Задача поиска ассоциативных свойств данных в реляционных системах

Филатов В.А. Харьковский национальный университет радиоэлектроники

В современных технологиях обработки информации, в которых преобладает разделение информационных единиц на данные и команды, создалась ситуация, при которой данные пассивны, а команды активны. То есть, все протекающие процессы инициируются командами, а данные используются этими командами только в случае необходимости. Целью проводимых исследований является анализ особенностей информационных единиц и структур данных, которые влияют на технологию извлечение знаний.

Агрегатными функциями называются функции, которые определяют групповые свойства данных. К агрегатным функциям относятся функции COUNT, SUM, MAX, MIN, AVG и возможно другие, предложенные разработчиком [1].

Пусть имеется база данных, для доступа к которой, реализован набор транзакций $T = \{T_1,\ldots, T_n\}$, $D = \{d_1,..., d_n\}$ – множество элементов из которых определяет транзакцию из T, то есть $T_i \subseteq D$ и $\Omega = \{COUNT, SUM, MAX, MIN, AVG,\ldots\}$ – набор агрегатных функций. Каждая транзакция представляет собой бинарный вектор, где $T_i = 1$, если элемент d_i присутствует в транзакции и $T_i = 0$ в противном случае. Транзакция T_i содержит набор элементов $X \subseteq D$, если $X \subset T_i$. Тогда продукцией будем называть функциональное ассоциативное правило – $\{P; X \Rightarrow \varpi (Y)\}$, если $X \subset D$, $Y \subset D$, $X \cap Y = \varnothing$ и $\varpi \in \Omega$, где P – условие активизации ядра правила [2].

Обратим внимание на условие активизации правила. Для реализации секвенции «$\Rightarrow$» этого правила необходимо выполнение условия применимости ядра. В теории реляционных баз данных отсутствие значения в атрибуте недопустимо. С другой стороны, если схема данных содержит несколько связных отношений, то возможна ситуация когда значение связного атрибута в данный момент времени неопределенно [3].

Пусть $\Re(R_1(a,b,\underline{c}), R_2(c,\underline{d})\ R_3(d,\underline{e}))$ – реляционная база данных (подчеркнутые атрибуты являются ключами), в схеме которой определены связи $R_1 \xleftarrow{1:M} R_2$, $R_2 \xleftarrow{1:M} R_3$.

Утверждение. Пусть $\Re\ (R_1(A, \underline{B}),\ R_2(B, \underline{C}))$ – схема реляционной базы данных. Функциональное ассоциативное правило вида $\{C \neq \varnothing;\ A \Rightarrow \varpi(C)\}$ существует в том случае, если между отношениями R_1 и R_2 установлена связь типа 1:M.

Доказательство. Основываясь на определении типа связи «один-ко-многим» и исключив ситуацию, при которой связный элемент отсутствует (условие $C \neq \varnothing$ в утверждении), покажем, что всегда можно найти множество различных элементов одного множества, соответствующих одному элементу другого множества, то есть построить функциональное ассоциативное правило.

Пусть заданы множества $A = \{a_1,\ldots, a_n\}$, $B = \{b_1,\ldots, b_m\}$ и $C = \{c_1,\ldots, c_k\}$ и пусть заданы отношения $R_1 \subseteq A \times B$ такие, что b_i не повторяются и $R_2 \subseteq B \times C$,

где не повторяются c_i (согласно ключевым атрибутам, определенным в схеме исходной БД). Запишем кортежи произведений в следующем виде

$$R_1 = \{<(a_1,\ldots, a_n), b_1>\} \quad R_2 = \{<(b_1,\ldots, b_m), c_1>\}$$
$$\{<(a_1,\ldots, a_n), b_2>\} \quad \{<(b_1,\ldots, b_m), c_2>\}$$
$$\ldots \quad \ldots$$
$$\{<(a_1,\ldots, a_n), b_m>\} \quad \{<(b_1,\ldots, b_m), c_k>\}$$

Такая запись показывает, что в R_1 каждому значению из $\{b_1,\ldots, b_m\}$ может соответствовать одно любое значение из $\{a_1,\ldots, a_n\}$, а также в R_2 каждому значению из $\{c_1,\ldots, c_m\}$ может соответствовать одно любое значение из $\{b_1,\ldots, b_n\}$.

Рассмотрим возможное состояние базы данных

R_1

A	B
a_1	b_1
…	…
a_n	b_m

R_2

B	C
b_1	c_1
…	…
b_1	c_k

В общем виде соответствие значений атрибутов A и C можно записать в виде $a_i \rightarrow (c_1,\ldots, c_k)$. Таким образом, можно применить агрегатную функцию для вычисления по атрибуту C сгруппированного по атрибуту A и построить соответствующее функциональное ассоциативное правило $\{C \neq \varnothing; A \Rightarrow \varpi(C)\}$.

Необходимо отметить, что на практике для однозначной идентификации значений не ключевого атрибута в ядре правила необходимо использовать значение ключа. Для рассмотренного в утверждении примера функциональное ассоциативное правило $\{C \neq \varnothing; A \Rightarrow C\}$ примет вид $\{C \neq \varnothing; A, B \Rightarrow C\}$[4].

Дальнейшие исследования процедур поиска и анализа продукций могут быть направлены на разработку методов логического вывода ассоциативных правил на основании системы аксиом. Кроме этого, при поиске правил можно использовать и другие свойства реляционной модели данных, такие как функциональные и другие виды зависимостей, в частности математический аппарат реляционного исчисления.

Список литературы:

1. Codd E.F. Relational completeness of data base sublanguages. – Ibid. 1972, p. 65-98.
2. R. Srikant, R. Agrawal. "Mining quantitative association rules in large relational tables". In Proceedings of the ACM SIGMOD Conference on Management of Data, Montreal, Canada, June 1996, p. 1-12.
3. Бениаминов Е.М. Алгебраические методы в теории баз данных и представлении знаний. - М.: Научный мир, 2003, 184 с
4. Пономаренко Л. А. Програмні агентні технології в адмініструванні баз даних / Л. А. Пономаренко, В. О. Філатов. // Вісник Київського торговельно-економічного університету. – 2001. – №3. – С. 68–73.