Computer Science Journal of Moldova, vol.**17**, no.1(49), 2009

An approach for testing the primeness of attributes in relational schemas

Cotelea Vitalie

Abstract

In this paper there is proposed a method of partition the attributes of relation scheme in equivalence classes and in nonredundant equivalence classes. Several properties of these equivalence classes are proved. Their properties serve as the basis for an algorithm with a polynomial complexity, which determines the prime attributes of a database schema.

Keywords: Relation scheme, functional dependencies, equivalence classes, prime attributes, polynomial complexity tasks.

1 Introduction

The scope of this paper is to propose a solution to the problem that arises during design and analysis of database, that is determination of prime attributes (attributes that are contained in schema's possible keys). This problem is known to be NP-complete, due to the fact that the solution to this problem was reached through keys searching. But a schema can have an exponential number of keys with respect to number of functional dependencies [1].

In the current paper a different approach is taken for the searching of prime attributes that avoids the necessity of keys determination. Namely, the notion of contribution graph (Definition 1) of a reduced set of functional dependencies is proposed. The strongly connected components are computed, where each component represents a vertex of condensed graph (Definition 2). Over vertices of condensed graph a

©2009 by Vitalie Cotelea

strict partial order is defined. Then it's presented how the inferred dependencies are reflected in contribution graph (Lemma 1 and Corollary 1).

Obviously, the strongly connected components of contribution graph split the set of attributes of relation scheme into equivalence classes. The notion of nonredundant equivalence classes of attributes is given (Definition 3). In section 4 several lemmas and theorems (Lemmas 2-3, Theorems 1-4) are proved that reflect the properties of equivalence classes of attributes.

It should be mentioned that redundant attributes represent the set of nonprime attributes of scheme (Corollary 4), and nonredundant equivalence classes of attributes consist only of prime attributes (Corollary 3). Proved properties in section 4, allow the determination of prime and nonprime attributes without scheme's keys finding.

In section 5 it is shown that the determination of prime and nonprime attributes can be performed in a polynomial time. This approach can be a part of the database analysis and design toolset.

2 Some basic concepts

In order to facilitate exposure of this paper's material, some preliminary notions are presented [2].

Let $Sch(R, F)$ be a relation scheme, where F is a set of functional dependencies defined on set R of attributes. Given a set F of functional dependencies on R, the *closure* of F, written as F^+, consists of all functional dependencies that are logically implied by F, that is $F^+ = = \{V \to W | F| = V \to W\}$.

Given a set F of functional dependencies on set R of attributes and a subset X of R, the *closure* of the set X under the set F, written as X^+, contains all attributes, each of which is functionally dependent on Xunder F, that is $X^+ = \{A | X \to A \in F^+\}$.

Let X and Y be two sets of attributes, where $X, Y \subseteq R$. The set X is a determinant for Y, under the set F of functional dependencies, if $X \to Y \in F^+$ and for every proper subset X' of the set X, the expression $X' \to Y \notin F^+$ takes place.

A subset K of R is a key for a relation scheme $Sch(R, F)$, if K is a determinant of the set R under the set F of dependencies. A relation scheme can have more than one key, but it always has at least one.

An attribute A in R is *prime* if A belongs to some key, and *nonprime* otherwise.

In this paper, it is considered that the set F of functional dependencies is reduced. Let $Sch(R, F)$ be a relation scheme. The set of functional dependencies F is reduced [2], if there is no attribute A in R and no dependency $X \to Y$ in F, so that they satisfy the following conditions:

1. $A \in X$ and $F \equiv F - \{X \to Y\} \bigcup \{(X - \{A\}) \to Y\}$,
2. $A \in Y$ and $F \equiv F - \{X \to Y\} \bigcup \{X \to (Y - \{A\})\}$.

For functional dependencies an inference tool, named maximal derivation [3], will be used. Maximal derivation of the set X of attributes under the set F of dependencies, is a sequence of sets $H =< X_0, X_1, ..., X_n >$ of attributes, where

1. $X_0 = X$;
2. $X_i = X_{i-1} \bigcup Z$, where $Z = \bigcup_j W_j$ for $\forall V_j \to W_j \in F$ that satisfy $V_j \subseteq X_{i-1}$ and $W_j \not\subset X_{i-1}$;
3. Nothing else is in X_i.

The last term of maximal derivation X_n is, in fact, the closure of the set X of attributes under the set F of dependencies, that is $X_n = X^+$.

Claim 1. [3]. *$X \to Y \in F^+$, if and only if there exists a derivation $H =< X_0, X_1, ..., X_k >$ for $X \to Y$ under F, where X_k is the first term that contains the set of attributes Y.*

Claim 2. [3]. *If $X \to Y \in F^+$ and X is a determinant for Y under F, then for every attribute A in $X - Y$ there exists in F a dependency $V \to W$ used in derivation $H =< X_0, X_1, ..., X_k >$ for $X \to Y$ under F, such that $A \in V$.*

3 Graphical representation of functional dependencies

Given a set F of functional dependencies on the set R of attributes, that are part of the relation scheme $Sch(R, F)$, a contribution graph is drawn, in order to represent F.

Definition 1. *Contribution graph $G = (S, E)$ of set F is a graph that:*

- *$\forall A \in R$ there exists in S a vertex labeled with attribute A;*
- *$\forall X \to Y \in F$ and $\forall A \in X$ and $\forall B \in Y$ there exists in E an edge $a = (A, B)$, that is directed from vertex A to vertex B.*

Example 1. *If $F = \{C \to B, AD \to B, AB \to DC, B \to E\}$ and $R = \{A, B, C, D, E\}$ then the contribution graph of set F of dependencies is presented in Figure 1.*

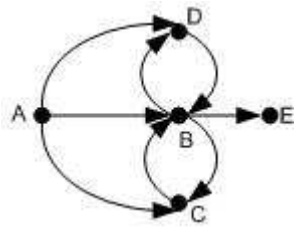

Figure 1. A contribution graph for set F

Two vertices $A, B \in S$ are strongly connected, if and only if there exists in graph G a path from A to B and backwards, from B to A. It is obvious that the relation of strong connectivity is an equivalence relation. So, there is a partition of set of vertices S into pairwise disjoint subsets. That is, $S = \bigcup_{i=1}^{n} S_i$ and all vertices in S_i, $i = \overline{1, n}$, are strongly connected, and every two vertices from different subsets are not strongly connected.

In accordance with this partition, subgraphs $G_i = (S_i, E_i)$, $i = \overline{1,n}$ are called strongly connected components [4] of the graph G, where E_i represents the set of edges that connect pairs of vertices in S_i.

Example 2. *The set of vertices of the graph represented in Figure 1 are split into three equivalence classes* $S_1 = \{A\}$, $S_2 = \{B, C, D\}$ *and* $S_3 = \{E\}$.

Definition 2. *Let* G^* *be the condensed graph of the graph* G. *Set of vertices of graph* G^* *represents set* $\{G_1, ..., G_n\}$ *of all strongly connected components of graph* G *and there is an edge from vertex* G_i *to vertex* G_j *of graph* G^*, *if there exists in* G *at least one edge that connects one vertex from component* G_i *to one vertex from component* G_j.

Obviously the graph G^* is an acyclic one.

Example 3. *The condensed graph of graph from Figure 1 has three vertices and two edges, as shown in Figure 2.*

Figure 2. Condensed graph of the graph from Figure 1

Over the set of vertices of graph G^* a strict partial order is defined. Vertex G_i precedes vertex G_j, if G_j is accessible from G_i. Now, the equivalence classes $S_1, ..., S_n$ will be sorted based on the corresponding order graph's G^* vertices.

Lemma 1. *If* $X \to Y \in F^+$ *and* X *is a determinant of set* Y *under* F, *then for every attribute* $A \in (X - Y)$ *there is an attribute* $B \in Y$ *so that in the contribution graph* G *there exists a path from vertex* A *to vertex* B *and for every attribute* $B \in (Y - X)$ *there exists in* X *an attribute* A, *from which the vertex* B *can be reached.*

Proof. Let attribute $B \in (Y - X)$ and let the subset X' of set X be determinant for B under F. Because $X' \to B \in F^+$, according to

Claim 1, there is a derivation $H =< X_0^{'}, X_1^{'}, ..., X_m^{'} >$ for dependency $X' \rightarrow B$ under F. Then, based on Claim 2, there exists a sequence of dependencies $V_1 \rightarrow W_1, ..., V_q \rightarrow W_q$ in F, where $A \in V_1$, $B \in W_q$ and $W_i \bigcap V_{i-1} \neq \emptyset$, for $i = \overline{1, q-1}$.

Contribution graph has a structure, such that for every dependency $V_j \rightarrow W_j$ in F, from each vertex labeled with an attribute in V_j an edge leaves to every vertex labeled with an attribute in W_j. So, there exists a path from every vertex $A \in X'$ to vertex B.

It must be mentioned that, if $\overline{X}$ is considered the union of all determinants of attributes in $Y - X$, then $\overline{X} \bigcup X \bigcap Y = X$. Indeed, if we suppose that the set $\overline{X} \bigcup X \bigcap Y$ is a proper subset of set X, this will contradict the supposition that X is a determinant for Y under F.

Corollary 1. *If reduced dependency $V \rightarrow W$ is used nonredundantly in building the derivation H for dependency $X \rightarrow Y$ under F, then in contribution graph G there exists a path from every vertex labeled with an attribute in V to every vertex labeled with an attribute in Y.*

4 Properties of equivalence classes of attributes

Theorem 1. *If X is a determinant under F of set $S_1 \bigcup ... \bigcup S_j$, where $j = \overline{1, n}$, then $X \subseteq S_1 \bigcup ... \bigcup S_j$.*

Proof. Let $X \not\subset S_1 \bigcup ... \bigcup S_j$. Then there exists an equivalence class S_t, where $t = \overline{j, n}$, such that $X \bigcap S_t \neq \emptyset$. By Lemma 1, in the contribution graph G, from every attribute $A \in X \bigcap S_t$ there is a path towards B, where $B \in S_1 \bigcup ... \bigcup S_j$. But this fact contradicts the supposition that the sets $S_1, ..., S_j$ precede the set S_t.

Corollary 2. *If X is a determinant of set $S_1 \bigcup ... \bigcup S_n$ under F, then $X \bigcap S_1 \neq \emptyset$.*

Proof. Indeed, for every attribute B in S_1 or $B \in X$, or, according to Lemma 1, there is in X an attribute A from which vertex B is accessible in contribution graph G. But then A is also a member of equivalence class S_1.

Definition 3. *Equivalence class S_j is called nonredundant, if and only if for every attribute A in S_j, the expression $(\bigcup_{i=1}^{n} S_i - S_j) \rightarrow A \notin F^+$ holds.*

Considering Lemma 1, it can be concluded that set S_j is nonredundant, if and only if for every attribute A in S_j, the expression $(\bigcup_{i=1}^{j-1} S_i) \rightarrow A \notin F^+$ holds.

From the ordered sequence of sets $S_1, ..., S_n$ a sequence of ordered nonredundant sets can be built $T_1, ..., T_n$, where $T_1 = S_1$ and $T_j = = S_j - (\bigcup_{i=1}^{j-1} T_i)_F^+$ for $j = \overline{2, n}$. As a result of this process, some sets T_j can become empty. These empty sets can be excluded from the sequence and a sequence of nonempty sets $T_1, ..., T_m$ will be obtained, keeping the precedence of prior sets.

Proposition 1. $T_1 = S_1$.

Proposition 2. $(T_1 \bigcup ... \bigcup T_m) \rightarrow (S_1 \bigcup ... \bigcup S_n) \in F^+$.

Example 4. *Sequence of equivalence classes of attributes $S_1 = \{A\}$, $S_2 = \{B, C, D\}$ and $S_3 = \{E\}$ turns into the following sequence of non redundant equivalence classes of attributes: $T_1 = \{A\}$, $T_2 = \{B, C, D\}$.*

Theorem 2. *Set X is a determinant of set $S_1 \bigcup ... \bigcup S_n$ under F, if and only if X is determinant of set $T_1 \bigcup ... \bigcup T_m$ under F.*

Proof. Necessity. Because X is a determinant of set $S_1 \bigcup ... \bigcup S_n$ and $T_1 \bigcup ... \bigcup T_m \subseteq S_1 \bigcup ... \bigcup S_n$, then $X \rightarrow (T_1 \bigcup ... \bigcup T_m) \in F^+$. Supposing X is not a determinant of set $T_1 \bigcup ... \bigcup T_m$ under F, thus there exists at least one attribute A in X for which the expression $(X - \{A\}) \rightarrow (T_1 \bigcup ... \bigcup T_m) \in F^+$ holds. Then, according to Proposition 2, the expression $(X - \{A\}) \rightarrow (S_1 \bigcup ... \bigcup S_n) \in F^+$ holds, fact that contradicts the hypothesis that X is a determinant of set $S_1 \bigcup ... \bigcup S_n$ under F.

Sufficiency. Let X be a determinant of set $T_1 \bigcup ... \bigcup T_m$ under F. Since $(T_1 \bigcup ... \bigcup T_m) \rightarrow (S_1 \bigcup ... \bigcup S_n) \in F^+$ and $T_1 \bigcup ... \bigcup T_m \subseteq S_1 \bigcup ... \bigcup S_n$, then X is a determinant for $S_1 \bigcup ... \bigcup S_n$ under F.

Lemma 2. *If X is a determinant under F of set $S_1 \bigcup ... \bigcup S_n$, then Z, where $Z = X \bigcap (S_1 \bigcup ... \bigcup S_j)$ and $j = \overline{1,n}$, is a determinant for $S_1 \bigcup ... \bigcup S_j$ under F.*

Proof. According to Theorem 1, the expression $X \subseteq S_1 \bigcup ... \bigcup S_n$ takes place. First it will be shown that $Z \rightarrow (S_1 \bigcup ... \bigcup S_j) \in F^+$. Lets suppose the contrary: $Z \rightarrow (S_1 \bigcup ... \bigcup S_j) \notin F^+$. Then there exists a set Z', where $Z' \subseteq X$, which is a determinant of set $S_1 \bigcup ... \bigcup S_j$ and $Z' \bigcap (\bigcup_{i=j+1}^{n} S_i) \neq \emptyset$. Considering Lemma 1, there is a path from every vertex labeled with A in $Z' \bigcap (\bigcup_{i=j+1}^{n} S_i)$ that leads to a vertex B in $\bigcup_{i=1}^{j} S_i$. A contradiction has been encountered. Therefore, $Z \rightarrow (S_1 \bigcup ... \bigcup S_j) \in F^+$.

To complete the proof of this lemma, it will be shown that Z is a determinant under F of set $S_1 \bigcup ... \bigcup S_j$. Indeed, if it is considered that Z is not a determinant of F under F, then there must exist in Z an attribute A, such that $(Z - \{A\}) \rightarrow (S_1 \bigcup ... \bigcup S_j) \in F^+$. But then $(Z - \{A\}) \rightarrow Z \in F^+$ takes place, fact that implies $(X - \{A\}) \rightarrow X \in F^+$. So, a contradiction has been encountered, that X is a determinant of set $S_1 \bigcup ... \bigcup S_n$ under X.

Theorem 3. *If set $Z = X \bigcap (T_1 \bigcup ... \bigcup T_j)$ of attributes is a determinant of set $S_1 \bigcup ... \bigcup S_n$, then $X \subseteq T_1 \bigcup ... \bigcup T_m$.*

Proof. Let S_j be the first set of attributes that doesn't coincide with T_j and assume that there is an attribute A in X, such that $A \in S_j$ and $A \notin T_j$. Lemma 2 implies that $(X \bigcap (S_1 \bigcup ... \bigcup S_j)) \rightarrow (S_1 \bigcup ... \bigcup S_j) \in F^+$. Since $A \notin T_j$, then $(X \bigcap (S_1 \bigcup ... \bigcup S_j)) \rightarrow A \in F^+$. So $(X - \{A\}) \rightarrow X \in F^+$, thus X is not a determinant of set $S_1 \bigcup ... \bigcup S_n$ under F.

Corollary 3. *If an attribute A in R is prime in scheme M, then $A \in \bigcup_{i=1}^{m} T_i$.*

Corollary 4. *If an attribute A in $O(||F||)$ is nonprime in scheme $Sch = (\bigcup_{i=1}^{n} S_i, F)$, then $A \in (\bigcup_{i=1}^{n} S_i - \bigcup_{i=1}^{m} T_i)$.*

Example 5. *Considering Corollaries 3 and 4, and Example 4, for the scheme $Sch(R, F)$, where $F = \{C \rightarrow B, AD \rightarrow B, AB \rightarrow DC, B \rightarrow E\}$*

and $R = \{A, B, C, D, E\}$, $\{A, B, C, D\}$ *is set of prime attributes and* E *is nonprime attribute.*

Theorem 3 and Lemma 2 can be paraphrased for nonredundant equivalence classes of attributes.

Lemma 3. *If* X *is a determinant under* F *of set* $T_1 \bigcup ... \bigcup T_m$*, then* Z*, where* $Z = X \bigcap (T_1 \bigcup ... \bigcup T_j)$ *and* $j = \overline{1, m}$*, is a determinant for* $T_1 \bigcup ... \bigcup T_j$ *under* F*.*

Proposition 3. *If set of attributes* X *is a determinant of set* $T_1 \bigcup ... \bigcup \bigcup T_j$*, then* $X \subseteq T_1 \bigcup ... \bigcup T_j$*, where* $j = \overline{1, m}$*.*

The soundness of this affirmation follows from theorems 1, 2 and 3.

Theorem 4. *If set of attributes* X *is a determinant of set* $T_1 \bigcup ... \bigcup T_m$*, then* $X \bigcap T_i \neq \emptyset$*, where* $i = \overline{1, m}$*.*

Proof. Let for a set T_j, where $j = \overline{1, m}$, the equality $X \bigcap T_j = \emptyset$ holds. From Corollary 2 and Proposition 1, follows that $X \bigcap T_1 \neq \emptyset$. According to Lemma 3 set Z, where $Z = X \bigcap (T_1 \bigcup ... \bigcup T_j)$ and $j = \overline{1, m}$, is a determinant for $T_1 \bigcup ... \bigcup T_j$ under F. From the fact that $X \bigcap T_j = \emptyset$ it follows that $Z \subseteq T_1 \bigcup ... \bigcup T_{j-1}$ and then $(T_1 \bigcup ... \bigcup T_{j-1}) \rightarrow T_j \in F^+$. But this contradicts the assumption that set $T_1 \bigcup ... \bigcup T_m$ is nonredundant.

5 Algorithmic aspects

From the algorithmic point of view, the problem of testing the primeness of attributes consists of two parts, construction of equivalence classes of scheme's attributes and elimination of the redundancy in these classes. In other words, being given a relation scheme $Sch(R, F)$, the sets $S_1 \bigcup ... \bigcup S_n = R$ and $T_1 \cup ... \cup T_m$ are to be build, respectively.

The method for determination of equivalence classes of attributes consists in the fact that for every attribute A in R, the list of attributes that label accessible vertices from A on the contribution graph is computed. So, accessibility matrix M will be computed, that will consist

of 0 and 1, with a dimension $|R| \times |R|$, where $|R|$ is cardinality of set R. The element $M(i,j) = 1$ if and only if there exists a path from vertex i to vertex j. Based on matrix M the set of equivalence classes of attributes R is constructed.

In the speciality literature (for example, in [5]) it is described an algorithm of finding the strongly connected components of a directed graph with a complexity $O(\max(|S|, |E|))$, where $|S|$- number of vertices, and $|E|$- number of edges. But, it is easy to observe that, using this algorithm is non suitable, because the computing of the contribution graph (for example its representation in form of adjacency lists) for a set F of functional dependencies requires $O(|R| \cdot ||F||)$ operations and the graph will have a number of edges proportionally to $|R|^2$. Where $||F||$ is the number of attributes involved in F, when duplicates are also considered. As $||F|| > |R|$, algorithm of computing the equivalence classes of attributes needs $O(|R| \cdot ||F||)$ operations.

Because the closure of a set of attributes under a set of functional dependencies is computed in a time $O(||F||)$ [2], then for equivalence classes of attributes the elimination of redundancies requires $O(|EquivClasses| \cdot ||F||)$, where $|EquivClasses|$ represents the number of equivalence classes of attributes. Since $|EquivClasses| \leq |R|$, this algorithm requires a time proportionally to $|R| \cdot ||F||$.

References

[1] Yu C.T., Johnson D.T. *On the complexity of finding the set of candidate keys for a given set of functional dependencies.* Information Processing Letters, 1978, V.5, N.4, p.100-101.

[2] Maier, D. *The theory of relational database.* Computer Science Press, 1983, 637 p.

[3] Cotelea, Vitalie. *Relational databases: logical design.* ASEM, Chisinau, 1997, 290 p. (in Romanian)

[4] Even, Shimon. *Graph Algorithms.* Computer Science press, 1979, 250 p.

[5] Aho, Alfred V. Hopcroft, John E. and Ullman, Jeffrey D. *The Design and Analysis of Computer Algorithms.* Addison-Wesley, placeCityReading, StateMA, 1974, 470 p.

Vitalie Cotelea

Received April 28, 2009

Vitalie Cotelea
Academy of Economic Studies of Moldova
Phone: (+373 22) 40 28 87
E–mail: *vitalie.cotelea@gmail.com*