MTA Számitástechnikai és Automatizálási Kutató Intézete, **Közlemények 24/1980.**

ГЕНЕРИРОВАНИЕ ФУНКЦИОНАЛЬНЫХ ЗАВИСИМОСТЕЙ И ИХ ПРЕДСТАВЛЕНИЕ С ПОМОЩЬЮ РЕЛЯЦИИ

Я. Деметрович, Дь. Дьепеши

Реляционная модель данных введенная Е.Ф. Коддом [6, 7], осуществляет хранение данных в форме матриц. Ряды этих матриц являются записями данных, столбцы - так называемыми атрибутами, т.е. теми свойствами, которые характеризуют отдельные данные.

Дадим теперь строгое определение.

Пусть $\Omega = \{a_1, \ldots, a_n\}$ непустое конечное множество. Конечное множество R называется реляцией над Ω, если элементы R являются функциями, определенными на Ω. Функции, составляющие множество R мы называем строками R.

Реляцию $R = \{f_1, \ldots, f_k\}$ над Ω можно представить в виде двухмерной таблицы:

	a_1	...	a_n
f_1	$f_1(a_1)$	...	$f_1(a_n)$
$\vdots$	$\vdots$	...	$\vdots$
f_k	$f_k(a_1)$	...	$f_k(a_n)$

В тесной связи с информационным обеспечением баз данных находится понятие функциональной зависимости [4,7]. Пусть R реляция над Ω и A, B - подмножества Ω. Мы говорим, что B функционально зависит от A в R $(A \underset{R}{\overset{f}{\rightarrow}} B)$, если для любых двух строк f и g реляции R :

$$(\forall a\in A)(f(a)=g(a)) \rightarrow (\forall b\in B)(f(b)=g(b)),$$

/т.е. если для каждого ряда R величины из B однозначно определяются величинами из A./

Обозначим через $\mathbf{F_R}$ множество функциональных зависимостей реляции R, т.е.

$$\mathbf{F_R} = \{(A,B) : A \underset{R}{\overset{f}{\rightarrow}} B\}.$$

Если R реляция над Ω, то $\mathbf{F_R} \subseteq P(\Omega) \times P(\Omega)$. Назовем <u>полными семействами</u> такие подмножества $\mathbf{Y}$ множества $P(\Omega) \times P(\Omega)$, которые являются функциональными зависимостями некоторой реляции над Ω. Точнее, некоторое подмножество $\mathbf{Y} \subseteq P(\Omega) \times P(\Omega)$ является полным семейством, если существует реляция R над Ω, для которой $\mathbf{Y} = \mathbf{F_R}$.

В ходе исследования функциональных зависимостей первой задачей было абстрактное описание полных семейств. Впервые это удалось В.В. Армстронгу [1]. Он доказал, что некоторое $\mathbf{Y} \subseteq P(\Omega) \times P(\Omega)$ тогда и только тогда является полным семейством, когда для него выполняются следующие аксиомы: $(\forall A,B,C,D \subseteq \Omega)$:

(F1) $(A,A)\in\mathbf{Y}$;

(F2) $(A,B)\in\mathbf{Y},\ (B,C)\in\mathbf{Y} \Rightarrow (A,C)\in\mathbf{Y}$;

(F3) $(A,B)\in\mathbf{Y},\ C \supseteq A,\ D \subseteq B \Rightarrow (C,D)\in\mathbf{Y}$;

(F4) $(A,B)\in\mathbf{Y},\ (C,D)\in\mathbf{Y} \Rightarrow (A \cup C,\ B \cup D)\in\mathbf{Y}$.

Совокупность аксиом (F1), (F2), (F3), (F4) мы называем f-аксиомами.

В этой статье мы занимаемся двумя проблемами теории полных семейств. Эти проблемы связаны со следующим вопросом: какое количество "информации" достаточно для описания некоторого полного семейства?

Дадим теперь строгую формулировку упомянутых проблем.

Проблема 1. Пусть $|\Omega| = n$. Каково наибольшее натуральное число $S(n)$, для которого справедливо следующее утверждение: существует такое полное семейство $F \subseteq P(\Omega) \times P(\Omega)$, что если R - реляция над Ω и $F_R = F$, то R содержит по крайней мере $S(n)$ строк?

Частный случай этой проблемы касающийся системы обозначенных ключей полного семейства, был рассмотрен в [9]:

Проблема 1'. Пусть $|\Omega| = n$. Каково наибольшее натуральное число $s(n)$, для которого справедливо следующее: существует такая система Шпернера $K \subseteq P(\Omega)$, что если R - реляция над Ω и $K = K(F_R)$, то R содержит по крайней мере $s(n)$ строк?

Проблема 2. Пусть $F \subseteq P(\Omega) \times P(\Omega)$ - полное семейство. Требуется охарактеризовать системы минимальной мощности, генерирующие F.

§1. Проблема 1. В этом параграфе будут даны оценки для $S(n)$ и $s(n)$. Проблема оценки $s(n)$ была рассмотрена в [9] где было доказано, что

$$\sqrt{2\binom{n}{[n/2]}} \leq s(n) \leq 2\binom{n}{[n/2]}.$$

В теореме 1.3 будет дано усиление этой оценки.

Дадим несколько определений.

Определение 1. Пусть $F \subseteq P(\Omega) \times P(\Omega)$ - полное семейство и пусть $(A, B) \in F$. В этом случае (A, B) называется максимальной зависимостью для F, если

$$B' \supseteq B,\ (A,B') \in F \rightarrow B' = B.$$

Через $M(F)$ будем обозначать множество максимальных зависимостей семейства F.

Множество $B \subseteq \Omega$ называется максимальной правой частью для F, если существует такое $A \subseteq \Omega$, что $(A, B) \in M(F)$. Через $I(F)$ будем обозначать множество максимальных правых частей семейства F.

Лемма 1.1. Пусть $F \subseteq P(\Omega) \times P(\Omega)$. Тогда

$$I(F) = \{B \subseteq \Omega : (\forall (A,C) \in F)(A \subseteq B \to C \subseteq B)\}.$$

Доказательство нетрудное, представляем его читателю.

Следствие 1.2. Пусть $F \subseteq P(\Omega) \times P(\Omega)$. Тогда $I(F)$ замкнуто относительно пересечения /т.е. $B, B' \in I(F) \to B \cap B' \in I(F)$/.

Доказательство. Согласно лемме 1.1,

$$I(F) = \{B \subseteq \Omega : (\forall (A,C) \in F)(A \subseteq B \to C \subseteq B)\}.$$

Пусть $B, B' \in I(F)$ и $(A, C) \in F$, тогда: если $A \subseteq B \cap B'$, то $A \subseteq B$ и $A \subseteq B'$, откуда $C \subseteq B$ и $C \subseteq B'$, т.е. $B \cap B' \in I(F)$.

В дальнейшем понадобится также

Определение 1.2. Пусть $F \subseteq P(\Omega) \times P(\Omega)$ и пусть $A \subseteq \Omega$; множество A называется ключом F, если $(A,\Omega) \in F$.

Множество A называется обозначенным ключом F, если A является ключом F и $A' \subseteq A, (A', \Omega) \in F \to A = A'$. Обозначим через $K(F)$ множество обозначенных ключей семейства F.

Легко заметить, что если $K_i, K_j \ (K_i \neq K_j) \in K(F)$, то $K_i \not\subset K_j$. Систему обладающей этим свойством назовем системой Шпернера.

Если F -полное семейство, то $K(F)$ является системой Шпернера и значит $|K(F)| \leq \binom{n}{[n/2]}$.

Теорема 1.3.

$$\frac{1}{n^2}\binom{n}{[n/2]} < s(n) \leq \binom{n}{[n/2]} + 1.$$

Доказательство: Сначала покажем справедливость верхней оценки

Пусть $|\Omega| = n$ и $K \subseteq P(\Omega)$ - система Шпернера. Обозначим через L систему таких подмножеств Ω, которые не содержат ни одного элемента K и максимальны в этом смысле. Тогда L является системой Шпернера и, следовательно, $|L|=k \leq \binom{n}{[n/2]}$. Пусть $L = \{X_2,\ldots,X_{k+1}\}$. Построим теперь реляцию R, содержащую $k+1$ строку : $h_1,\ldots, h_{k+1}$. Пусть $h_1(a) = 0$ для всех $a\in\Omega$ и пусть для $1 < i \leq k+1$

$$h_i(a) = \begin{cases} 0 \text{ , если } a\in X_i ; \\ i \text{ , если } a\notin X_i . \end{cases}$$

Если $A\in K$, то $A \nsubseteq X_i$ для любого i , и значит любые две строки реляции R отличаются друг от друга на множестве A, т.е. <u>A является ключом для F_R</u> .

Если $B\subset A$, то $B \notin K$, поскольку K - система Шпернера и B не содержит ни одного элемента K. Тогда существует такое i , что $B \subseteq X_i$. Отсюда следует, что h_o и h_i совпадают на множестве B, так как они совпадают на X_i . С другой стороны $X_i \neq \Omega$, так как K $P(\Omega)$. По определению h_o и h_i существует такое $a\in\Omega$, что $h_o(a) \neq h_i(a)$. Следовательно B не является ключом для F_R . Таким образом A является обозначенным ключом для F_R .

Поскольку A было произвольным элементом множества K , то таким образом мы доказали, что $K \subseteq K(F_R)$.

Пусть теперь $B \in P(\Omega)\setminus K$, и предположим, что $B\in K(F_R)$. Поскольку $K\subseteq K(F_R)$, то B не содержит ни одного элемента K, и значит существует такое i , что $B\subseteq X_i$. Выше однако мы видели, что в этом случае B не является ключом для F_R , а следовательно, не может быть и обозначенным ключом F_R , что является противоречием.

Таким образом $K = K(F_R)$.

Приведем теперь доказательство справедливости нужной оценки, данное Л. Роняи.

Сначала сделаем два тривиальных замечания.

1. Если R-реляция, содержащая s строк, то существует реляция R', которая использует не более s символов и для которой $F_R = F_{R'}$ и $K(F_R) = K(F_{R'})$. /Нетрудно догадаться, что для F_R не имеет значения, совпадают ли или различаются символы, встречающиеся в различных столбцах реляции R. Поскольку в реляции R количество строк $\leq s$, то и число символов, встречающихся в одном столбце не превышает s./

2. Если R содержит s строк и $s' > s$, то существует такая реляция R', содержащая s' строк, что $F_R = F_{R'}$ /и конечно $K(F_R) = K(F_{R'})$ /.

Из утверждений 1 и 2 следует, что число систем Шпернера, каждая из которых является системой обозначенных ключей некоторой реляции, содержащей максимум $s(n)$ строк, - не превышает $(s(n))^{s(n)\cdot n}$. Таким образом $(s(n))^{s(n)\cdot n} > 2^{\binom{n}{[n/2]}}$, откуда путем несложных вычислений получаем, что

$$s(n) > \frac{1}{n} \binom{n}{[n/2]} \qquad \square$$

Следствие 1.4.

$$\frac{1}{n^2} \binom{n}{[n/2]} \leq s(n).$$

Теорема 1.5.

$$s(n) \leq \frac{3}{2} \binom{n}{[n/2]} + 1.$$

Доказательство. В.В. Армстронг [1] доказал, что если множество $I \subseteq \Omega$ замкнуто относительно пересечения, то существует только одно полное семейство F, для кторого $I = I(F)$.

Таким образом достаточно доказать, что если множество $I \subseteq \Omega$ замкнуто относительно пересечения, то существует реляция R, содержащая не более $\frac{3}{2}\binom{n}{[n/2]} + 1$ строк, для которой $I = I(F_R)$.

Пусть $I \subseteq \Omega$ замкнуто относительно пересечения. Пусть $N = \{Y \in I : Y \neq \cap\{Y' \in I : Y' \supset Y,\ Y' \neq Y\}\}$. Тогда N таково, что если $Y \neq Y' \in N$, то $Y \cap Y' \notin N$. Согласно теореме Д. Клейтмана [10], в этом случае $|N| = k \leq \frac{3}{2}\binom{n}{[n/2]}$. В лемме 2.1. будет доказано, что наименьшей системой, замкнутой относительно пересечения и содержащей N, является I. Значит достаточно построить такую реляцию R, которая содержит не более $\frac{3}{2}\binom{n}{[n/2]} + 1$ строк и для которой $N \subseteq I(F_R) \subseteq I$.

Пусть $N = \{Y_2, \ldots, Y_{k+1}\}$ и пусть R - та реляция, которую мы построили при доказательстве теоремы 1.3.; для $i = 2, \ldots, k+1$ X_i следует заменить на Y_i. Очевидно, что в этом случае

$N \subseteq I \subseteq I(F_R)$. □

§ 2. Проблема 2.

Перейдем теперь к Проблеме 2, которую впервые исследовал В.В. Армстронг [2]. Упомянем также, что с этой проблемой тесно связано понятие выводимости / см. [2], [11] /, а также исследование числа функциональных зависимостей, которые могут быть выведены из данной системы ключей [9].

Определение 2.1. Пусть $F \subseteq P(\Omega) \times P(\Omega)$ полное семейство. Мы говорим, что некоторое $F' \subseteq F$ генерирует F /или, что F' является генератором F /, если для реляции R над Ω, для которой $F' \subseteq F_R$, справедливо также, что $F \subseteq F_R$.

В [2] дана логическая характеристика систем минимальной мощности, генерирующих полные семейства. В теореме 2.2. мы дадим комбинаторный эквивалент этой характеристики, а также приведем оценку "минимальной мощности".

Определение 2.2. Пусть $F \subseteq P(\Omega) \times P(\Omega)$. Обозначим через $N(F)$ семейство элементов $I(F)$ замкнутое относительно пересечения, то есть $N(F)=\{Y \in I(F): Y \neq \cap\{Y' \in I(F): Y' \supset Y, \ Y' \neq Y\}\}$
Мы говорим, что $M \subseteq I(F)$ генерирует $I(F)$, если $I(F) = \{\cap M' : M' \subseteq M\}$.

Лемма 2.1. Подмножество $M \subseteq I(F)$ тогда и только тогда генерирует $I(F)$, когда $N(F) \subseteq M$.

Доказательство. Приводимое ниже доказательство хорошо известно в теории сетей.

Если M генерирует $I(F)$, то очевидно, что $N(F) \subseteq M$. Остается доказать, что $N(F)$ генерирует $I(F)$. Предположим, что это неверно. Пусть $X \in I(F)$ - множество наименьшей мощности, для которого $X \neq \cap\{Y \in N(F): Y \supseteq X\}$. Тогда разумеется, $X \notin N(F)$, то есть

(1) $X = \cap\{X' \in I(F): X' \supset X, X' \neq X\}$.

Если $X' \supset X$, $X' \neq X$, то $|X'| > |X|$, и значит

(2) $X' = \cap\{Y \in N(F): Y \supseteq X'\}$.

На основе (1) и (2) имеем:

$$X = \cap\{Y \in N(F): Y \supseteq X\} ,$$

что является противоречием. □

Теорема 2.2. Пусть F - полное семейство и $F' \subseteq F$. Утверждается, что F' тогда и только тогда является генератором F с минимальной мощностью, когда

(*) $(\forall Y \in N(F))(\exists A_Y \subseteq \Omega)(F' = \{(A_Y, Y) : Y \in N(F)\})$

Доказательство. Если F' представимо в виде (*) , то очевидно, что оно генерирует F .

Предположим теперь, что F' генерирует F . Для $(A,B) \in F'$ обозначим через B' максимальное множество, для которого $B' \supseteq B$ и $(A,B') \in F$. Тогда семейство $F'' = \{(A,B') : (A,B) \in F'\}$ также генерирует семейство F , причем $|F''| \leqq |F'|$. Очевидно, что множество $\{B' : \exists (A,B') \in F''\}$ генерирует $I(F)$, и значит по лемме 2.1

$$|N(F)| \leqq |F''| \leqq |F'| .$$

Далее, если $|N(F)| = |F''| = |F'|$, то

$$N(F) = \{B' : \exists (A,B') \in F'\} \qquad \square$$

Следствие 2.3. Если $F \subseteq P(\Omega) \times P(\Omega)$ - полное семейство и $|\Omega| = n$, то существует такое F', генерирующее F, для которого $|F'| \leqq \frac{3}{2} \binom{n}{[n/2]}$.

ЛИТЕРАТУРА

[1] ARMSTRONG, W.W.: Dependency structures of data base relationships, Information Processing 74, N.-H. Publ. Co.(1974), 580-583.

[2] ARMSTRONG, W.W.: On the generation of dependency stuctures of relational data bases, Publication 272, Université de Montréal (1977)

[3] ARMSTRONG, W.W., DELOBEL, C.: Decomposition and functional dependencies in relations, Publication 271, Departement D´ Informatique et de Recherche Operationelle, Université de Montréal (1979).

[4] BEERI, C., FAGIN, R., HOWARD, I.H.: A complete axiomatization for functional and multivalued dependencies in database relations, Proc. ACM SIGMOD Int. Conf. on Management of Data, Toronto, (1977), 47-61.

[5] BÉKÉSSY, A. - DEMETROVICS J.: Contribution to the theory of data base relations, Discrete Math. 27(1979), 1-10.

[6] CODD, E.F.: A relational model of data for large shared data banks, Comm. ACM, 13(1970) 377-387.

[7] CODD, E.F.: Further normalization of the data base relational model, Courant Computer Science Symposia 6 Data Base System, Prentice Hall, Englewood Cliffs, N.I. (1971) 33-64.

[8] CZÉDLI, G.: Függőségek relációs adatbázis modellben, Alk. Mat. Lapok (1980)

[9] DEMETROVICS, J.: Candidate keys and antichains, SIAM J. on Algebraic and Discrete Methods 1 (1980)

[10] KLEITMAN, D.: On a combinatorial problem of Erdős Proc. AMS (1966, February), 139-141.

[11] 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, 5(1976), 100-101.

ÖSSZEFOGLALÁS

Funkcionális függőségek generálása és relációkkal való reprezentálása

J. Demetrovics - Gy. Gyepesi

Ebben a cikkben a funkcionális függőségek elméletének két kombinatórikus problémáját vizsgáljuk.

Először becslést adunk az [9]-ben definiált s(n)-re és S(n)-re, ahol s(n) az a legkisebb természetes szám, amelyre egy n elemü halmaz minden Sperner rendszere /részhalmazaiból álló antilánc/ reprezentálható s(n) soros reláció kulcsrendszereként; /S(n) ugyanez "Sperner rendszer" helyett "teljes család"-dal/.

Ezután kombinatórikusan jellemezzük a teljes családok minimális számosságu generátorrendszereit és becslést adunk erre a számosságra.

Generation of functional dependencies and their representation by relations

In this paper we deal with two combinatorial problems of the theory of functional dependencies.

Firstly we estimate s(n), defined in [9] and S(n), where s(n) is the minimal natural number such that every Sperner system of an n-element set may be represented as the set of candidata keys of a relation with s(n) rows; and S(n) is the same with "full fmaily" instead of "Sperner system".

Secondly we give a combinatorial characterization for the generating sets with minimal cardinality of full families and estimate this cardinality.