УДК 004.655

# ФОРМАЛЬНА СЕМАНТИКА АГРЕГАТНИХ ОПЕРАЦІЙ МУЛЬТИМНОЖИННОЇ ТАБЛИЧНОЇ АЛГЕБРИ

*І.М. Глушко*

Ніжинський державний університет імені Миколи Гоголя,
16600, Ніжин, вул. Кропив'янського 2,
glushkoim@gmail.com

Розглянуто мультимножинну табличну алгебру. Сигнатура мультимножинної табличної алгебри поповнена агрегатними операціями. Задано формальну математичну семантику цих операцій та наведено приклади їх застосування.

Multiset table algebra is considered. The signature of multiset table algebra is filled up with aggregate operations. A formal mathematical semantics of these operations is defined.

## Вступ

Реляційна модель даних на сьогодні широко використовується як у наукових дослідженнях в базах даних, так і на практиці. У формальному визначенні, запропонованому Е. Коддом [1], дана модель базується на множинах кортежів, тобто не дозволяє дублікати кортежів у відношенні. Багато мов, орієнтованих на роботу з базами даних, вимагають реляційну модель даних з мультимножинною семантикою (multi-set semantics) тому, що, по-перше, відношення, які дозволяють дублікати, корисні в багатьох прикладних областях, де об'єкти-дублікати можуть існувати; по-друге, в реляційній моделі даних видалення дублікатів після виконання операцій проекції та об'єднання передбачає злиття однакових елементів або здійснення інших трудомістких дій.

Питанню використання мультимножин в базах даних приділяли увагу Paul W.P.J. Grefen та Rolf A. de By [2], G. Lamperti, M. Melchiori, M. Zanella [3], Г. Гарсіа-Моліна, Дж. Ульман, Дж. Уидом [4], A. Silbeschatz, H. Korth, S. Sudarshan [5], Д.Б. Буй, С.А. Поляков, Ю.Й. Брона, В.Н. Редько [6]. Разом з тим, в жодній із зазначених робіт не приділяється увага агрегатним операціям над таблицями мультимножинної табличної алгебри.

## Мультимножини: основні поняття

Наведемо основні поняття мультимножин в термінах роботи [6]. Зафіксуємо деяку множину $U$. Під мультимножиною $\alpha$ з основою $U$ будемо розуміти відображення вигляду $\alpha : U \to N$, де $N = \{1,2,...\}$ – множина натуральних чисел.

Нехай $D$ – універсум елементів основ мультимножин, тоді булеан $P(D)$ – універсум основ мультимножин. Під характеристичною функцією мультимножини $\alpha$ розуміємо функцію вигляду $\chi_\alpha : D \to Z_+$, значення якої задається наступною кусковою схемою:

$$\chi_\alpha(d) = \begin{cases} \alpha(d), \text{ якщо } d \in \operatorname{dom}\alpha, \\ 0, \text{ інакше;} \end{cases}$$

для всіх $d \in D$, де $Z_+$ – множина цілих невід'ємних чисел..

Мультимножина називається порожньою і позначається як $\varnothing_m$, якщо її основа – порожня множина.

Мультимножини, областю значень яких є порожня множина або одноелементна множина вигляду $\{1\}$, називаються 1-мультимножинами. Дані мультимножини є аналогами звичайних множин.

Домовимося мультимножину $\alpha$ з основою $\{d_1,...,d_k\}$ записувати як $\{d_1^{n_1},...,d_k^{n_k}\}$, де $n_i$ – кількість дублікатів (екземплярів) елемента $d_i$ у мультимножині $\alpha$, $i = 1,...,k$.

Під рангом скінченної мультимножини $\alpha$ розуміємо суму дублікатів елементів її основи $\|\alpha\| = \sum_{d \in U} \chi_\alpha(d)$, де $U$ – основа мультимножини $\alpha$. Зрозуміло, що $\|\varnothing_m\| = 0$.

Мультимножина $\beta$ включається у мультимножину $\alpha$ ($\beta \preceq \alpha$), якщо:

$$\beta \preceq \alpha \Leftrightarrow U_\beta \subseteq U_\alpha \;\&\; \forall d\big(d \in U_\beta \Rightarrow \beta(d) \le \alpha(d)\big).$$

Якщо $\beta \preceq \alpha$, то мультимножина $\beta$ називається підмультимножиною мультимножини $\alpha$, а мультимножина $\alpha$ – надмультимножиною мультимножини $\beta$. Очевидно, що порожня мультимножина є підмультимножиною будь-якої мультимножини ($\varnothing_m \preceq \alpha,\ \forall \alpha$), а будь-яка мультимножина є своєю ж підмультимножиною ($\alpha \preceq \alpha,\ \forall \alpha$).

## Мультимножинна таблична алгебра

Розглянемо дві множини: $\boldsymbol{A}$ – множину атрибутів і $\boldsymbol{D}$ – універсальний домен. Довільну скінченну множину атрибутів $R \subseteq \boldsymbol{A}$ назвемо схемою.

Рядком схеми $R$ називається іменна множина на парі $R$, $\boldsymbol{D}$, проекція якої за першою компонентою рівна $R$ (тобто по суті розглядається функція вигляду $s : R \to \boldsymbol{D}$). Множину всіх рядків схеми $R$ позначимо $S(R)$, а множину всіх рядків – $S$. Для позначення відсутніх значень у таблиці використовується особливий елемент універсального домену $NULL$. Позначимо $s_R^{NULL}$ – константний рядок схеми $R$, тобто $s_R^{NULL} : R \to \{NULL\}$.

Задамо поняття таблиці як пари $\langle \psi, R \rangle$, де перша компонента $\psi$ – це довільна мультимножина, зокрема, нескінченна, а друга компонента $R$ – схема таблиці.

Таким чином, кожній таблиці приписується певна схема. Множину всіх таблиць схеми $R$ позначимо $\Psi(R)$, а множину всіх таблиць $\Psi = \bigcup\limits_{R \subseteq \mathrm{A}} \Psi(R)$.

Позначимо $Occ(s, \psi)$ – кількість дублікатів (екземплярів) рядка $s$ у мультимножині $\psi$. Домовимося мультимножину $\psi$ записувати як $\{s_1^{n_1}, ..., s_k^{n_k}, \ldots\}$, де $n_i = Occ(s_i, \psi)$, $i = 1,2,\ldots$, а $\Theta(\psi) = \{s_1, ..., s_k, \ldots\}$ – основа мультимножин $\psi$.

Під мультимножинною табличною алгеброю розуміємо алгебру $\langle \Psi, \Omega_{P,\Xi} \rangle$, де $\Psi$ – множина всіх таблиць, $\Omega_{P,\Xi} = \left\{ \bigcup_{All}^{\Psi,R}, \bigcap_{All}^{\Psi,R}, \setminus_{All}^{\Psi,R}, \sigma_{p,R}, \pi_{X,R}, \underset{R_1,R_2}{\otimes}, Rt_{\xi,R}, \sim_R \right\}_{X,R,R_1,R_2 \subseteq \mathrm{A}}^{p \in P, \xi \in \Xi}$ – сигнатура, $P$, $\Xi$ – множини параметрів.

Задамо операції. Під об'єднанням $\bigcup_{All}^R$ (перетином $\bigcap_{All}^R$, різницею $\setminus_{All}^R$) таблиць схеми $R$ розуміється бінарна (параметрична) операція, отримана обмеженням однойменних операцій $\bigcup_{All}$, $\bigcap_{All}$, $\setminus_{All}$ над мультимножинами на множину всіх таблиць схеми $R$.

Розглянемо кожну операцію окремо. Позначимо через $\Theta(\psi_1)$ і $\Theta(\psi_2)$ – основи мультимножин $\psi_1$ та $\psi_2$ відповідно. Таким чином,

$$\bigcup_{All}^R : \Psi(R) \times \Psi(R) \to \Psi(R), \quad \langle \psi_1, R \rangle \bigcup_{All}^R \langle \psi_2, R \rangle = \langle \psi_1 \bigcup_{All} \psi_2, R \rangle.$$

Основа мультимножини $\psi_1 \bigcup_{All} \psi_2$ дорівнює об'єднанню основ мультимножин таблиць-аргументів:

$$\Theta(\psi_1 \bigcup_{All} \psi_2) = \Theta(\psi_1) \bigcup \Theta(\psi_2).$$

Дублікати рядків, які з'явилися після виконання операції, не вилучаються. Кількість дублікатів кожного рядка визначається за формулою:

$$Occ(s, \psi_1 \bigcup_{All} \psi_2) = \begin{cases} Occ(s, \psi_1), \text{ якщо } s \in \Theta(\psi_1) \setminus \Theta(\psi_2), \\ Occ(s, \psi_2), \text{ якщо } s \in \Theta(\psi_2) \setminus \Theta(\psi_1), \\ Occ(s, \psi_1) + Occ(s, \psi_2), \text{ якщо } s \in \Theta(\psi_1) \bigcap \Theta(\psi_2); \end{cases}$$

де $s \in \Theta(\psi_1) \bigcup \Theta(\psi_2)$.

$\bigcap_{All}^R : \Psi(R) \times \Psi(R) \to \Psi(R)$, причому $\langle \psi_1, R \rangle \bigcap_{All}^R \langle \psi_2, R \rangle = \langle \psi_1 \bigcap_{All} \psi_2, R \rangle$.

Основа мультимножини $\psi_1 \bigcap_{All} \psi_2$ дорівнює перетину основ мультимножин таблиць-аргментів:

$$\Theta(\psi_1 \bigcap_{All} \psi_2) = \Theta(\psi_1) \bigcap \Theta(\psi_2),$$

а кількість дублікатів рядка визначається як

$$Occ(s, \psi_1 \bigcap_{All} \psi_2) = \min\big(Occ(s, \psi_1), Occ(s, \psi_2)\big), \text{ де } s \in \Theta(\psi_1) \bigcap \Theta(\psi_2).$$

$\setminus_{All}^R : \Psi(R) \times \Psi(R) \to \Psi(R)$, причому $\langle \psi_1, R \rangle \setminus_{All}^R \langle \psi_2, R \rangle = \langle \psi_1 \setminus_{All} \psi_2, R \rangle$, де $\langle \psi_1, R \rangle$, $\langle \psi_2, R \rangle \in \Psi(R)$.

Основа мультимножини $\psi_1 \setminus_{All} \psi_2$ визначається як $\Theta(\psi_1 \setminus_{All} \psi_2) = \Theta(\psi_1) \setminus \Theta(\psi_2) \bigcup C^{>}(\psi_1, \psi_2)$, де $C^{>}(\psi_1, \psi_2) = \{ s \mid s \in \Theta(\psi_1) \bigcap \Theta(\psi_2) \wedge Occ(s, \psi_1) > Occ(s, \psi_2) \}$.

Кількість дублікатів знаходиться так:

$$Occ(s, \psi_1 \setminus_{All} \psi_2) = \begin{cases} Occ(s, \psi_1), \text{ якщо } s \in \Theta(\psi_1) \setminus \Theta(\psi_2), \\ Occ(s, \psi_1) - Occ(s, \psi_2), \text{ якщо } s \in C^{>}(\psi_1, \psi_2), \end{cases}$$

де $s \in (\Theta(\psi_1) \setminus \Theta(\psi_2)) \bigcup C^{>}(\psi_1, \psi_2)$.

Нехай $p : S \tilde{\to} \{true, false\}$ – частковий предикат на множині рядків. Під селекцією за предикатом $p$ таблиць схеми $R$ розуміється унарна часткова параметрична операція $\sigma_{p,R}$, яка таблиці зіставляє її підтаблицю, що містить рядки, на яких предикат $p$ істинний (крім того, предикат $p$ має бути визначеним на всіх рядках таблиці-аргументу).

Отже,

$$\sigma_{p,R} : \Psi(R) \tilde{\to} \Psi(R),\ dom\,\sigma_{p,R} = \{\langle \psi, R\rangle \mid \Theta(\psi) \subseteq dom\ p\},\ \sigma_{p,R}(\langle \psi, R\rangle) = \langle \psi', R\rangle,$$

де $\langle \psi, R\rangle \in \Psi(R)$.

Областю означеності операції селекції є таблиці $\langle \psi, R\rangle \in \Psi(R)$, які містять рядки $s \in S(R)$, на яких предикат-параметр селекції $p$ визначений. Надалі розглядаємо тільки предикати-параметри вигляду $p(s) = true \Leftrightarrow \mathrm{p}(s(A_1),...,s(A_m)) = true$, де $\boldsymbol{p}$ – предикат на універсальному домені, а атрибути $A_1,...,A_m$ належать схемі рядка $s$.

Основа мультимножини результуючої таблиці визначається як: $\Theta(\psi') = \{s \mid s \in \Theta(\psi) \wedge p(s) \simeq true\}$, де $\simeq$ – узагальнена рівність (тобто обидві частини або одночасно не визначені або одночасно визначені і рівні).

В залежності від значення предиката на рядку $s$ усі дублікати цього рядка або входять до мультимножини отриманої таблиці, або ні: $Occ(s, \psi') = Occ(s, \psi)$, де $s \in \Theta(\psi')$. Зазначимо, що операція селекції не породжує нові дублікати рядка.

Нехай $X \subseteq \boldsymbol{A}$ – скінченна множина атрибутів. Під проекцією за множиною атрибутів $X$ таблиць схеми $R$ розуміється унарна параметрична операція $\pi_{X,R}$, значеннями якої є таблиці схеми $R \bigcap X$, що складаються з обмежень за $X$ рядків вихідних таблиць.

Отже, $\pi_{X,R} : \Psi(R) \to \Psi(R \bigcap X)$, $\pi_{X,R}(\langle \psi, R\rangle) = \langle \psi', R \bigcap X\rangle$, де $\langle \psi, R\rangle \in \Psi(R)$.

Основа мультимножини $\psi'$ визначається як $\Theta(\psi') = \{s \mid X \mid s \in \Theta(\psi)\}$.

Дублікати рядків, які з'явилися після виконання операції, не вилучаються. Кількість дублікатів кожного рядка визначається за формулою:

$$Occ(s', \psi') = \sum_{\substack{s \in \Theta(\psi), \\ s|X = s'}} Occ(s, \psi),\ \text{де}\ s' \in \Theta(\psi').$$

Під з'єднанням таблиць схем $R_1$, $R_2$ розуміється бінарна параметрична операція $\underset{R_1,R_2}{\otimes}$, значеннями якої є таблиці схеми $R_1 \bigcup R_2$, що складаються з усяких об'єднань сумісних рядків вихідних таблиць.

Отже,

$$\underset{R_1,R_2}{\otimes} : \Psi(R_1) \times \Psi(R_2) \to \Psi(R_1 \bigcup R_2),\ \langle \psi_1, R_1\rangle \underset{R_1,R_2}{\otimes} \langle \psi_2, R_2\rangle = \langle \psi', R_1 \bigcup R_2\rangle,$$

де $\langle \psi_1, R_1\rangle \in \Psi(R_1)$, $\langle \psi_2, R_2\rangle \in \Psi(R_2)$. Змістовно кажучи, кожний рядок з $\psi_1$ з'єднується з кожним рядком із $\psi_2$, незалежно від того – дублікат це чи ні.

Основою мультимножини $\psi'$ є множина рядків

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

Кількість дублікатів знаходиться так: $Occ(s', \psi') = Occ(s' \mid R_1, \psi_1) \cdot Occ(s' \mid R_2, \psi_2)$, де $s' \in \Theta(\psi')$.

Введемо операцію перейменування. Під перейменуванням таблиць схеми $R$ розуміється унарна параметрична операція $R\psi_{\xi,R}$, де $\xi : \boldsymbol{A} \tilde{\to} \boldsymbol{A}$ – ін'єктивне відображення на множині атрибутів, що здійснює перейменування атрибутів вихідних таблиць згідно з відображенням-параметром $\xi$. Перейменувати таблицю означає перейменувати атрибути її схеми, тобто перейменувати рядки таблиці. Перейменування рядків будемо здійснювати відповідно до [6].

Нехай $\eta : \boldsymbol{A} \to \boldsymbol{A}$ – функція перейменування атрибутів. Під перейменуванням рядків, що відповідає функції перейменування атрибутів $\eta$, розуміється відображення $Rs_\eta : S \to S'$, $Rs_\eta(s) = \{\langle \eta(A), s(A)\rangle \mid A \in \pi_1^2 s\}$, причому $\eta = \xi \bigcup \mathrm{id}_{\mathrm{A} \setminus dom\xi}$.

Схема $R$ називається $\xi$-допустимою, якщо $\xi[R] \bigcap (R \setminus dom\,\xi) = \varnothing$. Тут $\xi[R]$ – повний образ множини $R$ відносно функції $\xi$. Множину таблиць схеми $R$, де схема $R$ $\xi$-допустима позначимо $\Psi_\xi(R)$.

Під перейменуванням таблиць схеми $R$, що відповідає ін'єктивній частковій функції перейменування атрибутів $\xi : \boldsymbol{A} \xrightarrow{\sim} \boldsymbol{A}$, розуміється унарна параметрична операція $R\psi_{\xi,R}$ з областю означеності $\Psi_{\xi}(R)$, значення якої задаються таким чином: $R\psi_{\xi,R}\left(\langle\psi, R\rangle\right) = \left\langle Rs_{\eta}[\psi], \eta[R]\right\rangle$, де $\psi \in \Psi_{\xi}(R)$, $\eta = \xi \bigcup \mathrm{id}_{\mathrm{A}\setminus dom\,\xi}$, $Rs_{\eta}[\psi]$ – повний образ мультимножини $\psi$ з основою $\Theta(\psi)$ відносно функції $Rs_{\eta}$.

Основою мультимножини $Rs_{\eta}[\psi]$ є повний образ множини $\Theta(\psi)$ відносно функції $Rs_{\eta,R}$. А кількість дублікатів рядка у результуючій таблиці задається рівністю

$$Occ\left(s', Rs_{\eta}[\psi]\right) = Occ(s, \psi),$$

де $s \in Rs_{\eta}^{-1}(s')$, $s' \in \Theta(\psi')$.

Введемо операцію активного доповнення. Для цього введемо декілька допоміжних понять.

Активним доменом атрибута $A \in R$ відносно таблиці $\langle\psi, R\rangle$ називається таблиця

$$D_{A,\psi} = \pi_{\{A\},R}\left(\langle\psi, R\rangle\right).$$

Насиченням таблиці $\langle\psi, R\rangle$ є таблиця

$$C\left(\langle\psi, R\rangle\right) = \pi_{\{A_1\},R}\left(\langle\psi, R\rangle\right) \underset{\{A_1\},\{A_2\}}{\otimes} \ldots \underset{\{A_1,\ldots,A_{n-1}\},\{A_n\}}{\otimes} \pi_{\{A_n\},R}\left(\langle\psi, R\rangle\right).$$

Під активним доповненням таблиці схеми $R$ розуміється унарна параметрична операція $\sim_R$, яка таблиці зіставляє доповнення в її насиченні.

Отже, $\sim_R$:

$$\Psi(R) \to \Psi(R),\ \sim_R \left(\langle\psi, R\rangle\right) = C\left(\langle\psi, R\rangle\right) \setminus^{R}_{All} \langle\psi, R\rangle,\ \text{де } \langle\psi, R\rangle \in \Psi(R).$$

**Твердження.** Мають місце наступні твердження:

1) будь-який вираз мультимножинної табличної алгебри можна замінити еквівалентним йому виразом, який використовує лише операції селекції, з'єднання, проекції, об'єднання, різниці та перейменування;

2) будь-який вираз мультимножинної табличної алгебри, який містить лише скінченні таблиці можна замінити еквівалентним йому виразом, який використовує сталі таблиці з єдиним атрибутом і єдиним рядком, операції селекції, з'єднання, проекції, об'єднання, різниці й перейменування.

## Агрегатні операції

Широко використовуваними агрегатними операціями є $Sum$, $Avg$, $Min$, $Max$, $Count$. Так, операція $Sum$ розраховує суму значень у відповідному стовпці заданої таблиці, при цьому значення $NULL$ ігноруються. Операція $Avg$ визначає середнє арифметичне значень у відповідному стовпці заданої таблиці, при цьому значення $NULL$ ігноруються. Операції $Min$ та $Max$ знаходять найменше та найбільше значення у відповідному стовпці заданої таблиці, при цьому значення $NULL$ так само ігноруються. Операція $Count$ визначає кількість значень, відмінних від $NULL$, у відповідному стовпці заданої таблиці. Операція $Count(*)$ визначає кількість рядків у заданій таблиці.

Нехай $\langle\psi, R\rangle \in \Psi(R)$, причому $\psi$ – скінченна мультимножина і $A$ – деякий атрибут схеми $R$, $A \in R$. Позначимо через $\alpha_A$ – мультимножину, яка містить всі елементи стовпця з атрибутом $A$ таблиці $\langle\psi, R\rangle$. Тоді $\alpha_A = D_{A,\psi}$, де $D_{A,\psi} = \pi_{A,R}\left(\langle\psi, R\rangle\right)$ – активний домен атрибута $A$ відносно таблиці $\langle\psi, R\rangle$. Нехай $2^{\mathrm{D}'}_m = \{\alpha \mid \Theta(\alpha) \in 2^{\mathrm{D}'}\}$ – сім'я всіх мультимножин, основи яких є скінченними підмножинами множини $\boldsymbol{D}'$; тут $\boldsymbol{D}' \subseteq \boldsymbol{D}$ – підмножина універсального домену.

Нехай $Num$ – числова підмножина універсального домену $\boldsymbol{D}$, замкнена відносно додавання. Множина $Num$ розширена включенням особливого елемента $NULL$, але при цьому операція додавання на випадок, коли хоча б один з аргументів є $NULL$ не розширюється.

Задамо агрегатні операції $Sum$, $Avg$, $Min$, $Max$, $Count$. Їхніми аргументами є скінченні таблиці, а значеннями – одноатрибутні таблиці з одним рядком. Загальна схема: спочатку на скінченній мультимножині визначаються функції сумування, взяття найменшого та найбільшого значення, визначення середнього арифметичного і кількості елементів, а потім ці функції переносяться на таблиці.

Під операцією агрегування $Sum_{A,R}$ за атрибутом $A$ (скінченних) таблиць схеми $R$, $A \in R$, розуміється унарна параметрична операція вигляду

$$Sum_{A,R} : \Psi(R) \to \Psi(\{A\}),\ Sum_{A,R}\left(\langle \psi, R\rangle\right) = \left\langle \left\{\left\langle A, Sum(\alpha_A)\right\rangle\right\}^{1}, \{A\}\right\rangle,$$

де $\langle \psi, R\rangle \in \Psi(R)$, а $Sum$ – функція, що повертає суму значень стовпця з атрибутом $A$ таблиці $\langle \psi, R\rangle$ (ці значення можуть повторюватися), які відрізняються від значення $NULL$, крім того, цей стовпець містить лише дані числового типу. Отже,

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

Верхній індекс 1 вказує на те, що рядок $\{\langle A, Sum(\alpha_A)\rangle\}$ входить у результуючу таблицю лише один раз, тобто кількість дублікатів даного рядка дорівнює одним.

Таким чином,

$$Sum(\varnothing_m) = NULL,\ Sum(\{NULL^n\}) = NULL,\ Sum\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \sum_{i=1}^{k} d_i n_i,$$

в припущенні, що всі елементи $d_i$ відрізняються від елемента $NULL$.

Для випадку порожньої таблиці $\langle \psi_\varnothing, R\rangle$ операція агрегування $Sum_{A,R}$ застосовується так:

$$Sum_{A,R}\left(\langle \psi_\varnothing, R\rangle\right) = \left\langle \left\{\left\langle A, NULL\right\rangle\right\}^{1}, \{A\}\right\rangle.$$

*Приклад 1.* Серед студентів фізико-математичного факультету було проведено анонімне опитування на визначення рівня самооцінки. Проаналізовано відповіді 6 учасників вибраних довільним чином з кожного факультету. Дані відображені в таблиці $\langle \Phi M, R\rangle$, де $R = \{A, B, C, D\}$ (див. таблицю). Атрибути схеми $R$ – це запитання, дані в таблицях – це відповіді студентів.

Запитання (атрибути):

A – чи турбуєтеся Ви за свій психічний стан?

B – чи турбуєтеся Ви про своє майбутнє?

C – чи боїтеся Ви виступати з промовою перед незнайомими людьми?

D – чи часто Ви робите помилки?

Варіанти відповіді:

1 – часто;

2 – іноді;

3 – рідко;

4 – ніколи.

Таблиця

| ***A*** | ***B*** | ***C*** | ***D*** |
|---|---|---|---|
| 4 | 2 | 4 | 3 |
| 4 | 2 | 4 | 3 |
| 4 | 2 | 4 | 3 |
| 4 | 3 | 3 | 2 |
| 4 | 3 | 3 | 2 |
| 3 | 3 | 3 | 1 |

Позначимо рядки:

$$s_1 = \{<A,4>, <B,2>, <C,4>, <D,3>\},\ s_2 = \{<A,4>, <B,3>, <C,3>, <D,2>\},$$

$$s_3 = \{<A,3>, <B,3>, <C,3>, <D,1>\},$$

тоді $\Phi M = \{s_1^3, s_2^2, s_3^1\}$.

Тоді в результаті застосування операції агрегування $Sum_{A,R}$ до таблиці $\langle \Phi M, R \rangle$ за атрибутами $A, B, C, D$ отримаємо таблиці:

$$Sum_{A,R} = \langle \{\{A, 23\}\}, \{A\} \rangle,\ Sum_{B,R} = \langle \{\{B, 15\}\}, \{B\} \rangle,$$

$$Sum_{C,R} = \langle \{\{C, 21\}\}, \{C\} \rangle,\ Sum_{D,R} = \langle \{\{D, 14\}\}, \{D\} \rangle.$$

Нехай $\leq$ – лінійний порядок на універсальному домені $\boldsymbol{D}$. Під операцією агрегування $Min_{A,R}$ за атрибутом $A$ (скінченних) таблиць схеми $R$, $A \in R$, розуміється унарна параметрична операція вигляду:

$$Min_{A,R} : \Psi(R) \to \Psi(\{A\}),\ \text{де}\ Min_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \langle A, Min(\alpha_A) \rangle \right\}^{1} \right\}, \{A\} \right\rangle,$$

причому $\langle \psi, R \rangle \in \Psi(R)$, а $Min$ – функція, що повертає найменше значення серед значень стовпця з атрибутом $A$ таблиці $\langle \psi, R \rangle$, відмінних від значення $NULL$, тобто $Min : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$,

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

Таким чином, $Min(\varnothing_m) = NULL$, $Min(\{NULL^n\}) = NULL$, $Min\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \min\{d_1, \ldots, d_k\}$, в припущенні, що всі елементи $d_i$, $i = \overline{1,k}$, відрізняються від елемента $NULL$.

Для випадку порожньої таблиці $\langle \psi_\varnothing, R \rangle$ операція агрегування $Min_{A,R}$ застосовується так:

$$Min_{A,R}(\langle t_\varnothing, R \rangle) = \left\langle \left\{ \left\{ \langle A, NULL \rangle \right\}^{1} \right\}, \{A\} \right\rangle.$$

*Приклад 2.* Розглянемо таблицю із приклада 1. Застосуємо операцію агрегування $Min_{A,R}$ до таблиці $\langle \Phi M, R \rangle$ за атрибутами $A, B, C, D$ отримаємо таблиці:

$$Min_{A,R} = \langle \{\{A, 3\}\}, \{A\} \rangle,\ Min_{B,R} = \langle \{\{B, 2\}\}, \{B\} \rangle,$$

$$Min_{C,R} = \langle \{\{C, 3\}\}, \{C\} \rangle, Min_{D,R} = \langle \{\{D, 1\}\}, \{D\} \rangle.$$

Під операцією агрегування $Max_{A,R}$ за атрибутом $A$ (скінченних) таблиць схеми $R$, $A \in R$, розуміється унарна параметрична операція вигляду:

$$Max_{A,R} : \Psi(R) \to \Psi(\{A\}),\ Max_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\{ \langle A, Max(\alpha_A) \rangle \right\}^{1} \right\}, \{A\} \right\rangle,$$

де $\langle \psi, R \rangle \in \Psi(R)$, а $Max$ – функція, що повертає найбільше значення серед значень стовпця з атрибутом $A$ таблиці $\langle \psi, R \rangle$, відмінних від $NULL$, тобто $Max : 2_m^{\boldsymbol{D}} \to \boldsymbol{D}$,

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

Таким чином, $Max(\varnothing_m) = NULL$, $Max(\{NULL^n\}) = NULL$, $Max\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = \max\{d_1, \ldots, d_k\}$, в припущенні, що всі елементи $d_i$, $i = \overline{1,k}$, відрізняються від значення $NULL$.

Для випадку порожньої таблиці $\langle \psi_\varnothing, R \rangle$ операція агрегування $Max_{A,R}$ застосовується так:

$$Max_{A,R}(\langle \psi_\varnothing, R \rangle) = \left\langle \left\{ \left\{ \langle A, NULL \rangle \right\}^{1} \right\}, \{A\} \right\rangle.$$

*Приклад 3.* Застосуємо операцію агрегування $Max_{A,R}$ до таблиці $\langle \Phi M, R \rangle$ за атрибутами $A, B, C, D$ отримаємо таблиці:

$$Max_{A,R} = \langle \{\{A, 4\}\}, \{A\} \rangle,\ Max_{B,R} = \langle \{\{B, 3\}\}, \{B\} \rangle,$$

$$Max_{C,R} = \langle \{\{C, 4\}\}, \{C\} \rangle,\ Max_{D,R} = \langle \{\{D, 3\}\}, \{D\} \rangle.$$

Зазначимо, що функції *Min* та *Max* визначають найменший або найбільший елемент основи мультимножини $\alpha_A$ серед елементів основи, відмінних від *NULL*, тому порівнянність особливого елемента *NULL* з рештою елементів універсального домену в даному випадку неістотна.

Під операцією агрегування $Count_{A,R}$ за атрибутом $A$ (скінченних) таблиць схеми $R$, $A \in R$, розуміється унарна параметрична операція вигляду:

$$Count_{A,R} : \Psi(R) \to \Psi(\{A\}),\ Count_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Count(\alpha_A) \right\rangle \right\}^1, \{A\} \right\rangle,$$

де $\langle \psi, R \rangle \in \Psi(R)$, а *Count* – функція, що повертає кількість значень, відмінних від *NULL*, з урахуванням дублікатів, у стовпці з атрибутом $A$ таблиці $\langle \psi, R \rangle$, тобто:

$$Count : 2_m^{\mathrm{D}} \to Z_+,\ \ Count(\alpha_A) = \sum_{d \in \Theta(\alpha_A) \setminus \{NULL\}} \alpha_A(d);$$

покладається за означенням, що сума порожньої множини доданків дорівнює нулю.

Таким чином, $Count(\varnothing_m) = 0$, $Count(\{NULL^n\}) = 0$, $Count\left(\left\{d_1^{n_1}, \ldots, d_k^{n_k}\right\}\right) = n_1 + \ldots + n_k$, в припущенні, що всі елементи $d_i$, $i = \overline{1,k}$, відрізняються від елемента *NULL*.

Для випадку порожньої таблиці $\langle \psi_\varnothing, R \rangle$ операція агрегування $Count_{A,R}$ застосовується так:

$$Count_{A,R}(\langle \psi_\varnothing, R \rangle) = \left\langle \left\{ \left\langle A, 0 \right\rangle \right\}^1, \{A\} \right\rangle.$$

*Приклад 4.* Застосуємо операцію агрегування $Count_{A,R}$ до таблиці $\langle ФМ, R \rangle$ за атрибутами $A, B, C, D$ отримаємо таблиці:

$$Count_{A,R} = \langle \{\{A, 6\}\}, \{A\} \rangle,\ Count_{B,R} = \langle \{\{B, 6\}\}, \{B\} \rangle,$$

$$Count_{C,R} = \langle \{\{C, 6\}\}, \{C\} \rangle,\ Count_{D,R} = \langle \{\{D, 6\}\}, \{D\} \rangle.$$

Припустимо, що числова підмножина *Num* універсального домену замкнена відносно (часткової операції) ділення $/ : Num \times Num \xrightarrow{\sim} Num$. Довизначимо операцію ділення так, що коли перший агрумент дорівнює *NULL*, то функція приймає значення *NULL*.

Під операцією агрегування $Avg_{A,R}$ за атрибутом $A$ (скінченних) таблиць схеми $R$, $A \in R$, розуміється унарна параметрична операція вигляду:

$$Avg_{A,R} : \Psi(R) \to \Psi(\{A\}),\ Avg_{A,R}(\langle \psi, R \rangle) = \left\langle \left\{ \left\langle A, Avg(\alpha_A) \right\rangle \right\}^1, \{A\} \right\rangle,$$

де $\langle \psi, R \rangle \in \Psi(R)$, а *Avg* – функція, що повертає середнє арифметичне значення елементів стовпця з атрибутом $A$ таблиці $\langle \psi, R \rangle$, які відрізняються від значення *NULL*, з урахуванням дублікатів, тобто:

$$Avg : 2_m^{Num} \to Num,\ Avg(\alpha_A) = Sum(\alpha_A) / Count(\alpha_A).$$

Таким чином, з означення випливають рівності

$$Avg(\varnothing_m) = Sum(\varnothing_m) / Count(\varnothing_m) = NULL / 0 = NULL,$$

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

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

в припущенні, що всі елементи $d_i$ відрізняються від елемента *NULL*.

Для випадку порожньої таблиці $\langle \psi_\varnothing, R \rangle$ операція агрегування $Avg_{A,R}$ застосовується так:

$$Avg_{A,R}\left(\left\langle \psi_{\varnothing}, R\right\rangle\right)=\left\langle \left\{\left\{\left\langle A, NULL\right\rangle\right\}^{1}\right\}, \{A\}\right\rangle.$$

*Приклад 5.* Застосуємо операцію агрегування $Avg_{A,R}$ до таблиці $\langle \Phi M, R\rangle$ за атрибутами $A,B,C,D$ отримаємо таблиці

$$Avg_{A,R}=\left\langle \{\{A, \frac{23}{6}\}\}, \{A\}\right\rangle,\ Avg_{B,R}=\left\langle \{\{B, \frac{15}{6}\}\}, \{B\}\right\rangle,$$

$$Avg_{C,R}=\left\langle \{\{C, \frac{21}{6}\}\}, \{C\}\right\rangle,\ Avg_{D,R}=\left\langle \{\{D, \frac{14}{6}\}\}, \{D\}\right\rangle.$$

Під операцією агрегування $Count_{A,R}(*)$ (скінченних) таблиць схеми $R$ розуміється унарна параметрична операція вигляду:

$$Count_{A,R}(*): \Psi(R)\to\Psi(\{A\}),\ Count_{A,R}(*)\left(\left\langle \psi, R\right\rangle\right)=\left\langle \left\{\left\{\left\langle A, \left\|\psi\right\|\right\rangle\right\}^{1}\right\}, \{A\}\right\rangle,$$

при цьому $\langle \psi, R\rangle\in\Psi(R)$, а $\left\|\psi\right\|$ – це ранг мультимножини $\psi$.

Для випадку порожньої таблиці $\langle \psi_{\varnothing}, R\rangle$ операція агрегування $Count_{A,R}(*)$ застосовується так:

$$Count_{A,R}(*)\left(\left\langle \psi_{\varnothing}, R\right\rangle\right)=\left\langle \left\{\left\{\left\langle A, \left\|\varnothing_{m}\right\|\right\rangle\right\}^{1}\right\}, \{A\}\right\rangle=\left\langle \left\{\left\{\left\langle A, 0\right\rangle\right\}^{1}\right\}, \{A\}\right\rangle.$$

*Приклад 6.* Застосуємо операцію агрегування $Count_{A,R}(*)$ до таблиці $\langle \Phi M, R\rangle$ за атрибутами $A,B,C,D$ отримаємо таблиці:

$$Count_{A,R}(*)=\left\langle \{\{A, 6\}\}, \{A\}\right\rangle,\ Count_{B,R}(*)=\left\langle \{\{B, 6\}\}, \{B\}\right\rangle,$$

$$Count_{C,R}(*)=\left\langle \{\{C, 6\}\}, \{C\}\right\rangle,\ Count_{D,R}(*)=\left\langle \{\{D, 6\}\}, \{D\}\right\rangle.$$

## Висновки

Визначено формальну математичну семантику агрегатних операцій над таблицями як мультимножинами рядків однієї схеми, яка проілюстрована прикладами їх використання. Для заданих агрегатних операцій параметром виступає лише один атрибут, проте отримані результати можна розширити на випадки, коли параметром є деяка функція над рядком. Результати роботи можуть бути використані в теорії узагальнених табличних алгебр.


1. *Codd E.F.* A Relational Model of Data for Large Shared Data Banks / E.F. Codd // Comm. of ACM. – 1970.– 13, N 6. – P. 377–387.
2. *Grefen Paul W.P.J., Rolf A. De By* A Multi-Set Extended Relational Algebra. A Formal Approach to a Practical Issue // 10th International Conference on Data Engineering, ICDE, February 14-18, 1994, Houston, TX, USA. – 1994. – P. 80–88.
3. *Lamperti G., Melchiori M., Zanella M.* On Multisets in Database Systems // 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.
4. *Garcia-Molina H.* Database Systems: The Complete Book: [2nd Edition] / H. Garcia-Molina, J.D. Ullman, J. Widom. – Prentice Hall, 2008. – 1119 p.
5. *Silbeschatz A., Korth H., Sudarshan S.* Database System Concepts – McGraw-Hill, 2011. – 1376 p.
6. *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.