ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A PERSONALIZED SCHEME FOR INCOMPLETE AND DUPLICATE INFORMATION HANDLING IN RELATIONAL DATABASES

Dr. K. Sathesh Kumar*, Dr. S. Ramkumar*
*Assistant Professor, Department of Computer Science and Information Technology,Kalasalingam University, Krishnankoil, Virudhunagar (Dt). INDIA
*Assistant Professor, Department of Computer Applications, Kalasalingam University, Krishnankoil, Virudhunagar (Dt). INDIA

DOI: 10.5281/zenodo.154206

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 work proposes the renovated concept of partial information handling with complete prediction model, which 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 work 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 operator also allows them to state a strategy 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 present work also analyze how relational algebra operators and PIP operators interact with one another. This also handles the COALESCE function using the CIP operator.

KEYWORDS:Knowledge personalization and customization, Database semantics, Duplicate data, Missing value operator.

INTRODUCTION

Estimation of Missing data has the main goal in providing or determining of missing values by reasoning from experimental data. Because of missing values it can result in unfairness that impacts on the worthiness of the patterns or/and the classification performance, missing data replacement has been a main problem in learning from incomplete data. In the datasets with the homogeneous attributes which is defined as the independent attributes, the several techniques are developed for the missing values estimation. [1] Though, those techniques cannot be applied to many real datasets, for example equipment maintenance databases, gene databases, and industrial datasets, since these datasets are often with both continuous and discrete independent attributes. The heterogeneous datasets are called as mixed-attribute datasets [18]. In these heterogeneous datasets it has the independent attributes which is called as mixed independent attributes. In this dissertation proposes an enhanced operator for partial data handling and de-duplication at the linkage method, This also aimed at performing fully customized policies for partial information handling and duplicate information handling and also matches data in different types [2].

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

Missing value in database can be filling by three ways of proposed work

- *Aggregate PIP operator*: candidate values are aggregated by means of an aggregate operator v specified as a parameter of the PIP operator so that a single value v is determined. Finally, every occurrence of the considered null is replaced by v in the result. The families of single-valued PIP operators defined below essentially differ in the way that candidate values are determined.
- *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. The following example shows that the sufficient conditions above are not necessary conditions.
- *PIP operator base on another attribute*: A PIP operator based on another attribute is of the form$\rho^{att}(\mu, u, v, A, B, X)$. 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 β is determined; then, only those tuples having B-values closest to β 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*. In our WB/ UNESCO data set, a user may believe that missing primary school enrollment data for a country in a given year can be inferred by looking at the secondary school enrollment trends (for the same country) and using those trends to infer the missing data.

LITERATURE SURVEY

A number of methods have been developed for dealing with missing data, though most of these have focused on continuous variables. This section deals with various researchers concepts. Finch, W. Holmes, et al [3] studies 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,[19] as well as the calculation of scores difficult. They were many approaches were developed, it is not clear that these techniques for imputation are appropriate for the categorical items. However, methods of imputation specifically designed for categorical data are either limited in terms of the number of variables they can accommodate, or have not been fully compared with the continuous data approaches used with categorical variables. This concept compare 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 (i.e., using listwise deletion) 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 did the results for any of the imputation methods. Badia, Antonio et al [4], propose 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. They both support Armstrong's axioms and provide realizable null markers: at any time, some or all of the null markers can be replaced by actual values without causing an anomaly. Here this analysis the concept of FD in the presence of null markers, and have specified a set of properties that we believe 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 satisfy our axioms.

Calı, Andrea, et al [5] studies the issue of dealing with integrity constraints over the global schema in data integration. On the one hand, integrity constraints can be used to extract more information from incomplete sources, similarly to the case of databases with incomplete information. Data integration is 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. On the other hand, integrity constraints raise the problem of dealing with the inconsistency of the whole system, due to contradictory 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 theoretical issues and problems are suggested by the 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 that we wish to pose. The potential sources for the difficulty vary. These include examples such as 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]. Propose a statistical relational learning approach for estimating and replacing missing categorical data. First, for a given data set, all categorical attributes are classified as a proper number of groups, and these groups are independent of each other. Second, principles for ordering attributes in one group are proposed and the attribute sequence of the group could be indexed by the principles. Third, a hidden Markov model for estimating missing categorical value is represented. 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. Finally, the implement process of the proposed approach is illustrated by an example. Chiu and Sedransk [8] proposed a Bayesian method for estimating and replacing missing data based on some prior knowledge about the distributions of the data. However, this method primarily was applied to univariate missing data and some special multivariate cases and developed another Bayesian method for estimating and replacing missing categorical data, using the 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 the hypothesis lacks reasonable support. Martinez et al [10], studied Inconsistency management policies allow a relational database user to express customized ways for managing inconsistency according to his 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 problem that we address in this work is that of determining a subset of these policies that are suitable for application w.r.t. the set of constraints and user preferences. We propose 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 [14].

PROPOSED SYSTEM

The work proposes a unified framework for analysis on incomplete data, duplicate data using special operators, which captures existing approaches as a special case and provides an easy basis for the proposed system [13]. The study introduces a new approach named as customized information prediction policies which allow users to get suggestion about missing data in a database, taking into account their own knowledge of how the data was collected, their attitude to risk, and their mission needs.

Contributions of the work:

- The current works contribute 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.
- This proposes an effective index structures to support the efficient evaluation of CIP operators and show how to maintain them incrementally as the database is updated.
- This also concentrated 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.
- This study also deals with the duplicate value identification using the same strategy; this provides a single solution for multiple problems.

- The experiments help to assess the effectiveness of the proposed index structures with a real-world airline and census data set. Specifically, this compares 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.
- The system performs experiments to evaluate the effect of the Priority queue index structures when CIP operators are combined with classical relational algebra operators and study whether evaluating a CIP operator before or after a relational algebra operator, under the conditions which guarantee the same result, may lead to better performance.

METHODOLOGY

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 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**

The 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 was 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. As per the discussion above, the result of evaluating a PIP operator over a database D is a set of complete databases.

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=*f*A1, *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. This helps to the current predictive model to an observed data set of A1 and *A2* values. This helps to find the data even the specified constraint is not satisfied. 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 *remove the maximum* and another one is *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 appearing 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. The families of single-valued PIP operators defined below essentially differ in the way that candidate values are determined.

B. **Autoregressive Hidden Marko Models (ARHMM)**

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$\pi = \pi_i$.
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 depended 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. This is exactly the reason why ARHMM are superior to conventional HMM in this application. Conventional HMM assumes 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 bases 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**

The proposal deals with the duplicate entry finding with the same customized policies using linear functions and de-duplication functions. 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 form 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

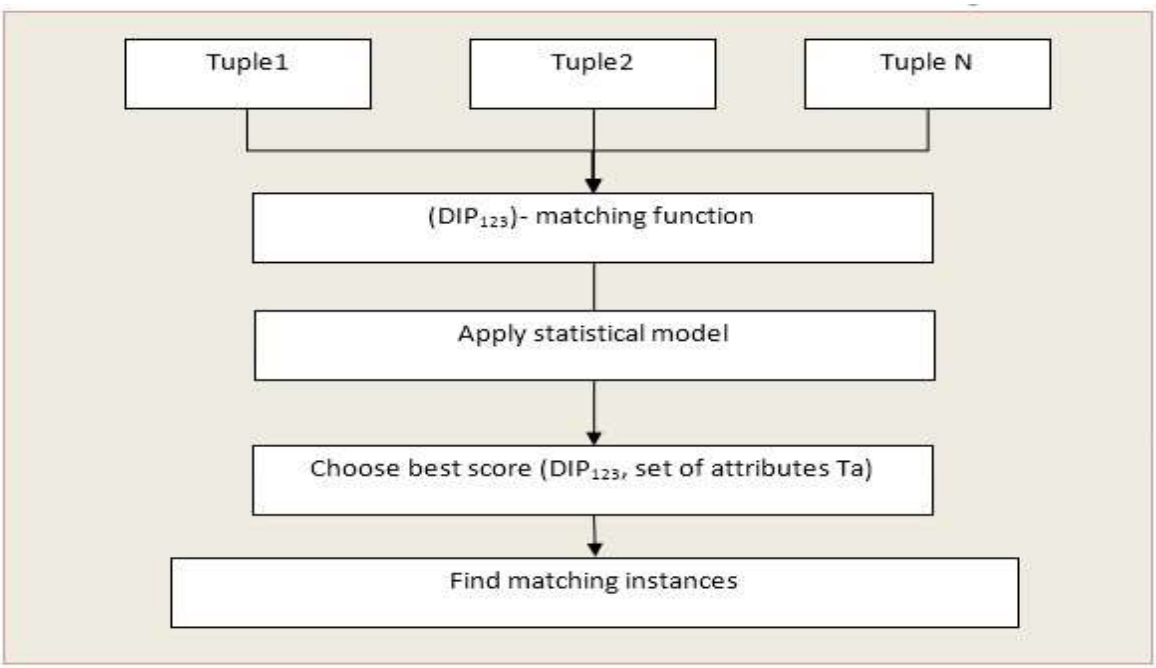


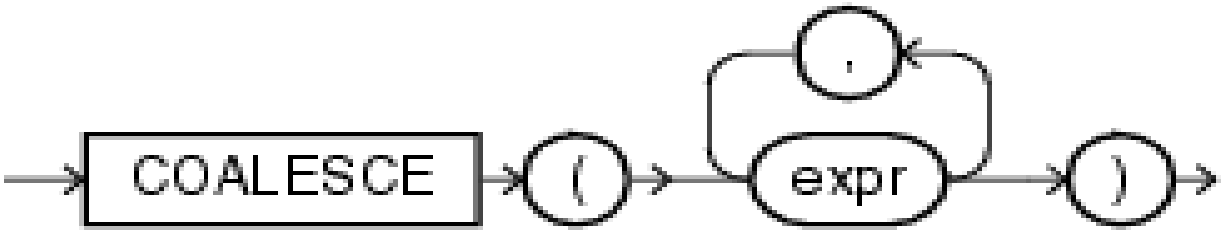
Figure. 1. Steps for data linkage and DIP

The system perfroms the above fig 1 shows steps for data linkage and DIP. 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.

D. About COALESCE:

COALESCE returns NULL if all its arguments evaluate to null. Otherwise, it returns the value of the first non-null argument in the scalar_expression list. COALESCE is a shorthand expression for the following full CASE expression:

CASE
WHEN scalar_expression_1 IS NOT NULL
THEN scalar_expression_1
...
WHEN scalar_expression_n IS NOT NULL
THEN scalar_expression_n
ELSE NULL
END



Purpose

COALESCE returns the first non-null *expr* in the expression list. You must specify at least two expressions. If all occurrences of*expr* evaluate to null, then the function returns null. Oracle Database uses **short-circuit evaluation**. The database evaluates each *expr* value and determines whether it is NULL, rather than evaluating all of the *expr* values before determining whether any of them is NULL. If all occurrences of *expr* are numeric datatype or any nonnumeric datatype that can be implicitly converted to a numeric datatype, then Oracle Database determines the argument with the highest numeric precedence, implicitly converts the remaining arguments to that datatype, and returns that datatype [15][16][17].

This function is a generalization of the NVL function.

You can also use COALESCE as a variety of the CASE expression. For example,
COALESCE (expr1, expr2) is equivalent to:
CASE WHEN expr1 IS NOT NULL THEN expr1 ELSE expr2 END Similarly,
COALESCE (expr1, expr2, ..., expr*n*), for *n*>=3 is equivalent to:
CASE WHEN expr1 IS NOT NULL THEN expr1 ELSE COALESCE (expr2, ..., expr*n*) END

RESULT AND DISCUSSIONS

To compare the proposed system with the existing schemes such as Naive, Index and PQ_ Index several data set and attribute has been created. 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.

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 15 thousand tuples and

the 10 percentage of rows with a null value by randomly selecting tuples and inserting nulls (of different kinds) in them.

From the Fig 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

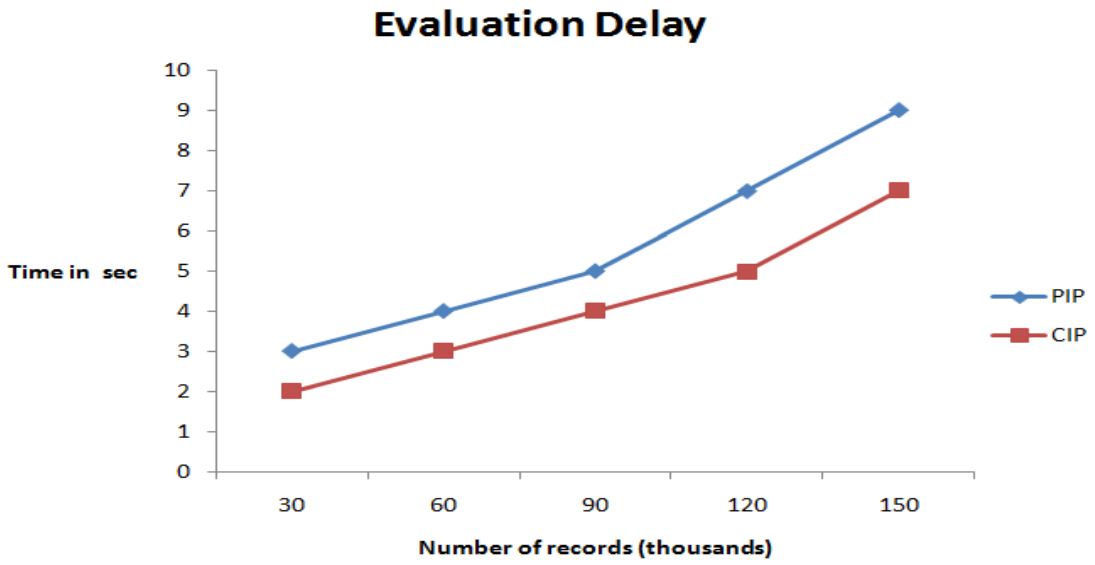


Figure2. Performance comparison of proposed CIP using PQ_Index with existing approaches based on the Evaluation Delay

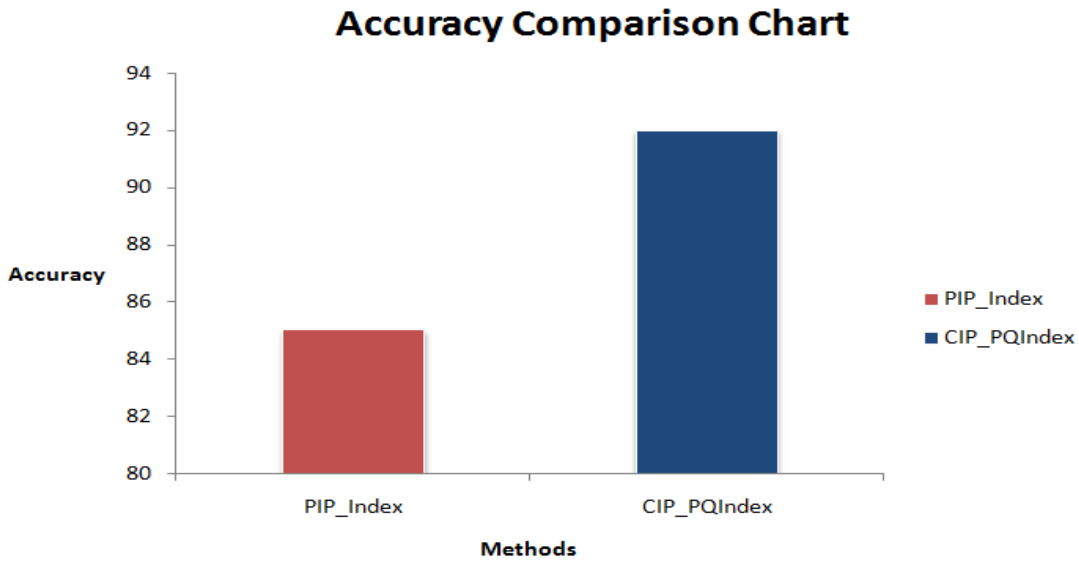


Figure3. Performance comparison of proposed CIP using PQ_Index with existing approaches based on the Result accuracy

ISSN: 2277-9655
Impact Factor: 4.116

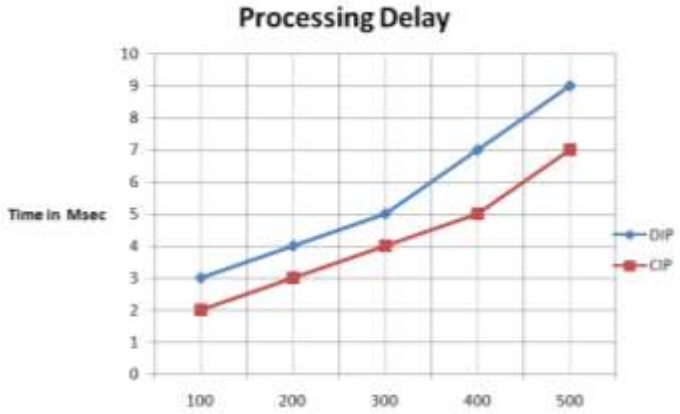


Figure4. Performance verification of proposed DIP and CIP using PQ_Index based on the processing delay. The chart describes the time taken for both DIP and CIP

CONCLUSION

This research work dealing with the management of imperfect databases, the DBMS dictates how incomplete or imperfect information should be handled. This study proposes the concept of a Customized information prediction policy CIP operator and Duplicate information policy operator. 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. It is also presented PQ_Index structures for evaluating CIP operators.The experimental study in this work showing that the PQ_Index structures used to efficiently manage very large database.

REFERENCES

[1] Jiawei Han and Micheline Kamber,.:Data Mining Concepts and Techniques‖, ISBN: 978-1-55860-901-3,2006
[2] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar," A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering,Vol.1,no .3,pp:114-121.
[3] Finch, W. Holmes. "Imputation methods for missing categorical questionnaire data: A comparison of approaches." *Journal of Data Science* 8.3 (2010): 361-378.
[4] Badia, Antonio, and Daniel Lemire. "Functional dependencies with null markers." *The Computer Journal* (2014): bxu039.
[5] Cal, Andrea, et al. "On the role of integrity constraints in data integration." *Bull. Of the IEEE Computer Society Technical Committee on Data Engineering* 25.3 (2002): 39-45.
[6] Wong, Eugene. "A statistical approach to incomplete information in database systems." *ACM Transactions on Database Systems (TODS)* 7.3 (1982): 470-488.
[7] Cao, Jianjun, Et Al. "An Approach Using Hidden Markov Model For Estimating And Replacing Missing Categorical Data."
[8] Chiu, H. Y. and Sedransk, J., A Bayesian Procedure for Imputing Missing Values in Sample Surveys. *J. Amer. Statist. Assoc.*, 81(3905), 1986, pp.5667-5676.
[9] Li X. B., A Bayesian Approach for Estimating and Replacing Missing Categorical Data. ACM Journal of Data and Information Quality, 1(1), 2009, pp.1-11.
[10] Martinez, Maria Vanina. "Contributions to personalizable knowledge integration." *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. Vol. 22. No. 3. 2011.
[11] Martinez, Maria Vanina, and Anthony Hunter. "Incorporating Classical Logic Argumentation into Policy-based Inconsistency Management in Relational Databases." *AAAI Fall Symposium: The Uses of Computational Argumentation*. 2009.
[12] Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533.

[13] Parmar, J., & Jain, P. (2013, December). A different approach of intrusion detection and Response System for Relational Databases. In *Green Computing, Communication and Conservation of Energy (ICGCE), 2013 International Conference on* (pp. 894-899). IEEE
[14] Gavankar, S., & Sawarkar, S. Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility.2015.IEEE
[15] http://www.ibmcourses.com/coalesce-function-in-esql-db2-and-oracle-database/
[16] Database SQL Reference http://docs.oracle.com/cd/B19306_01/server.102/b14200/functions023.htm
[17] Oracle database SQL Language Reference 11g Release 1(11.1) part number B28286-01 http://isu.ifmo.ru/docs/doc111/server.111/b28286/functions023.htm
[18] Shohdy, S., Su, Y., & Agrawal, G. (2015). Accelerating data mining on incomplete datasets by bitmaps-based missing value imputation.
[19] Hazan, E., Livni, R., & Mansour, Y. (2015, January). Classification with low rank and missing data. In *Proceedings of The 32nd International Conference on Machine Learning* (pp. 257-266).