SCHEMA TRANSFORMATIONS AND DEPENDENCY PRESERVATION

TAUQEER HUSSAIN
Faculty of Information Technology
University of Central Punjab (UCP)
31-Main Gulberg, Lahore-54660,
PAKISTAN
www.ucp.edu.pk/tauqeer

ABSTRACT: - For developing good quality information systems, the need of developing good quality conceptual models cannot be over emphasized. To improve the quality of a conceptual model, schema transformation rules have been proposed in our previous research. For applications developed using relational database management systems, conceptual models are translated into relational schema and the quality of a relational schema is determined in terms of the normal form satisfied by each relation in the database schema. Though various normal forms have been proposed in the literature, database designers, in practice, usually normalize a relational database schema up to third normal form (3NF) or Boyce-Codd Normal Form (BCNF). Deciding between 3NF and BCNF has always been an important design issue because a BCNF decomposition of a relation schema may lose dependency preservation. In this paper, we identify properties of the relations which are in 3NF and not in BCNF. Some of these properties have been informally stated in text books or in the literature without any formal proofs. We state these properties clearly and also present their formal proofs. This research facilitates in devising schema transformation rules that can further improve the quality of a conceptual model.

Keywords: - Schema transformations, ER model, Relational schema, Dependency preservation, 3NF, Conceptual model, BCNF

1 Introduction

Conceptual modeling is one of the most demanding and challenging steps in the database design methodology. The quality of an information system depends upon the quality of its conceptual model [1], and the quality of a conceptual model can be determined by measuring the proximity and accuracy with which it represents the problem domain. For conceptual modeling of database applications, the Entity-Relationship (ER) model [2-4] has been widely used in the industry for its ease of use and representation. However, the quality of an ER model is usually discussed subjectively. Also, to improve the quality of an ER model, a structured methodology does not exist. Therefore, appropriate metrics to measure the quality of an ER model and an approach to improve the quality have been proposed in the research [5-10]. The proposed approach [8] uses the set of functional dependencies (FDs) identified from a problem domain and a set of proposed schema transformation rules in order to transform a given ER model to a better quality ER model. As a result, every relation of the relational schema produced from the transformed model is always in third normal form (3NF) or higher [6]. As an example of schema transformations, Fig. 1 shows an ER diagram designed for a company database. Using the schema transformation rules [8] and the following set of functional dependencies identified from the problem domain, Fig. 1 is then transformed to Fig. 2.

FD1: P#, LOC → Cost
FD2: LOC → Super
FD3: ID → Name
FD4: ID → Dept#

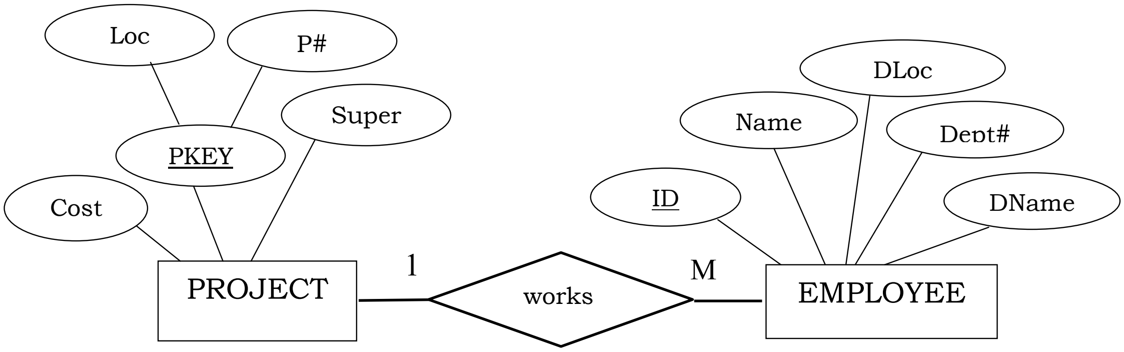

Fig. 1: An ER Diagram for a Company Database

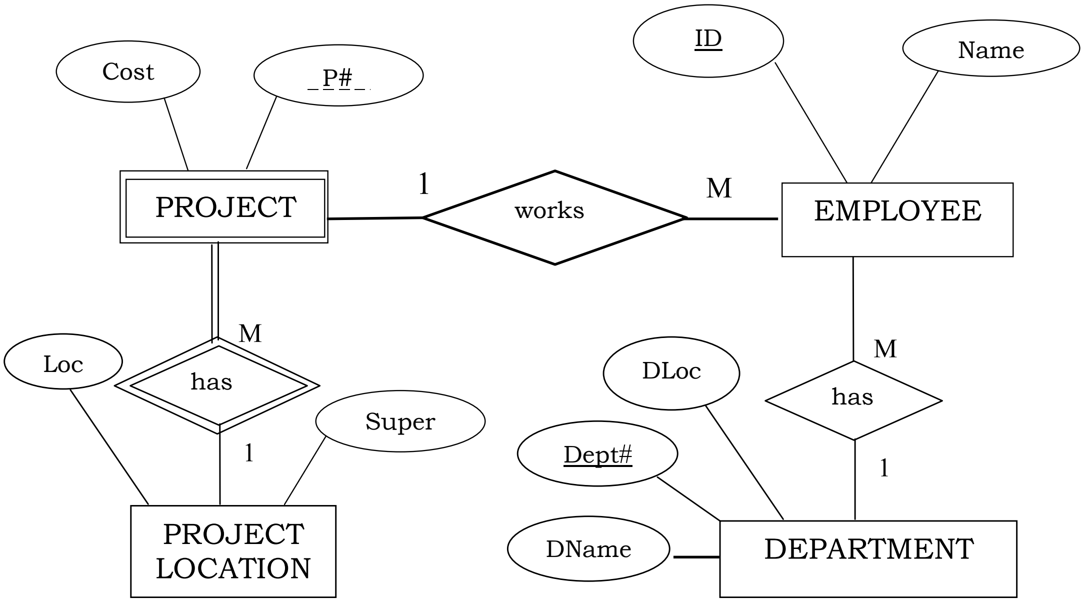

Fig. 2: An ER Diagram after Schema Transformations of Fig. 1

FD5: Dept# → DName
FD6: Dept# → DLoc
FD7: ID → P#
FD8: ID → LOC

Though the relational schema generated for Fig. 2 is guaranteed to be in 3NF at least, a higher normal form is always desirable from quality perspective. In this paper, we intend to establish that the resulting relational schema also satisfies Boyce-Codd Normal Form (BCNF) under certain conditions, and if it does not then any attempt to convert it to BCNF is not possible without losing dependency preservation. This work is important to devise schema transformation rules in future which can possibly generate relational schema in BCNF.

The relational database design approach requires the process of normalization in order to minimize data redundancy and update anomalies in the relational schema. Various normal forms evolved over time starting from first normal form (1NF) to 6NF or Domain-Key Normal Form (DKNF). However, in practice, database designers usually normalize a relational schema up to 3NF or BCNF. Deciding between 3NF and BCNF has always been an important design issue, and there is significant debate on it in the research. During the process of normalization, a relation schema is decomposed into a number of relation schemas and consequently the decomposition may lose one or more functional dependencies. Dependency preservation is a desirable property for good database design. Therefore, we are interested in schema transformations that can generate relational schema with dependency preservation. In this paper, we study the possibility of producing such schema by first stating and then proving certain properties which can provide a basis for further research. These properties are presented as Theorems and Lemma in sections 3 and 4.

This paper is organized into five sections. For ease of reference, definitions of the terms in relational model and normalization theory which are referred in this paper are reviewed in section 2. In section 3, we state and prove properties related to the relations which are in 3NF. In section 4, we state and prove an important theorem which discusses property of a relation which is in 3NF but not in BCNF. Finally, section 5 concludes our work.

2 Definitions Revisited

For ease of reference, we present definitions of the terms related to 3NF and BCNF. These definitions can be found in any text book on database systems e.g. [4,11] and in the literature.

Definition 1: **Superkey of a Relation Schema**

A set of attributes X is a superkey of a relation schema R ($A_1, A_2, \ldots, A_n$) if $X \subseteq R$ and a constraint on R states that there cannot be more than one tuple with a given X-value in any legal relation instance r(R). Mathematically,

$$t_i[X] \neq t_j[X] \quad \forall\ i \neq j$$

where $t_i[X]$ refers to an ordered set of values corresponding to X in i^{th} tuple of R.

Definition 2: **Key (or Candidate Key) of a Relation Schema**

A set of attributes $K = \{A_1, A_2, \ldots, A_k\}$ is a key (or a candidate key) of a relation schema R if and only if:

a) K is a superkey of R, and

b) $K - \{A_i\}$ is not a superkey of R $\forall i=1,2,\ldots,k$

Definition 3: **Prime Attribute of a Relation Schema**

An attribute A of a relation schema R is a prime attribute if $A \in K$ where K is any key of R. Mathematically,

$$A \in \bigcup_{i=1}^{m} K_i$$

where $K_1, K_2, \ldots, K_m$ is a set of all keys of R.

Definition 4: **Third Normal Form (3NF)**

A relation schema R is in 3NF if, whenever a nontrivial FD: $X \rightarrow Y$ holds in R, either

a) X is a superkey of R, or

b) Y is a prime attribute of R.

Definition 5: **Boyce-Codd Normal Form (BCNF)**

A relation schema R is in BCNF if, whenever a nontrivial FD: $X \rightarrow Y$ holds in R, then X is a superkey of R.

3 Properties of Relations in 3NF and BCNF

Consider a relation schema R($A_1, A_2, \ldots, A_n$) which is in 3NF and let K be the set of attributes such that $K \subseteq$ Attr (R) and $K \rightarrow A_1, A_2, \ldots, A_n$ where Attr(R) refers to a set of all attributes of R. Hence, K is a key of R and it cannot be null by entity integrity rule [4]. Now, there exists only one of the two possibilities:

1. K is a singleton set, or
2. K is composite

The first case where K is a singleton set is presented as Theorem 1 whereas the second case where K is composite is presented as Lemma 1. Proofs for the theorem and the lemma follow respectively. It should be noted that we have already presented and proved a theorem in [3] which deals only with the primary key of a relation schema R. Theorem 1 in this paper is a general form of the theorem in [3] as it is based on any key of R.

Theorem 1: A relation schema R($\underline{A_1}$, A_2, …, A_n) cannot violate BCNF if R satisfies 3NF and every key of R is a singleton set.

Proof: We prove this theorem by contradiction. Let there be a relation schema R which satisfies 3NF and violates BCNF. This implies that there exists a functional dependency FD: $X \rightarrow Y$ due to which R violates BCNF, and that,

X is not a superkey (1)

For the left hand side of the given FD, that is the determinant, there are following four possibilities that can exist:

a) $X = K$
b) $X \subset K$
c) $X \supset K$
d) $X \neq K$

Case a):

Since $X = K$, X is a superkey which negates (1) and our hypothesis.

Case b):

Impossible case because K is a singleton set.

Case c):

This case implies that X is a superkey which negates (1) and our hypothesis.

Case d):

As given in (1), X is not a superkey. So, Y has to be a prime attribute as R is in 3NF (given). That is, $Y \subseteq K'$ where K' is a key of R. But K' is a singleton set and so $Y = K'$ which implies $X \rightarrow K'$ or X is a superkey which negates (1) and our hypothesis.

Hence the proof.

Theorem 1 proves that if every key of a relation R is a singleton set and R satisfies 3NF, then R also satisfies BCNF and no decomposition or transformation is required. But this may not be the case, if at least one of the keys of R is composite. This is presented as Lemma 1.

Lemma 1: If R (A_1, A_2, …, A_n) is a relation schema which is in 3NF and has at least one composite key, then R may not satisfy BCNF.

Proof: Let K be a composite key of R. Now, there may exist a FD: $X \rightarrow Y$ such that $Y \subset K$ and X is not a superkey. In every such case, BCNF is violated (though 3NF is satisfied).

Lemma 2 (Converse of Lemma 1): If R (A_1, A_2, …, A_n) is a relation schema which is in 3NF but not in BCNF, then R has at least one composite key.

For its proof, we refer to [12]. In [12],Vincent & Srinivasan proved that a relation R which is in 3NF but not in BCNF has at least two overlapping candidate keys. Lemma 2 is its direct consequence and need not be proved here.

4 Dependency Preservation and BCNF

As stated earlier, dependency preservation is a desirable property for any decomposition of a relation schema R. If R is in 3NF and not in 3NF, then Theorem 2 proves that a dependency preservation decomposition of R cannot be achieved.

Theorem 2: If R is a relation schema which is in 3NF but not in BCNF then there does not exist a dependency preservation decomposition of R in which every relation satisfies BCNF.

Proof: By Lemma 2, let K be a composite candidate key of the relation R. Also, let FD: $X \rightarrow Y$ be a FD due to which R violates BCNF. Since R violates BCNF,

X cannot be a superkey. (2)

Since R is in 3NF, Y has to be a prime attribute. So,

$$Y \subset K \quad (3)$$

A decomposition D of R based upon the given FD: $X \rightarrow Y$ such that every relation in D is in BCNF can be written as:

$D = (R_1, R_2)$ where
$R_1 = R - Y$, and
$R_2 = X \cup Y$

We now prove that at least one FD is lost in D, that is, it exists neither in R_1 nor in R_2. Since $K \rightarrow (R_1 \cup R_2)$ being a key, it is a valid FD and we show that it is lost in D. For this purpose, it is sufficient to prove that K is present neither in R_1 nor in R_2.

For R_1, let $y \in Y$, then using (3), $y \in K$.

But, $y \notin (R-Y)$. Hence:

$$K \not\subseteq (R-Y)$$

For R_2, let us contradict by saying that K is present in R2. So,

$$K \subseteq (X \cup Y)$$

It implies:

$$(K\text{-}Y) \subseteq X \Rightarrow X \rightarrow (K\text{-}Y) \Rightarrow X \rightarrow K$$

which means X is a superkey contradicting (2).

Hence the proof.

As a result, if an ER model is translated to a relation schema in which a relation is in 3NF and not in BCNF, then its normalization to BCNF deprives us of the dependency preservation. This gives us the motivation to explore alternate designs in which such an ER model can be transformed to another ER model that generates a relational schema having every relation in BCNF. In our future work, we intend to devise schema transformation rules that can provide ER models of better quality based upon the above observations.

5 Conclusion

In this paper we have stated and proved the properties exhibited when a relation is in 3NF. We have noted that:

a) A relation schema that satisfies 3 NF and has no composite key also satisfies BCNF. Therefore, the schema transformation rules need no revision for this case.

b) A relation schema which is in 3NF and has at least one composite key may not be in BCNF. Hence, a transformation rule should be devised which can check this condition and then improve the ER model accordingly.

c) If a relation schema is in 3NF and not in BCNF, then it has at least one composite key. This finding is important for schema transformations as it can lead to defining entity types with split attributes of the composite key.

d) If a relation schema which is in 3NF and not in BCNF then there does not exist a dependency preservation decomposition of R in which every relation satisfies BCNF. This is probably the most important property in order to explore the possibility of developing an alternate ER model which can generate a different relational schema which satisfies BCNF and is dependency preserving.

References

[1] Moody, D. L. (2005): Theoretical and practical issues in evaluating the quality of conceptual models: current state and future directions. *Data & Knowledge Engineering* 55 (2005)

[2] Chen, P.P. (1976): The entity relationship model: towards a unified view of data. *ACM Transactions on Database Design*, 1(1), pp. 9-36

[3] Engels, G., Gogolla, M., Hohenstein, U., Hülsmann, K., Löhr-Richter, P., Saake, G., Ehrich D. (1992) Conceptual modelling of database applications using an extended ER model. *Data and Knowledge Engineering*, 9(2), pp. 157-204

[4] Elmasri, R., and Navathe, S. (2004) *Fundamentals of database systems,* 4th Ed. Pearson Education, Inc.

[5] Moody, D. L. (1998) "Metrics for Evaluating the Quality of Entity Relationship Models". *Proceedings of the 17th International Conference on Conceptual Modeling*, 1998, pp. 213-225

[6] Hussain, T., Shamail, S. and Awais, M. (2003) Eliminating Process of Normalization in Relational Database Design. *Proceedings of 7th IEEE International Multi-Topic Conference (INMIC 2003)*, December 8-9, 2003, Islamabad, Pakistan, pp. 408-413

[7] Hussain, T., Shamail, S. and Awais, M. (2004) Improving Quality in Conceptual Modeling. *Proceedings Companion 19th Annual ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA)*, October 24-28, 2004, Vancouver, British Columbia, Canada.

[8] Hussain, T., Shamail, S. and Awais, M. (2004) Schema Transformations – A Quality Perspective. *Proceedings of 8th IEEE International Multi-topic conference (INMIC 2004)*, December 24-26, 2004, Lahore.Pakistan.

[9] T. Hussain, Mian. M. Awais, S. Shamail (2005). A Fuzzy Based Approach to Measure

Completeness of an Entity-Relationship Model. Proceedings of ER 2005 workshops published as *Lecture Notes in Computer Science*, Volume 3770/2005, Springer, pp 410-422.

[10] Tauqeer Hussain, Mian M. Awais, Shafay Shamail (2007) Applying fuzzy logic to measure completeness of a conceptual model. *Applied Mathematics and Computation* 185(2): 1078-1086 (2007)

[11] Date, C. J. (2003) An Introduction to Database Systems, 8th Ed. Pearson Education, Inc.

[12] Vincent, W., Srinivasan, B. (1993) A Note on Relation Schemes which are in 3NF but not in BCNF. *Information Processing Letters* 48(6), December 1993, pp 281-283