

ЗАВИСИМОСТИ В СОСТАВНЫХ БАЗАХ ДАННЫХ

База данных (БД) — это довольно сложная структура, состоящая из данных (элементов) и связей между ними. Структуру БД определяет тип ее элементов, так называемых записей, и способ образования допустимых связей между ними. Модели данных в БД обычно классифицируют по способу образования допустимых связей. Будем различать две модели данных:

— сетевую модель, в которой между типами записей можно задать связи с помощью произвольных ссылок;

— реляционную модель, в которой между типами записей нельзя осуществить никакой связи; в реляционных БД единственным средством, осуществляющим связь на множестве свойств БД, является тип записи.

Задание типа записи есть не что иное, как определение области определения и области значения. Задаваемый таким образом тип записи представляет собой функцию из области определения в область значения. Элементы области значения называют атрибутами.

Представляет интерес определение разных типов зависимостей между типами записей, а также тех аксиом, с помощью которых можно задать эти зависимости в реляционных БД. Этот вопрос рассматривается в [1].

В [2] исследовалась задача определения минимального числа строк в БД с фиксированной системой ключей. В данной работе частично рассматривается эта задача, а также исследуется задача определения минимального числа строк в случае задания в БД системы зависимостей.

Часто БД естественным образом строятся из многих БД. Иногда ее искусственным образом пытаются разделить на отдельные БД. В обоих случаях удобно, если некоторые параметры сложной БД можно вычислить или оценить по соответствующим параметрам частичных БД. Попытаемся это сделать для минимального числа строк, что осуществимо только тогда, когда большая БД построена из независимых баз.

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ И ПОНЯТИЯ

Одним из важных понятий БД является функциональная зависимость, определенная Э. Ф. Коддом [3]. Из сказанного выше следует, что в некоторый момент времени содержание реляционной БД можем рассматривать как матрицу, поэтому понятие функциональной зависимости определим для матриц.

Введем основные определения (те, которые не приводятся в этой статье, можно найти в [1]). БД рассмотрим как матрицу, состоящую из n столбцов и m строк, значениями которой являются неотрицательные целые числа, а все строки разные. Строки соответствуют объектам, столбцы — атрибутам. Множество столбцов представляет собой ключ, если не найдутся две такие разные строки, элементы которых совпадают в этих столбцах. Иначе говоря, значения, стоящие в столбцах ключа, однозначно определяют строку. Это можно сформулировать и более общим образом. Если A и B — множество столбцов и две строки, которые совпадают в A , совпадают также и в B , то говорим, что из A следует B (в обозначениях $A \rightarrow B$). Такие пары (A, B) называем зависимостью. Ясно, что если A — ключ и X — множество всех столбцов, то $A \rightarrow X$. Ключ A назовем минимальным ключом, если нет такого ключа B , для которого выполняется $B \subset A$, $B \neq A$. Систему минимальных ключей обозначим через \mathcal{K} . Известно [4], что \mathcal{K} есть система ключей некоторой БД тогда и только тогда, когда \mathcal{K} является непустой системой Спернера, т. е. в ней нет двух разных A, B , для которых выполнялось бы $A \subset B$.

Пусть матрице M некоторой БД поставлена в соответствие функция, отображающая подмножество X в подмножество Y :

$$\mathcal{L}_M(A) = \{b : A \rightarrow \{b\}\}. \quad (1)$$

Легко убедиться, что, с одной стороны, \mathcal{L}_M однозначно определяет полную систему зависимостей ($A \rightarrow B \Leftrightarrow B \subseteq \mathcal{L}_M(A)$), а с другой — обладает следующими свойствами:

$$A \subseteq \mathcal{L}(A), \quad (2)$$

$$A \subseteq B \Rightarrow \mathcal{L}(A) \subseteq \mathcal{L}(B), \quad (3)$$

$$\mathcal{L}(\mathcal{L}(A)) = \mathcal{L}(A). \quad (3a)$$

Функции \mathcal{L} , обладающие такими свойствами, называем замыканиями. Известно, что для любого замыкания \mathcal{L} существует такая матрица M БД, в которой функция \mathcal{L}_M , порожденная зависимостями, совпадает с \mathcal{L} .

Пусть \mathcal{K} является системой Спернера подмножеств n -элементного множества. Как было отмечено, существует такая БД (точнее, ее матрица), в которой \mathcal{K} представляет собой систему минимальных ключей. Пусть среди этих БД имеется БД, состоящая из $s(\mathcal{K})$ минимальных строк. В [2] можно найти результаты, относящиеся к $s(\mathcal{K})$ в случае некоторых простых систем \mathcal{K} . Аналогично

они принадлежат одному классу, это свойство выполнится для них и в N' . Следовательно, $a \notin \mathcal{L}_{N'}(A)$.

Пусть $\gamma = (\gamma_1, \dots, \gamma_u)$ и $\delta = (\delta_1, \dots, \delta_u)$ — строки из первого и второго класса N' . Исключим из матрицы элемент γ и вместо каждого γ_i , присутствующего в i -м столбце первого класса, напишем δ_i . Для матрицы N'' , получающейся таким образом, докажем утверждение

$$\mathcal{L}_{N''} = \mathcal{L}_{N'} \quad (8)$$

Предположим, что $a \in \mathcal{L}_{N'}(A)$, и пусть две строки N'' в A совпадают. Если эти строки находятся в одном классе N' , то их тождество следует и в a , если нет, — они из прежнего первого и второго класса, μ'' и γ'' . Здесь i -й элемент этих двух строк совпадает, если они оба равны δ_i . Но тогда в соответствующей строке из N' в μ' стоит γ_i на i -м месте. Иначе говоря, часть μ' , попадающая в A , совпадает с γ . Из $a \in \mathcal{L}_{N'}(A)$ следует, что a -й элемент этой строки в N' есть γ_a , а в N'' — δ_a . Аналогично можно убедиться, что a -й элемент строки ν' , соответствующей ν'' , есть δ_a , т. е. μ'' и ν'' совпадают на a -м месте. Следовательно, $a \in \mathcal{L}_{N''}(A)$.

Наоборот, предположим, что $a \notin \mathcal{L}_{N'}(A)$, т. е. у N' существуют две строки, которые в A совпадают, а в a — нет. Естественно, они попадут в одинаковые классы. Эти две строки без изменения совпадают и в N'' . Тем самым $a \in \mathcal{L}_{N''}(A)$. Из этого, с учетом (8), следует, что для N' предположения леммы уже выполняются при k , значит, можно применить индуктивное предположение

$$\begin{aligned} (\text{число строк } N) - 1 &= \text{число строк } N'' \geq \\ &\geq s(\mathcal{L}_{N''}) + k - 1. \end{aligned}$$

В правой части неравенства, в силу (7) и (8), можно записать $s(\mathcal{L}_N) + k - 1$, т. е.
число строк $N \geq s(\mathcal{L}_N) + k$,

что и требовалось доказать.

Вернемся к доказательству теоремы. Пусть для какого-то $M \mathcal{L}_M = \mathcal{L}_1 \times \mathcal{L}_2$ и пусть M_1 и M_2 — подматрицы, соответствующие подмножеству \mathcal{L}_1 и \mathcal{L}_2 . Согласно лемме 1 $\mathcal{L}_{M_1} = \mathcal{L}_1$. Если $a \in \mathcal{L}_1(A)$ ($A \subset X_1$), то $a \notin \mathcal{L}_1(A) \cup X_2 = (\mathcal{L}_1 \times \mathcal{L}_2)$ ($A \cup X_2$), т. е. у M существуют две такие строки, которые в $A \cup X_2$ совпадают, а в a — нет. Иначе говоря, в M_1 можно найти две такие строки, что 1) часть из них, принадлежащих M_2 , совпадает, 2) они совпадают в A , но не совпадают в a . Тем самым строки M_1 (их число совпадает с числом разных строк M_2) разделили на такие классы, которые удовлетворяют условиям леммы 2. Таким образом, на основе (6)

$$\begin{aligned} \text{число строк } M_1 &\geq s(\mathcal{L}_1) \\ + (\text{число разных строк } M_2) &- 1. \end{aligned}$$

Согласно лемме 1 число разных строк $M \geq s(\mathcal{L}_2)$. Теорема доказана.

Если \mathcal{K}_1 и \mathcal{K}_2 — системы минимальных ключей на непересекающихся множествах X_1 и X_2 , тогда $\mathcal{K}_1 \times \mathcal{K}_2$ состоит из таких множеств $X_1 \cup X_2$, которые представляются в виде $A \cup B$, $A \in \mathcal{K}_1$, $B \in \mathcal{K}_2$. Легко можно убедиться, что $\mathcal{K}_1 \times \mathcal{K}_2$ — система Спернера, т. е. существует такое замыкание, в котором они являются минимальными ключами.

Теорема 2. Справедливо

$$s(\mathcal{K}_1 \times \mathcal{K}_2) \leq s(\mathcal{K}_1) + s(\mathcal{K}_2) - 1. \quad (9)$$

Доказательство. Неравенство доказывается аналогично первой части доказательства теоремы 1. Практика показывает, что в (9) почти всегда выполняется неравенство, подобно теореме 1, однако следующий пример доказывает, что это не всегда верно.

Пусть $X_1 = \{1, 2, 3, 4, 5\}$, $X_2 = \{6, 7, 8, 9, 10\}$, $\mathcal{K}_1 = \{\{1, 2\}, \{3, 4\}, \{1, 5\}, \{2, 5\}, \{3, 5\}, \{4, 5\}\}$ и \mathcal{K}_2 изоморфно этим совокупностям на X . Максимальная система, не являющаяся ключами на X_1 , — $\mathcal{K}^{-1} = \{\{5\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}\}$.

Вначале покажем, что $s(\mathcal{K}_1) \geq 5$. Предположим от противного, что существует такая четырехстрочная матрица M_1 , система ключей которой есть \mathcal{K}_1 . Будем использовать следующую простую лемму.

Лемма 3.¹ Если система ключей $M - \mathcal{K}$, то для любого элемента A из \mathcal{K}^{-1} существует пара строк в M , которые совпадают именно в A , и разным элементам \mathcal{K}^{-1} соответствуют разные пары строк.

Доказательство. Так как каждому A , не являющемуся ключом, соответствует пара строк, совпадающих в A , но не полностью, то и для $A \in \mathcal{K}^{-1}$ существует такая пара строк. Если эти две строки совпадали бы не только на A , то у A существовало бы такое действительное расширение, которое находилось бы в \mathcal{K}^{-1} , что противоречит $A \in \mathcal{K}^{-1}$.

Из этого сразу следует, что $s(\mathcal{K}_1) \geq 4$, так как $\binom{3}{2} = 3 < 5 = |\mathcal{K}_1^{-1}|$. Но если число строк M — четыре, то и тогда множество совпадений любой пары строк, кроме одной, будет каким-то элементом из \mathcal{K}_1^{-1} . Используем следующую лемму.

Лемма 4. Для трех множеств совпадения A_1, A_2, A_3 трех строк некоторой матрицы выполняется

$$(A_i \cap A_j) - A_k = \emptyset, \quad (10)$$

где i, j, k — некоторая перестановка чисел 1, 2, 3. Доказательство очевидно.

Из шести пар, образованных из четырех строк M , пяти соответствует множество из \mathcal{K}_1^{-1} (исключением является пара строк (3, 4)). Эти пять пар определяют две тройки строк, тогда выполняется (10). Ввиду того, что в \mathcal{K}_1^{-1} нет трех множеств, пересечение которых не пусто, три множества

¹ Утверждение леммы неявно указано в [1].

должны быть попарно непересекающимися. Таким образом, три множества (без учета симметрии) можно выбрать только одним способом: $\{1, 3\}$, $\{2, 4\}$, $\{5\}$. Но тогда выбор второй тройки однозначен: $\{1, 4\}$, $\{2, 3\}$, $\{5\}$, т. е. у двух троек не может быть более одного общего множества. Значит, в силу симметрии можно предположить, что общее множество строк 1· и 2· есть $\{5\}$, 1· и 3· — $\{1, 3\}$, 2· и 3· — $\{2, 4\}$, наконец, 1· и 4· — $\{2, 3\}$. Но тогда строки 3· и 4· совпадают в $\{1, 2\}$, а в $\{3, 4\}$ — нет, т. е. у \mathcal{K}_1^{-1} должен существовать элемент, содержащий $\{1, 2\}$. Это противоречие доказывает, что $s(\mathcal{K}_1) \geq 5$.

Если бы в (9) выполнялось равенство, то было бы $s(\mathcal{K}_1 \times \mathcal{K}_2) = s(\mathcal{K}_1^2) = 9$. С помощью одной конструкции докажем, что $s(\mathcal{K}_1^2) \leq 8$.

Используем следующие обозначения:

$$\begin{aligned} \alpha &= 0, 0, 0, 0, 0, \\ \beta &= 1, 1, 1, 1, 0, \\ \gamma &= 0, 1, 0, 1, 1, \\ \delta &= 1, 0, 0, 1, 1. \end{aligned}$$

Матрица, которая доказывает $s(\mathcal{K}_1^2) \leq 8$, имеет вид

$$M = \begin{pmatrix} \alpha & \alpha \\ \alpha & \beta \\ \alpha & \gamma \\ \beta & \alpha \\ \beta & \beta \\ \beta & \delta \\ \gamma & \alpha \\ \delta & \beta \end{pmatrix}.$$

Покажем, что множество минимальных ключей $M - \mathcal{K}_1^2$. Для этого вначале надо доказать, что множества из \mathcal{K}_1^2 являются ключами в M , иначе говоря, у M нет двух таких строк, которые различаются, но в одном \mathcal{K}_1^2 или в каком-то его расширении совпадают. А именно: две строки либо в правой, либо в левой части совпадают или на множестве из \mathcal{K}_1 , или на его расширении. Легко можно проверить, что среди $\alpha, \beta, \gamma, \delta$ только одна пара, γ и δ , совпадает в \mathcal{K}_1 или в его расширении. Сравнение двух частей показывает, что, действительно, нет двух таких целых строк, которые в обеих частях из \mathcal{K}_1 или в его расширении совпадают.

В то же время надо доказать, что элементы не являются ключами. Легко можно убедиться, что $(\mathcal{K}_1^2)^{-1}$ состоит из таких множеств, которые на одной части состоят из множеств из \mathcal{K}_1^{-1} , а на другой — из полного основного множества. Для таких множеств надо найти пару строк из M , которые в нем совпадают, но не тождественны. В силу симметрии обеих частей можно предположить, что искомые две строки в первой половине совпадают. Легко убедиться, что тогда во второй половине для каждого элемента из \mathcal{K}_1^{-1} можно найти

две такие строки, которые совпадают в нем, а первая половина, относящаяся к нему, будет одинаковой.

ПОЛНЫЕ СИСТЕМЫ СПЕРНЕРА И ЗАМКЯНИЯ

Систему Спернера назовем полной, если ее невозможно расширить новым множеством так, чтобы свойство Спернера сохранилось. Если \mathcal{L} — некоторое замыкание на основном множестве X , систему

$\{K : \mathcal{L}(K) = X, \nexists K' \subset K, \mathcal{L}(K') = X\}$, являющуюся системой минимальных ключей \mathcal{L} , обозначим через $\mathcal{K}(\mathcal{L})$.

Теорема 3. Если \mathcal{K} — полная система Спернера, то равенство $\mathcal{K}(\mathcal{L}) = \mathcal{K}$ однозначно определяет замыкание \mathcal{L} .

Доказательство. Покажем, что если \mathcal{K} полное и $\mathcal{K}(\mathcal{L}) = \mathcal{K}$, то в любом случае $A \mathcal{L}(A)$ определено однозначно. В силу полноты \mathcal{K} для каждого множества A имеется $K \in \mathcal{K}$, для которого выполняется: 1) $A \supseteq K$ или 2) $A \subset K, A \neq K$. В случае 1) $\mathcal{L}(A) = X$, значит, замыкание действительно однозначно. В случае 2) докажем, что $\mathcal{L}(A) = A$. Предположим от противного, что $a \in \mathcal{L}(A) \setminus A$. В силу полноты имеем множество $K_1 \in \mathcal{K}$, для которого выполнено либо 2а) $A \cup \{a\} \subseteq K_1$, либо 2б) $A \cup \{a\} \supset K_1$. В случае 2а), согласно (2), $\mathcal{L}(K_1 - \{a\}) \supseteq K_1 - \{a\}$, а на основе (3) (так как $K_1 - \{a\} \supseteq A$) можно утверждать, что $\mathcal{L}(K_1 - \{a\}) \supseteq \mathcal{L}(A) \supseteq A \cup \{a\}$. Суммируя, получаем $\mathcal{L}(K_1 - \{a\}) \supseteq K_1$, из чего с учетом (2а) следовало бы $\mathcal{L}(K_1 - \{a\}) = \mathcal{L}(K_1 - \{a\}) \supseteq \mathcal{L}(K_1) = X$, что противоречит минимальности \mathcal{K}_1 . В случае 2б) $\mathcal{L}(A) \supseteq A \cup \{a\} \supseteq K_1$, или $\mathcal{L}(A) \subset \mathcal{L}(\mathcal{L}(A)) \supseteq \mathcal{L}(K_1) = X$, т. е. опять получено противоречие. Таким образом, полная система Спернера однозначно определяет \mathcal{L} .

Докажем обратное утверждение теоремы. Если \mathcal{K} не полное, то существует множество различных \mathcal{L} , которые удовлетворяют условию $\mathcal{K}(\mathcal{L}) = \mathcal{K}$. Построение этих \mathcal{L} кажется трудным. Одну конструкцию построить довольно просто. Пусть $\mathcal{L}(A) = X$, если $A \supset K$ для какого-то K , иначе $\mathcal{L}(A) = A$. Легко доказать, что это — замыкание, т. е. оно удовлетворяет условию (2), (3) и (3а). Однако общей конструкции еще найти не удалось.

Теорема 4. Пусть $l(n)$ означает число всех замыканий n -элементного множества. Существует такое замыкание \mathcal{L} , для которого выполняется

$$s(\mathcal{L}) = \frac{\log l(n)}{n \log \log l(n)}.$$

Доказательство теоремы эквивалентно доказательству из [5].

Окончание см. на стр. 115.

СПИСОК ЛИТЕРАТУРЫ

1. Деметровиц Я., Дьепешн Д. Аксиоматизирование обобщенных функциональных зависимостей в реляционных базах данных.— Кибернетика, 1982, № 2, с. 42—48.
2. Demetrovics J., Katona G. O. H. Extremal combinatorial problems in relational data base: Fundamentals of computation th. — Lect. Notes Comput. Sci., 1981, 117, p. 110—119.
3. Codd E. F. A relational model of data for large shared data banks.— Commun ACM, 1970, 13, p. 377—387.
4. Demetrovics J. On the equivalence of candidate keys with Sperner systems.— Acta cybernet., 1979, N 4, p. 247—252.
5. Demetrovics J., Gyepesi Gy. On the functional dependency and some generalization of it.— Ibid., 1981, N 5, p. 296—305.

Поступила 20.01.83