IJCRT.ORG ISSN : 2320-2882
INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Frequent Itemsets Mining With Privacy Over Large Scale Data

1 Ch VinodKumar, 2 Dr. M Arathi 1 Mtech Scholar 2 Associate Professor 1 Computer Science,2 Department of Computer Science
1,2 School of Information Technology - JNTUH, Hyderabad, India

ABSTRACT: Mining assets with varying privacy mean mining problems are all commonly supported beyond the scope provided in the given database, with the impediment that mineral effects should not violate the privacy of any single transaction. Data Mining plays a vital role in gaining an understanding of this big data. Similar solutions to this problem cannot properly balance the functionality, privacy and use of data over large scale data. this model big data is divided into smaller parts and then the data algorithms will be used.To protect the sensitive data that is proposed the algorithm uses the validation process and finally uses the TKU algorithm to retrieve the remaining Top-k objects.

1. INTRODUCTION

In recent years, with the eruptive growth of data and the ever growing of information technology, various industries have gained a wealth of information through various channels. To utilize this large data for additional applications (e.g. business decisions, potential customer analysis, etc.), data mining is quickly built. It has produced a good impact in many areas such as business care and therapy. Along with the huge benefits of this development, heaps of data also contains sensitive privacy information, which can be hacked if not properly managed. For example, Smartphone applications record users' location using GPS sensors andsend information to their servers medical records are also key.

Medical records also maintain relationships between different diseases and information. Mines in user location data or medical record data both provide important information; however, they may review the user's privacy. So mining information under reliable privacy guarantee is highly anticipated. This model investigates how to generate frequently occurring items that guarantee the confidentiality of big data. A company (such as an information consulting firm) has very big data. The company prefers to make the database public and therefore allows the public to create virtual reality mines to find partnerships or profits. But due to privacy considerations, the company cannot provide the original data directly. Therefore, privacy mechanisms are needed to process the information.

To make sure that the privacy of data mining, traditional methods are based on anonymity and in their extended model. These approaches require some consideration; It is difficult to protect privacy . The deficiency of k-anonymity and its extended models is that there is no robust description of the attack model, and that the attacker's information cannot be quantitatively described. In pursuit of a strong privacy analysis, Dwork proposed a robust privacy protection model called privacy difference.

2. LITERATURE SURVEY

The security issue of incessant itemsets mining is a primary focal point of exploration endeavors. We sort important work dependent on the basic strategies - from secrecy to differential protection. Namelessness Approaches. For dispersed datasets, Clifton et al. proposed a safe multi- party security ensuring affiliation rule mining calculation [8]. The thought is to change the issue into a protected multi-party calculation issue under flat dissemination. Vaidya et al. proposed a privacypreserving affiliation decide calculation that utilizations secure scalar estimation technique to locate all incessant itemsets under vertical conveyance [9]. In [10], Z Teng et al. proposed a cross breed security safeguarding calculation under vertical dissemination. For unified datasets, Wong et al. proposed to utilize 1-to-n encryption technique to change unique itemsets to ensure information protection when redistributing incessant itemsets mining [11]. Ling et al. proposed a calculation that changes business data into extremely long parallel vector and a seriesn of arbitrary planning capacities dependent on sprout channels. Afterward, Tai et al. proposed a k-uphold namelessness based successive itemsets mining calculation [13]. Every one of these strategies above penance the accuracy of mining result. Differential Privacy Approaches. Since customary methodologies depend on heuristics, a strong security ensure is missing. Subsequently, analysts started to research successive itemsets mining with differential security. Bhaskar et al. introduced two mining calculations [14], which are delegates of continuous itemsets mining with differential protection. Afterward, to comprehend the high dimensional test of dataset,Li et al. proposed the PrivBasis calculation that joins premise and planning method to accomplish top-k regular itemsets mining [16]. Zeng et al. proposed a ravenous strategy for exchange truncation approach by restricting the greatest length of exchanges of dataset [17].

3. SYSTEM ANALYSIS

System analysis consists of the various existing systems, drawbacks in them, and proposed system with advantages

3.1 Existing System

With the growth of more data and the evolving of information technology privacy becomes the major concern . In Existing system the Apriori algorithm is applied to the entire data set so it takes more time to process the results.

3.2 Disadvantages

1. It takes more time
2. Lack of privacy
3. It does not support for decision making

3.3 Proposed system

In the proposed program a separate division algorithm for private assets used for big data by combining ideas, with better performance due to new sample and better Truncation strategies.First Create an algorithm in FP-Tree for regular asset mines. To solve the problem of building FP-Tree with big data, it uses the concept of sampling to obtain representative data to include the most commonly closed items, which are later used to find the most common end-to- end items in big data

3.4 Advantages

1. High Accurate and Efficient.
2. Returns the top k always for good
3. Provide privacy sensitive information

4. IMPLEMENTATION

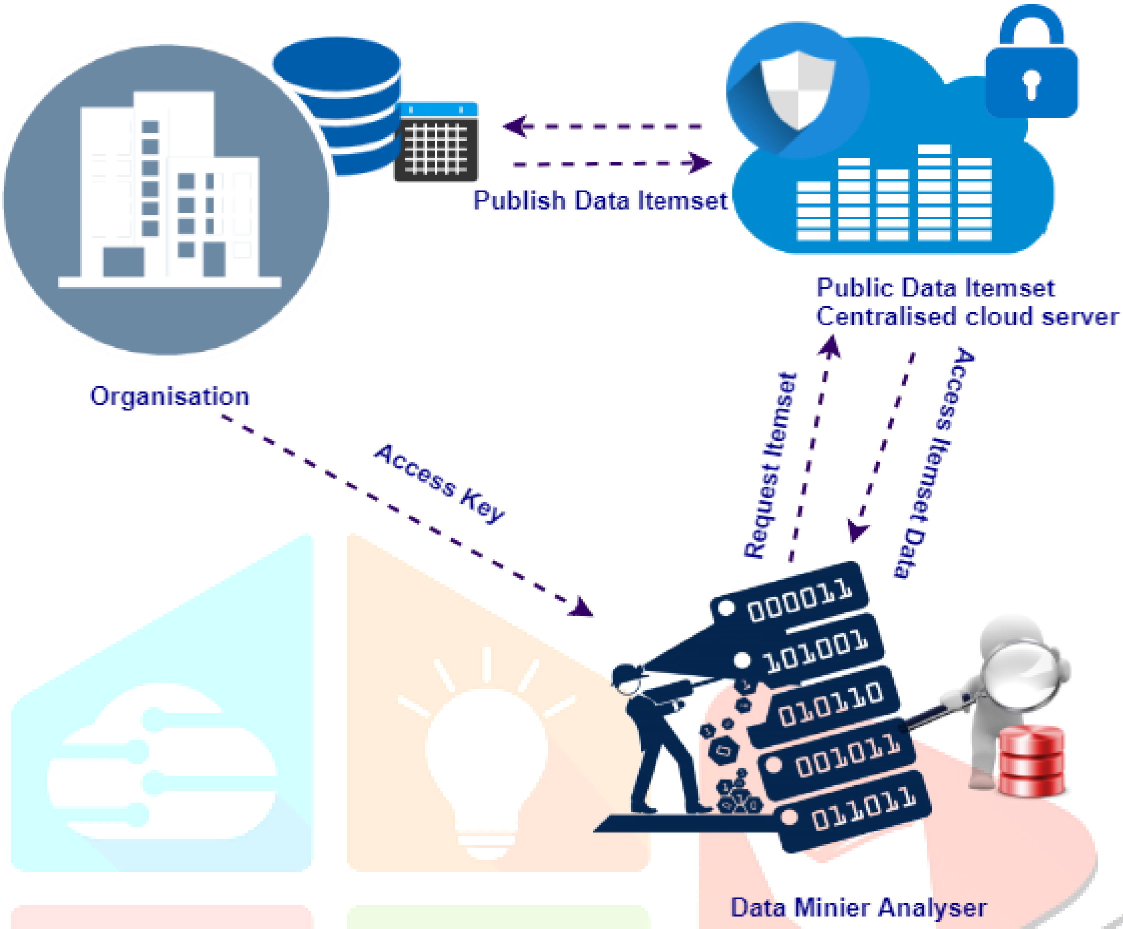
Frequent itemsets mining, be a transactional dataset consisting ofN transactions, be a set of different items, and X be a subset of I such that X U I. If X is containedin a transaction and X has k items, X is called a k-itemset. The support of an itemset is defined as the total number of transactions that contains the itemset. frequent itemsets mining is to find all itemsetsthat have support greater than a given threshold. Frequent itemsets is employed for finding association rules for a group of data items. Given the large-scale dataset, we first sample the dataset and then compute the closed frequent itemsets in the smaller sample using a traditional frequent itemsets mining algorithm. We later estimate the length distribution of the sampled dataset and obtain the maximum length constraint, which is later used to shrink the dataset. Some elements out of the closed frequent itemsets are removed from the source dataset if their supports are below the support threshold. We then employ string matching ideas to cut off the transactions in the dataset; in this step, the purpose of converting the dataset is to shrink the data size and simultaneously retain the potential frequent items.

We first sample the dataset to have a roughestimation of the dataset using the central limit theorem. Wefirst compute the sample size and then use SAS data analysissoftware for random sampling. The samples can reduce thecomputational intensity of the constructed FP-Tree and findthe potential frequent itemsets of the source dataset. We obtain a maximum length constraint lmax to shrinkthe transactions in the dataset.We deduce the sample size now. Fix an item modelled asa binomial distribution with occurring probability p.After thepreprocessing phase, we get the shrinking dataset which hassmaller number of transactions and smaller dimension to builda noisy FP-Tree. Because computing support directly destroysthe privacy.

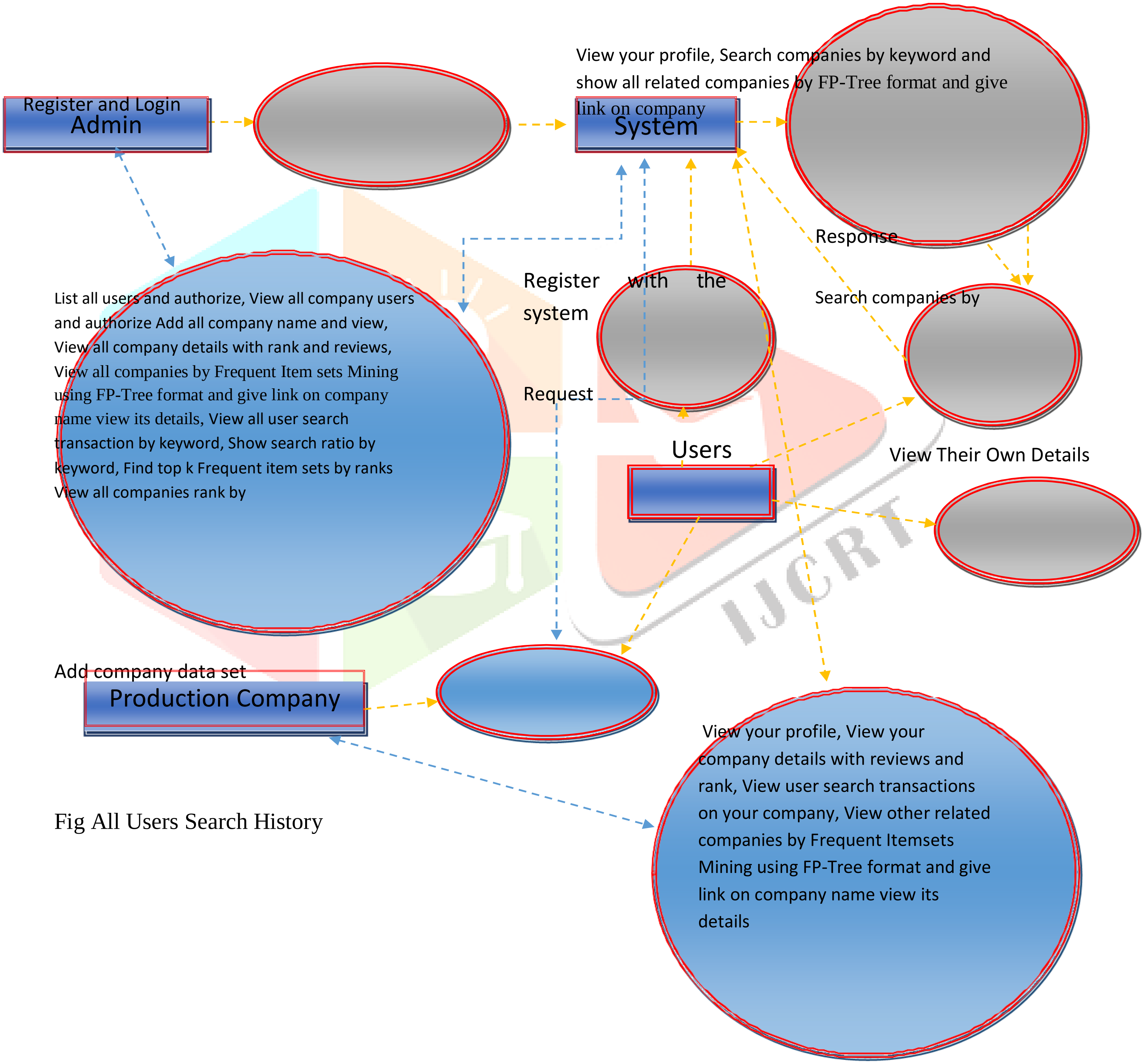
Algorithm

Newly proposed algorithm, called DPFIM, which merges the ideas of, but employs adifferent(better) truncation scheme and boosts computation efficiency using both sampling and truncation. Compared with previous work using random truncation, our new stringsimilarity- matching-based truncation mechanism has better performance than previous work, which is because string-similarity-matching-based truncation preserves more useful frequent itemset candidates. The experimental results also confirm the better performance. The algorithm is differentially private; it takes a threshold value and outputs the frequent itemsets with support at least. The basic idea is as follows: first, compute a noisy support for the threshold,then truncate the original database noisily, finally construct a noisy FP-Tree for mining frequent itemsets.

5. ARCHITECTURE



6. DATAFLOW DIAGRAM



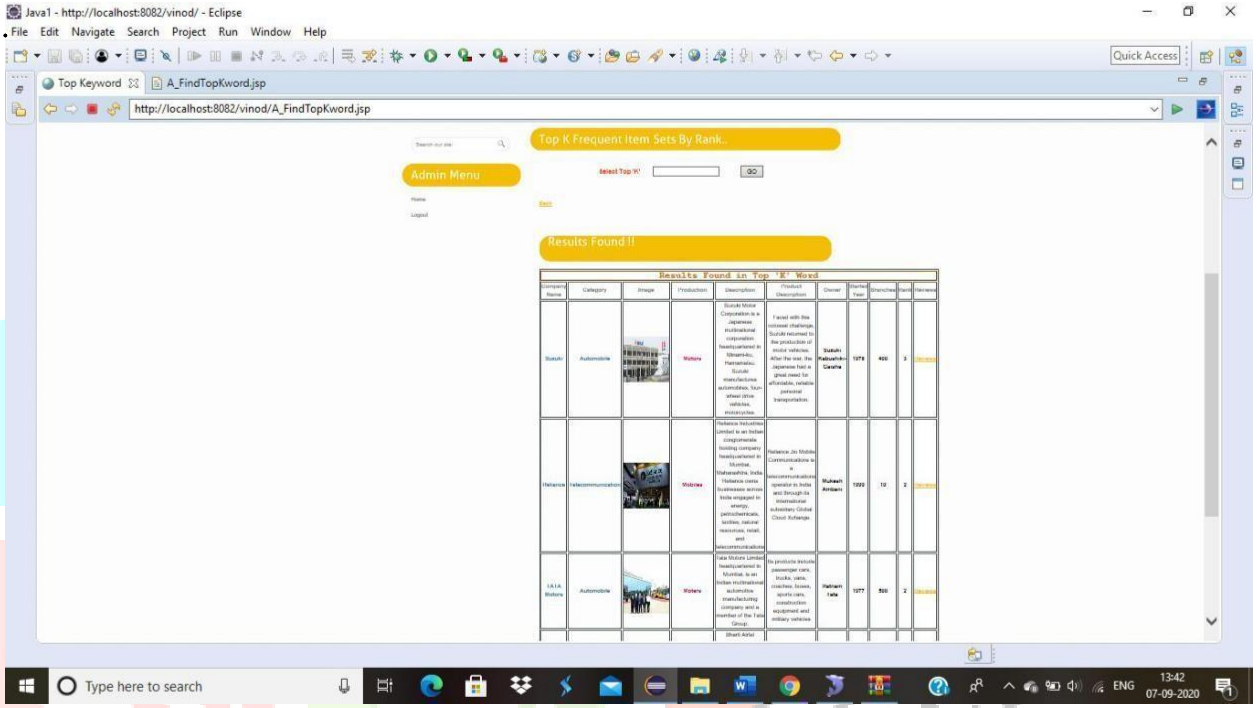
Register and Login
Admin
View your profile, Search companies by keyword and show all related companies by FP-Tree format and give link on company
System
Response
Register with the system
Search companies by
List all users and authorize, View all company users and authorize Add all company name and view,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

Fig All Users Search History

7. RESULTS

Top k frequent itemsets by rank

After placing k value (where k is integer) in the above search box top k frequent itemsets will be displayed as shown in the above image.further it helps in decision making

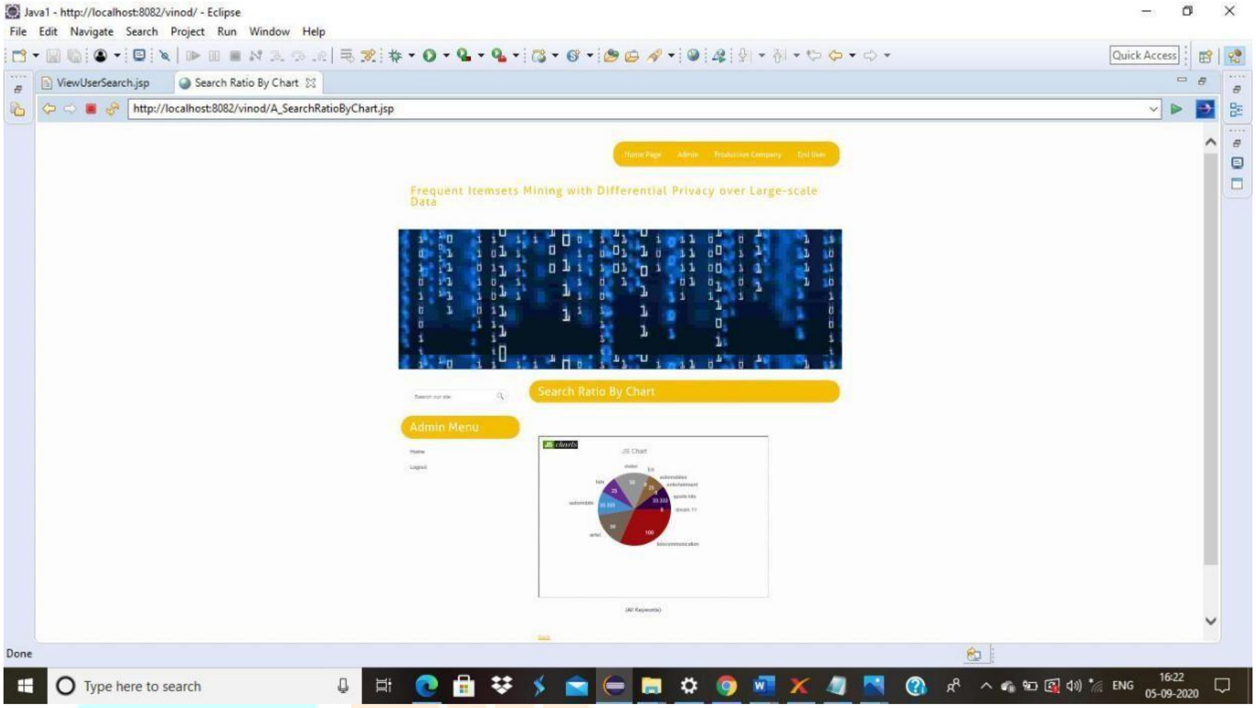

Java1 - http://localhost:8082/vinod/ - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
Top Keyword
A_FindTopKword.jsp
http://localhost:8082/vinod/A_FindTopKword.jsp
Top K Frequent Item Sets By Rank..
Admin Menu
Home
Logout
Select Top 'K'
GO
Back
Results Found !!
Type here to search
ENG
13:42
07-09-2020

Results Found in Top 'K' Word

Company Name	Category	Image	Production	Description	Product Description	Owner	Started Year	Branches	Rank	Reviews
Suzuki	Automobile		Motors	Suzuki Motor Corporation is a Japanese multinational corporation headquartered in Minami-ku, Hamamatsu. Suzuki manufactures automobiles, four-wheel drive vehicles, motorcycles	Faced with the colossal challenge, Suzuki returned to the production of motor vehicles. After the war, the Japanese had a great need for affordable, reliable personal transportation.	Suzuki Kabushiki-Gaisha	1979	400	3	Reviews
Reliance	Telecommunication		Mobiles	Reliance Industries Limited is an Indian conglomerate holding company headquartered in Mumbai, Maharashtra, India. Reliance owns businesses across India engaged in energy, petrochemicals, textiles, natural resources, retail, and telecommunications	Reliance Jio Mobile Communications is a telecommunications operator in India and through its international subsidiary Global Cloud Xchange.	Mukesh Ambani	1999	10	2	Reviews
TATA Motors	Automobile		Motors	Tata Motors Limited headquartered in Mumbai, is an Indian multinational automotive manufacturing company and a member of the Tata Group.	Its products include passenger cars, trucks, vans, coaches, buses, sports cars, construction equipment and military vehicles.	Ratnam Tata	1977	500	2	Reviews
				Bharti Airtel						

search ratio by chart

This page will helps the users by showing all the search ratio by representing in pictorial representation i.e(by chart).thus it helps the user in decision making



Java1 - http://localhost:8082/vinod/ - Eclipse
File Edit Navigate Search Project Run Window Help
Quick Access
ViewUserSearch.jsp
Search Ratio By Chart
http://localhost:8082/vinod/A_SearchRatioByChart.jsp
Home Page Admin Production Company End User
Frequent Itemsets Mining with Differential Privacy over Large-scale Data
Search Ratio By Chart
Admin Menu
Home
Logout
JS charts
JS Chart
100
(All Keywords)
Done
Type here to search
ENG
16:22
05-09-2020

8. CONCLUSION

This paper makes sures the company data is processed securely with help of the admin page and anyone can publish their data publicly without worrying about privacy .In this application end users and production company users have separate login pages so there is no issue of any ambiguity.This application is user friendly and there is no need to worry about any background knowledge. Finally it is providing top-k frequent itemsets thus end users can get benefit out of it and it also helps the production company which product customers are looking for thus it helps scalability and raising the profit of the company.

FUTURE ENHANCEMENTS

With the increase in the explosion of information privacy which is a major problem an advanced application that works for locals in the future can be upgraded to android and the algorithm used in this split privacy to have safer algorithms that will soon be developed to address advanced technology.Frequent itemsets with privacy play a game changing role in market basket analysis.

REFERENCES

1 . Z. Teng and W. Du, "A hybrid multi-group approach for privacy preserving data mining," Knowledge And Information Systems, vol. 19,no. 2, pp. 133–157, 2009

2.R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(_, k)-anonymity:an enhanced k-anonymity model for privacy preserving data publishing,"in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge

3.SQL complete reference by Livion

4.JAVA Complete Reference

5.Java Script Programming by Yehuda Shiran

6. Mastering JAVA Security

7.JAVA server pages by Larne Pekowsky

[8] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, pp. 1026–1037,2004.

[9] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,pp. 639–644, ACM, 2002.

[10] Z. Teng and W. Du, "A hybrid multi-group approach for privacypreserving data mining," Knowledge And Information Systems, vol. 19, no. 2, pp. 133–157, 2009.

[11] W. K. Wong, D. W. Cheung, E. Hung, B. Kao, and N. Mamoulis,"Security in outsourcing of association rule mining," in Proceedings of the 33rd International Conference on Very Large Databases, pp. 111–122, VLDB Endowment, 2007.

[12] Q. Ling, L. Yingjiu, and W. Xintao, "An approach to outsourcing data mining tasks while protecting business intelligence and customer privacy," in Sixth IEEE International Conference on Data Mining Workshops,pp. 551–558, IEEE, 2006.

[13] C.-H. Tai, P. S. Yu, and M.-S. Chen, "k-support anonymity based on pseudo taxonomy for outsourcing of frequent itemset mining," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 473–482, ACM, 2010.

[14] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining

of association rules," Information Systems, vol. 29, no. 4, pp. 343–364, 2004.

[15] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta, "Discovering frequent patterns in sensitive data," in Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 503–512, ACM, 2010.

[16] N. Li, W. Qardaji, D. Su, and J. Cao, "Privbasis: Frequent itemset mining with differential privacy," Proceedings of the VLDB Endowment, vol. 5, no. 11, pp. 1340–1351, 2012.

[17] C. Zeng, J. F. Naughton, and J.-Y. Cai, "On differentially private frequent itemset mining," Proceedings of the VLDB Endowment, vol. 6, no. 1,1pp. 25–36, 2012.