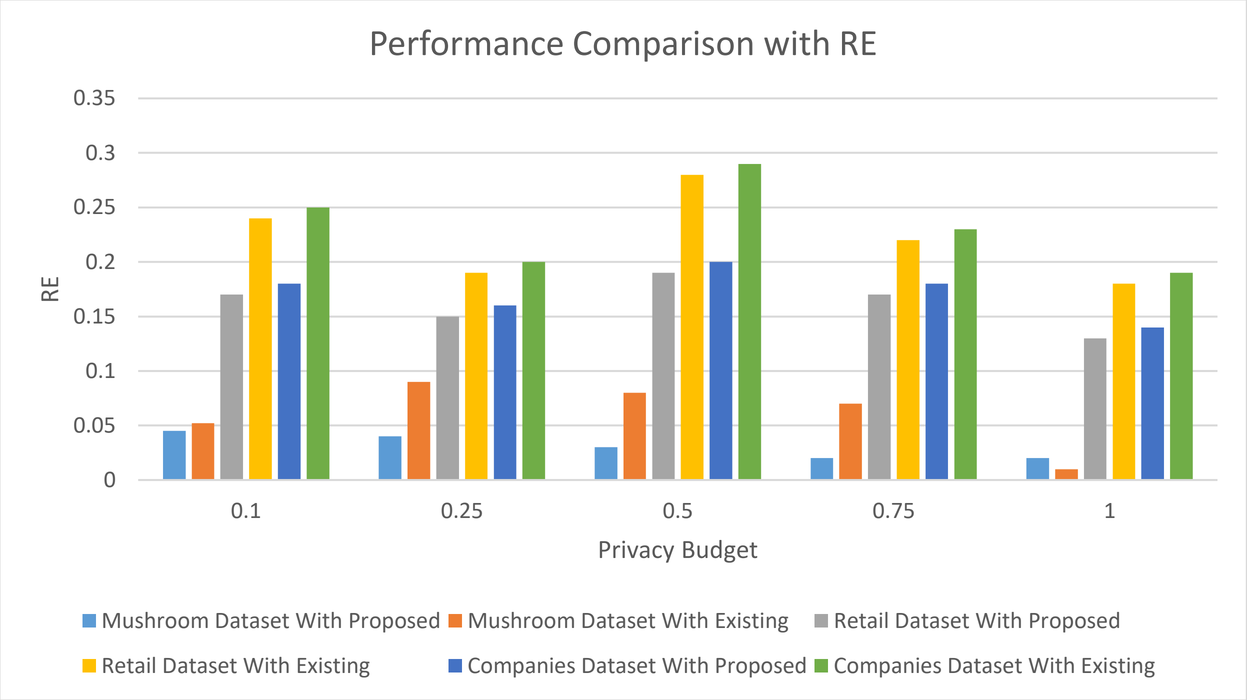

Figure 5: Privacy budget vs. RE

As shown in figure 5, the horizontal axis shows privacy budget. Vertical axis shows performance of the algorithms in terms of RE (a measure used to know performance of frequent item set mining). The privacy budget has its influence on the performance in terms of RE. The proposed system showed better performance over existing on different datasets.

5. CONCLUSION AND FUTURE WORK

Frequent item set mining with privacy preserved and efficiency is an important research are considered in this paper. POC tree is used for faster computations that led to speed in mining item sets that are frequent. In addition to this differential privacy concept is used in order to have privacy preserving mining that prevent data leakage while discovering frequent item sets. The algorithm proposed towards this end is Fast and Privacy Preserving Item set Mining (FPPIM). The algorithm takes different inputs such as dataset, privacy parameter, support and confidence and generate frequent item sets faster. It is employed on the synthetic dataset containing companies' data and two datasets taken from UCI. The results are evaluated and found the proposed algorithm performs better than the existing. The metrics used for comparison are RE and F-score. In future we intend to improve it with more fine grained approach in generating frequent item sets.

References

[1] Z. John Lu, "The elements of statistical learning: data mining, inference, and prediction," Journal of the Royal Statistical Society: Series A(Statistics in Society), vol. 173, no. 3, pp. 693–694, 2010.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining toknowledge discovery in databases," AI magazine, vol. 17, no. 3, p. 37,1996.

[3] H. Yang, K. Huang, I. King, and M. R. Lyu, "Localized support vectorregression for time series prediction," Neurocomputing, vol. 72, no. 10-12, pp. 2659–2669, 2009.

[4] C. Romero and S. Ventura, “Educational data mining: A review of thestate of the art,” IEEE Transactions on Systems, Man, and Cybernetics,Part C (Applications and Reviews), vol. 40, pp. 601–618, Nov 2010.

[5] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques.Elsevier, 2011.

[6] X. Fang, Y. Xu, X. Li, Z. Lai, and W. K. Wong, “Robust semi-supervisedsubspace clustering via non-negative low-rank representation,” IEEETransactions on Cybernetics, vol. 46, pp. 1828–1838, Aug 2016.

[7] M. Pe˜na, F. Biscarri, J. I. Guerrero, I. Monedero, and C. Le´on, “Rulebasedsystem to detect energy efficiency anomalies in smart buildings,a data mining approach,” Expert Systems with Applications, vol. 56,pp. 242–255, 2016.

[8] Deng, Z and Lv, S. (2014). Fast mining frequent item sets using Nodesets. Elsevier, Expert Systems with Applications, 41, p4505-4512.

[9] Agrawal, R., &Srikant, R. (1998). Fast algorithm for mining association rules. In VLDB’94 (pp. 487–499).

[10] Deng, Z. H., Wang, Z. H., & Jiang, J. J. (2012). A new algorithm for fast mining frequent itemsets using N-lists. Science China Information Sciences, 55(9), 2008–2030.

[11] Han, J., Cheng, H., Xin, D., & Yan, X. (2007). Frequent itemset mining: current status and future directions. DMKD Journal, 15(1), 55–86.

[12] Deng, Z. H., Wang, Z. H., & Jiang, J. J. (2012). A new algorithm for fast mining frequent itemsets using N-lists. Science China Information Sciences, 55(9), 2008–2030.

[13] Deng, Z. H. (2014). Fast mining Top-Rank-K frequent patterns by using node-lists. Expert Systems with Applications, 41(4–2), 1763–1768.

[14] UCI (2016). UCI Machine Learning Repository. Available online at: <https://archive.ics.uci.edu/ml/datasets.html>

[15] Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., & Yang, D. (2001). H-mine: hyper-structure mining of frequent itemsets in large databases. In ICDM’01 (pp. 441–448).

[16] Wang, J. Y., Han, J., & Pei, J. (2003). CLOSET+: searching for the best strategies formining frequent closed itemsets. In SIGKDD’03 (pp. 236–245).

[17] Lee, A. J. T., Wang, C. S., Weng, W. Y., Chen, Y. A., & Wu, H. W. (2008). An efficientalgorithm for mining closed inter-transaction itemsets. Data and Knowledge Engineering, 66(1), 68–91.