数理解析研究所講究録
第353巻 1979年 11-19

11

RELATIONAL DATABASE DESIGN BY A SYNTHETIC APPROACH

Yahiko Kambayashi
Department of Information Sci.
Kyoto University
Sakyo, Kyoto, 606, Japan

1. Introduction

For relational database design two approaches(the decomposition approach and the synthetic approach) are known and both approaches have different advantages and problems. This paper gives a new design procedure based on the synthetic approach which solves the major problems of these two approaches. The characteristic features of the procedure are as follows. (1)The closure of the functional dependencies (FDs) realized by the designed relations is the same as the closure of the given set of FDs. (2)The universal relation (the relation consisting of all attributes) can be generated by lossless joins, thus the problem pointed out by Arora and Carlson is handled. (3)Multiple-key relations are utilized, since their use reduces the number of relations and thus the number of necessary joins in processing queries. A problem of Bernstein's algorithm is also solved. (4)First relations which reflect all FDs are designed and then decomposition with multi-valued dependencies (MVDs) is applied. Conditions for meaningful MVD decompositions on normalized relations are given. (5)Before applying the MVD decomposition, relations in Boyce-Codd normal form and third normal form are generated. At this stage, the number of relations in third normal form but not in Boyce-Codd normal form is minimized.

2. Basic concepts

A relation R_j on the set of attributes $A_1,\ldots,A_n$ is denoted by $R_j(A_1,\ldots,A_n)$. K_i, W_i, X_i, Y_i, Z_i are used to represent attribute

sets. A functional dependency(FD), a full functional dependency (FFD) and a multivalued dependency(MVD) are denoted by $\rightarrow$, $\Rightarrow$ and $\twoheadrightarrow$, respectively. Delobel and Casey [3] used Boolean variables a_i, b_i corresponding to attributes A_i, B_i, respectively. Each FD $g: A_1 \ldots A_k \rightarrow B_j$ corresponds to a logical product $\hat{g} = a_1 \ldots a_k \overline{b}_j$. For a given set F of FDs, the corresponding Boolean function $\hat{F}$ is defined as a logical sum of all logical products representing the FDs. In [3] it is shown that g is an FD in the closure F^+ of F if and only if $\hat{g}$ is an implicant of $\hat{F}$ and g is an FFD in F^+ if and only if $\hat{g}$ is a prime implicant of $\hat{F}$. Join is represented by * and the projection of R on X is represented by R[X]. Relations in first, second, third, Boyce-Codd and fourth nornal forms are defined in [2][4][8][12]. If R(X,Y,Z) is decomposed by the existence of an MVD $X \twoheadrightarrow Y \mid Z$ (an FD $X \rightarrow Y$), then the decomposition is called an MVD decomposition (an FD decomposition, respectively). The inverse of the decomposition defines MVD-based (FD-based, respectively) lossless join. Keys in a relation are specified by underlining when necessary. Other definitions in [8] are also used in this paper.

3. Problems of the database design

For the relational database design two approaches - decomposition and synthetic - are known [10].

In this section we will show that both of these approaches are insufficient for the database design by showing several problems.

P1: Relations generated by joins and projections require careful handling, even if the joins are lossless.

(1) Even in the decomposition approach, FDs which are not reflected by the decomposition must by stored explicitly in the table form instead of just the information of the existence of the FDs (the latter approach is used in [11]). (2) By the reason stated in (1), all FDs must be stored explicitly in the table form. In order to minimize storage space for storing all FDs, a procedure exactly the same as the synthetic approach must be used.

P2: Even if the closure of FDs of the relations are the same as the closure of the given set of FDs, some FD may not be obtained from the relation. This problem is pointed out by Arora and Carlson[18].

P3: MVDs must be handled.

Our design procedure solves these problems by the following approaches.

For P1: The base of our approach is the synthetic approach.

For P2: Joinability of the relation is considered. A procedure to examine whether or not the given set of FDs generates the universal relation by lossless joins is given. One relation is added when the given set cannot generate the universal relation.

For P3: Decomposability of a BCNF relation by MVDs is discussed. Sufficient conditions for nondecomposable BCNF relations are given.

4. Basic considerations on the design procedure

The best algorithm known based on the synthetic approach is the algorithm developed by Bernstein [8].

[Bernstein's algorithm to obtain a minimum schema in third normal form] [8]

(1) Eliminate extraneous attributes.
(2) Find a nonredundant cover H.
(3) Partition H into groups such that all of the FDs in each group have identical left sides.
(4) Merge two groups with left sides X and Y if there exists bijection $X \leftrightarrow Y$ in H^+. Repeat this step until no such X, Y are found (X and Y are equivalent keys).
(5) Eliminate transitive dependencies.
(6) Construct relations.

Problems of Bernsteins algorithm are that (a) the result is affected by the nonredundant set selected in step(2), and (b) the relation may not be in BCNF, but normally the third normal form property (there are no restrictions among prime attributes) is not used for the minimization. Because of these problems procedures for designing a minimum set of relations for several normalization classes are given in [16]. The design procedure given in this paper handles problems P1, P2 and P3. The outline is as follows.

Algorithm 1: Outline of the design procedure

(1) Generate all FFDs for the given set F of FDs.

(2) Obtain single-key relations in second normal form.

(3) Generate single-key relations in BCNF (Algorithm 2).

(4) Generate multiple-key relations in BCNF (Algorithm 3).

(5) Generate additional relations in third normal form if necessary in order to make the closure of FDs realized by the relations be F^+ (Algorithm 4).

(6) Remove redundant relations. Let the set of relations be S_1.

(7) Examine whether or not the universal relation is obtained by lossless join of relations in S_1. If the universal relation is obtained, find a minimum set S_2 of relations which will be used to form the universal relation such that the closure of FDs in S_2 is F^+ (Algorithm 5). If the universal relation cannot be obtained, find a minimum set of relations such that the closure of FDs is F^+, and add one relation in order to have the universal relation by lossless joins (Algorithm 6).

(8) Remove redundant attributes from the relations. Let S_3 be the resulting set of relations.

(9) Apply MVD decompositions to the relations in S_3 (Algorithm 7).

All FFDs in F^+ are obtained by calculating all prime implicants of $\hat{F}$. Classify FFDs such that each class contains FFDs of the same left-hand side set. For each class form a relation consisting of all attributes of the FFDs in the class. The key of this relation is the left-hand side set. By this way all single-key relations in second normal form can be obtained. Note that the relations may not be in second normal form for other keys (the hidden key problem).

Algorithm 2: Generation of single-key relations in BCNF.

(1) Obtain a set of single-key relations in second normal form.

(2) Modify the relations by the following steps.

(2-1) In relation R(K,Q) (K is the key and Q is the set of attributes such that $K \rightarrow Q$). if there exists an FFD $X \Rightarrow Y$ (Y is the maximum set) such that $X \neq K$ and $Y \neq K \cup Q$ (X is not an equivalent key), then add the following relations to the

relation list S and remove R(K,Q) from S.
(i) If $K \cap Y = \phi$, produce R'(K,Q-Y). $X \rightarrow Y$ is erased since Y is removed.
(ii) Let Z=X-K, then for every non-empty subset Z' of Z, produce R''(K,Q-Z'). Only a proper subset of X is permitted in R''. Note that none of these relations have the FFD $X \Rightarrow Y$.
(2-2) Repeat (2-1) until there are not such relations.

Algorithm 3: Generation of multiple-key relations in BCNF
(1) Obtain a set of single-key relations in BCNF by Algorithm 2.
(2) Obtain multiple-key relations by repeated applications of the following steps.
(2-1) Find two relations R_1 and R_2 satisfying the following conditions.

$$P_1 \cup Q_1 \supseteq P_2, \quad P_2 \cup Q_2 \supseteq P_1.$$

Here, P_i is a set of prime attributes in R_i, and Q_i is a set of nonprime attributes in R_i.
(2-2) Form a relation R_3, whose key sets are the union of the key sets of R_1 and R_2 and $P_3 = P_1 \cup P_2$, $Q_3 = Q_1 \cap Q_2 - P_3$.

Algorithm 4: Generation of additional relations in third normal form in order to make the closure of FDs realized by the relations be F^+ (F is the set of given FDs).
(1) Let F_1 be the set of FDs realized by the relations obtained by Algorithm 3. Calculate the corresponding logic function $\hat{F}_1$. If $\hat{F}_1 = \hat{F}$, then the closure of the FDs realized by the BCNF relations equeals the closure of the given FDs and no additional relations are required.

(2) If $\hat{F}_1 \neq \hat{F}$, calculate all prime implicants of $\bar{\hat{F}}_1 \cdot \hat{F}$. Obtain the set of relations in third normal form by FFDs corresponding to these prime implicants.

5. Joinable property of relations

Definition 1: A join graph G=(V,E,L) for a given set of relations is defined as follows. V is the set of vertices, each of which corresponds to a relation. E is the set of directed edges. L is a mapping from V to subsets of attributes. Here, E and L are determined as follows.

(1) $L(v_i)$ is initially the set of attributes of R_i which corresponds to v_i.
(2) If at least one key of R_i is contained in $L(v_j)$, a directed edge from v_j to v_i is created.
(3) After the creation of an edge in (2), $L(v_j)$ is replaced by the union of $L(v_i)$ and current $L(v_j)$.
(4) If $L(v_j)$ is modified and there is a directed edge from v_k to v_j, $L(v_k)$ may be required to be modified, since $L(v_k)$ is defined as the union of all possible $L(v_i)$'s and the attribute set of R_k, where v_i can be accessed from v_k through one directed edge. Repeat this process until no updates of L are required.
(5) Repeat steps (2), (3) and (4) until no additional edges are created.

Algorithm 5: Generation of a minimum set of relations such that the universal relation can be formed and the closure of FDs realized by the relations is the closure F^+ of the given set F of FDs, when the original set of relations can form the universal relation.
(1) Find a minimal subgraph of the join graph which still has the vertex whose L(v) is the set of all attributes.
(2) Calculate the closure F_o^+ of the set of F_o of FDs realized by the relations corresponding to the subgraph vertices. If $F_o^+ = F^+$, no additional relation is required, otherwise we add the minimum number of relations by the following steps.
(3) Rename relations as $R_1, R_2, \ldots, R_k$, which do not have the corresponding vertex in the subgraph. For each R_i $(i=1,\ldots,k)$ Boolean variable x_i is assigned. Obtain the solution for the following equality which can be formulated by the 0-1 Integer Programming.

$$\text{Minimize} \quad \sum_{i=1}^{k} x_i$$

Under the constraint:

$$\hat{F}_o + x_1\hat{F}_1 + x_2\hat{F}_2 + \ldots + x_k\hat{F}_k = \hat{F}.$$

where F_i is the set of FDs in relation R_i and x_i is 1 or 0.
(4) To the set obtained in (2), add every relation R_i whose corresponding value x_i is 1 in the solution. In the solution the number of non-BCNF relations should be minimized.

(5) For the minimization of the number of the relations, it is better to try all possible minimal subgraphs obtained in (1).

Theorem 1: If a set of relations $R_i(X_i,Y_i)$ (X_i is one of the keys)without the joinable property is given, we can have a set of relations with the joinable property if the relation with the following set of attributes is added.

$$X_1(X_2\text{-}X_1\text{-}Y_1)(X_3\text{-}X_1\text{-}X_2\text{-}Y_1\text{-}Y_2)\ \text{--}\ (X_n\text{-}\overset{n-1}{\underset{i=1}{\cup}}X_i\ \text{-}\overset{n-1}{\underset{i=1}{\cup}}Y_i)$$

Algorithm 6: Generation of a minimum set of relations such that the universal relation can be formed and the closure of FDs realized by the relations is the closure F^+ of the given set F of FDs, when the original set of relations cannot form the universal relation.

(1) Find a minumum set of relations satisfying that the closure of FDs realized by the relations is F^+ by applying the same procedure of step (3) of Algorithm 5, except that $\hat{F}_o = 0$ and that Boolean variables are assigned to all relations.

(2) Find an additional relation by Theorem 1 in order to form the universal relation.

6. MVD decompostions of BCNF relations

Definition 2: Decomposition of R(X,Y,Z) by an MVD : $X \twoheadrightarrow Y \mid Z$ is said to be non-effective if there exists an FD : $X \rightarrow YZ$.

Definition 3: Decomposition of R(X,Y,Z) by an MVD $X \twoheadrightarrow Y \mid Z$ is said to be key preserving if every key of R is contained in at least one of the decomposed relations, that is for every key K_i of R, $K_i \subseteq X \cup Y$ or $K_i \subseteq X \cup Z$.

Theorem 2: In a BCNF relation $R(X_1,X_2,X_3,Y)$ with at least one FD and without any constant-value attributes, only MVDs of the form $X_1Y \twoheadrightarrow X_2 \mid X_3 (Y \neq \phi,\ X_2 \neq \phi,\ X_3 \neq \phi)$ can be used for the effective decomposition of R, where Y is the set of nonprime attributes, X_1, X_2, X_3 are prime attributes and X_2, X_3 are attributes contained in every key of R.

Corollary 1: In a BCNF relation $R(X_1,X_2,X_3,Y)$ with at least one FD and with no constant-value attributes, there exists

no MVDs which will realize key preserving and effective decomposition.

Corollary 2: For a BCNF relation with at least one FD and without constant value attributes, there exists no effective MVD decomposition if at least one of the following conditions is satisfied.

(1) All attributes in R are prime.

(2) There exists a key consisting of one attribute.

(3) There exists two keys, each of which consists of two attributes.

(4) In general at most one attribute is contained in common in all keys.

Algorithm 7: Generation of fourth normal form relations.

(1)Find BCNF relations for which we cannot apply MVD decompositions using Corollary 2.

(2)For other BCNF relations find an MVD satisfying the condition of Theorem 2 and apply the MVD decomposition.

(3)For third normal form relations (added by Algorithm 4) and a all-key relation(added by Algorithm 6),apply MVD decompositions.

Acknowledgment

The author wishes to thank Professor Shuzo Yajima and Mr.Katsumi Tanaka for valuable discussions and comments on the subject.

References

[1] E.F.Codd,"A relational model of data for large shared data banks," CACM, vol.13, no.6, pp.377-387,June 1970.

[2] E.F.Codd,"Further normalization of the data base relational model," in Data Base Systems, Prentice-Hall, 1972.

[3] C.Delobel and R.G.Casey,"Decomposition of a data base and the theory of Boolean switching functions," IBM J. Res. Develop., vol.17, no.5, pp.374-386, Sept. 1972.

[4] E.F.Codd,"Recent investigation into relational data base systems," IFIP 74, pp.1017-1021, Aug. 1974.

[5] W.W.Armstrong,"Dependency structures of data base relationships," IFIP 74, pp.580-583, Aug. 1974.

[6] C.P.Wang and H.H.Wedekind,"Segment synthesis in logical data base design," IBM J. Res. Develop., vol.19, no.1, pp.71-77, Jan. 1975.

[7] C.Zaniolo,"Analysis and design of relational schemata for database systems," Computer Science Dept., UCLA, Technical Report, UCLA-ENG-7769, July 1976.

[8] P.A.Bernstein,"Synthesizing third normal form relations from function dependencies," ACM TODS, vol.1, no.4, pp. 277-298, Dec. 1976.

[9] C.Beeri, R.Fagin and J.H.Howard,"A complete axiomatization for functional and multivalued dependencies," ACM-SIGMOD Conf., pp.47-61, Aug. 1977.

[10] R.Fagin,"The decomposition versus the synthetic approach to relational database design," 3rd VLDB, pp.441-446, Oct. 1977.

[11] Y.Tanaka and T.Tsuda,"Decomposition and composition of a relational database," 3rd VLDB, pp.454-461, Oct. 1977.

[12] R.Fagin,"Multivalued dependencies and a new normal form for relational databases," ACM TODS, vol.2, no.3, pp.242-278, Sept. 1977.

[13] R.Fagin,"Dependendy in a relational database and propositional logic," IBM J. Res. Develop., vol.21, no.6, pp.534-544, Nov. 1977.

[14] J.Rissanen,"Independent component of relations," ACM TODS, vol.2, no.4, pp.317-325, Dec. 1977.

[15] E.A.Lewis, L.C.Sekino and P.D.Ting,"A canonical representation for the relational schema and logical data independence," First COMPSAC, pp.276-280, Nov. 1977.

[16] Y.Kambayashi,"Equivalent key problem of the relational database model," International Conf. on Mathematical Studies of Information Processing, Kyoto, pp.155-181, Aug. 1978 (to appear in Lecture Notes in Computer Science).

[17] C.Beeri, P.A.Bernstein and N.Goodman,"A sophisticate's introduction to database normalization theory," 4th VLDB, pp.113-124, Sept. 1978.

[18] A.K.Arora and C.B.Carlson,"The information preserving properties of relational database transformations," 4th VLDB, pp.352-359, Sept. 1978.