УДК 004.655

©2014. **И. Н. Глушко**

ФОРМАЛЬНАЯ СЕМАНТИКА АГРЕГАТНЫХ ОПЕРАЦИЙ В ТАБЛИЧНЫХ АЛГЕБРАХ

В работе проводится расширение табличной алгебры (введенной В.Н. Редько и Д.Б. Буем и являющейся обобщением классической реляционной алгебры Кодда), которое предполагает пополнение универсального домена специальным элементом NULL и расширение сигнатуры табличной алгебры конечных таблиц агрегатными операциями нахождения суммы, наибольшего (наименьшего) значений, среднего арифметического, количества строк и количества элементов, отличных от NULL. Задана формальная математическая семантика этих операций, которая проиллюстрирована содержательными примерами. При определении агрегатных операций используется понятие мультимножества. Общая схема задания агрегатных операций: сначала операции задаются на конечных мультимножествах, а затем переносятся на таблицы, в частности, на пустую таблицу.

Ключевые слова: *реляционные базы данных, табличная (реляционная) алгебра, расширенная табличная алгебра, агрегатные операции.*

1. Введение. Традиционно считается, что классическая реляционная (табличная) алгебра лежит в основе большинства СУБД и языков запросов, которые поддерживают реляционную модель. Реляционная алгебра была разработана в [1] в виде совокупности операторов над таблицами. Было предложено 8 операций реляционной алгебры: традиционные операции над множествами (объединение, пересечение, разность) и специальные операции над таблицами (проекция, декартово соединение, theta-, equi-соединения, деление, селекция). Этот набор операций со временем был расширен в соответствии с потребностями языков запросов. Кроме указанных выше операций к сигнатуре реляционной алгебры сейчас также относят операции переименования и активного дополнения [2, 3]. В ходе развития коммерческих реляционных СУБД возникла потребность в использовании агрегатных функций, которые позволяют находить суммарные, средние, максимальные, минимальные и другие значения элементов в столбце таблицы. В [4] реляционная алгебра и реляционное исчисление расширены такими агрегатными функциями. Доказана эквивалентность полученных при этом двух формальных языков. Даны точные определения агрегатных функций, которые не используют понятие «дубликаты». Реляционная алгебра пополнена новой операцией агрегатного образования (aggregate formation). В [5–7] операции агрегирования рассматриваются как множественно-ориентированные: сначала разбивают таблицу на подмножества в соответствии со значениями атрибута (или множества атрибутов), затем выполняют функциональные вычисления для каждого подмножества и, наконец, строят исходную таблицу, формируя одну строку для каждого подмножества. Такая схема вычислений используется для запросов третьего типа (с группировкой) языка SQL [3]. Отметим, что для семантики таких конструкций надо вводить в рассмотрение совокупности с повторениями, т.е. мультимножества, что и сделано в данной работе.

2. Основные определения. Все неопределенные понятия понимаем также, как и в [3]. Рассмотрим два множества: $\boldsymbol{A}$ – множество атрибутов и $\boldsymbol{D}$ – универсальный домен, содержащий специальное значение $NULL$. Под табличной алгеброй конечных таблиц понимаем алгебру $\langle \boldsymbol{T'}, \Omega_{P,\Xi} \rangle$, где $\boldsymbol{T'}$ – множество всех конечных таблиц, $\Omega_{P,\Xi} = \{\bigcup_R, \bigcap_R, \setminus_R, \sigma_{p,R}, \pi_{X,R}, \underset{R_1,R_2}{\otimes}, \div_{R_2}^{R_1}, Rt_{\xi,R}, \sim_R\}$ – сигнатура, $p \in P, \xi \in \Xi$, а $X, R, R_1, R_2 \subseteq \boldsymbol{A}$ (P и Ξ – множества параметров).

Под таблицей схемы R понимаем пару $\langle t, R \rangle$, где $t \in T(R)$[1)] – конечное множество строк схемы R. Тогда $\boldsymbol{T'}(R) = \{\langle t, R \rangle \,|t \in T(R)\}$ – множество всех конечных таблиц схемы R, а $\boldsymbol{T'} = \bigcup\limits_{R \subseteq \mathrm{A}} \boldsymbol{T'}(R)$ – множество всех конечных таблиц.

В соответствии с [3,8] под мультимножеством α с основой U понимаем функцию вида $\alpha : U \to \mathbb{N}$. Пусть $\Theta(\alpha)$ – основа мультимножества α, а $2_m^{\boldsymbol{D'}} = \{\alpha | \Theta(\alpha) \in 2^{\boldsymbol{D'}}\}$ – семейство всех мультимножеств, основы которых являются конечными подмножествами множества $\boldsymbol{D'}$ ($\boldsymbol{D'} \subseteq \boldsymbol{D}$ – подмножество универсального домена). Мультимножество α с основой $\{d_1, ..., d_k\}$ будем записывать как $\{d_1^{n_1}, ..., d_k^{n_k}\}$, где n_i – количество дубликатов (экземпляров) элемента d_i в мультимножестве α, $i = 1, ..., k$.

3. Основные результаты. Широко используемыми агрегатными операциями являются Sum, Avg, Min, Max, $Count$. Их аргументы – это конечные таблицы, а значения – одноатрибутные таблицы с одной строкой. Так, операция Sum рассчитывает сумму значений в соответствующем столбце заданной таблицы, при этом значения $NULL$ игнорируются. Операция Avg определяет среднее арифметическое значений в соответствующем столбце заданной таблицы, при этом значения $NULL$ игнорируются. Операции Min и Max находят наименьшее и наибольшее значения в соответствующем столбце заданной таблицы, при этом значения $NULL$ также игнорируются. Операция $Count$ определяет количество значений, отличных от $NULL$, в соответствующем столбце заданной таблицы. Операция $Count(*)$ определяет количество строк в заданной таблице.

Пусть Num – числовое подмножество универсального домена $\boldsymbol{D}$, замкнутое относительно сложения. Зададим агрегатные операции. Общая схема: на конечном мультимножестве определяются функции суммирования, взятие наименьшего и наибольшего значений, определение среднего арифметического и количества элементов, а затем эти функции переносятся на таблицы. Заметим, что функции суммирования и нахождения среднего арифметического определены на конечном числовом мультимножестве.

Рассмотрим таблицу $\langle t, R \rangle \in \boldsymbol{T'}(R)$ и пусть $A \in R$. Обозначим через α_A мультимножество, которое содержит все элементы столбца с атрибутом A таблицы $\langle t, R \rangle$. Тогда $\Theta(\alpha_A) = D_{A,t}$, где $D_{A,t} = \{d | \exists s(s \in t \wedge \langle A, d \rangle \in s)\}$ – активный домен атрибута A относительно таблицы $\langle t, R \rangle$ [2, 3].

Для определения количества дубликатов элемента d в мультимножестве α_A зададим отображение $\varphi : t \to D_{A,t}$, где $\varphi(s) = s(A)$ ($s \in t$). Тогда количество дубликатов элемента основы $d \in D_{A,t}$ мультимножества α_A равно $\alpha_A(d) = \left|\varphi^{-1}(d)\right|$, где

[1)] Множество $T(R)$ понимается в смысле [3].

$\varphi^{-1}(d) = \{s | s \in t \wedge \varphi(s) = d\}$ – прообраз элемента $d \in D_{A,t}$ относительно отображения φ, а $|X|$ – мощность множества X.

Агрегированием $Sum_{A,R}$ по атрибуту A (конечных) таблиц схемы R назовем такую унарную параметрическую операцию $Sum_{A,R} : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Sum_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, Sum(\alpha_A)\rangle\}\}, \{A\}\rangle,$$

где $\langle t, R\rangle \in \boldsymbol{T}'(R)$, а Sum[2)] – функция, возвращающая сумму значений столбца с атрибутом A таблицы $\langle t, R\rangle$ (значения могут повторяться), которые отличаются от значения $NULL$. Кроме того, предполагается, что этот столбец содержит только числовые данные. Таким образом, $Sum : 2_m^{Num} \to Num$, где

$$Sum(\alpha_A) = \begin{cases} NULL, \text{ если } \Theta(\alpha_A) = \emptyset; \\ NULL, \text{ если } \Theta(\alpha_A) = \{NULL\}; \\ \sum\limits_{d \in \Theta(\alpha_A)\setminus\{NULL\}} d\alpha(d), \text{ если } \Theta(\alpha_A)\setminus\{NULL\} \neq \emptyset. \end{cases}$$

Из определения следует, что $Sum(\emptyset_m) = NULL$[3)], $Sum(\{NULL^n\}) = NULL$, $Sum\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \sum\limits_{i=1}^{k} d_i n_i$, в предположении, что все элементы d_i $(i = \overline{1,k})$ отличаются от элемента $NULL$, $Sum_{A,R}\left(\langle t_\emptyset, R\rangle\right) = \langle\{\{\langle A, NULL\rangle\}\}, \{A\}\rangle$, где $\langle t_\emptyset, R\rangle$ – пустая таблица.

Проиллюстрируем применение операции агрегирования $Sum_{A,R}$ на примере.

Пример 1. Для таблицы

$$\langle t, R\rangle = \begin{array}{ccc} \mathbf{A} & \mathbf{B} & \mathbf{C} \\ 1 & 2 & 3 \\ 2 & 1 & 1 \\ 2 & 3 & NULL \end{array}$$

получим $Sum_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 5\rangle\}\}, \{A\}\rangle$, $Sum_{B,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 6\rangle\}\}, \{B\}\rangle$ и $Sum_{C,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 4\rangle\}\}, \{C\}\rangle$.

Агрегированием $Count_{A,R}$ по атрибуту A (конечных) таблиц схемы R назовем такую унарную параметрическую операцию $Count_{A,R} : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Count_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, Count(\alpha_A)\rangle\}\}, \{A\}\rangle,$$

где $\langle t, R\rangle \in \boldsymbol{T}'(R)$, а $Count$ – функция, возвращающая количество значений, которые отличаются от значения $NULL$ с учетом дубликатов в столбце с атрибутом A таблицы $\langle t, R\rangle$. Таким образом, $Count : 2_m^{\boldsymbol{D}} \to \mathbb{Z}_+$, где

$$Count(\alpha_A) = \sum_{d \in \Theta(\alpha_A)\setminus\{NULL\}} \alpha_A(d).$$

[2)] Операция $Sum_{A,R}$ определена на конечных таблицах, а функция Sum – на конечных мультимножествах чисел.

[3)] $\emptyset_m$ – пустое мультимножество.

По определению считаем, что сумма пустого множества слагаемых равна нулю. Итак, $Count(\emptyset_m) = 0$, $Count(\{NULL^n\}) = 0$ и $Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \sum\limits_{i=1}^{k} n_i$, в предположении, что все элементы d_i $(i = \overline{1,k})$ отличаются от элемента $NULL$. Для случая пустой таблицы $\langle t_\emptyset, R\rangle$ имеем $Count_{A,R}\left(\langle t_\emptyset, R\rangle\right) = \langle\{\{\langle A, 0\rangle\}\}, \{A\}\rangle$.

Пример 2. Для таблицы из примера 1 получим

$$Count_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 3\rangle\}\}, \{A\}\rangle,$$
$$Count_{B,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 3\rangle\}\}, \{B\}\rangle,$$
$$Count_{C,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 2\rangle\}\}, \{C\}\rangle.$$

Агрегированием $Count_{A,R}(*)$ (конечных) таблиц схемы R назовем такую унарную параметрическую операцию $Count_{A,R}(*) : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Count_{A,R}(*)\left(\langle t, R\rangle\right) = \langle\{\{\langle A, |t|\rangle\}\}, \{A\}\rangle,$$

что $\langle t, R\rangle \in \boldsymbol{T}'(R)$. Содержательно говоря, операция $Count_{A,R}(*)$ определяет количество строк заданной таблицы. Атрибут-параметр служит только для формирования схемы таблицы-результата. Для случая пустой таблицы $\langle t_\emptyset, R\rangle$ имеем $Count_{A,R}(*)\left(\langle t_\emptyset, R\rangle\right) = \langle\{\{\langle A, 0\rangle\}\}, \{A\}\rangle$.

Пример 3. Для таблицы из примера 1 получим

$$Count_{A,R}(*)\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 3\rangle\}\}, \{A\}\rangle,$$
$$Count_{B,R}(*)\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 3\rangle\}\}, \{B\}\rangle,$$
$$Count_{C,R}(*)\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 3\rangle\}\}, \{C\}\rangle.$$

Допустим, что числовое подмножество Num универсального домена замкнуто относительно (частичной операции) деления $/ : Num \times Num \tilde{\to} Num$. Доопределим операцию деления так, что когда первый аргумент равен $NULL$, то функция принимает значение $NULL$. Это связано с тем, что мы будем осуществлять суперпозиции и вместо первого аргумента подставлять значение функции Sum, а вместо второго – значение функции $Count$, учитывая, что функция $Count$ в качестве значения не может выдать значение $NULL$.

Агрегированием $Avg_{A,R}$ по атрибуту A (конечных) таблиц схемы R назовем такую унарную параметрическую операцию $Avg_{A,R} : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Avg_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, Avg\left(\alpha_A\right)\rangle\}\}, \{A\}\rangle,$$

где $\langle t, R\rangle \in \boldsymbol{T}'(R)$, а Avg – функция, которая возвращает среднее арифметическое значение элементов столбца с атрибутом A таблицы $\langle t, R\rangle$, которые отличаются от значения $NULL$, причем с учетом дубликатов, т.е. $Avg : 2_m^{Num} \to Num$ и

$$Avg(\alpha_A) = \frac{Sum(\alpha_A)}{Count(\alpha_A)}.$$

Из определения следуют равенства

$$Avg(\emptyset_m) = \frac{Sum(\emptyset_m)}{Count(\emptyset_m)} = \frac{NULL}{0} = NULL,$$

$$Avg(\{NULL^n\}) = \frac{Sum(\{NULL^n\})}{Count(\{NULL^n\})} = \frac{NULL}{0} = NULL,$$

$$Avg\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \frac{Sum\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right)}{Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right)} = \frac{1}{(n_1 + \ldots + n_k)} \sum_{i=1}^{k} d_i n_i$$

в предположении, что все элементы d_i $(i = \overline{1,k})$ отличны от $NULL$.

Пример 4. Для таблицы из примера 1 получим

$$Avg_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 5/3\rangle\}\}, \{A\}\rangle,$$

$$Avg_{B,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 2\rangle\}\}, \{B\}\rangle,$$

$$Avg_{C,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 2\rangle\}\}, \{C\}\rangle.$$

Пусть $\leq$ – линейный порядок на универсальном домене $\boldsymbol{D}$. Агрегированием $Min_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$ назовем такую унарную параметрическую операцию $Min_{A,R} : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Min_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, Min(\alpha_A)\rangle\}\}, \{A\}\rangle,$$

где $\langle t, R\rangle \in \boldsymbol{T}'(R)$, а Min – функция, которая возвращает наименьшее значение среди значений столбца с атрибутом A таблицы $\langle t, R\rangle$, которые отличаются от значения $NULL$. Таким образом $Min : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$, где

$$Min(\alpha_A) = \begin{cases} NULL, \text{ если } \Theta(\alpha_A) = \emptyset; \\ NULL, \text{ если } \Theta(\alpha_A) = \{NULL\}; \\ \min\{d | d \in \Theta(\alpha_A) \backslash \{NULL\}\}, \text{ если } \Theta(\alpha_A) \backslash \{NULL\} \neq \emptyset. \end{cases}$$

Из определения следует, что $Min(\emptyset_m) = NULL$, $Min(\{NULL^n\}) = NULL$, а $Min\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \min\{d_1, \ldots, d_k\}$ в предположении, что все элементы d_i $(i = \overline{1,k})$ отличны от элемента $NULL$. Для случая пустой таблицы $\langle t_\emptyset, R\rangle$ имеем $Min_{A,R}\left(\langle t_\emptyset, R\rangle\right) = \langle\{\{\langle A, NULL\rangle\}\}, \{A\}\rangle$.

Пример 5. Для таблицы из примера 1 получим

$$Min_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 1\rangle\}\}, \{A\}\rangle,$$

$$Min_{B,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 1\rangle\}\}, \{B\}\rangle,$$

$$Min_{C,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 1\rangle\}\}, \{C\}\rangle.$$

Агрегированием $Max_{A,R}$ по атрибуту A (конечных) таблиц схемы R назовем такую унарную параметрическую операцию $Max_{A,R} : \boldsymbol{T}'(R) \to \boldsymbol{T}'(\{A\})$, что

$$Max_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, Max(\alpha_A)\rangle\}\}, \{A\}\rangle,$$

где $\langle t, R\rangle \in \boldsymbol{T}'(R)$, а Max – функция, которая возвращает наибольшее значение среди значений столбца с атрибутом A таблицы $\langle t, R\rangle$, которые отличаются от $NULL$. Таким образом $Max : 2^{\boldsymbol{D}}_m \to \boldsymbol{D}$, где

$$Max(\alpha_A) = \begin{cases} NULL, \text{ если } \Theta(\alpha_A) = \emptyset; \\ NULL, \text{ если } \Theta(\alpha_A) = \{NULL\}; \\ \max\{d | d \in \Theta(\alpha_A)\backslash\{NULL\}\}, \text{если } \Theta(\alpha_A)\backslash\{NULL\} \neq \emptyset. \end{cases}$$

Из определения следует, что $Max(\emptyset_m) = NULL$, $Max(\{NULL^n\}) = NULL$, $Max\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \max\{d_1, \ldots, d_k\}$ в предположении, что все элементы d_i ($i = \overline{1,k}$) отличны от значения $NULL$. Для случая пустой таблицы $\langle t_\emptyset, R\rangle$ имеем $Max_{A,R}\left(\langle t_\emptyset, R\rangle\right) = \langle\{\{\langle A, NULL\rangle\}\}, \{A\}\rangle$.

Пример 6. Для таблицы из примера 1 получим

$$Max_{A,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle A, 2\rangle\}\}, \{A\}\rangle,$$

$$Max_{B,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle B, 3\rangle\}\}, \{B\}\rangle,$$

$$Max_{C,R}\left(\langle t, R\rangle\right) = \langle\{\{\langle C, 3\rangle\}\}, \{C\}\rangle.$$

Отметим, что функции Min и Max определяют наименьший или наибольший элементы основы мультимножества, которые отличаются от значения $NULL$, поэтому сопоставимость особого элемента с остальными элементами универсального домена в данном случае несущественна. В конкретных реализациях SQL элемент $NULL$ может быть как наименьшим, так и наибольшим элементом [4)].

4. Выводы. В статье определена формальная математическая семантика агрегатных операций, которая проиллюстрирована примерами их применения. Результаты работы могут быть использованы в теории обобщенных табличных алгебр. Полученные результаты можно расширить на таблицы, рассматриваемые как мультимножества строк. Кроме того, параметром агрегатной операции может выступать не только отдельный атрибут, но и некоторая функция над строкой.

1. *Codd E.F.* A Relational model of data for large shared data banks // Comm. of ACM. – 1970. – № 6. – P. 377–387.
2. *Мейер Д.* Теория реляционных баз данных. – М.: Мир, 1987. – 608 с.
3. *Редько В.Н., Брона Ю.Й., Буй Д.Б., Поляков С.А.* Реляційні бази даних: табличні алгебри та SQL-подібні мови. – Київ: Видавничий дім «Академперіодика», 2001. – 198 с.
4. *Klug A.* Equivalence of relational algebra and relational calculus query languages having aggregate functions // J. ACM. – 1982. – № 3. – P. 699–717.

[4)] Сопоставимость особого элемента важна при интерпретации фразы ORDER BY, которая предназначена для «упорядочения» результата запроса (подробности см. в [3]).

5. *Silbeschatz A., Korth H., Sudarshan S.* Database system concepts. – NY: McGraw-Hill, 2011. – 1376 p.
6. *Garcia-Molina H., Ullman J.D., Widom J.* Database systems: the complete book. – NY: Prentice Hall, 2008. – 1119 p.
7. *Дейт К.Дж.* Введение в системы баз данных. – М.: Издательский дом «Вильямс», 2005. – 1328 с.
8. *Петровский А.Б.* Основные понятия теории мультимножеств. – М.: Едиториал УРСС, 2002. – 80 с.

I. M. Glushko

A formal semantics of aggregate operations.

The paper deals with the extension of table algebra (this algebra is introduced by V.N. Redko and D.B. Bui and is a generalization of well-known classic Codd's relational algebra), involving the completion of a special element NULL of the universal domain and expansion of signature finite table algebra with such aggregate operations: finding the sum, the largest (smallest) value, the average value, the number of table rows and number of elements different from NULL. The formal mathematical semantics of these operations illustrated by meaningful examples is given. The concept of multiset is used under defining the aggregate operators. The general scheme of the aggregate operations is following: at first, operations are defined on finite multisets, then they are extended to the tables, in particular, on an empty table.

Keywords: *relation databases, table (relation) algebra, extending table algebra, aggregate operations.*

Нежинский государственный ун-т им. Николая Гоголя
`glushkoim@gmail.com`

Получено 21.01.14