International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 02 | Feb-2016 www.irjet.net p-ISSN: 2395-0072

SEMANTIC MANAGEMENT OF DEDUPLICATE TUPLES IN THE RELATIONAL DATABASES

V. BALAJI KUMAR[1], S. RAMAIAH[2],

[1]M.Tech Student, Department of CSE, KMM Institute of Technology and Sciences, Tirupati, India

[2] S. RAMAIAH, Assistant Professor, Department of CSE, KMM Institute of Technology and Sciences, Tirupati

---

Abstract:

Relational database is a collection of relations. Duplicate tuple existence is common in many real time relational databases. In a relational database, if the same real-world entity is represented by more than one tuple, then such tuples are called duplicate tuples. Finding duplicate tuples and then replacing them by one best tuple is called a fusion operation. Whenever duplicate tuples are found in the relations of any database, those tuples must be replaced with one special best approximate tuple that represents the joint information of all the duplicate tuples. Present study proposes new techniques to find duplicate tuples and then remove those duplicate tuples with the correct real world tuples. In the first step duplicate tuples in the relation are classified based on the class label and in the second step then for each set of duplicate tuples functional dependency method or union method is applied to replace duplicate tuples with the corresponding correct real world single tuple. One possibility is to replace one set of duplicate tuples with one correct real world tuple. Another possibility is to replace two or more sets of duplicate tuples in the relation by one set of correct real world tuples. Sometimes the removal of duplicate tuples in the relations of any relational database can create referential integrity violations. All such violations must be controlled and coordinated syntactically as well as semantically in relations.

Key Words: *De-duplication of tuples, propagation, Integrity constraints, Relational databases*

I. INTRODUCTION

In many real time applications relations in the databases are inherently associated with duplicate tuples. Finding and then removing of duplicate data tuples in the relational database is the most important and latest research topic. Data duplication is also known as entity resolution or record linkage. Duplicate data tuples are present in one or more relational databases when there exit multiple descriptions of the same real world entity. The presence of duplicate tuples causes many database maintenance problems. Some of the reasons for the existence of duplicate tuples are presence of missing attribute values, data entry errors, typing errors and not following standards in data entry and data maintenance. Finding and then removing of duplicate data tuples in the relational database is the most important and latest research topic. In general, data tuples are duplicated in one or more relations of any relational databases when there exit multiple descriptions of the same real world entity (record). Duplicate tuple detection and replacement with correct tuple is inevitable in many relations of the relational database. A special procedure is needed to take care of integrity constraint violations that occur when duplicate tuples are removed from the relations.

Integrity constraints imposed on the relational database must be satisfied before and after deleting the original duplicate data tuples. First determine all such duplicate tuples in the relations of any relational database and then replace all such duplicate tuples by a single correct tuple. Particularly referential integrity must be considered and controlled in propagation of data fusion. Several integrity constraints management strategies such as on delete cascade, on update cascade, set null, set not null, restrict are available in database modifications. These techniques are syntactically correct but semantically incorrect.

Present study proposes a new method to eliminate duplicate tuples in the relations of a relational database. This new technique is called union fusion function technique that is applicable for attribute values.

Present study also proposes another duplicate tuple replacing technique using functional dependency approach, which is more generalized approach. This generalized functional dependency approach covers both the partial preservative and also complete preservative functions.

Present study also proposes another new technique to model duplicate tuples. For example, consider a very big relation called TEMP, in order to find duplicate tuples in the given relation, TEMP, initially we apply a classification technique to classify all the tuples, and then based on the class labels, and duplicate tuples are identified and then these duplicate tuples are replaced by the correct real world tuple. Decision tree is the probably most important and highly interpretable classification technique to the data. Decision tree is used as a benchmark technique before applying any classification technique. Also the time complexity of decision tree is (n x number of attributes x log n) where n is the number of tuples in the relation.

II. PROBLEM DEFINITION

Data duplication is common in many real time applications particularly in the relations of any relational databases. Finding duplicate tuples and then replacing them by one best tuple is called a fusion operation. During fusion operation integrity constraint violations must be controlled carefully and relational database must be managed in a consistence way before and after database modifications as well as after removal of duplicate tuples in the relations of relational databases.

In the present research paper, a sample set of three relations viz, 1.COLLEGE, 2.CONFERENCE and 3.CONDUCTED_CONFERENCES is considered as running example for understanding purpose. In the relation COLLEGE tuples 1 and 2 are duplicated because of some reasons such as typographic errors, missing of values and lack of standard data representation procedures etc.

Both one and two duplicate tuples describe the same real world entity. These two duplicate tuples are identified and consequently replaced by one equivalent real and correct tuple. Finding and then removing these duplicate tuples with one correct and real world tuple is called a fusion operation. Present study also proposes a new fusion operation called union. Union fusion operation accepts a set of duplicate tuples and then replaces with one correct real world entity. Working principle of union fusion function operation is explained below:

Union of College_Code = {3G}$\cup${3G} = {3G}

Union of College_Name = {KMM}$\cup${KMM} = {KMM}

Union of Principle_Name = {Rama}$\cup${Rama} = {Rama}

Union of Affiliated_University = {null}$\cup${ JNT University} = {JNT University}

In the COLLEGE relation duplicate tuples 1 and 2 are replaced by the following single tuple using proposed new union fusion function technique. The replacing function may be either partially preservative or complete preservative function. Partially preservative function is defined as follows: There exists t$\in$DupSet such that t[A] = REP(Dup)[A]

When A $\subseteq$ DupSet, it is called partial preservative and when A = DupSet, it is called complete preservative replace function. For example, let A = {SNo, College_Code, College_Name, Principle_Name}

Here t[A] = Rep(Dup)[A] and Let B = {JNT University}, then t[B] = Rep[B]

In this particular example, replacing function is partial preservative but not complete (full) preservative. Hence, 1 and 2 duplicate tuples in the COLLEGE relation are mapped with one correct real world tuple. In the COLLEGE relation, tuples 5, 6, and 7 are duplicate tuples. This is an example for complete preservative. These three duplicate tuples are shown in the FIGURE 6 and then they are replaced by the single tuple shown in the FIGURE 7.

Here, t[all attributes] = RepDup[all attributes]. Complete preservative replacing function replaces a set of tuples with another equivalent and simplified set of tuples.

Consider once again the relation COLLEGE1 with the functional dependency that holds on it, College_Code → {College_Name,Principal_Name,Affiliatedto}. The functional dependency states that when two values on different tuples are same on the attribute College_Code then all values of the three attributes in the right side of the functional dependency are also same. That is, if t_1[College_Code] = t_2[College_Code] then t_1[College_Name,Principal_Name,Affiliatedto] = t_2[College_Name,Principal_Name,Affiliatedto]. Therefore duplicate tuples 1 and 2 in the COLLEGE1 relation are replaced by tuple 1 by applying functional dependency constraints.

First proposed method takes union among the attributes. Second proposed method takes union among the tuples. Third proposed functional dependency method is more generalized version of the above two methods. Third proposed method also takes care of partial and complete preservative functions also.

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
1	3G	KMM	Rama	Null
2	3G	KMM	Null	JNT University
3	3C	Vidyanikethan	CSReddy	Null
4	2D	SHREE	Null	JNT University
5	6E	Annamacharya	Dr.MuniSwamy	JNT University
6	6E	Annamacharya	Dr.MuniSwamy	JNT University
7	6E	Annamacharya	Dr.MuniSwamy	JNT University

FIG-1 COLLEGE

Coference_Id	Conference_Type
Con1	IEEE
Con2	SPRINGER
Con2	ACM

FIG-2 CONFERENCE

SNo	Confrence_Id	Numberof_days	Start_Date
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006
1	Con3	4	3/9/2012
2	Con1	3	8/6/2009
2	Con2	1	10/12/2006
3	Con1	6	3/5/2013
5	Con1	4	29/12/2010
6	Con1	5	29/12/2010
7	Con1	6	29/12/2010

FIG-3 CONDUCTED_CONFERENCES

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
1	3G	KMM	Rama	Null
2	3G	KMM	Null	JNT University

FIG-4

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
1	3G	KMM	Rama	JNT University

FIG-5

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
5	6E	Annamacharya	Dr.MuniSwamy	JNT University
6	6E	Annamacharya	Dr.MuniSwamy	JNT University
7	6E	Annamacharya	Dr.MuniSwamy	JNT University

FIG-6

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
5	6E	Annamacharya	Dr.MuniSwamy	JNT University

FIG-7

SNo	Confrence_Id	Numberof_days	Start_Date
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006
1	Con3	4	3/9/2012
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006
3	Con1	6	3/5/2013
5	Con1	4	29/12/2010
5	Con1	5	29/12/2010
5	Con1	6	29/12/2010

FIG-8 CONDUCTED_CONFERENCES AFTER PROPAGATION

SNo	Confrence_Id	Numberof_days	Start_Date
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006
1	Con3	4	3/9/2012
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006

FIG-9

SNo	Confrence_Id	Numberof_days	Start_Date
3	Con1	6	3/5/2013

FIG-10

SNo	Confrence_Id	Numberof_days	Start_Date
5	Con1	4	29/12/2010
5	Con1	5	29/12/2010
5	Con1	6	29/12/2010

FIG-11

SNo	Confrence_Id	Numberof_days	Start_Date
1	Con1	3	18/6/2009
1	Con2	1	10/12/2006
1	Con3	4	3/9/2012
3	Con1	6	3/5/2013
5	Con1	4	29/12/2010

FIG-12 FINAL CONDUCTED_CONFERENCES_CORRECTED

SNo	College_Code	College_Name	Principal_Name	Affiliated_University
1	3G	KMM	Rama	JNT University
3	3C	Vidyanikethan	CSReddy	Null
4	2D	SHREE	Null	JNT University
5	6E	Annamacharya	Dr.MuniSwamy	JNT University

FIG-13 FINAL COLLEGE_CORRECTED relation

III. ALGORITHM

Let R be a relation of tuples and assume that set of duplicate tuples are denoted by delta. That is, delta ⊆ R. Let R' be the child relation corresponding to the parent relation, R. this algorithm will be executed in two steps. In the first step duplicate records are identified and then in the second step identified duplicated records are replaced with the correct real world records and also these changes are propagated to the dependent (referenced) relations in a semantically correct way in addition to the syntactic correctness of relations with respect to many database operations such as insert, delete, and update.

Assume that sample parent relation R = COLLEGE, and the dependent child relation of the parent relation is taken as R' = CONDUCTED_CONFERENCES. Also assume that tuples t ∈ delta ⊆ R and tuples t^* ⊆ R'. The relationship between parent and child relations is one to many from COLLEGE to CONDUCTED_CONFERENCES. In the COLLEGE relation tuples 1 and 2 are duplicated and this type of duplication is deleted using union operation of between or among the attributes. Tuples 5, 6, and 7 are also duplicated and these types of duplication of records are removed by taking the union operation among the tuples but not among the attributes. In the second case sets of duplicate records are identified and then replaced with the one or more sets of real world and original or correct records.

INPUT:

Relations with duplicated tuples

OUTPUT:

Relations with duplicate tuples removed

1. For each tuple t ∈ delta do
2. In the relation R' find a set off tuples whose foreign
3. key matches with the primary key of the tuple t in
4. R. Let S_t be the set of such tuples
5. For all $t^* \in S_t$ replace foreign key values in R' with
6. the respective primary key of the tuple t ∈ R
7. End for
8. End for
9. For each set s_t find projected set of tuples based on
10. their primary key
11. End for
12. Now apply union operation for all S_t sets

IV. CONCLUSIONS

Data duplication is common in many real life applications. Records are duplicated in many relational databases because of many reasons such as inclusions of null values, non-standard method representation, and typographic errors. There is no standard method for identification of duplicated records in the relations of relational databases. When there exist no specific standard method for detecting duplicate records it is very difficult to find duplicate records. Hence, there is a scope for formulating specific standard methods for duplicate record detection.

REFERENCES

[1] Antoon Bronselaer, Daan Van Britsom, and Guy De Tr_e Propagation of Data Fusion IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 5, MAY 2015

[2] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, third ed. Morgan Kaufmann, 2011.

[3] Arun K. Pujari – Data Mining

[4] Machine Learning, Ethem Alpaydin
[5]machine Learning, Michel