数理解析研究所講究録
458巻 1982年 162-171

RECOVERY OF INCOMPLETE TABLES UNDER DATA DEPENDENCIES

SATORU MIYANO, Dept. of Math., Kyushu Univ.

MAKOTO HARAGUCHI, Res. Inst. Fund. Inf. Sci., Kyushu Univ.

INTRODUCTION

In this paper we deal with the problems arising from incomplete information recovery and study their computational complexity. In general, incomplete information recovery is to recover incomplete information so that the resulting information will be consistent with the knowledge which is supposed for the information. We focus our attention on the relational model of data and we use various data dependencies as the knowledge in recovery. As a representation of relations we use tables. Incomplete tables are introduced and the complexities of recovery problems are analyzed in various situations. We consider two types of incomplete tables. The one is called finite type and the other infinite type. A finite type incomplete table has entries with finite sets of values. On the other hand, for infinite type incomplete tables, each entry is either known or unknown.

For the knowledge in recovery, we mainly concentrate on functional dependency. With respect to the complexity of recovery, in order to elucidate the difference between functional dependency and several another known dependencies, we consider multivalued dependency (Codd (1971)), two-fold first order hierarchical decomposition (Delobel (1978)) and mutual dependency (Nicolas (1978)).

It is likely that there are several recoveries which are consistent with the given knowledge. In a practical sense, such situation is not preferable. So, we also pay attention to unique recoverability.

From these points, various complexity results will be presented. For example, we show that the recoverability problem for finite type incomplete tables under functional dependencies is NP-complete but the unique recoverability problem is co-NP-hard. On the other hand, for infinite type incomplete tables, these problems are shown to have polynomial time solutions and moreover the completeness for P is shown. We also show the difference when another dependencies are used as the knowledge. As a model of tables which varies from time to time, we introduce time variant tables. We consider dynamic recovery problem under functional dependencies and show that this problem is PSPACE-complete.

1. INCOMPLETE TABLES AND RECOVERABILITY

DEFINITION 1.1. U is a set called an attribute set and elements in U are called attributes. It is assumed that a set D(A) called the attribute domain of A is associated with each attribute A in U. An incomplete table T over the attribute set U is a matrix indexed by $\{1,\ldots,n\} \times U$ for some $n \geq 0$ such that for each attribute A in U the entries in the column of A are nonempty subsets of D(A). An incomplete table is said to be complete if each table entry is a sigleton, that is, a set consisting of a single element.

DEFINITION 1.2. Let T be an incomplete table over an

(1) T is said to be of finite type if each table entry of T is a finite set.

(2) If each attribute domain D(A) of A in U is infinite and each table entry is a singleton or the attribute domain itself, then T is said to be of infinite type.

We now define functional dependencies, multivalued dependencies (Codd (1970,1971)), two-fold first order hierarchical decompositions (Delobel

(1978)) and mutual dependencies (Nicolas (1978)).

NOTATION. Let T be a complete table over an attribute set U and let X be a subset of U. For a row t in T, we denote by t[X] the tuple obtained by projecting t on X.

DEFINITION 1.3. Let T be a complete table over an attribute set U and let X and Y be nonempty subsets of U. Then we say that the functional dependency (FD) $X \to Y$ holds in T if for all rows t_1 and t_2 in T, $t_1[Y]=t_2[Y]$ whenever $t_1[X]=t_2[X]$.

DEFINITION 1.4. Let X, Y and Z be nonempty disjoint subsets of an attribute set U. We say that the two-fold first order hierarchical decomposition (2HD) $X \to Y|Z$ holds in the complete table T over U if there exists a row t in T such that t[X]=x, t[Y]=y and t[Z]=z whenever there are rows t_1 and t_2 IN T with $t_1[X]=t_2[X]$, $t_1[Y]=y_2$ and $t_2[Z]=z$. If we choose Z=U-X-Y in 2HD, then instead of $X \to Y|U-X-Y$ we say that the multivalued dependency (MVD) $X \twoheadrightarrow Y$ holds in T.

DEFINITION 1.5. Consider nonempty disjoint subsets X and Y of U and let Z=U-X-Y. Then we say that the mutual dependency (MD) $X \leftrightarrow Y$ holds in the complete table T over U if there exists a row t in T such that t[X]=x, t[Y]=y and t[Z]=z whenever there are rows t_1, t_2 and t_3 in T with $t_1[X]=t_3[X]=x$, $t_1[Y]=t_2[Y]=y$ and $t_2[Z]=t_3[Z]=z$.

DEFINITION 1.6. Let F be a family of FDs (MVDs, 2HDs or MDs) over U and let T be a complete table over U. Then T is said to be consistent with F if all constraints in F hold in T.

DEFINITION 1.7. An incomplete table T' is said to be an extension of an incomplete table T if T and T' have the same attribute set and the same number of rows and each table entry of T' is a subset of the corresponding entry of T.

DEFINITION 1.8. Let T be an incomplete table over U and let F be a family of FDs (MVDs, 2HDs or MDs).

(1) The incomplete table T is said to be recoverable under F if there is an extension of T which is complete and consistent with F.

(2) The incomplete table T is said to be uniquely recoverable under F if there is exactly one extension of T which is complete and consistent with F.

2. RECOVERY OF FINITE TYPE INCOMPLETE TABLES

This section is devoted to the study of computational complexity of finite type incomplete table recovery under functional dependencies. Problems are analyzed from two points. The one is whether incomplete tables with which we are concerned are of finite type or of infinite type. The other point is the difference between recoverability and unique recoverability. The following two problems will be analyzed:

FINITE TYPE RECOVERABILITY UNDER FD

INSTANCE: A finite type incomplete table T and a finite collection F of functional dependencies.

QUESTION: Is T recoverable under F?

FINITE TYPE UNIQUE RECOVERABILITY UNDER FD

INSTANCE: A finite type incomplete table T and a finite collection F of functional dependencies.

QUESTION: Is T uniquely recoverable under F?

THEOREM 2.1. FINITE TYPE RECOVERABILITY UNDER FD is NP-complete.

Proof: Reduction from 3SATISFIABILITY (Garey and Johnson (1979)).

THEOREM 2.2. FINITE TYPE UNIQUE RECOVERABILITY UNDER FD is co-NP-hard.

Proof: Reduction from UNIQUE 4SATISFIABILITY, which is shown to be co-NP-hard.

UNIQUE 4SATISFIABILITY (U4SAT)

INSTANCE: A finite collection C of clauses over some variable set V such that each clause c in C has $|c|=4$.

QUESTION: Is C uniquely satisfiable, i.e., is there exactly one truth assignment which satisfies all clauses in C?

LEMMA 2.3. UNIQUE 4SATISFIABILITY is co-NP-hard.

REMARK. It is not hard to see that FINITE TYPE UNIQUE RECOVERABILITY UNDER FD is in Δ_2^p, the class of problems solvable by deterministic polynomial time orcle Turing machines with oracles in NP.

3. RECOVERY OF INFINITE TYPE INCOMPLETE TABLES

We now consider the case of infinite type incomplete tables. In contrast with the case of finite type incomplete tables, where all problems are computationally hard, we see that infinite type incomplete table recovery problems are tractable.

INFINITE TYPE RECOVERABILITY UNDER FD

INSTANCE: An infinite type incomplete table T and a finite collection F of functional dependencies.

QUESTION: Is T recoverable under F?

INFINITE TYPE UNIQUE RECOVERABILITY UNDER FD

INSTANCE: An infinite type incomplete table T and a finite collection F of functional dependencies.

QUESTION: Is T uniquely recoverable under F?

THEOREM 3.3. INFINITE TYPE RECOVERABILITY UNDER FD is complete for P under log space reductions.

THEOREM 3.4. INFINITE TYPE UNIQUE RECOVERABILITY UNDER FD is complete for P under log space reductions.

4. REDUNDANCY MINIMIZATION PROBLEM

Let T be a complete table which is consistent with a family F of FDs, where all attribute domains are assumed to be infinite. If an incomplete table T' is uniquely recoverable to T under the family F of FDs, then the values in the entries of T which are occupied by variable symbols x[i] in T' may be considered as redundant information. There are possibly many such incomplete tables which are uniquely recoverable to T. Since the redundant information is represented by variable symbols, the larger number of variable symbols means less redundancy. Then the problem to find an incomplete table with maximum number of variables which is still uniquely recoverable to the original complete table may receive attention. We consider the complexity of the following problem:

REDUNDANCY MINIMIZATION UNDER FD

INSTANCE: A complete table T, a finite family F of FDs such that all FDs in F hold in T, and a nonnegative integer K.

QUESTION: Is there an infinite type incomplete table T' such that the number of variables in T' is not less than K and T' is uniquely recoverable under F to the original complete table T when all attribute domains are assumed to be infinite?

THEOREM 4.1. REDUNDANCY MINIMIZATION UNDER FD is NP-complete.

By Theorem 4.1, to minimize the redundancy under FDs is computational-

ly hard. However, if the families of FDs are restricted to the acyclic families, which is defined below, then a polynomial solution is feasible.

DEFINITION 4.1. Let U be an attribute set and F be a finite set of FDs over U. We define a directed graph $G_F=(V_F,E_F)$ called the dependency graph of F in the following way:

(1) The set V_F of nodes consists of all attributes in U.

(2) (A,B) is an edge of G_F if there is an FD X Y such that A is in X and B is in Y.

The family F of FDs is said to be acyclic if the dependency graph G is acyclic.

THEOREM 4.2. Let T be a complete table over an attribute set U and let F be a finite family of FDs. Assume that for each attribute A in U the attribute domain D(A) is infinite and all FDs in F holds in T. Then if F is an acyclic family of FDs, then we can compute in polynomial time an infinite type incomplete table with maximum number of variables which is uniquely recoverable to T under F.

5. RECOVERABILITY UNDER ANOTHER DEPENDENCIES

This section presents some complexity results concerning recoverability under multivalued dependencies (MVDs), two-fold first order hierarchical decompositions (2HDs) and mutual dependencies (MDs). The problems dealt with in this section are the same problems considered in Section 2 except that MVDs, 2HDs and MDs are are used instead of FDs. Formal definitions of these problems are omitted. Proofs of the results stated below requires rather complicated reductions.

THEOREM 5.1. FINITE TYPE RECOVERABILITY UNDER MVD is NP-complete.

THEOREM 5.2. FINITE TYPE UNIQUE RECOVERABILITY UNDER MVD is co-

NP-hard.

Since MVD is a special case of 2HD and FINITE TYPE RECOVERABILITY UNDER 2HD is easily seen to be in NP, we have the following corollaries of Theorems 5.1 and 5.2:

COROLLARY 5.3. FINITE TYPE RECOVERABILITY UNDER 2HD is NP-complete.

COROLLARY 5.4. FINITE TYPE UNIQUE RECOVERABILITY UNDER 2HD is co-NP-hard.

For MVDs we do not know the complexity of recoverability problems for infinite type incomplete tables. However, for 2HDs, in contrast with the case of FDs, we can show that both of recoverability and unique recoverability are hard problems.

THEOREM 5.5. INFINITE TYPE RECOVERABILITY UNDER 2HD is NP-complete.

THEOREM 5.6. INFINITE TYPE UNIQUE RECOVERABILITY UNDER 2HD is co-NP-hard.

THEOREM 5.7. FINITE TYPE RECOVERABILITY UNDER MD is NP-complete.

THEOREM 5.8. FINITE TYPE UNIQUE RECOVERABILITY UNDER MD is co-NP-hard.

6. RECOVERY OF INCOMPLETE TIME VARIANT TABLES

This section introduces the notion of time variant tables and discuss the complexity of recoverability. FDs are used as the knowledge in recovery.

DEFINITION 6.1. Let $U = \{A_1,\ldots,A_n\}$ be a finite set of attributes. For each integer T we define $U^{(t)} = \{A_i(t) : 1 \le i \le n\}$ and the attribute set U^{∞} is defined as the union of $U^{(t)}$, where t ranges over all integers. For each attribute A_i, we assume that $D(A_i)=D(A_i(t))$ for all t.

DEFINITION 6.2. Let U and U^{∞} be as above. An incomplete time variant table over U is an incomplete table over U^{∞}. We define complete time

variant tables over U in the same way.

DEFINITION 6.3. Let $X = \{A_{i_1}(0),\ldots,A_{i_k}(0),A_{i_{k+1}}(1),\ldots,A_{i_p}(1)\}$ and let $Y = \{A_{j_1}(0),\ldots,A_{j_m}(0),A_{j_{m+1}}(1),\ldots,A_{j_q}(1)\}$. Given an FD $X \to Y$, we define the dynamic expansion of $X \to Y$ as the family of FDs $\{X^{(t)} \to Y^{(t)} : t \text{ is an integer}\}$, where for each t we define $X^{(t)} = \{A_{i_1}(t),\ldots,A_{i_k}(t),A_{i_{k+1}}(t+1),\ldots,A_{i_p}(t+1)\}$ and $Y^{(t)} = \{A_{j_1}(t),\ldots,A_{j_m}(t),A_{j_{m+1}}(t+1),\ldots,A_{j_q}(t+1)\}$. In the same way we define the dynamic expansion F^{∞} when we are given a family F of FDs over $U^{(0)} \cup U^{(1)}$.

DEFINITION 6.4. Given an incomplete table T over U, we define an incomplete time variant table T^{∞} over U called the dynamic expansion of T by repeating T at each time. More formally, T has the same number of rows and for each attribute A_i in U, (r,A_i)-entry of T is equal to $(t,A_i(t))$-entry of T for each T and each row r.

DEFINITION 6.5. Let T be an incomplete table T over U. Given a family F of FDs over $U^{(0)} \cup U^{(1)}$, T is said to be dynamically recoverable under F if the dynamic expansion T^{∞} of T is recoverable under the dynmaic expansion F^{∞} of F.

We consider the following recovery problem:

DYNAMIC RECOVERABILITY UNDER FD

INSTANCE: A finite type incomplete table T over an attribute set U and a finite collection F of FDs over $U^{(0)} \cup U^{(1)}$.

QUESTION: Is T dynamically recoverable?

We obtained the following theorem:

THEOREM 6.1. DYNAMIC RECOVERABILITY UNDER FD is PSPACE-complete.

REFERENCES

Codd, E.F. (1970), A relational model of data for large shared data banks, Comm. ACM 13, 377-397.

Codd, E.F. (1971), Further normalization of the database relational model, Courant Computer Science Symposia 6, "Data Base Systems," New York, Prentice-Hall, 33-64.

Cook, S.A. (1974), An observation on time-storage trade-off, J. Comput. Systems Sci. 9, 308-316.

Delobel, C. (1978), Normalization and hierarchical dependencies in the relational data model, ACM Trans, on Database Systems 3, 201-222.

Haraguchi, M. and S. Miyano (1982), On minimizing redundant data in relations using functional dependencies, Research Institute of Fundamental Information Science, Research Report No. 100.

Garey, M.R. and D.S. Johnson (1979), Computers and Intractability, W.H. Freeman and Company, San Francisco.

Lingas, A. (1978), A PSPACE-complete problem related to a pebble game, Proc. of the Fifth Colloquium on Automata, Languages and Programming (Lecture Notes in Computer Science 62) G. Ausiello and C. Bohm, Ed., 300-321.

Miyano, S. and M. Haraguchi (1982), Recovery of incomplete information, to appear in Bull. Informatics and Cybernetics 20.

Miyano, S. and M. Haraguchi, The complexity of incomplete relation recovery, in preparation.

Nicolas, J. (1978), Mutual dependencies and some results on undecomposable relations, Proc. of the Fourth Int. Conf. on Very Large Data Bases, 360-367.