MINING PROBABILISTICALLY FREQUENT SEQUENTIAL PATTERNS IN LARGE UNCERTAIN DATABASE

Mrs.NITHYA.T (AP/Sl.Gr)[1]
RANGANAYAKI.R[2], NAVEENA.M[3], SURENDH KEERTHAN.P[4]
Department of Information Technology
Velalar College of Engineering and Technology
Erode.
tnithya27@gmail.com[1]
ranganayaki226@gmail.com[2], naveenamoorthi2207@gmail.com[3],
surendarsurendar097@gmail.com[4]

ABSTRACT:

Proposed concept is to measure pattern frequentness based on the possible word semantics. We establish two uncertain sequence data models abstracted from many real-life applications involving uncertain sequence data, and formulate the problem of mining probabilistically frequent sequential patterns (p-FSPs) from data that conform to our models. However, the number of possible words is extremely large, which makes the mining prohibitively expensive. Inspired by the famous EARM (Enhanced Association Rule Mining) algorithm, developed new algorithm, collectively called Enhanced Association Rule Mining that effectively avoids the problem of "possible words explosion", and when combined with our four pruning and validating methods, achieves even better performance. We also propose a fast validating method to further speed up our EARM algorithm. The efficiency and effectiveness of EARM are verified through extensive experiments on both real and synthetic datasets.

INDEX TERMS: Association rule mining, hadoop, frequent patterns, occurences.

INTRODUCTION:

DATA uncertainty is inherent in many real-world applications such as sensor data monitoring, RFID localization and location-based services, due to environmental factors, device limitations, privacy issues, etc. As a result, uncertain data mining has attracted a lot of attention in recent research. The problem of mining Frequent Sequential Patterns (FSPs) from deterministic databases has attracted a lot of attention in the research community due to its wide spectrum of real life applications. For example, in mobile tracking systems, FSPs can be used to classify or cluster moving objects; and in biological research, FSP mining helps discover correlations among gene sequences. In this work, we consider the problem of mining FSPs in the context of uncertain sequence data. In contrast to previous work that adopts expected support to measure pattern frequentness, we propose to define pattern frequentness based on the possible word semantics. This approach leads to more effective mining of high quality patterns with respect to a formal probabilistic data model. We develop two uncertain sequence data models abstracted from many real-life applications involving uncertain sequence data. Based on the models we define the problem of mining probabilistically frequent sequential patterns.

PROPOSED SYSTEM:

Proposed work is to develop two new algorithms, collectively called EARM, for p-FSP mining. EARM effectively avoids the problem of "possible words explosion", and when combined with our four pruning and validating methods, achieves even better performance and also propose a fast validating method to further speed up our EARM algorithm. The efficiency and effectiveness of EARM are verified through extensive experiments on both real and synthetic datasets. EARM

adopts the prefix-projection recursion framework of the EARM algorithm in a new algorithmic setting, and effectively avoids the problem of "possible words explosion".

Our contributions are summarized as follows:

To our knowledge, this is the first work that attempts to solve the problem of p-FSP mining, the techniques of which are successfully applied in the frequent items application for trajectory pattern mining.

We consider two general uncertain sequence data models that are abstracted from many real-life applications involving uncertain sequence data: the sequence-level uncertain model, and the element-level uncertain model.

Based on the prefix-projection method of EARM, we design two new EARM algorithms that mine p-FSPs from uncertain data conforming to our models.

Pruning techniques and a fast validating method are developed to further improve the efficiency of U-EARM, which is verified by extensive experiments.

LITRATURE REVIEW:

1. PREFIX-PROJECTED PATTERN GROWTH

Sequential pattern mining is an important data mining problem with broad applications. It is challenging since one may need to examine a combinatorial explosive number of possible subsequence patterns. Most of the previously developed sequential pattern mining methods follow the methodology of Apriori which may substantially reduce the number of combinations to be examined. However, Apriori still encounters problems when a sequence database is large and/or when sequential patterns to be mined are numerous and/or long. In this work, we propose a novel sequential pattern mining method, called Prefix-Span (i.e., Prefix-projected Sequential pattern mining), which explores prefix-projection in sequential pattern mining. Prefix-Span mines the complete set of patterns but greatly reduces the efforts of candidate subsequence generation. Moreover, prefix-projection substantially reduces the size of projected databases and leads to efficient processing. Our performance study shows that Prefix-Span outperforms both the Apriori-based GSP algorithm and another recently proposed method, Free-Span, in mining large sequence databases. Sequential pattern mining, which discovers frequent subsequences as patterns in a sequence database, is an important data mining problem with broad applications, including the analysis of customer purchase behavior, Web access patterns, scientific experiments, disease treatments, natural disasters, DNA sequences, and so on. The sequential pattern mining problem was first introduced by Agrawal and Srikant: Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min support threshold, sequential pattern mining is to find all of the frequent subsequences, i.e., the subsequences whose occurrence frequency in the set of sequences is no less than min support.

2. MINING SEQUENTIAL PATTERNS

We are given a large database of customer transactions, where each transaction consists of customer-id, transaction time, and the items bought in the transaction. We introduce the problem of mining sequential patterns over such databases. We present three algorithms to solve this problem, and empirically evaluate their performance using synthetic data. Two of the proposed algorithms, AprioriSome and AprioriAll, have comparable performance, albeit AprioriSome performs a little better when the minimum number of customers that must support a sequential pattern is low. Scale-up experiments show that both AprioriSome and AprioriAll scale linearly with the number of customer transactions. They also have excellent scale-up properties with respect to the number of transactions per customer and the number

of items in a transaction. The problem of discovering what items are bought together in a transaction" over basket data was introduced. While related, the problem of finding what items are bought together is concerned with finding intra-transaction patterns, whereas the problem of finding sequential patterns is concerned with inter-transaction patterns. A pattern in the first problem consists of an unordered set of items whereas a pattern in the latter case is an ordered list of sets of items. Discovering patterns in sequences of events has been an area of active research in AI. However, the focus in this body of work is on discovering the rule underlying the generation of a given sequence in order to be able to predict a plausible sequence continuation. We another hand are interested in finding all common patterns embedded in a database of sequences of sets of events.

3. EARM ALGORITHM

Data uncertainty is inherent in many real-world applications such as environmental surveillance and mobile tracking. As a result, mining sequential patterns from inaccurate data, such as sensor readings and GPS trajectories, is important for discovering hidden knowledge in such applications. Previous work uses expected support as the measurement of pattern frequentness, which has inherent weaknesses with respect to the underlying probability model, and is therefore ineffective for mining high-quality sequential patterns from uncertain sequence databases. In this work, we propose to measure pattern frequentness based on the possible world semantics. We establish two uncertain sequence data models abstracted from many real-life applications involving uncertain sequence data, and formulate the problem of mining probabilistically frequent sequential patterns (or p-FSPs) from data that conform to our models. Based on the prefix-projection strategy of the famous EARM algorithm, we develop two new algorithms, collectively called U-EARM, for p-FSP mining. UEARM effectively avoids the problem of “possible world explosion”, and when combined with our three pruning techniques and one validating technique, achieves good performance. The efficiency and effectiveness of EARM are verified through extensive experiments on both real and synthetic datasets. Example,. In this work, we propose to define pattern frequentness based on the possible world semantics. We develop two uncertain sequence data models (sequence-level and element-level models) abstracted from many real-life applications involving uncertain sequence data, based on which we define the problem of mining probabilistically frequent sequential patterns (or p-FSPs). We now introduce our data models through the following examples.

Consider a wireless sensor network (WSN) system, where each sensor continuously collects readings about environmental conditions, such as temperature and humidity, within its detection range. In such a case, the readings of a sensor are inherently noisy, and can be associated with a confidence value determined by, for example, the stability of the sensor. A possible set of readings from a WSN application that monitors the temperature. Let us assume that each sensor reports temperature ranges, and reading B represents and a new reading is appended to the sequence of already reported readings whenever the temperature range changes. We also assume that each region is associated with a group of sensors. For example, s1 is the reading sequence detected by a sensor in one region within a time period, and s21 and s22 are the reading sequences detected by two sensors in another region within that time period.

4. RADIO FREQUENCY IDENTIFICATION (RFID) TECHNOLOGIES

Radio Frequency Identification (RFID) technologies are used in many

applications for data collection. However, raw RFID readings are usually of low quality and may contain many anomalies. An ideal solution for RFID data cleansing should address the following issues. First, in many applications, duplicate readings (by multiple readers simultaneously or by a single reader over a period of time) of the same object are very common. The solution should take advantage of the resulting data redundancy for data cleaning. Second, prior knowledge about the readers and the environment (e.g., prior data distribution, false negative rates of readers) may help improve data quality and remove data anomalies, and a desired solution must be able to quantify the degree of uncertainty based on such knowledge. Third, the solution should take advantage of given constraints in target applications (e.g., the number of objects in a same location cannot exceed a given value) to elevate the accuracy of data cleansing. There are number of existing RFID data cleaning techniques. However, none of them support all the aforementioned features. In this work we propose a Bayesian inference based approach for cleaning RFID raw data. Our approach takes full advantage of data redundancy. To capture the likelihood, we design an *n*-state detection model and formally prove that the 3-state model can maximize the system performance. Moreover, in order to sample from the posterior, we devise a Metropolis-Hastings sampler with Constraints (MH-C), which incorporates constraint management to clean RFID raw data with high efficiency and accuracy. We validate our solution with a common RFID application and demonstrate the advantages of our approach through extensive simulations. In this work, we propose an innovative approach of cleaning RFID raw data which is able to take full advantage of duplicate readings and integrate prior knowledge as well as environmental constraints. We demonstrate the efficiency and effectiveness of our (Basiyan Inference) approach by comparing the performance of MH-C with the Sequential Importance Sampling (SIS) based solution through extensive simulations.

5. FREQUENT PATTERN MINING WITH UNCERTAIN DATA

This work studies the problem of frequent pattern mining with uncertain data. We will show how broad classes of algorithms can be extended to the uncertain data setting. In particular, we will study candidate generate-and-test algorithms, hyper-structure algorithms and pattern growth based algorithms. One of our insightful observations is that the experimental behavior of different classes of algorithms is very different in the uncertain case as compared to the deterministic case. In particular, the hyper-structure and the candidate generate-and-test algorithms perform much better than tree-based algorithms. This counter-intuitive behavior is an important observation from the perspective of algorithm design of the uncertain variation of the problem. We will test the approach on a number of real and synthetic data sets, and show the effectiveness of two of our approaches over competitive techniques.

One observation from our extensions to the uncertain case is that the respective algorithms do not show similar trends to the deterministic case. For example, in the deterministic case, the FP-growth algorithm is well known to be an extremely efficient approach. However, in our tests, we found that the extensions of the candidate generate-and-test as well as the hyper-structure based algorithms are much more effective. Furthermore, many pruning methods, which work well for the case of low uncertainty probabilities, do not work very well for the case of high uncertainty probabilities. This is because the extensions of some of the algorithms to the uncertain case are significantly more complex, and require different kinds of trade-offs in the underlying computations. Thus, in addition to the new efficient methods proposed by this work, an

important contribution of this work is the insight that natural extensions of deterministic algorithms may show counter-intuitive behavior. This work is organized as follows. The next section defines the uncertain version of the problem. We will also discuss the extension of candidate generate-and-test algorithms to the uncertain version of the problem. The remainder of the work discusses the extension of other classes of algorithms, and provides comparative experimental results.

Computing statistical information on probabilistic data has attracted a lot of attention recently, as the data generated from a wide range of data sources are inherently fuzzy or uncertain. In this work, we study an important statistical query on probabilistic data: finding the frequent items. One straightforward approach to identify the frequent items in a probabilistic data set is to simply compute the expected frequency of an item and decide if it exceeds a certain fraction of the expected size of the whole data set. However, this simple definition misses important information about the internal structure of the probabilistic data and the interplay among all the uncertain entities. Thus, we propose a new definition based on the possible world semantics that has been widely adopted for many query types in uncertain data management, trying to find all the items that are likely to be frequent in a randomly generated possible world. Our approach naturally leads to the study of ranking frequent items based on confidence as well. Finding likely frequent items in probabilistic data turn out to be much more difficult. We first propose exact algorithms for offline data with either quadratic or cubic time. Next, we design novel sampling-based algorithms for streaming data to find all approximately likely frequent items with theoretically guaranteed high probability and accuracy. Our sampling schemes consume sub linear memory and exhibit excellent scalability. Finally, we verify the effectiveness and efficiency of our algorithms using both real and synthetic data sets with extensive experimental evaluations. The exact algorithms perform as indicated in our analysis, either with quadratic, or cubic time complexity depending on x-tuples have single item or multiple items. To produce data sets with single-item x-tuples, for each of the data sets described above, we simply retain only the first item in each x-tuple and the running time respectively. Clearly, the exact algorithms are costly and do not scale when data sets increase, although they are able to return exact results. The wcday46 data set is the most expensive due to high number of unique items. We also observe that the pruning lemma indeed dramatically reduces the cost. Further illustrates the effectiveness of the pruning lemma, where for skewed data sets, more than 90% of the items are pruned. Obviously, the higher φ and τ are, the more items will be pruned. We also reported the memory usage of the two exact algorithms, clearly demonstrating either a linear or quadratic trend.

6. MINING METHODS AND ALGORITHMS:

In recent years, a number of indirect data collection methodologies have led to the proliferation of uncertain data. Such databases are much more complex because of the additional challenges of representing the probabilistic information. In this work, we provide a survey of uncertain data mining and management applications. We will explore the various models utilized for uncertain data representation. In the field of uncertain data management, we will examine traditional database management methods such as join processing; query processing, selectivity estimation, OLAP queries, and indexing. In the field of uncertain data mining, we will examine traditional mining problems such as frequent pattern mining, outlier detection, classification, and clustering. We discuss different methodologies to process and mine uncertain data in a variety of forms.

A given query over an uncertain database may require computation or aggregation over a large number of possibilities. In some cases, the query may be nested, which greatly increases the complexity of the computation. There are two broad semantic approaches used: Intentional semantics. This typically models the uncertain database in terms of an event model, and use treelike structures of inferences on these event combinations. This tree-like structure enumerates all the possibilities over which the query may be evaluated and subsequently aggregated. The tree-like enumeration results in an exponential complexity in evaluation time, but always yields correct results, Extensional semantics. Extensional semantics attempts to design a plan which can approximate these queries without having to enumerate the entire tree of inferences. This approach treats uncertainty as a generalized truth value attached to formulas, and attempts to evaluate the uncertainty of a given formula based on that of its sub formulas.

MODULE DESCRIPPTION:

1. CANDIDATE ITEM SET VALIDATION:

It has following steps,

Candidate generation: In the first iteration, size-1 item sets that can be 1-PFIs are obtained, using the PFIs discovered from D, as well as the delta database d. In subsequent iterations, this phase produces size (k + 1) candidate item sets, based on the k-PFIs found in the previous iteration. If no candidates are found, then the process halts.

Candidate pruning: With the aid of d and the PFIs found from D, this phase filters the candidate item sets that must not be a PFI.PFI testing- For item sets that cannot be pruned, they are tested to see whether they are the true PFIs. This involves the use of the updated database, as well as the spmfs of PFIs on D. The incremental mining can also applied be applied using the tree based approach. FP-tree is used to store the transaction data .whenever an increment is applied to the original database the tree is updated with the checking of frequent item sets in the new additional transactions. The major problem with fixed width bit strings is that they are not efficient representations at lower levels of the enumeration tree at which only a small number of items are relevant, and therefore most entries in these bit strings are 0. This module is to speed this up is to perform the item-wise projection only at selected nodes in the tree, when the reduction in the number of items from the last ancestor at which the item-wise projection was performed is at particular multiplicative factor.

2. FREQUENT PATTERNS IN GRAPHS AND STRUCTURED DATA:

The sliding window model is used here. The sliding window should be divided into two sub-windows. The entire window is denoted as ‘w’ and the sub-windows are ‘w0’ and ‘w1’. The sub-windows should be partitioned dynamically based on the inputs. it can derive all frequent induced sub graphs from both directed and undirected graph structured data having loops (including self-loops) with labeled or unlabeled nodes and links. Its performance is evaluated through the applications to Web browsing pattern analysis and chemical carcinogenesis analysis to avoid the problem of numerous database scans and candidate generate –and-test process. The corresponding algorithm is called FP Growth Algorithm. To obtain the information about the database, it requires two scans only. Frequent patterns are mined from the tree structure, since contents of the database are captured in a tree structure. Specifically, FP-growth starts by scanning the database once to find all frequent 1-itemsets. Afterwards, the algorithm makes a ranking table, in which items appear in descending frequency order.

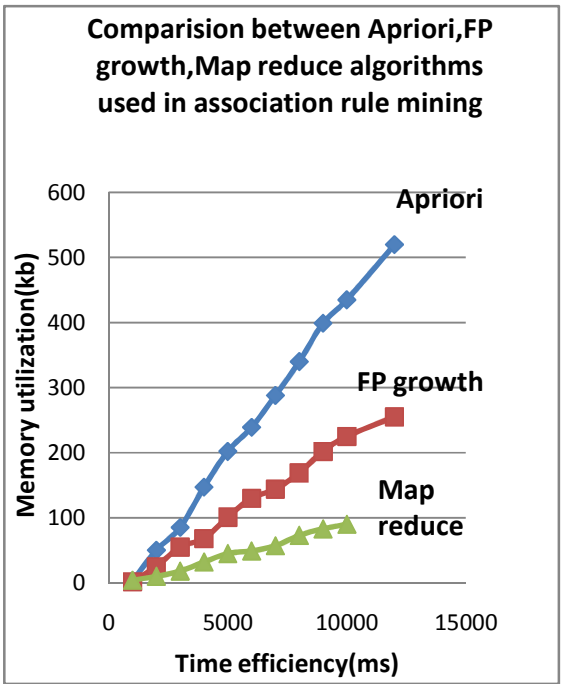
3. SEQUENTIAL PATTERN MINING:

It is to generate approximation count by dividing the database in a number of non-overlapping segments. After the first database scan, item set that are frequent locally in each segment can be found. For an item set to be globally frequent in the database, it must be locally frequent item set in at least one partition (or segment). So, after gathering all local frequent item set, the Partition algorithm scans the database for the second and last time to check which of those local frequent item set are actually frequent globally in the whole database. As a result, this technique reduces drastically the number of scans needed by Apriori-based algorithms to only two. So, Partition algorithm always depends on the data distribution and the number of segments. As the database is scanned, this counter is updated by subtracting the corresponding "over-estimate" for each item in the pattern. If the counter gets below the minimum support, any pattern containing that item cannot be frequent and hence can be pruned. DP with its two improvements is a very effective technique and it improves both runtime and memory requirements of EARM. Even though it is still bounded by generate and test approach limitations, the application of the decremental technique (known as U-Prefix Span algorithm) is a reasonable Apriori-based adaptation for uncertain data.

4. DATA TREE PROJECTION:

It is to generate the data report as Tree Structure. By using this structure, the algorithm tries to improve the mining time. Once the H-struct (Hybrid Tree Structure) is constructed, the EARM algorithm just needs to maintain and update the numerous links that point from one transaction to the next that contains the same set of items. Since EARM keeps all transactions that contain frequent items in memory, there is no need to read the database more than once. From that point on, all information is extracted from the H-struct. EARM outperformed Apriori by finding frequent patterns quicker and requiring less memory than FP-growth, especially with small minimum support threshold.

EFFICIENCY GRAPH:



CONCLUSION:

Proposed concept is to study the problem of mining probabilistically frequent sequential patterns (p-FSPs) in uncertain databases. Our study is founded on two uncertain sequence data models that are fundamental to many real-life applications. We propose two new EARM algorithms to mine frequent pattern from data that conform to our sequence level and element-level uncertain sequence models. We also design three pruning rules and one early validating method to speed up pattern frequentness checking. These rules are able to improve the mining efficiency. To further enhance the algorithmic efficiency, we devise two approximation methods to verify the probabilistic frequentness of the patterns based on Poisson and Normal distributions. The experiments conducted on synthetic and real datasets show that

our two EARM algorithms effectively avoid the problem of "possible world explosion" and the approximation methods PA and NA are very efficient and accurate.

REFERENCES:

[1] M. Muzammal and R. Raman, "Mining sequential patterns from probabilistic databases," in *Proc. 15th PAKDD*, Shenzhen, China, 2011.

[2] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *Proc. 13th ACM SIGKDD*, San Jose, CA, USA, 2007.

[3] D. Tanasa, J. A. López, and B. Trousse, "Extracting sequential patterns for gene regulatory expressions profiles," in *Proc. KELSI*, Milan, Italy, 2004.

[4] J. Pei *et al.*, "EARM: Mining sequential patterns efficiently by prefix-projected pattern growth," in *Proc. 17th ICDE*, Berlin, Germany, 2001.

[5] R. Agrawal and R. Srikant, "Mining sequential patterns," in *Proc. 11th ICDE*, Taipei, Taiwan, 1995.

[6] M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," *Mach. Learn.*, vol. 42, no. 1–2, pp. 31–60, 2001.

[7] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M. C. Hsu, "FreeSpan: Frequent pattern-projected sequential pattern mining," in *Proc. 6th SIGKDD*, New York, NY, USA, 2000.

[8] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *Proc. 5th Int. Conf. EDBT*, Avignon, France, 1996.

[9] Z. Zhao, D. Yan, and W. Ng, "Mining probabilistically frequent sequential patterns in uncertain databases," in *Proc. 15th Int. Conf. EDBT*, New York, NY, USA, 2012.

[10] C. Gao and J. Wang, "Direct mining of discriminative patterns for classifying uncertain data," in *Proc. 16th ACM SIGKDD*, Washington, DC, USA, 2010.

[11] N. Pelekis, I. Kopanakis, E. E. Kotsifakos, E. Frentzos, and Y. Theodoridis, "Clustering uncertain trajectories," *Knowl. Inform. Syst.*, vol. 28, no. 1, pp. 117–147, 2010.

[12] H. Chen, W. S. Ku, H. Wang, and M. T. Sun, "Leveraging spatiotemporal redundancy for RFID data cleansing," in *Proc. ACM SIGMOD*, Indianapolis, IN, USA, 2010.

[13] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein, and W. Hong, "Model-driven data acquisition in sensor networks," in *Proc. 13th Int. Conf. VLDB*, Toronto, ON, Canada, 2004.

[14] L. Sun, R. Cheng, D. W. Cheung, and J. Cheng, "Mining uncertain data with probabilistic guarantees," in *Proc. 16th ACM SIGKDD*, Washington, DC, USA, 2010.

[15] C. C. Aggarwal, Y. Li, J. Wang, and J. Wang, "Frequent pattern mining with uncertain data," in *Proc. 15th ACM SIGKDD*, Paris, France, 2009.

[16] Q. Zhang, F. Li, and K. Yi, "Finding frequent items in probabilistic data," in *Proc. ACM SIGMOD*, Vancouver, BC, Canada, 2008.

[17] T. Bernecker, H. P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic frequent itemset mining in uncertain databases," in *Proc. 15th ACM SIGKDD*, Paris, France, 2009.

[18] C. K. Chui, B. Kao, and E. Hung "Mining frequent itemsets from uncertain data," in *Proc. 11th PAKDD*, Yichang, China, 2007.

[19] C. C. Aggarwal, and P. S. Yu, "A survey of uncertain data algorithms and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 5, pp. 609–623, May 2008.

[20] J. Yang, W. Wang, P. S. Yu, and J. Han, "Mining long sequential patterns in a noisy environment," in *Proc. ACM SIGMOD*, Madison, WI, USA, 2002.

[21] S.Ramkumar, K.Sathesh Kumar," Querying Unintelligible Data on Geospatial Trajectory Database", International Journal of Scientific Research in Science and Technology (IJSRST), Vol. 2(4), pp.180-187, Aug-2016.

[22] S.Ramkumar and G.Ganesaperumal," A Feasibility Study on Big Data Integration and its Methodologies for Hadoop Techniques Using Map Reduce Model", International Journal of Modern Trends in Engineering and Research, Vol. 3(9), pp. 230-238, Sep-2016.

[23] S.Ramkumar, K.Sathesh Kumar," A Personalized Scheme For Incomplete And Duplicate Information Handling In Relational Databases", International Journal Of Engineering Sciences & Research Technology, Vol. 5(9), pp.360-369, Sep-2016

[24] V.Vasanthi , S.Akram Saeed Aglan Alhammadi, S.Ramkumar and Sathish Kumar," Achieving security for data access control using cryptography techniques", International Journal of Scientific Research in Science and Technology (IJSRST), Vol. 3(5), pp.172-182, Aug-2016.

[25] S.Ramkumar, K.Sathesh Kumar, A Hybrid Approach for Horizontal Aggregation Function Using Clustering", International Journal of Computer Science and Network, Vol. 6(5), pp. 551-558, Oct-2017.