УДК 004.655

Мультимножественная табличная алгебра: дополнительные операции

Д. Б. Буй[*], И. Н. Глушко[**]

[*] *Киевский национальный университет имени Тараса Шевченко, Киев, Украина*
[**] *Нежинский государственный университет имени Николая Гоголя, Нежин, Украина*

Сигнатура мультимножественной табличной алгебры пополнена новыми операциями: операциями внутренних и внешних соединений, операцией полусоединения, агрегатными операциями. Задана формальная математическая семантика указанных операций. Для задания внешних соединений введен особый элемент универсального домена NULL.

Ключевые слова: *реляционные базы данных, мультимножественная табличная алгебра, расширенная мультимножественная табличная алгебра.*

Сигнатура мультимножинної табличної алгебри поповнена новими операціями: операціями внутрішніх і зовнішніх з'єднань, операцією напівз'єднання, агрегатними операціями. Задано формальну математичну семантику вказаних операцій. Для задання зовнішніх з'єднань введено особливий елемент універсального домену NULL

Ключові слова: *реляційні бази даних, мультимножинна таблична алгебра, розширена мультимножинна таблична алгебра.*

The signature of multiset table algebra is filled up with new operations such as inner and outer joins, semijoin and aggregate operations. A formal mathematical semantics of these operations is defined. The special element NULL is introduced into the universal domain to define of outer join.

Key words: *relation databases, multiset table algebra, extending multiset table algebra.*

1. Общая постановка задачи и её актуальность

Реляционная модель данных в настоящее время широко используется как в научных исследованиях в базах данных, так и на практике. Данная модель основана на множествах кортежей, то есть не позволяет дубликаты кортежей в отношении [1]. Однако многие языки, ориентированные на работу с базами данных, требуют реляционную модель данных с мультимножественной семантикой (multiset semantics). Это предполагает понимание таблиц как мультимножеств, т.е. совокупностей с дубликатами. Вопросу использования мультимножеств в базах данных уделяли внимание G. Lamperti, M. Melchiori, M. Zanella [2], Г. Гарсиа-Молина, Дж. Ульман, Дж. Уидом [3], A. Silbeschatz, H. Korth, S. Sudarshan [4], а также отечественные ученые Д.Б. Буй, С.А. Поляков, Ю.Й. Брона, В.Н. Редько [5]. Обзор литературы об использовании мультимножеств в базах данных проведен в статье [6], которая насчитывает 9 источников по данной теме.

Вместе с тем, этот вопрос требует уточнения и расширения, поскольку ни в одной из указанных работ не уделяется достаточное внимание операциям внутренних и внешних соединений, операции полусоединения, внешним мультимножественным операциям, а также агрегатным операциям над таблицами мультимножественной табличной алгебры.

2. Основные понятия теории мультимножеств

Приведем основные понятия мультимножеств в терминах работ [5, 7]. Зафиксируем множество U. Под мультимножеством α с основой U понимаем отображение вида $\alpha : U \to \{1,2,...\}$. Пусть D – универсум элементов основ мультимножеств, тогда булеан $P(D)$ – универсум основ мультимножеств.

Под характеристической функцией мультимножества α понимаем функцию вида $\chi_\alpha : D \to \{0,1,2,...\}$, значение которой задается кусочной схемой :

$$\chi_\alpha(d) = \begin{cases} \alpha(d), \textit{если } d \in \mathrm{dom}\,\alpha, \\ 0, \textit{иначе}; \end{cases}$$

для всех $d \in D$, где $\mathrm{dom}\,\alpha$ – область определения мультимножества α, т.е. его основа.

Мультимножество называется пустым и обозначается как $\varnothing_m$, если его основа – пустое множество.

Мультимножества, областью значений которых является пустое множество или одноэлементное множество вида $\{1\}$ называются 1-мультимножествами. Эти мультимножества есть аналогами обычных множеств.

Договоримся мультимножество α с основой $\{d_1,...,d_k\}$ записывать как $\{d_1^{n_1},...,d_k^{n_k}\}$, где n_i – количество дубликатов (экземпляров) элемента d_i в мультимножестве α, т.е. $n_i = \alpha(d_i)$, $i = 1,...,k$.

Под рангом конечного мультимножества α понимаем количество дубликатов элементов его основы $\|\alpha\| = \sum_{d \in \mathrm{dom}\,\alpha} \alpha(d)$; при этом $\|\varnothing_m\| = 0$.

Скажем, что мультимножество β включается в мультимножество α ($\beta \preceq \alpha$), если: $\beta \preceq \alpha \Leftrightarrow U_\beta \subseteq U_\alpha \;\&\; \forall d\big(d \in U_\beta \Rightarrow \beta(d) \le \alpha(d)\big)$. Здесь U_α и U_β основы мультимножеств α и β соответственно.

Если $\beta \preceq \alpha$, то мультимножество β называется подмультимножеством мультимножества α, а мультимножество α – надмультимножеством мультимножества β.

В роботе [5] операции над мультимножествами определены в терминах характеристических функций. Авторы определяют операции объединения $\bigcup_1$, пересечения $\bigcap_1$, разности $\setminus_1$ мультимножеств, которые строят 1-мультимножества, основы которых получаются соответственноc теоретико-множественными объединением, пересечением и разницей основ мультимножеств-аргументов. Кроме того, вводятся операции объединения $\bigcup_{All}$, пересечения $\bigcap_{All}$, разности $\setminus_{All}$ мультимножеств, которые строят мультимножества общего вида. Также задано операцию декартового соединения мультимножеств $\otimes$ и операцию $Dist(\alpha)$, которая строит 1-мультимножество, основа которого совпадает с основой исходного мультимножества. Наконец, в этой работе вводится аналог полного образа (множества относительно функции) для мультимножеств.

3. Мультимножественная табличная алгебра

Рассмотрим два множества: $\boldsymbol{A}$ – множество атрибутов (имен) и $\boldsymbol{D}$ – универсальный домен (множество денотатов). Произвольное (конечное) множество атрибутов $R \subseteq \boldsymbol{A}$ назовём схемой. Под строкой схемы R понимаем именное множество на паре $\boldsymbol{A}$ и $\boldsymbol{D}$ [5], проекция которого по первой компоненте совпадает с R, т. е. строка схемы R – это функция вида $s : R \to \boldsymbol{D}$. Множество всех строк схемы R обозначим $S(R)$, а множество всех строк – S.

Под таблицей схемы R понимаем пару $\langle \psi, R \rangle$, где первая компонента ψ – это произвольное мультимножество, основой которого $\Theta(\psi)$ является произвольное множество, в частности, бесконечное, строк схемы R, а вторая компонента R – схема (таблицы).

Под мультимножественной табличной алгеброй понимаем алгебру $\langle \Psi, \Omega_{P,\Xi} \rangle$, где $\Psi = \bigcup\limits_{R \subseteq \boldsymbol{A}} \Psi(R)$ – множество всех таблиц, $\Psi(R)$ – множество всех таблиц схемы R,

$$\Omega_{P,\Xi} = \left\{ \bigcup\nolimits_{All}^{R}, \bigcap\nolimits_{All}^{R}, \setminus_{All}^{R}, \sigma_{p,R}, \pi_{X,R}, \underset{R_1,R_2}{\otimes}, Rt_{\xi,R}, \sim_R \right\}_{X,R,R_1,R_2 \subseteq \boldsymbol{A}}^{p \in P, \xi \in \Xi}$$ – сигнатура; P, Ξ – множества параметров. Операции мультимножественной табличной алгебры задано в [8].

Пополним сигнатуру мультимножественной табличной алгебры новыми операциями: операциям внутренних и внешних соединений, операцией полусоединения, агрегатными операциями.

4. Операции внутреннего соединения

Под декартовым соединением Cj (Cartesian Join) таблиц схем R_1 и R_2, причем $R_1 \cap R_2 = \varnothing$, понимаем бинарную параметрическую операцию вида $\underset{R_1,R_2}{Cj} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{Cj} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, где $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. Основой мультимножества ψ' является множество строк $\Theta(\psi') = \{ s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s = s_1 \cup s_2) \}$. Количество дубликатов определяется так: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, где $s \in \Theta(\psi')$ и $s = s_1 \cup s_2$ (очевидно, что это разложение строки s единственно).

Ниже используется бинарное отношение совместимости строк, которое вводится так: $s_1 \approx s_2 \overset{def}{\Leftrightarrow} s_1 \mid R' = s_2 \mid R'$, де $R' = R_1 \cap R_2$, а R_1, R_2 – схемы строк s_1, s_2 соответственно [5]. Основное свойство этого отношения состоит в следующем: $s_1 \cup s_2 \in S(R_1 \cup R_2) \Leftrightarrow s_1 \approx s_2$ [5].

Под внутренним естественным соединением (Inner Natural Join) таблиц схем R_1 и R_2 понимаем бинарную параметрическую операцию $\underset{R_1,R_2}{\otimes}$, значениями

которой являются таблицы схемы $R_1 \cup R_2$, которые, говоря содержательно, содержат все объединения совместных строк исходных таблиц. Таким образом, $\underset{R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2)$, $\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, где $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. Основой мультимножества ψ' является множество строк вида

$$\Theta(\psi') = \{ s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2 \wedge s = s_1 \cup s_2) \}.$$

Количество дубликатов определяется, как и ранее, так: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, где $s \in \Theta(\psi')$ и $s = s_1 \cup s_2$ (как и ранее, приведенное разложение единственно).

Под внутренним соединением по атрибутам $A_1, ..., A_n$ (Inner Join using $A_1, ..., A_n$), причем все $A_1, ..., A_n$ попарно различны, $n \geq 1$, таблиц схем R_1 и R_2, где $R_1 \cap R_2 \supseteq \{A_1, ..., A_n\}$ (подразумевается, что общие атрибуты, отличающиеся от атрибутов $A_1, ..., A_n$, перед соединением переименовываются) понимаем бинарную параметрическую операцию вида

$$\underset{A_1,...,A_n,R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \cup R_2),$$

причем $\langle \psi_1, R_1 \rangle \underset{A_1,...,A_n,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$, где $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. Основой мультимножества ψ' является множество строк

$$\Theta(\psi') = \left\{ s \mid \exists s_1 \exists s_2 \left(s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge \bigwedge_{i=1}^{n} s_1(A_i) = s_2(A_i) \wedge s = s_1 \cup s_2 \right) \right\}.$$

Количество дубликатов определяется, как и ранее, так: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, где $s \in \Theta(\psi')$ и $s = s_1 \cup s_2$ (как и ранее, приведенное разложение единственно).

Отметим, если таблицы-аргументы имеют еще и другие общие атрибуты, отличающиеся от атрибутов $A_1, ..., A_n$, то перед соединением их нужно переименовать.

Пусть $p : S \times S \xrightarrow{\sim} \{true, false\}$ – вообще говоря частичный бинарный предикат на множестве всех строк S, такой, что выполняется импликация $\forall s_1 \forall s_2 \left(\langle s_1, s_2 \rangle \in \operatorname{dom} p \wedge p(s_1, s_2) = true \Rightarrow s_1 \approx s_2 \right)$.

Под внутренним соединением по предикату p (Inner Join on p) таблиц схем R_1 и R_2 понимаем частичную бинарную параметрическую операцию вида $\underset{p,R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, $\operatorname{dom} \underset{p,R_1,R_2}{\otimes} = \{ \langle \langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle \rangle \mid \Theta(\psi_1) \times \Theta(\psi_2) \subseteq \operatorname{dom} p \}$, $\langle \psi_1, R_1 \rangle \underset{p,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \cup R_2 \rangle$. Основой

мультимножества ψ' является множество строк $\Theta(\psi') = \{s \mid \exists s_1 \exists s_2 (s_1 \in \Theta(\psi_1) \wedge s_2 \in \Theta(\psi_2) \wedge p(s_1, s_2) \simeq true \wedge s = s_1 \bigcup s_2)\}$. Количество дубликатов определяется, как и ранее, так: $Occ(s, \psi') = Occ(s_1, \psi_1) \cdot Occ(s_2, \psi_2)$, где $s \in \Theta(\psi')$ и $s = s_1 \bigcup s_2$. Выше предполагалось, что пара таблиц $\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle$ принадлежит указанной области определенности.

Отметим следующий очевидный факт. Операция естественного соединения $\underset{R_1,R_2}{\otimes}$ является расширением произвольной другой операции соединения в следующем смысле: $\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{Cj} \langle \psi_2, R_2 \rangle = \langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle$,

$\langle \psi_1, R_1 \rangle \underset{A_1,\ldots,A_n,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle = \langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle$,

$\left(\langle \psi_1, R_1 \rangle \underset{p,R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle \right)_1 \preceq \left(\langle \psi_1, R_1 \rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2 \rangle \right)_1$[1],

при условии, что значения операций в левых частях этих двух равенств и включения определены.

Под операцией полусоединения (Semijoin) двух таблиц схем R_1 и R_2 понимаем бинарную параметрическую операцию $\square_{R_1,R_2}$, значением которой является таблица схемы R_1, которая содержит те строки первой таблицы, которые входят в (естественное) соединение таблиц-аргументов.

Следовательно, $\square_{R_1,R_2} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1)$, $\langle \psi_1, R_1 \rangle \square_{R_1,R_2} \langle \psi_2, R_2 \rangle = \langle \psi', R_1 \rangle$, где $\langle \psi_1, R_1 \rangle \in \Psi(R_1)$, $\langle \psi_2, R_2 \rangle \in \Psi(R_2)$. Основой мультимножества ψ' является множество строк $\Theta(\psi') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 (s_2 \in \Theta(\psi_2) \wedge s_1 \approx s_2)\}$. Количество дубликатов определяется так: $Occ(s, \psi') = Occ(s, \psi_1)$, где $s \in \Theta(\psi')$.

5. Операции внешнего соединения

При применении операций внутреннего соединения возможна потеря информации, поскольку строки одной таблицы, которые не соединяются со строками другой таблицы, не будут включены в результирующую таблицу. В тех случаях, когда необходимо учесть строки таблиц-аргументов, которые не попали в результат исходного внутреннего соединения, используют операции внешнего соединения.

Для обозначения отсутствующих значений атрибутов строк результирующей таблицы используем особый элемент универсального домена $NULL$. Обозначим как $s_{R,NULL}$ константную строку схемы R вида $s_{R,NULL} : R \to \{NULL\}$, присваивающую всем атрибутам своей схемы значения $NULL$.

---

[1]Запись $(\langle \psi, R \rangle)_1$ обозначает первую компоненту пары $\langle \psi, R \rangle$, то есть мультимножество ψ.

Используем логическую схему определения операций внешнего соединения из [5], следуя которой четыре операции внешнего соединения вводятся как операции, подчиненные одной операции внутреннего соединения.

Пусть $\varphi : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$ – некоторая частичная бинарная операция на множестве таблиц, причем выполняется включение $\left(\varphi\left(\langle \psi_1, R_1\rangle, \langle \psi_2, R_2\rangle\right)\right)_1 \preceq \left(\langle \psi_1, R_1\rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2\rangle\right)_1$ для всех $\left\langle \langle \psi_1, R_1\rangle, \langle \psi_2, R_2\rangle \right\rangle \in \operatorname{dom}\varphi$.

Отметим, что операции внутреннего соединения $\underset{R_1,R_2}{Cj}$, $\underset{R_1,R_2}{\otimes}$, $\underset{A_1,\ldots,A_n,R_1,R_2}{\otimes}$, $\underset{p,R_1,R_2}{\otimes}$ именно такие.

Зафиксируем таблицы $\langle \psi_1, R_1\rangle$, $\langle \psi_2, R_2\rangle$ из области определенности операции φ. Тогда таблица $\langle \psi_1, R_1\rangle$ предполагает следующее представление: $\langle \psi_1, R_1\rangle = \left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle \cup_{All}^{R_1} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle$, где $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle = \langle \psi', R_1\rangle$, основой мультимножества ψ' является множество строк

$$\Theta(\psi') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \exists s_2 \left(s_2 \in \Theta(\psi_2) \wedge s_1 \cup s_2 \in \Theta\left(\left(\varphi\left(\langle \psi_1, R_1\rangle, \langle \psi_2, R_2\rangle\right)\right)_1\right)\right)\},$$

а количество дубликатов $Occ(s_1, \psi') = Occ(s_1, \psi_1)$, $s_1 \in \Theta(\psi')$ и $\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle = \langle \psi'', R_1\rangle$, основой мультимножества ψ'' является множество строк

$$\Theta(\psi'') = \{s_1 \mid s_1 \in \Theta(\psi_1) \wedge \forall s_2 \left(s_2 \in \Theta(\psi_2) \Rightarrow s_1 \cup s_2 \notin \Theta\left(\left(\varphi\left(\langle \psi_1, R_1\rangle, \langle \psi_2, R_2\rangle\right)\right)_1\right)\right)\},$$

а количество дубликатов определяется так: $Occ(s_1, \psi'') = Occ(s_1, \psi_1)$, где $s_1 \in \Theta(\psi'')$.

Говоря содержательно, строки таблицы $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle$ используются при формировании результата (внутреннего) соединения, а строки таблицы $\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle$ – не используются. Аналогичное представление таблицы $\langle \psi_2, R_2\rangle$ получим, поменяв роли таблиц $\langle \psi_1, R_1\rangle$ и $\langle \psi_2, R_2\rangle$ в представлении таблицы $\langle \psi_1, R_1\rangle$.

Отметим, что если операция φ совпадает с операцией $\underset{R_1,R_2}{\otimes}$, то $\left\langle \psi_1 \underset{\varphi}{\cap} \psi_2, R_1 \right\rangle = \langle \psi_1, R_1\rangle \square_{R_1,R_2} \langle \psi_2, R_2\rangle$, т.е. таблица в левой части последнего

равенства получается в результате применения операции полусоединения к исходным таблицам.

Ниже для упрощения записи будем считать, что операции соединения имеют больший приоритет, чем операции объединения.

Определим четыре операции внешнего соединения (Outer Join), индуцированные одной операцией внутреннего соединения φ. Для этого рассмотрим следующие естественные соединения:

$$\left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1, R_2 \setminus R_1}{\otimes} \left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle = \left\langle \psi', R_1 \cup R_2 \right\rangle,$$

где основой мультимножества ψ' является множество строк $\Theta(\psi') = \{s_1 \cup s_{R_2 \setminus R_1, NULL} \mid s_1 \in \Theta(\psi_1 \underset{\varphi}{-} \psi_2)\}$, а количество дубликатов $Occ(s', \psi') = Occ(s_1, \psi_1 \underset{\varphi}{-} \psi_2)$, $s' \in \Theta(\psi')$ и $s' = s_1 \cup s_{R_2 \setminus R_1, NULL}$ (очевидно, что здесь и далее аналогичные представления строк единственны), а также

$$\left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2, R_1 \setminus R_2}{\otimes} \left\langle \{s^1_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \right\rangle = \left\langle \psi'', R_1 \cup R_2 \right\rangle,$$

где основой мультимножества ψ'' является множество строк $\Theta(\psi'') = \{s_{R_1 \setminus R_2, NULL} \cup s_2 \mid s_2 \in \Theta(\psi_2 \underset{\varphi}{-} \psi_1)\}$, а количество дубликатов $Occ(s'', \psi'') = Occ(s_2, \psi_2 \underset{\varphi}{-} \psi_1)$, $s'' \in \Theta(\psi'')$ и $s'' = s_{R_1 \setminus R_2, NULL} \cup s_2$.

Выше верхний индекс 1 в записи $\left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle$ указывает на то, что строка $s_{R_2 \setminus R_1, NULL}$ входит в исходную таблицу только один раз, т.е. $\left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle$ – константная таблица схемы $R_2 \setminus R_1$, первая компонента которой $\{1\}$-мультимножество с основой $\{s_{R_2 \setminus R_1, NULL}\}$. Для таблицы $\left\langle \{s^1_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \right\rangle$ полностью аналогично.

Под внешним левым соединением (Outer Left Join), индуцированным операцией φ, понимаем частичную бинарную операцию вида $\varphi_l : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, где $\mathrm{dom}\,\varphi_l = \mathrm{dom}\,\varphi$, $\varphi_l\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) =$

$$= \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \cup^{R_1 \cup R_2}_{All} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1, R_2 \setminus R_1}{\otimes} \left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle.$$

Под внешним правым соединением (Outer Right Join), индуцированным операцией φ, понимаем частичную бинарную операцию вида

$\varphi_r : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, где $\mathrm{dom}\,\varphi_r = \mathrm{dom}\,\varphi$, $\varphi_r\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) =$

$= \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \bigcup_{All}^{R_1 \cup R_2} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2, R_1 \setminus R_2}{\otimes} \left\langle \{s^1_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \right\rangle$.

Под внешним полным соединением (Outer Full Join), индуцированным операцией φ, понимаем частичную бинарную операцию вида $\varphi_f : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, где $\mathrm{dom}\,\varphi_f = \mathrm{dom}\,\varphi$,

$$\varphi_f\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) = \varphi\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) \bigcup_{All}^{R_1 \cup R_2} \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1, R_2 \setminus R_1}{\otimes}$$

$$\underset{R_1, R_2 \setminus R_1}{\otimes} \left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle \bigcup_{All}^{R_1 \cup R_2} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2, R_1 \setminus R_2}{\otimes}$$

$$\underset{R_2, R_1 \setminus R_2}{\otimes} \left\langle \{s^1_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \right\rangle.$$

Под внешним соединением объединением (Outer Union Join), индуцированным операцией φ, понимаем частичную бинарную операцию вида $\varphi_\cup : \Psi(R_1) \times \Psi(R_2) \xrightarrow{\sim} \Psi(R_1 \cup R_2)$, где $\mathrm{dom}\,\varphi_\cup = \mathrm{dom}\,\varphi$, $\varphi_\cup\left(\langle \psi_1, R_1 \rangle, \langle \psi_2, R_2 \rangle\right) =$

$$= \left\langle \psi_1 \underset{\varphi}{-} \psi_2, R_1 \right\rangle \underset{R_1, R_2 \setminus R_1}{\otimes} \left\langle \{s^1_{R_2 \setminus R_1, NULL}\}, R_2 \setminus R_1 \right\rangle \bigcup_{All}^{R_1 \cup R_2} \left\langle \psi_2 \underset{\varphi}{-} \psi_1, R_2 \right\rangle \underset{R_2, R_1 \setminus R_2}{\otimes}$$

$$\underset{R_2, R_1 \setminus R_2}{\otimes} \left\langle \{s^1_{R_1 \setminus R_2, NULL}\}, R_1 \setminus R_2 \right\rangle.$$

6. Агрегатные операции

Широко используемыми (параметрическими) 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 $\langle \psi, R \rangle \in \Psi(R)$, где ψ – конечное мультимножество и $A \in R$. Обозначим как α_A – мультимножество, которое содержит все элементы с учетом дубликатов столбца с атрибутом A таблицы $\langle \psi, R \rangle$. Тогда $\Theta(\alpha_A) = \{d \mid \exists s(s \in \Theta(\psi) \wedge \langle A, d \rangle \in s)\} = \{d \mid \{\langle A, d \rangle\} \in \Theta\left(\left(\pi_{\{A\}, R}\left(\langle \psi, R \rangle\right)\right)_1\right)\}$ – аналог

активного домена атрибута A относительно таблицы [5, 9]. Количество дубликатов элемента основы $d \in \Theta(\alpha_A)$ в мультимножестве α_A определяется как $\alpha_A(d) = Occ(\{\langle A, d\rangle\}, (\pi_{\{A\},R}(\langle \psi, R\rangle))_1) = \sum_{\substack{s\in\Theta(\psi),\\ s(A)=d}} Occ(s,\psi)$. Пусть $2_m^{\boldsymbol{D}'} = \{\alpha \mid \Theta(\alpha) \in 2^{\boldsymbol{D}'}\}$ – семейство всех мультимножеств, основы которых являются конечными подмножествами множества $\boldsymbol{D}'$; здесь $\boldsymbol{D}' \subseteq \boldsymbol{D}$ – подмножество универсального домена.

Пусть Num – числовое подмножество универсального домена $\boldsymbol{D}$, замкнутое относительно сложения. Множество Num расширим включением особого элемента $NULL$, но операцию сложения на случай, когда хотя бы один из аргументов является $NULL$, расширять не будем.

Зададим агрегатные операции Sum, Avg, Min, Max, $Count$. Общая схема: на конечном мультимножестве определяются функции суммирования, взятия наименьшего и наибольшего значений, определения 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.

Под операцией агрегирования $Sum_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$, понимаем унарную параметрическую операцию вида $Sum_{A,R} : \Psi(R) \to \Psi(\{A\})$, $Sum_{A,R}(\langle \psi, R\rangle) = \left\langle \left\{ \left\langle A, Sum(\alpha_A)\right\rangle \right\}^1, \{A\} \right\rangle$, где $\langle \psi, R\rangle \in \Psi(R)$, а Sum – функция, которая возвращает сумму значений столбца с атрибутом A таблицы $\langle \psi, R\rangle$ (эти значения могут повторяться), которые отличаются от значения $NULL$, кроме того предполагается, что этот столбец содержит только данные числового типа. Следовательно, $Sum : 2_m^{Num} \to Num$,

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

Как и раньше верхний индекс 1 в записи $\left\langle \left\{ \left\langle A, Sum(\alpha_A)\right\rangle \right\}^1, \{A\} \right\rangle$ указывает на то, что строка $\{\langle A, Sum(\alpha_A)\rangle\}$ входит в исходную таблицу только один раз, т.е. $\{\{\langle A, Sum(\alpha_A)\rangle\}^1\}$ – $\{1\}$-мультимножество.

Таким образом, имеем $Sum(\varnothing_m) = NULL$, $Sum(\{NULL^n\}) = NULL$, $Sum\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \sum_{i=1}^{k} d_i n_i$, в предположении, что все элементы d_i отличны от элемента $NULL$.

Для случая пустой таблицы $\langle \psi_{\varnothing}, R \rangle$ операция агрегирования $Sum_{A,R}$ применяется так: $Sum_{A,R}\left(\langle \psi_{\varnothing}, R \rangle\right) = \left\langle \left\{ \left\langle A, NULL \right\rangle^{1} \right\}, \{A\} \right\rangle$, здесь $\psi_{\varnothing} = \varnothing_m$.

Пусть $\leq$ – линейный порядок на универсальном домене $\boldsymbol{D}$. Под операцией агрегирования $Min_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$, понимаем унарную параметрическую операцию вида $Min_{A,R} : \Psi(R) \to \Psi(\{A\})$, $Min_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{ \left\langle A, Min(\alpha_A) \right\rangle^{1} \right\}, \{A\} \right\rangle$, где $\langle \psi, R \rangle \in \Psi(R)$, а Min – функция, которая возвращает наименьшее значение среди значений столбца с атрибутом A таблицы $\langle \psi, R \rangle$, которые отличаются от значения $NULL$, т.е. $Min : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$,

$$Min(\alpha_A) = \begin{cases} NULL, \textit{если } \Theta(\alpha_A) = \varnothing; \\ NULL, \textit{ если } \Theta(\alpha_A) = \{NULL\}; \\ \min\{d \mid d \in \Theta(\alpha_A) \setminus \{NULL\}\}, \textit{если } \Theta(\alpha_A) \setminus \{NULL\} \neq \varnothing. \end{cases}$$

Таким образом, имеем $Min(\varnothing_m) = NULL$, $Min(\{NULL^n\}) = NULL$, $Min\left(\left\{ d_1^{n_1}, \ldots, d_k^{n_k} \right\}\right) = \min\{d_1, \ldots, d_k\}$, в предположении, что все элементы d_i, $i = \overline{1,k}$, отличны от элемента $NULL$.

Для случая пустой таблицы $\langle \psi_{\varnothing}, R \rangle$ операция агрегирования $Min_{A,R}$ применяется так: $Min_{A,R}\left(\langle \psi_{\varnothing}, R \rangle\right) = \left\langle \left\{ \left\langle A, NULL \right\rangle^{1} \right\}, \{A\} \right\rangle$.

Под операцией агрегирования $Max_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$, понимаем унарную параметрическую операцию вида $Max_{A,R} : \Psi(R) \to \Psi(\{A\})$, $Max_{A,R}\left(\langle \psi, R \rangle\right) = \left\langle \left\{ \left\langle A, Max(\alpha_A) \right\rangle^{1} \right\}, \{A\} \right\rangle$, где $\langle \psi, R \rangle \in \Psi(R)$, а Max – функция, которая возвращает наибольшее значение среди значений столбца с атрибутом A таблицы $\langle \psi, R \rangle$, которые отличаются от $NULL$, т.е. $Max : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$,

$$Max(\alpha_A) = \begin{cases} NULL, \textit{если } \Theta(\alpha_A) = \varnothing; \\ NULL, \textit{ если } \Theta(\alpha_A) = \{NULL\}; \\ \max\{d \mid d \in \Theta(\alpha_A) \setminus \{NULL\}\}, \textit{если } \Theta(\alpha_A) \setminus \{NULL\} \neq \varnothing. \end{cases}$$

Таким образом, имеем $Max(\varnothing_m) = NULL$, $Max(\{NULL^n\}) = NULL$, $Max\left(\left\{ d_1^{n_1}, \ldots, d_k^{n_k} \right\}\right) = \max\{d_1, \ldots, d_k\}$, в предположении, что все элементы d_i, $i = \overline{1,k}$, отличны от элемента $NULL$.

Для случая пустой таблицы $\langle \psi_{\varnothing}, R \rangle$ операция агрегирования $Max_{A,R}$ применяется так: $Max_{A,R}(\langle \psi_{\varnothing}, R \rangle) = \left\langle \left\{ \left\langle A, NULL \right\rangle \right\}^{1}, \{A\} \right\rangle$.

Отметим, что функции *Min* и *Max* определяют наименьший или наибольший элементы основы мультимножества, которые отличаются от значения *NULL*, поэтому сравнимость особого элемента с остальными элементами универсального домена в данном случае несущественна. Это свойство важно при задании семантики фразы ORDER BYоператора запросов SELECT, отвечающей за упорядочение строк.

Под операцией агрегирования $Count_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$, понимаем унарную параметрическую операцию вида $Count_{A,R} : \Psi(R) \to \Psi(\{A\})$, $Count_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Count(\alpha_A) \right\rangle \right\}^{1}, \{A\} \right\rangle$, где $\langle \psi, R \rangle \in \Psi(R)$, а *Count* – функция, которая возвращает количество значений, которые отличаются от значения *NULL*, с учетом дубликатов в столбце с атрибутом A таблицы $\langle \psi, R \rangle$, т.е. $Count : 2^{\boldsymbol{D}}_{m} \to N$, $Count(\alpha_A) = \sum\limits_{d \in \Theta(\alpha_A) \setminus \{NULL\}} \alpha_A(d)$; полагается по определению (и это естественно), что сумма пустого множества слагаемых равна нулю.

Таким образом, имеем $Count(\varnothing_m) = 0$, $Count(\{NULL^n\}) = 0$, $Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = n_1 + \ldots + n_k$, в предположении, что все элементы d_i, $i = \overline{1,k}$, отличны от элемента *NULL*.

Для случая пустой таблицы $\langle \psi_{\varnothing}, R \rangle$ операция агрегирования $Count_{A,R}$ применяется так: $Count_{A,R}(\langle \psi_{\varnothing}, R \rangle) = \left\langle \left\{ \left\langle A, 0 \right\rangle \right\}^{1}, \{A\} \right\rangle$.

Допустим, что числовое подмножество *Num* универсального домена замкнуто относительно (частичной операции) деления $/ : Num \times Num \xrightarrow{\sim} Num$. 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* не может дать *NULL*.

Под операцией агрегирования $Avg_{A,R}$ по атрибуту A (конечных) таблиц схемы R, $A \in R$, понимаем унарную параметрическую операцию вида $Avg_{A,R} : \Psi(R) \to \Psi(\{A\})$, $Avg_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Avg(\alpha_A) \right\rangle \right\}^{1}, \{A\} \right\rangle$, где $\langle \psi, R \rangle \in \Psi(R)$, а *Avg* – функция, которая возвращает среднее арифметическое значение элементов столбца с атрибутом A таблицы $\langle \psi, R \rangle$, которые

отличаются от значения $NULL$, с учетом дубликатов, т.е. $Avg : 2_m^{Num} \to Num$, $Avg(\alpha_A) = Sum(\alpha_A) / Count(\alpha_A)$.

Таким образом, из определений следуют равенства $Avg(\varnothing_m) = Sum(\varnothing_m) / Count(\varnothing_m) = NULL / 0 = NULL$,

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

$Avg\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = Sum\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) / Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \frac{1}{(n_1 + \ldots + n_k)} \sum_{i=1}^{k} d_i n_i$, в предположении, что все элементы d_i, $i = \overline{1,k}$, отличны от элемента $NULL$.

Для случая пустой таблицы $\langle \psi_\varnothing, R \rangle$ операция агрегирования $Avg_{A,R}$ применяется так: $Avg_{A,R}\left(\langle \psi_\varnothing, R \rangle\right) = \left\langle \left\{\left\{\langle A, NULL \rangle\right\}\right\}, \{A\} \right\rangle$.

Под операцией агрегирования $Count_{A,R}(*)$ (конечных) таблиц схемы R понимаем унарную параметрическую операцию вида $Count_{A,R}(*) : \Psi(R) \to \Psi(\{A\})$, $Count_{A,R}(*)\left(\langle \psi, R \rangle\right) = \left\langle \left\{\left\{\langle A, \|\psi\| \rangle\right\}^1\right\}, \{A\} \right\rangle$, где $\langle \psi, R \rangle \in \Psi(R)$, а $\|\psi\|$ – это введенный ранее ранг мультимножества ψ.

Для случая пустой таблицы $\langle \psi_\varnothing, R \rangle$ операция агрегирования $Count_{A,R}(*)$ применяется так: $Count_{A,R}(*)\left(\langle \psi_\varnothing, R \rangle\right) = \left\langle \left\{\left\{\langle A, \|\varnothing_m\| \rangle\right\}^1\right\}, \{A\} \right\rangle = \left\langle \left\{\left\{\langle A, 0 \rangle\right\}^1\right\}, \{A\} \right\rangle$.

7. Выводы по результатам и направления дальнейших исследований

Существует ряд прикладных задач, особенностью которых является множественность и повторяемость данных. Например, социологические опросы различных групп населения, вычисления на ДНК, уточнения таблиц с дубликатами строк и другие. Математическим уточнением совокупностей с повторениями выступают мультимножества. Естественно возникает потребность в расширении возможностей реляционных баз данных за счет использования мультимножеств.

В данной работе рассматривается мультимножественная табличная алгебра, сигнатура которой пополнена новыми операциями: операциям внутренних и внешних соединений, операцией полусоединения, агрегатными операциями. Для определения внешних операций введен особый элемент универсального домена $NULL$.

Следует также отметить, что параметром агрегатной операции может выступать не только отдельный атрибут, но и некоторая функция над строкой. Соответствующее уточнение не трудно провести на основе представленных в работе построений.

В дальнейшем планируется исследовать свойства операций мультимножественной табличной алгебры.

ЛИТЕРАТУРА

1. Codd E.F. Relational Completeness of Data Base Sublanguages / Codd E.F. // Data Base Systems. – New York: Prentice-Hall. – 1972. – P. 65-93.
2. Lamperti G. On Multisets in Database Systems / G. Lamperti, M. Melchiori, M. Zanella // Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View, number 2235 in Lecture Notes in Computing Since. – Berlin: Springer-Verlag, 2001. – P. 147-215.
3. Гарсиа-Молина Г. Системы баз данных: [полный курс: пер. с англ.] / Г. Гарсиа-Молина, Дж. Ульман, Дж. Уидом. – Москва: "Вильямс", 2004. – 1088 с.
4. Silbeschatz A., Korth H., Sudarshan S. Database System Concepts. – McGraw-Hill, 2011. –1376 p.
5. Реляційні бази даних: табличні алгебри та SQL-подібні мови / В.Н. Редько, Ю.Й. Брона, Д.Б. Буй, С.А. Поляков. – Київ: Видавничий дім "Академперіодика", 2001. – 198 с.
6. Буй Д.Б. Сучасний стан теорії мультимножин / Д.Б. Буй, Ю.О. Богатирьова // Вісник Київського університету. Сер.: фіз.-мат. науки. – 2010. – Вип. 1. – С. 51-58.
7. Петровский А.Б. Пространства множеств и мультимножеств / А.Б. Петровский. – Москва: "Едиториал УРСС", 2003. – 248 с.
8. Глушко І.М. Мультимножинна таблична алгебра / І.М. Глушко // Proceedings of the International Scientific Conference of Student and Young Scientists "Theoretical and Applied Aspects of Cybernetics" (TAAC'2011, Kyiv, February 21–25, 2011). – P. 77-79.
9. Мейер Д. Теория реляционных баз данных: [пер. с англ.] / Д. Мейер. – Москва: Мир, 1987. – 608 с.

Надійшла у першій редакції 15.05.2014, в останній - 17.07.2014.