DOI: http://dx.doi.org/10.26483/ijarcs.v9i2

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

REMOVAL OF DUPLICATES IN DATABASE RELATIONS AND THE ASSOCIATED PROPAGATION MANAGEMENT

Dr. L. Venkateswara Reddy
Professor
Dept. of IT,
Sree Vidyanikethan Engineering College, Tirupati, India

Dr. A. K. Damodaram
Professor
Dept. of Mech., Engineering,
Sree Vidyanikethan Engineering College, Tirupati, India.

Abstract: Removing duplicate records in the relations of a database is an essential operation and it is a crucial and a critical step in the data integration. If the record duplication problem is unmanaged or miss managed it leads to poor quality, consistency, integrity of data. The present paper reviewed the problem contexts of data duplication and the techniques available for the management of the problem. This work also proposed some improved techniques to deal with data duplication problem. A set of data fusion techniques are proposed. A new way of data propagation is presented that should follow the fusion result to maintain data consistency.

Keywords: database, relations, integration

1. INTRODUCTION

The occurrence of data duplication is common in the relations of a relational database. This data duplication if not managed earlier the problem becomes worse after the formation of many referenced relations. The record duplicate data problem is also called entity resolution or record linkage problem. Within the same relation the same real world objects are represented by using multiple descriptions leading to the confusion in understanding the object properties. The problem of representing the same object with multiple descriptions occurs due to data missing, data modification, data deletion, typographical errors and not following standard rules and procedures in data manipulation operations. There does not exist any standard and more generalized framework for accurate and effective data manipulations of databases. Also there does not exist any standard and deterministic methodology for finding duplicate tuples in the same relation without using primary key concept for identifying duplicate data descriptions in the tuples.

The duplication arises in two forms. Partial data duplication occurs when the data is duplicated for a subset of a tuple instance. This can be removed by simple database operations. When the data duplication is for the entire record this is called full duplication. Full or complete data duplication in two different tuples of the same relation is not allowed knowingly but it is allowed unknowingly. Whenever such unknown full data duplication occurs it must be identified and then removed with the best consolidated or fitted tuple. The aggregated tuple must match with all the duplicated tuples to the greater extent. For simplicity purpose duplicated tuple set is called bad tuple set. The number of bad tuples in the bad tuple set may be either two or more. The entire bad tuple set must be replaced with one better tuple that is more than 90% is similar to the all the tuples in the bad tuple set.

2. LITERATURE REVIEW

Data duplication generally occurs due to manual errors and misassumptions. To deal with the data duplication database related management tools are available such as the use of not null, null, on delete cascade, on update cascade and so on. Other intelligent management techniques were proposed by various authors in the literature of the data duplication management. In the database management literature many research people have identified that one solution for controlling referential integrity is by means of classical techniques already available in the database management system software using null, not null, default, on delete cascade, on update cascade, and restricting controls. These techniques control referential integrity but do not support semantically correctness of the relationships in the relation of the database after the completion of the data fusion operations. Second solution for better management of relationships semantically is to use generalized and semantic version of existing data referential integrity management techniques. M.A. Herna´ndez and S.J. Stolfo stated that, in the database community, the record linkage or record duplication problem is described as merge-purge [7].

Ahmed K. Elmagarmid et al. [1] said that duplicate record detection is the process of finding different or multiple records that refer to one unique real-world entity or object or record. Authors also said that for duplicate record detection they have implemented a variety of string similarity metrics, such as Jaro, edit distance, and q-gram distance. Ravi Kumar and Cohen [8] follow a similar approach and proposed a hierarchical, graphical model for learning matched record pairs.

B. Zhao et al. [2] proposed a Bayesian approach to perform data fusion operation. It learns the quality of data sources and incorporates the learned knowledge in the data fusion operation. Because of its stateless nature the proposed approach is not up to the mark in the online setting.

Web applications commonly require duplicate-free data and error-free representation of records. The goal of the former

is achieved through Record Linkage (RL) technique while the latter is achieved through Data Fusion technique. The two techniques - Record Linkage and Data Fusion are the two well-studied problems [1] [5]. While signficant effort has been dedicated to solve the above problems but a very little work has been conducted to apply them at the query execution time.

Hotham Altwaijry et al.[3] said that efficiency, scalability, performance and data quality are the main challenges of entity resolution and entity resolution can be computationally expensive.

Hairong Dong and David Evans [4] defined data fusion as a formal framework that express the means and tools for the alliance of data originating from diverse sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application. Kamakshi Lakshminarayan [6] explored explores the use of machine-learning based options for data imputation, in dealing with missing data. The authors proposed two well-known machine learning techniques. The first one is data clustering which is an unsupervised learning strategy that make use of a Bayesian approach to cluster the data into classes. The resultant groups of clustering were used to predict multiple choices for the attribute of interest. The second one is a supervised learning technique that models the missing variables by a supervised induction of a decision tree-based classifier. This model predicts the most likely value for the attribute of interest. Empirical tests have been performed in order to compare the two proposed techniques. These tests showed that both approaches are useful and have limitations too. Verykios et al. [10] proposed a set of techniques for reducing the complexity of record comparison. Sarawagi and Bhamidipaty [9] designed an efficient code called ALIAS, a learning-based duplicate detection system that uses the idea of a "reject region".

From the literature it is evident that there exist different types of data fusion functions such as

1. First order fusion function
2. Second order fusion function
3. Join fusion function
4. Set oriented fusion function
5. f-optimal fusion function
6. f-value functions
7. Random fusion functions
8. Maximal coherent fusion functions

3. PROBLEM CONTEXT

Incomplete data are everywhere in data sources and as a result, available data are inefficient and often biased. Sometimes database modifications result record duplication in the relations. Data duplication is always a challenging situation. Identification, removal and replacement of the undesired tuples from the relations of a database are called data fusion operation. For effective implementation of data fusion operation the present set of available techniques are not complete. There is a need for new techniques in a high level semantic manner.

Assume that there exists a parent relation and one or more referenced relations. Also assume that there exist duplicate tuples in the parent relation. Generally data fusion is performed as a first step and data propagation is performed as a second step. In the first step duplicate tuples are identified and replaced with correct tuples and in the second step modified details from the parent relations. Fusion functions are used in the data fusion step. The data propagation may be either backward or forward. On delete cascade is on solution to maintain referential integrity in database operations but this approach introduces randomness in the process of selecting and deleting and which tuples must keep them as it is. As a result of this these is no guarantee of maintaining data quality and semantically correctness of relationships after successful completion of data fusion operation.

4. PROPOSED METHODOLOGY

4.1 Objectives

For effective database management duplicate tuples must be identified using a more generalized framework and then removed two or more incorrect tuples with one or more correct tuples. Database must always satisfy data consistency property before and after database modifications. Database must always satisfy data integrity constraints in particular referential integrity constraints before and after data modifications. There exist many techniques such as set null, set not null, on delete cascade, on update cascade, restrict and so on for controlling and smooth management of referential integrity constraints. All these techniques are specialized techniques only but not generalized techniques to propagate database updating and deletions in a high level semantic procedure way. Existing methods do not provide quality relationship management in a semantic way. In modern very large database management systems there is a need to apply and use optimized quality of data relationships among the relations of a database.

Present study proposed a new semantic based framework for efficient, accurate, effective and optimal quality of relationships management in the database operations. This new data fusion technique is independent of another record duplication finding methods so that the new technique can be applied for very large and different varieties of data fusion operations as an independent, semantic, generalized and scalable approach in the domain of SQL data management. In the present paper a running example is taken for better understanding of data fusion operation on linked relations with the intention of preserving referential integrity as well as semantically correctness of the relationship in the relations of a specific database.

The proposed algorithm for controlling data fusion operations is well defined, designed and proposed a well-investigated data propagation algorithm which can manage, control, and coordinate the net impact of a fusion operation on joined relations with respect to both data preservation of referential integrity and the semantic correctness of the linked relationship after successful completion of the data fusion operation. The algorithm takes care of consistency of referential integrity after the data fusion of duplicate tuples in the main parent relationship of the selected database and the algorithm uses a standard framework of fusion functions that operate on multi-valued data.

4.2 The approach

The proposed methodology attempts to fuse the set of duplicate records into one record for maintaining database consistency against modifications of the database. Different

strategies are developed for this. In the first strategy the fusion function makes use of the attribute union. Union is applied either attribute value by attribute value basis or record by record basis whichever is convenient or possible. In the second strategy the attribute mean is used to update the missing value of the attribute in order to fuse the records. This is called mean imputation in machine learning terminology. In the third strategy majority value of the attribute is used to update the missing value of the attribute. Updating the missing value with majority value is particularly useful when values of the attribute are categorical or discrete only.

The first strategy is explained with the following work out. A database consisting of three relations are considered for explaining the fusion operation in the relations. The three relations are Establishment, Entrance, and Entrance details which are respectively are shown in TABLE-1, TABLE-2, and TABLE-3. For simplicity and easy understanding purpose only a limited set of missing values are taken and then fusion operation involving union is applied.

Example database one

TABLE-1 an Establishment relation

University-id	University-name	state	Establishment-data
1	Ou	TS	1916
2	Ou	-	-
3	-	TS	1916
4	Svu	AP	-
5	svu	-	1950

TABLE-2 Entrance

Entrance-id	Entrance -name	Eligibility
1	EAMCET	Inter pass
2	ICET	Degree pass
3	LAWCET	Degree pass
4	PGCET	B.tech pass

TABLE-3 Entrance Details

University-id	Entrance-id	Number of times - conducted
1	1	12
1	2	8
1	4	6
2	1	12
3	2	8
4	1	9
4	2	7
5	1	9

Union of tuples 1, 2 and 3 in the Establishment relation are computed below:

= t1[university-name] Union t2[university-name] Union t3[university-name]

= {OU} Union {OU} Union {null}

= {OU}

Similarly union values on the state attribute in the Establishment relation for tuples 1, 2, and 3 are computed below:

= t1[state] Union t2[state] Union t3[state]

= {TS} Union {null} Union {TS}

= {TS}

Similarly union values of tuples 1, 2 and 3 on attribute established-date are computed below:

= t1[Established-date] Union t2[Established-date] Union t3[Established-date]

= {1916} Union {null} Union {1916}

= {1916}

In the establishment relation, tuples 1, 2, and 3 are identified as one set of duplicate tuples and tuples 4 and 5 are identified as another set of duplicate tuples. First set of three duplicate tuples is replaced with one new correct tuple {1, OU, TS, 1916}. Second set of two duplicate tuples is replaced with one new tuple, {4, SVU, AP, 1950}.

= t4[university-name] Union t5[university-name]

= {SVU} U {SVU}

= {SVU}

Similarly for the state attribute

= t4[state] Union t5[state]

= {AP} Union {null}

= {AP}

Similarly for the Established_date attribute

= t4[Established_date] Union t5[Established_date]

= {null} Union {1950}

= {1950}

After successful completion of data fusion operation on the Establishment relation, the modified (fused) Establishment relation is shown in TABLE-4.

TABLE-4 Fused_Establishment

University_id	University_name	state	Established_date
1	OU	Ts	1916
4	SVU	AP	1950

Now based on the Fused Establishment relation the referenced/dependent relations must be updated through data propagation technique and updated data details of dependent relations are shown in the TABLE-5, Propagation. In the Propagation relation University_id values 1, 2, and 3 are replaced with 1 and the values 4 and 5 are replaced with the correct University_id, 4.

TABLE-5 Propagation

University_id	Entrance_id	Number of times conducted
1	1	12
1	2	8
1	4	6
1	1	12
1	2	8
4	1	9
4	2	7
4	1	9

TABLE-6: Modified_Propagation

University_id	Entrance_id	Number of times conducted
1	1	12
1	2	8
1	4	6
4	1	9
4	2	7

To remove duplicate tuples from the parent relation a fusion function is used. The first order fusion function takes a set of duplicate tuples and then maps them to one correct tuple. If the first order fusion function is true only for subset of attributes then it is called partial preservative and if it is true for all the attributes then it is called full preservation or simply preservation fusion function. Second order fusion function operates on multi-valued data. Second order fusion function takes multiple sets of duplicate tuples from the parent relation and then replaces them with one particular correct set of tuples. In the literature second order fusion function are called multi-valued fusion function replaces sets of input duplicate tuples with one particular and correct input set of tuples.

The second strategy fuse the duplicates by filling the missed value with attribute mean and it is called the mean imputation technique. When the attribute values are of categorical or nominal the mode of the attribute is considered to fill the missing value and this is called the mode imputation. This strategy is explained with the following example with the help of tables TABLE-7.

Table-7: Table with duplicates assumed

Att1	Att2	Att3	Att4
1	12	x	8
2	12	-	-
3	-	x	8
4	4	y	-
5	4	-	10

In the above table if the first three records are assumed as duplicates the mean imputation strategy is applicable to attribute “att2” and the mode imputation strategy is applicable to the attribute “Att3”.similarly the fusion of the last two records can be made. The resulted fusion is shown in table 8 and Table 9.

Table 8-The fusion in progress

Att1	Att2	Att3	Att4
1	12	x	8
2	12	x	8
3	12	x	8
4	4	y	10
5	4	y	10

Table 9-The fusion result

Att1	Att2	Att3	Att4
1	12	x	8
4	4	y	10

The third strategy fuses the duplicates by filling the missed value with majority attribute value and it is called the majority imputation technique. This is almost similar to the mode imputation strategy.

5. ALGORITHM

The algorithm used for data fusion and associated propagation is presented here.

5.1 Proposed Algorithm

Algorithm Data-Fusion-Propagation

Input

R, original master relation
R*, referenced relation
D, duplicate set of tuples such that $D \subseteq R$
F, selected fusion function

OUTPUT

Updated relations R and R*

1. for each tuple t in D do
2. S_t = referenced tuples in R* with respect to primary key in R
3. $S_t^* = S_t$
4. for all tuples t* in S_t^* do
5. $t^*[FK^*]$ in R* = Fusion Function[K]
6. end for
7. end for
8. S-projected=projected set of tuples with respect to primary key
9. B=multi valued fusion function of S-projected
10. B=filtered non key attributes (B)
11. for all tuples in D do
12. $R^* = R^* - S_t$
13. end for
14. $R^* = R^* \cup B$

5.2 Algorithm description

Steps 1 and 2 for each tuple t in duplicate set, D, a set of linked tuples in R* are constructed and stored in the set S_t.
Steps 3, 4, 5 and 6 tuples in S_t are assigned to S_t^* after replacing the foreign key with primary key.
Step 8 projected set of tuples with respect to primary key are stored in S-projected.
Step 9 tuples resulted after applying fusion function to S-Projected is stored in B
Step 10 for all tuples in B non key attributes are filtered
Steps 11 to 14 resolve any conflicts and store the final result in R*

6. CONCLUSION

In this paper a new framework is identified for removing duplicates from relations of data through data fusion. Data propagation is followed for making the data consistent. This framework is semantically correct and it is more generalized version of traditional methods such as null, not null, on delete cascade, on update cascade, and restrict and so on. DBMS must control referential integrity constraints wherever tuples are deleted from the parent table. New framework intelligently manages not only referential integrity problems but also semantically related details with modified data. Data fusion process uses data fusion functions. This work proposed three strategies for data fusion operation. All the strategies are explained with numerical examples. In the future there is a scope to find and use new fusion function.
The main disadvantage of the forward data propagation is that the linked datasets of tuples cannot be fused directly by a multi valued fusion function. Data fusion and data propagation operations must be considered separately and hence the memory required must be independent of the number of data propagations. Different multi valued fusion functions will give different accuracy results. In general, the accuracy depends on the size of the set of duplicate tuples. Many techniques will give better results than on delete cascade operation particularly when the duplicate dataset is very large. Data propagation technique improves the data quality.

REFERENCES

1. Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios, Member, “Duplicate Record Detection: A Survey”, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 19, NO. 1, JANUARY 2007.

2. B. Zhao, B. I. P. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. PVLDB, 2012.
3. H. Altwaijry, D. V. Kalashnikov, and S. Mehrotra. Query-driven approach to entity resolution. PVLDB, 2013.
4. Hairong Dong and David Evans "Data-Fusion Techniques and Its Application", IEEE Xplore: 18 December 2007, ISBN: 0-7695-2874-0
5. J. Bleiholder and F. Naumann. Data fusion. ACM Comput. Surv., 41(1), Jan. 2009.
6. Kamakshi Lakshminarayan, Steven A. Harp, Robert Goldman and Tariq Samad, "Imputation of missing data using machine learning techniques", KDD-96 Proceedings. Copyright © 1996.
7. M.A. Herna´ndez and S.J. Stolfo, "Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem," Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 9-37, Jan. 1998.
8. P. Ravikumar and W.W. Cohen, "A Hierarchical Graphical Model for Record Linkage," 20th Conf. Uncertainty in Artificial Intelligence (UAI '04), 2004.
9. S. Sarawagi and A. Bhamidipaty, "Interactive Deduplication Using Active Learning," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '02), pp. 269-278, 2002.
10. V.S. Verykios, A.K. Elmagarmid, and E.N. Houstis, "Automating the Approximate Record Matching Process," Information Sciences, vol. 126, nos. 1-4, pp. 83-98, July 2000.