A NOVEL TECHNIQUE FOR DUPLICATE AND INCOMPLETE INFORMATION HANDLING IN LARGE DATABASES

M. Janaki [1] T.Sumadhi[2]

ABSTRACT

Missing data replacement is a crucial process in most real world databases. Due to the tremendous improvement of data management, users of such database can effectively manage the incompleteness using their customized policies. This paper proposes the renovated concept of partial information handling with complete prediction model, which can handle incompleteness in relational databases. The incomplete data management brings a new challenge which is the data duplication. The Customized Information Prediction Policies with effective index method has been proposed in this paper for handling missing data. Different users in the real world have different ways in which they want to handle incompleteness. The CIP operators suggest the best match to replace the null value, and this also allows them to specify a policy that matches their attitude to risk and their knowledge of the application. Using the same strategy DIP operators has been introduced to handle duplicate data's in the relational database. Using the Autoregressive HMM the system improves the prediction method. The CIP manages all data and policies using PQ_ Index structures, which is known as Priority Queue based Index. The proposed work also studies how relational algebra operators and PIP operators interact with one analyzes another. This also handles the COALESCE function using the CIP operator.

Keywords: Knowledge personalization and customization, database semantics

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. The obtained information can be used to increase revenue, cuts, costs, or both. There are a number of data analysis tools available in the market. Data mining uses a number of analytical tools for analyzing data from different scope, classify it, and summarize the relationships among the data [1]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. The related terms like data phishing, data snooping and data dredging refers to the use of data mining techniques to group parts of a larger population data set that are too small for reliable statistical inferences to be made about the validity of any patterns to be discovered. These techniques can, however, be used in creating new hypotheses to test against the larger data populations [2].

1 Research scholar Dept. of Computer Science Karpagam University, mdjanaki@gmail.com

[2]Associate Professor, Dept.of Computer Science, Karpagam University, t_sumathi@yahoo.co.in

In this work, an enhanced operator for partial data handling and de-duplication using the linkage method. This also aims at performing fully customized policies for partial information handling and duplicate information handling and also matches data in different types.

The main goal of this work is to develop the necessary operator and implement these operators with the database support.. So, end users can bring both their application specific knowledge as well as their personalized risk to bear when resolving inconsistencies. This work proposes a new missing data handling operator with duplicate detection method which aims at performing null value replacement and duplicate entry detection along with the autoregressive HMM process for complete data analysis.

A. Missing value in database can be filling by three ways

Aggregate PIP operator: candidate values are aggregated by means of an aggregate operator v specified as a parameter to the PIP operator so that a single value v can be determined. Finally, every occurrence of the considered null is replaced by v in the result. The group of single-valued PIP operators defined below fundamentally vary in the way that candidate values are resoluted.

Regression oriented PIP operator: When a regression oriented PIP operator is evaluated, the loss of duplicates does not change the set of data used to build the regression model.

PIP operator base on another attribute: A PIP operator based on another attribute is of the form $\rho^{agg}(u, v, A, B, X)$. Intuitively, the basic idea of this family of PIP operators is the following: if the A-value of a tuple t is a null, an aggregate operator μ is applied to the B-values of the tuples having the same X-value as t so that a value g is determined; then, only those tuples having B-values closest to are considered and their A-values are used to determine a candidate value (this is done by applying an aggregate operator u). If several candidate values exist, one is chosen according to the third parameter v.

II. Literature Review

This section deals with various concepts and ideas proposed researchers for effectively handling missing data and duplicate values. A number of methods have been developed for dealing with missing data on continuous variables among which most relevant concepts have been listed for comparison with the proposed system.

Finch, W. Holmes, et al [3] in his work has analysed the problem of missing data in the situation of questionnaires. The respondents which do not respond to one or more items, they were in making the conduct of statistical analyses, as well as the calculation of scores were difficult. Many approaches has been developed, but they are not clear that these techniques for imputation are appropriate for the categorical items. This concept compares the performance of these explicitly categorical imputation approaches with the better established continuous method used with categorical item responses. This approaches prior studies on missing data with normally

distributed data; it seems clear that ignoring the missing values is inappropriate, whether the data are MCAR or MAR. In both cases, the standard errors and results of hypothesis tests for the slopes varied from the complete data case to a greater extent than the results for any of the imputation methods.

Badia, Antonio et al [4], has proposed a set of properties that any extension of functional dependencies over relations with null markers should possess. Two new extensions attempt to allow null markers where they make sense to practitioners. In this work the concept of FD is analysed in the presence of null markers, and have specified a set of properties that any definition of the concept should satisfy. And at the same time allows those null markers to be updated to real values consistently. These FDs can also be enforced efficiently in computational terms despite null markers. Both definitions have slightly different properties (LFDs enforce lossless join in addition to database consistency), but they both have satisfied their proposed axioms.

Calý, Andrea, et al [5] has proped the set of issue that deals with integrity constraints over the global schema in data integration. The integrity constraints can be used to extract more data from incomplete information. The problem of combining the data residing at different sources, and providing the user with a unified view of these data, called global (or mediated) schema, over which queries to the data integration system are expressed is known as Data integration. On the other hand, integrity constraints raise the problem of dealing with the discrepancy of the whole system, due to ambiguous data at the sources. The presence of such constraints in the global schema blurs the distinctions between GAV and LAV, even of a simple form, raises the need of dealing with incomplete information and possibly with inconsistencies.

Wong, Eugene [6] addresses the problems related to preliminary analysis. The existence of replicated data, especially in distributed systems, suggests the use of redundancy to reduce uncertainty. There are numerous situations in which a database cannot provide a precise and unambiguous answer to some of the queries. The problems analysed measurement and recording errors, missing data, incompatible scaling, obsolescence, and data aggregation of one kind or another. However, the cost of using more than one copy is large and must be kept to a minimum by strategies provided by sequential analysis. Another issue concerns how the a priori distribution information is to be acquired. In some cases it must be done empirically by sampling.

Cao, Jianjun, et al [7]. has proposed a statistical relational learning approach for estimating and replacing missing categorical data. Categorizing ,ordering and its estimation is done using hidden markov model. According to complete record samples, probabilities of missing value belonging to each possible value are estimated by the model. The missing value can be replaced through referring to the probabilities.

Chiu and Sedransk [8] has proposed a Bayesian method for estimating and missing data replacement based on some prior knowledge about the distributions of the data.

This proposed method uses univariate and multivariate cases for estimating missing data has been performed using uniform prior distribution and a Dirichlet posterior distribution. Their method performed very well when the missing data is missing at random, but it remains to be tested for cases where data is missing not at random.

Li [9] proposed a simple Bayesian approach for estimating and replacing missing categorical data. With this approach, the posterior probabilities of a missing attribute value belonging to a certain category are estimated. The approach is nonparametric and does not require prior knowledge about the distributions of the data. However, when the approach estimates missing values of any empty field, it must use all the other un-missing categorical values, and those fields which are irrelevant to the empty field are also included. For relational data, the hypothesis that the attributes are conditionally independent of each other under a given class value, is a basic precondition for computing estimate value. But this hypothesis lacks reasonable support.

Martinez et al [9] has studied Inconsistency management policies and allow a relational database user to express customized ways for managing inconsistency according to their need. For each functional dependency, a user has a library of applicable policies, each of them with constraints, requirements, and preferences for their application that can contradict each other. The problems addressed in this work is that of determining a subset of these policies that are suitable for application with reference to the set of constraints and user preferences.

A classical logic argumentation-based solution, which is a natural approach given that integrity constraints in databases and data instances are, in general, expressed in first order logic (FOL). An automatic argumentation-based selection process allows retaining some of the characteristics of the kind of reasoning that a human would perform in this situation.

Table 1: Merits and demerits of the existing system

Method	Description	Disadvantage
Single-Value Imputation	Replacing Missing values with a single value changes the distribution of that variable by decreasing the variance that is likely present.	Not suitable for different type of data. Not suitable when applying for multivariate parameters such as regression coefficients.
Recommendations Concerning Widely Used Methods(complete case analysis)		1. Properties are dependent on the magnitude of the correlations. 2. Applications are limited. 3. Invalid estimates can occur.
Complete-Case Analysis	Easy to implement. This segments the data into 2 groups, which are missing data and complete data. Missing data will be replaced by the complete data.	failed when there are large amounts of missing data
Available Case Analysis	Available case analysis, or pair wise deletion, uses all available data to estimate parameters of the model. Uses means and variances	Failed to focus on bi-variate or multivariate relationships
Model-Based Methods for Multivariate Normal Missing Data (Assumption-based) (EM Algorithm, and multiple imputation)	The major advantage of these two methods is that given the assumptions, the results obtained apply to a broader range of contexts with fewer conditions than the methods of the previous section.	Require more complex computations
Maximum Likelihood Methods Using the EM Algorithm	One method for estimating unknown parameters of a model is the use of maximum likelihood.	Does not result in values for individual missing variables.
Multiple Imputation with	Multiple imputation avoids two	This does not handle the

III. Proposed System

The work proposes a unified framework for analysis on incomplete data and duplicate data using special operators, which captures existing approaches as a special

case and provides an easy basis for the proposed system. The preliminary review introduces a new approach named as customized information prediction policies by which the users can get suggestion about the missing data in a particular database, considering their own knowledge of how the data was collected, their attitude to risk, and their mission needs the missing data has been replaced using the CIP and DIP operators.

Contributions in this proposed work:

The current work proposes two new operators named as CIP and DIP operators for resolving different kinds of incompleteness problems and data redundant elimination, and this gives several useful and spontaneous illustrations of CIP and DIP operators.

An effective index structures, to support the efficient evaluation of CIP operators and show how to maintain them incrementally as the database is updated.

The proposed method also concentrat on the duplicate information elimination policies using the linkage methods.

The system effectively utilizes the autoregressive HMM model for incomplete information handling.

This performs a study using the interaction between classical relational algebra operators and PIP operators.

The experiment all results obtained shows the effectiveness of the proposed index structures with a real-world airline and census data set. Specifically, the comparison with an algorithm exploiting the index structures with priority queue with normal index and a naive one not relying on them and shows that the former greatly outperforms the latter and is able to manage very large databases.

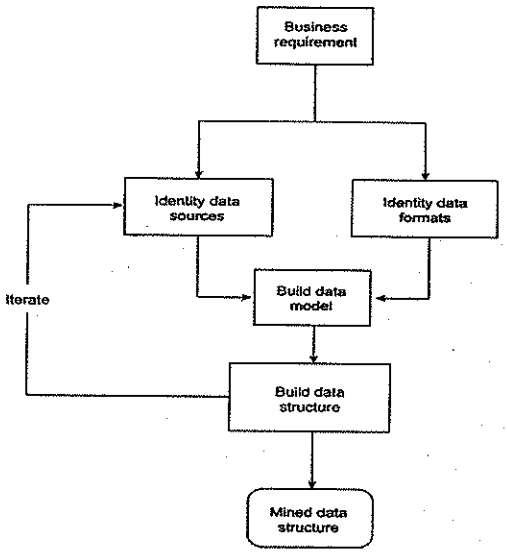

Figure1:Framework of the proposed work

IV. Methodologies

This section clearly defines the CIP operator implementation (customized information prediction policy) **with the autoregressive HMM model for accurate partial information handling when data linkage is performed. The system additionally implements the** DIP

operator **which is named as** Duplicate information handling policies **which is blended with the PIP Coalesce function which is a null handling function in SQL.**

A.Cip Operators

This section introduces CIP operators which allow users to make replacement about missing data in a database, taking into account their customized policies and existing knowledge of how the data has been collected, their attitude to risk, and their mission needs. A CIP operator maps a database to a subset of its completions that this call preferred completions. The CIP operator gives the approximation data to the user based on the prediction score.

Initially the system performs the following; the complete DB will be obtained by replacing every unknown value with an actual value using the CIP operator. Every user can express the preferred attribute to replace the data. The system will replace the data based on the given attribute.The completions chosen as preferred by the user and the set of attribute lists which are those where each unknown data is replaced with a value determined by linear regression technique; if other user specified any other attribute, then the system replaces the value based on the current user requirement.

The CIP operator performs the following steps to perform this operation.

Algorithm: 1 (CIP for Null replacement) Input: set of Constraints (C) , database D

Steps:

1. Read the database D. Find every attribute list (A_i) in the D.
2. Get Constraints C1,C2....Cn from the user constraint list C.
3. Find the Null values and missing values (M) from the D.
4. For each Null value (Mi from M) do

a. Find the attribute Ai for Mi.

b. For every Ai do the step i.

i. Perform the AHMM with the above database D.

ii. Given the AHMM , = (C1, C2, .Cn), What is the probability of generating a specific observation sequence Ai=*fA*1, *A*2, ... , *An*

iii. Add into the index with score.

5. Find priority Queue $_{PQ}$ for every indexed item from step 4.b.i
6. Perform statistics method for Ai to analyze the null value N

a. Compute P={Ai,C,μ}- where μ is the statistical function and P is the score value.

7. If (avg(Ai) ==P) find the Attribute value and do the step 8. Else do step 9

8. Replace the null value by P.

9. Replace the null value by avg(Ai).

10. Update the database (D).

The above algorithm 1describes the process included in the CIP operator. This effectively applies the priority queue on index and AHMM model for CIP operator.

For data prediction, forecasting or null value reduction in large data environment, linear regression model has been used. The current predictive model observed data set of A1 and *A2* values. And to finds the data and checks specified constraint. If an additional value of *A1* is then given without its accompanying value of *A2*, the fitted model can be used to make a prediction of the value of *A2*.

The null value replacement in certain applications requires items having keys in a certain order, however not necessarily in full sorted order and not necessarily all at once at a time. Frequently, the system collects a set of items and processes the one with the high priority value. The appropriate data type in this case supports two type of operations one is to *remove the maximum* and another one is to perform *insert operation*. This kind of process is known as priority queue. The CIP operator has been developed using the priority queue for evolving the data effectively.

The common idea in the CIP operators is the following. Each PIP operator tries to fill in unknown and no-information nulls appears in attribute A (which is one of the parameters of the CIP operator). Since a relation can contain multiple occurrences of the same unknown or null value, then the algorithm finds different candidate values for it. Those attribute values are aggregated by means of an aggregate operator; the aggregator operators such as avg(), min(), max(), count(), sum() etc., these operators are specified as a parameter of the CIP operator so that a single value P is established. At last, every occurrence of the considered null is replaced by v in the result.

B. Autoregressive Hidden Markovmodels

Autoregressive hidden Markov model is a combination of autoregressive time series and hidden Markov chains. Observations are generated by a few autoregressive time series while the switches between each autoregressive time series are controlled by a hidden Markov chain. A time series may sometimes consist of observations generated by different mechanisms at different times. When this happens, the time series observations would act like switching back and forth between couple of distinct states. When changing into a different state, the time series may have a significant change in their means or in their frequencies or breadths of their fluctuations. The *Autoregressive Hidden Markov model* (*ARHMM*) is often being used to deal with this kind of time series. As

indicated by the name, an ARHMM is the combination of an autoregressive time series model and a hidden Markov model. The autoregressive structure admits the existence of dependency amongst time series observations while the hidden Markov chain could capture the probability characteristics of the transitions amongst the underlying states. Actually, ARHMM is also referred as *time series with change in regime* (or *states*) by the econometricians.

To be more specific, let us see an example of ARHMM. As usual, Y = *{Y1, Y2... YT}* denote the observation sequence. Each Y_t is a observation vector with k component $Y_t = \{y1, y2 \ldots y_k\}'$.

X = $\{X_1, X2 \ldots X_T\}$ is a hidden state sequence with *N* possible states. X is assumed to be a Markov chain with transition matrix $A = [aij]$ and initial distribution vector.

But it should be mentioned that the ARHMM with heteroskedasticity (unequal variance) for distinct state *Xt* could also be developed with more complexity. In such cases, the error term *"t* will usually be replaced by *"Xt* which depends on the value of current state *Xt*.

E-M algorithm or segmental K-mean algorithms could only lead to a local maximum of the HMM likelihood function. For ARHMM, this is also true.

To get the parameter estimates with a global maximum likelihood, a grid search approach might be used. In grid search approach, the parameter space is seen as a grid with many small cells and all the vertices are used as the initial values of the parameters. Because the parameter space is so big in the case of ARHMM, the grid search method requires considerable computational power which is intractable for practical purposes.

Another notable feature of ARHMM estimation is the high autoregressive coefficients. Conventional HMM assumes that there are independency relations between the observations. But this is rarely the case for time series observations. As in this application, SST data are collected on a day-by-day basies and apparently the independency assumption is inappropriate. Comparatively, the autoregressive structure contributes the superiority of ARHMM in a way it prevents the frequent fluctuations of state path. Conventional HMM are very sensitive to the numerical swings of the current SST and hence mistakes several fluctuations of SST as the switches of states. While for the same data, ARHMM state path are more stable and close to reality.

C. Dip Operator

Duplicate information policy is the task of identifying the duplicate database records. In relational databases, accurate duplicate record finding is often dependent on the merge decisions made for records of other types. The above CIP operator helps to replace the null value by applying the most appropriate value from the database, but the database quality may reduced due to the duplicate entries in the database. All previous approaches have merged records of different types independently and replace the null values; this work facilitates these inter-dependencies explicitly to collectively de-duplicate

records of multiple data types with null value replacement. This effectively finds the duplicate entries based on the customized policies. The DIP operator performs the splitting score values to evaluate the similarity.

Algorithm 2: DIP

Input: set of Constraints (C) , database D, initial score value V

Steps:

1. Read the database D1 and D2. Find every attribute list (A_i) in the D1 and D2.
2. Get Constraints C1,C2....Cn from the user constraint list C.
3. Get initial score from default table D1 and D2.
4. For each attribute Ai do
5. Calculate statistics functions for each attribute and every rows.

a. Mean(Ai)={Ai1,Ai2....Ain} /no of items

b. Median (Ai)={Ai1,Ai2....Ain} and so on.

6. Find the best score for matching two instances
7. Apply the matching function MF(tuple1, tuple2)
8. Perform step 7 until EOF.
9. Highlight tuple t which having same data
10. Return t

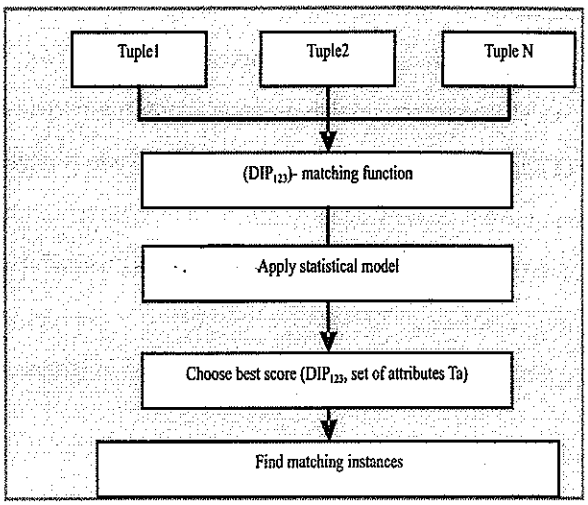

Figure 2: Framework of DIP architecture

The system performs the following steps for data linkage using DIP which is illustrated in figure 1. The system initially calculates the score with certain statistical parameters. Once the model is built (based on attributes from tuple T1), each item holds a data set containing the matching records from tuple t2. To create a compact representation of the DIP operator and for it to be more generalized each item is represented by a set of probabilistic models.

V. Result And Discussions

To compare the proposed system with the existing schemes some parameter lite, Naive, Index and PQ_ Index several data set and attribute has been implemented. The data set includes the airline records and census dataset. The proposed system concentrated on the PQ_ index scheme with CIP operator. The proposed system also provides an effective way to identify the duplicate tuples from the database tables. The system effectively

implements the AHMM for improving the accuracy of the prediction model.

A. Performance evaluation

In this section measure the performance of the existing PIP_Index then measure the results of the CIP_PQIndex. The efficiency is improved in the CIP with the use of PQIndex. First compared the times taken by the naïve, Index and PQ_index based approaches to evaluate a CIP operator. The varied the size of the DB up to 15o thousand tuples and the 10 percentage of rows with a null value by randomly selecting tuples and inserting nulls (of different kinds) in them.

Figure 2 shows the performance measure based on the evaluation delay and the proposed approach CIP took less time while comparing the existing PIP method. From the fig 3 shows the performance measure based on the accuracy of detected value and the proposed approach CIP took less time while comparing the other methods. From the Fig 4 shows Performance verification of proposed DIP and CIP using PQ_Index based on the processing delay. And also describes the time taken for both DIP and CIP

Figure 2, 3 and 4 represents Performance comparison of proposed CIP using PQ_Index with existing approaches based on the Evaluation Delay, Performance comparison of proposed CIP using PQ_Index with existing approaches based on the Result accuracy and Performance verification of proposed DIP and CIP using PQ_Index based on the processing delay. The chart describes the time taken for both DIP and CIP. The proposed system has been evaluated based on several metrics like evaluation delay, processing delay, accuracty, insertion and deletion time and execution time is given in the following tables below.

Table 2: Evaluation based on delay

Evaluation Delay		
Time(sec)	PIP	CIP
30	3	2
60	4	3
90	5	4
120	7	5
150	9	7

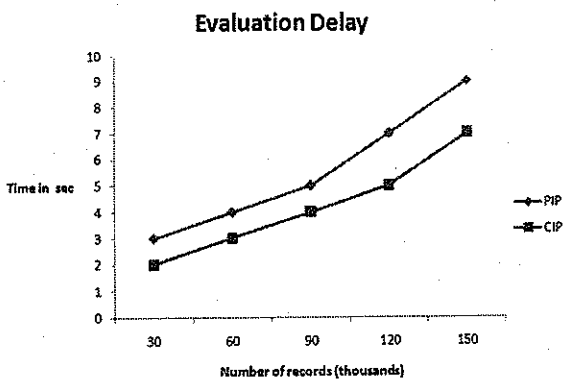

Figure 3: chart showing evaluation delay

Table 3: Evaluation based on Accuracy

Accuracy	PIP_Index	CIP_PQIndex
Percentage (%)	85	92

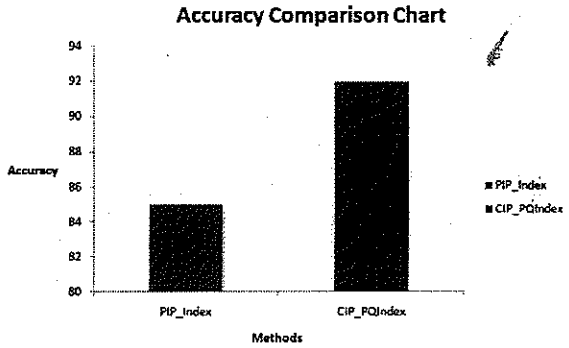

Figure 4: chart showing accuracy

Table 4: comparison based on Processing delay

Processing Delay		
Time (m.sec)	DIP	CIP
100	3	2
200	4	3
300	5	4
400	7	5
500	9	7

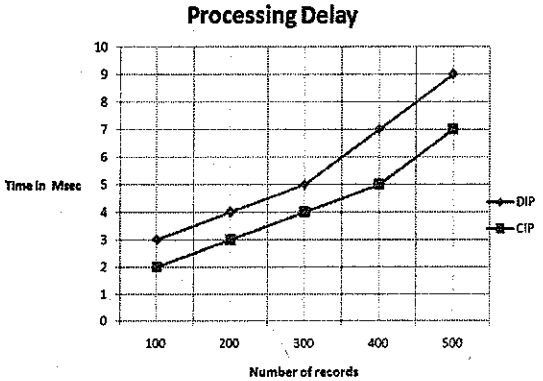

Figure 5: chart showing processing delay

Table 5: Insertion time comparison

Time to execute Tuple insertions(m.sec)			
DB Size	**Naïve**	**Index**	**PQ_Index**
125,000	9	7	5
250,000	11	9	8
500,000	13	11	10
750,000	15	13	12
1,000,000	19	17	15

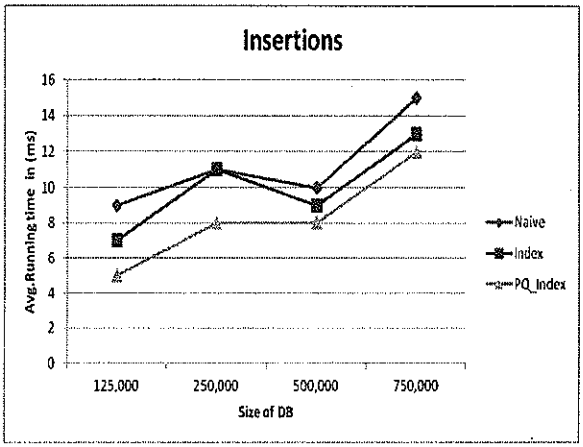

Figure 6 : Chart showing insertion time

Table 6: Deletion time comparison

Time to execute Tuple Deletion			
DB Size	**Naïve**	**Index**	**PQ_Index**
125,000	9	10	4
250,000	8	9	6
500,000	13	12	8
750,000	17	15	11
1,000,000	23	22	18

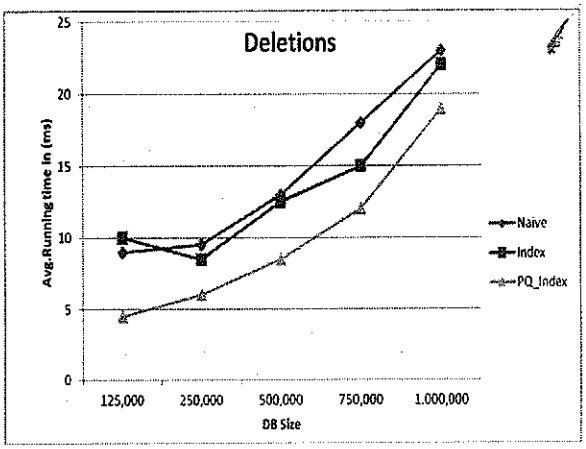

Figure 7: chart showing deletion time

Table 7: Comparison based on execution time

Time to execute Tuple Updates(m.sec)			
DB Size	**Naïve**	**Index**	**PQ_Index**
125,000	7	7	7
250,000	17	16	15
500,000	23	21	18
750,000	15	18	16
1,000,000	19	25	25

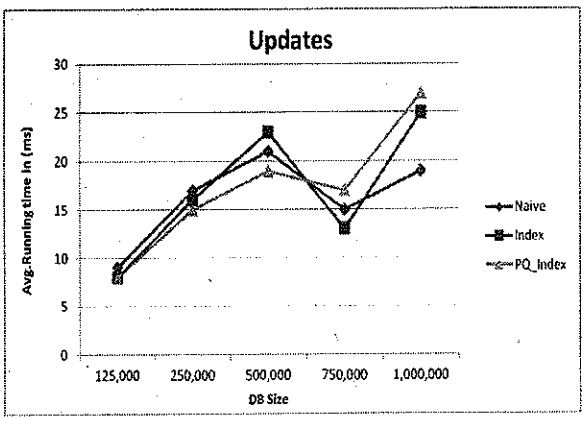

Figure 8: Comparison based on execution time

The results obtained clearly shows that the proposed system outperforms in all the metrics when compared with the existing system. The execution time is also less when the size of the database increases.

IV. Conclusion

This work deals with the management of imperfect databases, the DBMS dictates how incomplete or imperfect information should be handled. The literature study perform proposes the concept of a Customized information prediction policy(CIP) operator and Duplicate information policy operator(DIP). Using CIP operators, users can specify the prediction policy that they want to handle partial information. DIP operators will find duplicate tuples which are replaced by the CIP operators. PQ_Index structures has been used for evaluating CIP operators.The experimental results obtained show that the PQ_Index structures can be used to efficiently manage very large database.

References

1. Brakatsoulas, Sotiris, et al. "On map-matching vehicle tracking data." Proceedings of the 31st international conference on Very large data bases. VLDB Endowment, 2005.

2. Li, Xu, et al. "A practical map-matching algorithm for GPS-based vehicular networks in Shanghai urban area." Wireless, Mobile and Sensor Networks, 2007.(CCWMSN07). IET Conference on. IET, 2007.

3. Kuijpers, Bart, and Walied Othman. "Trajectory databases: Data models, uncertainty and complete query languages." Database Theory–ICDT 2007. Springer Berlin Heidelberg, 2006. 224-238.

4. Liu, Hechen, and Markus Schneider. "Querying moving objects with uncertainty in spatio-temporal databases." Database Systems for Advanced Applications. Springer Berlin Heidelberg, 2011.

5. Hajari, Hadi, and Farshad Hakimpour. "A Spatial Data Model for Moving Object Databases." arXiv preprint arXiv:1403.3304 (2014).

6. Karthikeyani, V., and I. ShahinaBegam. "Different Direction Flow Analysation Algorithm (DDFAA) Of Moving Object Using Spatial

7. Trajcevski, Goce, et al. "The geometry of uncertainty in moving objects databases." Advances in Database Technology—EDBT 2002. Springer Berlin Heidelberg, 2002. 233-250.

8. Pelekis, Nikos, et al. "Clustering uncertain trajectories." Knowledge and Information Systems 28.1 (2011): 117-147.

9. F. Diebold, T. Gunther, and A. Tay, "Evaluating density forecasts with applications to financial risk management," Int. Econ. Rev., vol. 39, no. 4, pp. 863–883, Nov. 1998.

10. Z. Ding, "UTR-Tree: An index structure for the full uncertain trajectories of network-constrained moving objects," in Proc. 9th Int. Conf. MDM, Beijing, China, 2008, pp. 33–40.

11. Z. Ding and R. H. Güting, "Uncertainty management for network constrained moving objects," in Proc. 15th Int. Conf. DEXA, Berlin, Heidelberg, Germany, 2004, pp. 411–421.

12. E. Frentzos, K. Gratsias, and Y. Theodoridis, "On the effect of location uncertainty in spatial querying," IEEE Trans. Knowl. Data Eng., vol. 21, no. 3, pp. 366–383, Mar. 2009.

13. J. Hightower and G. Borriello, "Location systems for ubiquitous computing," IEEE Comput., vol. 34, no. 8, pp. 57–66, Aug. 2001.

14. K. Hornsby and M. J. Egenhofer, "Modeling moving objects over multiple granularities," Ann. Math. Artif. Intell., vol. 36, no. 1–2, pp. 177–194, Sept. 2002.

15. Jain, E. Y. Chang, and Y.-F. Wang, "Adaptive stream resource management using Kalman filters," in Proc. 2004 ACM SIGMOD Int. Conf. Manage. Data, pp. 11–22.

16. A S. Jensen, H. Lahrmann, S. Pakalnis, and J. Runge, "The INFATI data." TIMECENTER, Tech. Rep. TR-79. 2008.

17. H. Jeung, H. T. Shen, and X. Zhou, "Mining trajectory patterns using hidden Markov models," in Proc. 9th Int. Conf. DaWaK, Regensburg, Germany, 2007, pp. 470–480.

18. B. Kanagal and A. Deshpande, "Online filtering, smoothing and probabilistic modeling of streaming data," in Proc. 2008 IEEE 24th ICDE, Cancun, Mexico, pp. 1160–1169.

19. D. E. Knuth, The Art of Computer Programming, Vol. 3, Sorting and Searching, 2nd ed. Redwood City, CA, USA: Addison Wesley Longman Publishing Co., Inc., 1998.

20. B. Kuijpers, B. Moelans, W. Othman, and A. A. Vaisman, "Analyzing trajectories using uncertainty and background information," in Proc. 11th Int. Symp. SSTD, Aalborg, Denmark, 2009, pp. 135–152.

21. B. Kuijpers and W. Othman, "Trajectory databases: Data models, uncertainty and complete query languages," J. Comput. Syst. Sci., vol. 76, no. 7, pp. 538–560, Nov. 2010.

22. A. LaMarca and E. de Lara, Location Systems: An Introduction to the Technology behind Location Awareness. San Rafael, CA, USA: Morgan and Claypool Publishers, 2008.

23. H. Liu and M. Schneider, "Querying moving objects with uncertainty in spatio-temporal databases," in Proc. 16th Int. Conf. DASFAA, Hong Kong, China, 2011, pp. 357–371.

24. H. Lu, B. Yang, and C. S. Jensen, "Spatio-temporal joins on symbolic indoor tracking data," in Proc. IEEE 27th ICDE, Hannover, Germany, 2011, pp. 816–827.

25. H. J. Miller, "A measurement theory for time geography," Geographical Anal., vol. 37, no. 1, pp. 17–45, 2005

AUTHORS' BIOGRAPHY

Dr. Sumadhi.T is an Associate Professor, Department of Computer Science in Karpagam Univesity, Coimbatore. She received her B.Sc.(CS), in 1996 and MCA in 1999 from Bharathiar University, Coimbatore. She obtained her M.Phil. in the area of Image Processing from Periyar University, Salem in 2006. Her research interest lies in the area of Image Processing and Data mining. . She obtained her Ph.D. in the area of Image Processing from Karpagam University, Coimbatore in 2013. She has published about 20 papers in international and national journals. She has attended about 15 international and national conference held at various places in Tamil Nadu.

M. Janaki is a Student in karpagam University. She received her BCA in 2004 from Karpagam University,Coimbatore and M.sc in 2006 from Annamalai University,Coimbatore . She Finished M.phil in Data Minining from Karpagam University, Coimbatore in 2015. Her Area of interest in Data mining. She has attended International Conference held in Coimbatore.