электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0421100025. ISSN 1994-0408

Алгоритм синтеза частично оптимальной схемы реляционной базы данных

77-30569/294486

01, январь 2012

Григорьев Ю. А.

УДК 004.654

МГТУ им. Н.Э. Баумана
grigorev@iu5.bmstu.ru

Благодаря своей простоте и ясным концептуальным основам, реляционная модель данных получила широкую поддержку среди поставщиков коммерческих СУБД. Начиная с исторической работы Кодда [1], вокруг этой модели развернулись активные теоретические исследования. В теории проектирования реляционных баз данных одной из центральных является проблема синтеза оптимальной логической схемы базы данных на основе множества функциональных зависимостей (ФЗ) атрибутов универсального отношения [2]. Как показано в [3], задача поиска оптимальной схемы является NP-полной, т.е. относится к классу труднейших по вычислительной сложности. Соответственно, алгоритм ее решения имеет экспоненциальную временную сложность. Поэтому на практике часто проектируют неоптимальную схему, имеющую полиномиальный алгоритм синтеза, гарантирующий получение заданных свойств.

В традиционной постановке задача синтеза неоптимальной схемы сформулирована в следующем виде [2, 4]. Пусть задана схема S = (R; F), где R - множество атрибутов универсального отношения, F - множество ФЗ атрибутов из R. Необходимо получить схему T как множество подсхем (схем отношений) вида $S_i = (R_i; K_i)$, где R_i - множество атрибутов подсхемы, $K_i = \{X_{i1},\ldots, X_{iq}\}$ - множество ключей подсхемы ($X_{ij} \subseteq R_i$, $i = \overline{1,p}$, $j = \overline{1,q}$). При этом схема T должна удовлетворять следующим требованиям.

1. Декомпозиция $\{R_1,\ldots, R_p\}$ обладает свойством естественного соединения результирующих отношений без потерь информации.

2. Обеспечивается сохранение множества всех ФЗ из замыкания F^+ (т.е. из объединения всех ФЗ подсхем логически следуют все зависимости, принадлежащие F).

3. Все подсхемы S_i находятся в третьей нормальной форме (3НФ).

4. Число подсхем в схеме минимально, т.е. не существует схемы, удовлетворяющей требованиям 1-3 и содержащей менее p подсхем.

Свойства сохранения информации и множества ФЗ, отраженные в пунктах 1 и 2, имеют большое значение, так как позволяют восстановить исходную схему S по декомпозиции T. Соответствие подсхем S_i 3НФ (пункт 3) позволяет избежать значительной части аномалий включения, удаления и модификации кортежей базы данных. Требование, изложенное в пункте 4, позволяет обеспечить минимальный объем хранения базы данных.

Оптимальная схема базы данных, кроме выполнения требований 1-4, должна также иметь возможно меньшее суммарное число атрибутов в подсхемах и минимальное множество ключевых атрибутов.

Решение задачи синтеза (неоптимальной) схемы, удовлетворяющей требованиям 1-4, сформулировано Бернштейном в виде следующего алгоритма [4].

Алгоритм А.

Шаг 1: (устранение лишних атрибутов и поиск неизбыточного покрытия). Для каждого отображения f: X→A∈F и каждого B∈X, если f*: (X - {B}) → A), заменить f на f*. Затем найти неизбыточное покрытие G для F.

Шаг 2: (разбиение). Разделить G на группы так, что все ФЗ в каждой группе имеют одинаковые левые части.

Шаг 3: (объединение эквивалентных ключей). Пусть J = 0. Для каждой пары групп, скажем, G1 и G2 с левыми частями X и Y соответственно, объединить G1 и G2 вместе, если существует X ↔ Y в G+ (т.е. для X → Y∃Y→X). Для каждого A∈Y добавить f1: X→A к J и если f1 имеется в G, то удалить ее из G. Аналогично, для каждого B∈X добавить f2: Y→B к J и если f2 имеется в G, то удалить ее из G.

Шаг 4: Найти G*⊆G такое, что (G* + J)+ = (G + J)+ и никакое собственное подмножество G* не обладает этим свойством. Добавить каждую ФЗ из J в соответствующую группу G*.

Шаг 5: (конструирование отношений). Для каждой группы построить отношение, состоящее из атрибутов, появляющихся в этой группе. Каждое множество атрибутов, появляющихся в левой части ФЗ в этой группе, являются ключом отношения (каждый ключ, определенный таким образом, называется синтезированным). Множество созданных отношений образует схему T для данного множества ФЗ.

Недостатком этого алгоритма является сложность машинной реализации шага 4, что затрудняет использование данного алгоритма в системах автоматизированной поддержки проектирования баз данных. Использование при реализации шага 4 условия $(G^* + J)^+ = (G + J)^+$ вызывает трудности вычисления, так как получение покрытия любого множества ФЗ F^+ может экспоненциально зависеть от размера F.

В настоящей работе предлагается алгоритм решения задачи синтеза частично оптимальной схемы базы данных, обеспечивающий выполнение требований 1-3 для T и относительно простую возможность машинной реализации [8].

Алгоритм Б.

Шаг 1: Пусть T = Ø.
Шаг 2: Построить G - минимальное покрытие для F.
Шаг 3: Каждую зависимость X→A из G заменить на XA (запись вида XA означает объединение множества атрибутов X и атрибута A). Получившееся таким образом множество подсхем обозначить через Q.
Шаг 4: Если A1A2…Am ∈Q, то добавить в T подсхему A1A2…Am и выйти из алгоритма.
Щаг 5: Добавить в T в качестве подсхем те атрибуты, которые не входят ни в какие подсхемы из Q.
Шаг 6: Добавить в T все подсхемы из Q.
Шаг 7: Если ни одна из подсхем, входящих в T, не содержит ключ универсального отношения R, то добавить в T любой ключ в качестве подсхемы.

Покрытие G будем называть минимальным, если оно содержит минимальное число ФЗ и минимальное число атрибутов в левой и правой частях каждой ФЗ. Развернутое определение минимального покрытия и алгоритм его построения представлен в [6].

Замечание. Предложенный алгоритм Б не гарантирует выполнения пункта 4 требований к множеству T, так как не предполагает объединения ФЗ с эквивалентными левыми частями. Однако пункт 4 не является определяющим при проектировании "хорошей" схемы базы данных. Более того, соблюдение требований данного пункта в случае распределенного хранения таблиц базы данных приводит к значительным издержкам выполнения операций соединения. Этот и ряд других факторов [5] приводят к отказу от повсеместного и всеобязательного принципа исключения избыточности.

В [6] приведены теоремы 5.7 и 5.8, подтверждающие, что алгоритм Б обеспечивает выполнение требований 1-3 для T.

Пример. Рассмотрим гипотетическую базу данных учебного отдела ВУЗа, имеющую следующее универсальное отношение:

R = (A - дисциплина, B - преподаватель, C - час начала занятия, D - номер аудитории, E - студент, K - оценка)

Пусть среди атрибутов данного отношения существуют следующие ФЗ:

F = {

A→B - каждую дисциплину ведет только один преподаватель,

CD→A - в аудитории одновременно может читаться только одна дисциплина,

CB→D - преподаватель может одновременно находится только в одной аудитории,

AE→K - по каждой дисциплине каждый студент имеет только одну оценку,

AC→D - каждая дисциплина может одновременно читаться только в одной аудитории,

CE→D - студент может одновременно находится только в одной аудитории

}

Необходимо построить схему базы данных, отвечающую условиям 1-3.

Шаг 1: T = Ø

Шаг 2: G = { A→B, CD→A, CB→D, AE→K, AC→D, CE→D}

а1) A→B, G - {A→B} = {CD→A, CB→D, AE→K, AC→D, CE→D}

$A^+ = A \Rightarrow B \notin A^+ \Rightarrow A \to B \notin (G - \{A \to B\})^+$

б1) CD→A, G - {CD→A}= { A→B, CB→D, AE→K, AC→D, CE→D }

$(CD)^+ = CD \Rightarrow A \notin (CD)^+ \Rightarrow CD \to A \notin (G - \{CD \to A\})^+$

в1) CB→D, G - {CB→D} = { A→B, CD→A, AE→K, AC→D, CE→D}

$(CB)^+ = CB \Rightarrow D \notin (CB)^+ \Rightarrow CB \to D \notin (G - \{CB \to D\})^+$

г1) AE→K, G - {AE→K} = { A→B, CD→A, CB→D, AC→D, CE→D}

$(AE)^+ = AEB \Rightarrow K \notin (AE)^+ \Rightarrow AE \to K \notin (G - \{AE \to K\})^+$

д1) AC→D, G - {AC→D} = { A→B, CD→A, CB→D, AE→K, CE→D}

$(AC)^+ = ACBD \Rightarrow D \in (AC)^+ \Rightarrow AC \to D \in (G - \{AC \to D\})^+ \Rightarrow$ ФЗ AC→D может быть исключена из множества G, т.е. теперь

G = { A → B, CD→A, CB→D, AE→K, CE→D}

е1) CE→D, G - {CE→D} = { A→B, CD→A, CB→D, AE→K}

$(CE)^+ = CE \Rightarrow D \notin (CE)^+ \Rightarrow CE \to D \notin (G - \{CE \to D\})^+$

Далее рассматриваются ФЗ из G, имеющие 2 и более атрибутов в левой части, и анализируются собственные подмножества левых частей:

а2) CD→A

C→A, ниже замыкания множества атрибутов берутся для G,

$C^+ = C \Rightarrow C \to A \notin G^+$

D→A,

$D^+ = D \Rightarrow D \to A \notin G^+$

Аналогично можно показать, что

б2) для CB→D ФЗ C→D и $B \to D \notin G^+$

в2) для AE→K ФЗ A→K и $E \to K \notin G^+$

г2) для CE→D ФЗ C→D и $E \to D \notin G^+$

Таким образом G = { A→B, CD→A, CB→D, AE→K, CE→D} – это минимальное покрытие для множества исходных функциональных зависимостей F.

Шаг 3: Q = {AB, CDA, CBD, AEK, CED}

Шаг 4: $ABCDEK \notin Q$

Шаг 5: Все атрибуты принадлежат хотя бы одной подсхеме из Q

Шаг 6: T = {AB, CDA, CBD, AEK, CED}

Шаг 7:

$X_0 = ABCDEK$

$(BCDEK)^+ = BCDEKA = S \Rightarrow X_1 = BCDEK$

$(CDEK)^+ = CDE\,KAB = S \Rightarrow X_2 = CDEK$

$(DEK)^+ = DEK \neq S$

$(CEK)^+ = CEKDAB = S \Rightarrow X_3 = CEK$

$(EK)^+ = EK \neq S$

$(CK)^+ = CK \neq S$

$(CE)^+ = CEDABK = S \Rightarrow X_4 = CE$

$C^+ = C \neq S$

$E^+ = E \neq S$

Следовательно, X = CE – ключ универсального отношения. Так как $CE \subseteq CED$, то схема базы данных T = {AB, CDA, CBD, AEK, CED} обладает

1) свойством соединения без потерь,
2) свойством сохранения зависимостей,
3) каждая подсхема находится в 3НФ.

Литература

1. Codd E.F. A relational model of data for large shared data banks. Comm. ACM. 1970. V. 13. № 6.
2. Мейер Д. Теория реляционных баз данных. - М.: Мир, 1987. – 608 с.
3. Lucchesi C., Osborn S. Candidate keys for relations// J. Computer and System Sciences. 1978. V. 17. № 2.
4. Bernstein P. Synthesizing third normal form relations from functional dependencies// ACM Trans. on Database Systems. 1976. V. 1. № 4.
5. Зиндер Е. Проектирование баз данных: новые требования, новые подходы// СУБД. - 1996. - № 3.
6. Ульман Дж. Основы систем баз данных. - М.: Финансы и статистика, 1983. – 334 с.
7. Преснякова Г.В. Проектирование интегрированных реляционных баз данных. – М.: КДУ: СПб.: Петроглиф, 2007. – 224 с.
8. Григорьев Ю.А., Плутенко А.Д. Теория и практика проектирования систем на основе баз данных: Учебное пособие. – Благовещенск: Амурский гос. ун-т, 2007. – 396 с.

electronic scientific and technical periodical
SCIENCE and EDUCATION
EL № FS 77 - 30569. №0421100025. ISSN 1994-0408

Algorithm of synthesis of suboptimal scheme of relational database

77-30569/294486

 # 01, January 2012

Grigor'ev Yu.A.

Bauman Moscow State Technical University
grigorev@iu5.bmstu.ru

This article deals with algorithm of synthesis of database scheme based on the given set of functional dependences. Bernshtein's algorithm was described; it was shown that usage of (G* + J)+ = (G + J)+ condition caused calculation problems, because covering of FD F+ set could exponentially depend on the size of F. Algorithm, based on Ullman's ideas, was proposed; it provided relatively simple implementation. Generated database scheme had the property of lossless connection, maintenance of functional dependences and each sub-scheme of this database was in the third normal form. Example of database scheme synthesis using the proposed algorithm was included in this article

Publications with keywords: functional dependence, normal form, scheme of database, sub-scheme of database, relation scheme, attribute, lossless connection, locking of functional dependences, maintenance of functional dependences
Publications with words: functional dependence, normal form, scheme of database, sub-scheme of database, relation scheme, attribute, lossless connection, locking of functional dependences, maintenance of functional dependences

Reference

1. Codd E.F., A relational model of data for large shared data banks, Comm. ACM 13 (6) (1970) 377-387.
2. Meier D., The theory of relational databases, Moscow, Mir, 1987, 608 p.
3. Lucchesi C., Osborn S., Candidate keys for relations, J. Computer and System Sciences 17 (2) (1978) 270-279.
4. Bernstein P., Synthesizing third normal form relations from functional dependencies, ACM Trans. on Database Systems 1 (4) (1976) 277-298.
5. Zinder E., Database Design: new challenges, new approaches, SUBD 3 (1996).
6. Ul'man Dzh., Fundamentals of database systems, Moscow, Finansy i statistika, 1983, 334 p.

7. Presniakova G.V., Design of integrated relational databases, Moscow, KDU, SPb., Petroglif, 2007, 224 p.

8. Grigor'ev Iu.A., Plutenko A.D., Theory and practice of designing systems on the basis of databases, Blagoveshchensk, Amurskii gos. un-t, 2007, 396 p.