УДК 004.65 http://doi.org/10.15407/pp2024.02-03.028

І.М. Лисенко

ЧИСЛЕННЯ РЯДКІВ ДЛЯ МУЛЬТИМНОЖИННОЇ ТАБЛИЧНОЇ АЛГЕБРИ

Дана робота становить логічне продовження досліджень, присвячених актуальній проблемі розробки теоретичних основ реляційних (табличних) баз даних. Питання використання мультимножин у табличних базах даних є важливим і актуальним, зважаючи на те, що багато мов запитів, орієнтованих на роботу з базами даних, потребують реляційну модель, яка передбачає мультимножинну семантику. Це обумовлено наявністю прикладних задач, особливістю яких є множинність і повторюваність даних. Наприклад, це соціологічні опитування різних груп населення, розрахунки на ДНК та ін. У цьому контексті розглядається питання побудови числення рядків для мультимножинної табличної алгебри, в якій поняття таблиці уточнюється, використовуючи поняття мультимножини. У роботі проведено формалізацію числення рядків для мультимножинної табличної алгебри: визначено алфавіт, синтаксис термів, атомів і формул числення рядків; введено множину дозволених формул, використовуючи концепцію вільних і зв'язаних рядків; також вводяться поняття схеми та множини атрибутів, з якими рядок зустрічається у формулах. Дано означення виразу числення рядків для мультимножинної табличної алгебри, згідно з яким – це мультимножина рядків, що задовольняють визначену дозволеною формулою умову. У статті наведено правила, за якими визначається кількість дублікатів рядків у результуючій мультимножині. Інший важливий результат полягає у доведенні того, що побудоване числення рядків є не менш виразним, ніж мультимножинна таблична алгебра. Це дослідження відкриває нові можливості для розвитку теорії баз даних і може бути корисним для спеціалістів у сфері інформаційних технологій та баз даних. Воно сприяє глибшому розумінню принципів побудови запитів, що є важливим аспектом у сучасній комп'ютерній науці та індустрії.
Ключові слова: реляційні бази даних мультимножина, мультимножинна таблична алгебра, числення рядків.

I.M. Lysenko

TUPLE CALCULUS FOR MULTISET TABLE ALGEBRA

This paper is a logical continuation of research devoted to the actual problem of developing the theoretical foundations of table (relational) databases. The issue of using multisets in table databases is important and relevant. Many database-oriented languages require a relational model with multiset semantics. There are many applied problems, the feature of which is multiplicity and repeatability of data. For example, these are sociological polls of different population groups, calculations on DNA, and others. In this context, the question of constructing a tuple calculus for a multiset table algebra is considered, in which the concept of a table is refined using the concept of a multiset. In the article, the formalization of tuple calculus for multiset table algebra is carried out. The alphabet, and the syntax of terms, atoms, and formulas are defined. A set of legal formulas is introduced through the concept of the free and bound variable. The concept of a scheme and set of attributes with which a tuple variable occurs in a formula are also introduced. The definition of tuple calculus expression for multiset table algebra is given, according to which it is a multiset of tuples that satisfy the condition defined by the legal formula. The article provides rules for determining the number of tuple duplicates in the resulting multiset. Another important result consists in proving that constructed tuple calculus is as expressive as multiset table algebra. This research opens up new possibilities for database theory development and may be useful for information technology and database professionals. It contributes to a deeper understanding of construction query principles, an important aspect of modern computer science and industry.
Key words: relation databases, multiset, multiset table algebra, tuple calculus.

Вступ

Реляційне числення є основою більшості реляційних мов запитів, оскільки орієнтоване лише на очікуваний результат. Водночас реляційна алгебра передбачає побудову реляційного виразу та виконання операцій. Існує два основних підходи до реляційного числення:

1) числення кортежів (Е. Кодд) – оперує рядками таблиць [1];

2) числення доменів (М. Лакруа, А. Піротт) – фокусується на доменах таблиць [2].

У роботі [3] розглядається таблична алгебра нескінченних (скінченних) таблиць, яка суттєво узагальнює та розширює класичну реляційну алгебру Кодда. Для даної алгебри побудовано числення на кортежах (доменах) та доведено еквівалентність табличної алгебри та даних числень. Методологічну основу вказаних досліджень складає композиційний підхід до програмології.

Крім того у роботі [3] побудовано мультимножинну табличну алгебру, яка розширює можливості баз даних за рахунок використання мультимножин. Природно постає питання про розробку числення для даної алгебри.

Мета дослідження – побудувати числення рядків для мультимножинної табличної алгебри та показати, що воно є не менш виразним, ніж мультимножинна таблична алгебра.

1. Мультимножинна таблична алгебра

Почнемо з розгляду ключових термінів мультимножин на основі джерел [4, 5]. Нехай U – довільна множина. Під мультимножиною α з основою U розумітимемо відображення вигляду $\alpha : U \to N$, де $N = \{1, 2, ...\}$ – множина натуральних чисел.

Позначимо через $\boldsymbol{D}$ – універсум елементів основ мультимножин. Характеристична функція мультимножини α – це функція вигляду $\chi_\alpha : \boldsymbol{D} \to Z_+$, значення якої задаються кусковою схемою:

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

для всіх $d \in \boldsymbol{D}$.

Порожньою мультимножиною $\varnothing_m$ називається мультимножина, основа якої – це порожня множина.

1-мультимножинами називаються мультимножини, областю значень яких є одноелементна множина вигляду $\{1\}$.

Основні поняття та твердження мультимножинної табличної алгебри розглянемо, спираючись на монографію [3].

Розглянемо дві множини: $\boldsymbol{A}$ – множина атрибутів та $\boldsymbol{D}$ – універсальний домен. Схемою будемо називати довільну скінченну множину атрибутів $R \subseteq \boldsymbol{A}$. Рядком схеми R будемо називати іменну множину на парі R, $\boldsymbol{D}$, проекція якої за першою компонентою рівна R. Використаємо наступні позначення: $S(R)$ – множина всіх рядків схеми R, а S – множина всіх рядків.

Пара $\langle \psi, R \rangle$, де ψ – довільна (можливо нескінченна) мультимножина, а R – схема таблиці, називається таблицею мультимножинної табличної алгебри.

Множина всіх таблиць схеми R позначається як $\Psi(R)$, а множина всіх таблиць – $\Psi = \bigcup_{R \subseteq \boldsymbol{A}} \Psi(R)$.

Через $Occ(s,\psi)$ позначають кількість дублікатів рядка s у мультимножині ψ. Мультимножину ψ також записують 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 ψ.

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

Має місце наступне твердження.

Твердження 1. Будь-який вираз мультимножинної табличної алгебри можна замінити еквівалентним йому виразом, який використовує лише операції селекції, з’єднання, проєкції, об’єднання, різниці та перейменування.

2. Побудова числення рядків для мультимножинної табличної алгебри

Основою реляційного числення є числення предикатів першого порядку. Побудову числення рядків для мультимножинної табличної алгебри почнемо з визначення алфавіту. Алфавіт числення рядків для мультимножинної табличної алгебри складають:

- множина атрибутів $\boldsymbol{A}$ (тобто множина імен атрибутів) і універсальний домен $\boldsymbol{D}$;
- множина предметних змінних, тобто змінних рядків $x_1, x_2, ...$;
- множина предметних констант $d_1, d_2, ...$;
- множина функціональних символів $f_1^{n_1}, f_2^{n_2}, ..., n_i \geq 1$;
- множина предикатних символів $p_1^{m_1}, p_2^{m_2}, ..., m_i \geq 1$;
- множина символів сталих таблиць – $\langle \alpha, R \rangle$;
- множина символів змінних таблиць – $\langle X, R \rangle$;
- знаки логічних операцій $\neg$, $\wedge$, $\vee$ та квантори $\exists, \forall$;
- знаки пунктуації – дужки та кома $(,)$.

Область інтерпретації предметних констант – це універсальний домен $\boldsymbol{D}$, а область інтерпретації предметних змінних – це множина всіх рядків на домені $\boldsymbol{D}$.

Надалі в тексті використовуємо наступні позначення: $\boldsymbol{x}$ – синтаксична змінна, областю зміни якої є змінні: $\boldsymbol{f}$ ($\boldsymbol{p}$) – синтаксична змінна, областю зміни якої є функціональні (предикатні) символи; $\boldsymbol{d}$ – синтаксична змінна, областю зміни якої є константи; $\mathcal{A}$ – синтаксична змінна, областю зміни якої є атрибути.

Серед слів, записаних за допомогою символів алфавіту, виділяють терми та формули. Сформулюємо означення цих синтаксичних об'єктів індукцією за їхньою довжиною.

Дамо означення *термів* числення рядків для мультимножин:

a) $\boldsymbol{d}$ – терм;

b) $\boldsymbol{x}(\mathcal{A})$ – терм;

c) якщо $\boldsymbol{f}$ – n-арний функціональний символ, а $\boldsymbol{u}_1, ..., \boldsymbol{u}_n$ – терми, то $\boldsymbol{f}(\boldsymbol{u}_1, ..., \boldsymbol{u}_n)$ – терм;

d) ніяких інших термів, крім зазначених в пунктах a)-c) немає.

Надалі в тексті $\boldsymbol{u}$ – це синтаксична змінна, областю зміни якої є терми. Областю інтерпретації термів є універсальний домен $\boldsymbol{D}$.

Дамо означення *формул* числення рядків для мультимножин. Почнемо з *атомарних формул* (атомів), які бувають трьох типів:

a1') нехай $\langle \alpha, R \rangle$ – стала таблиця, а $\boldsymbol{x}$ – змінний рядок, тоді $\alpha_R(\boldsymbol{x})$ – атом, який (за інтерпретації) означає, що рядок $\boldsymbol{x} \in \langle \alpha, R \rangle$;

a1") нехай $\langle X, R \rangle$ – змінна таблиця, а $\boldsymbol{x}$ – змінний рядок, тоді $X_R(\boldsymbol{x})$ – атом, який (за інтерпретації) означає, що рядок $\boldsymbol{x} \in \langle X, R \rangle$;

a2) нехай $\boldsymbol{p}$ – m-арний предикат на універсальному домені $\boldsymbol{D}$, а $\boldsymbol{u}_1, ..., \boldsymbol{u}_m$ – терми, тоді $\boldsymbol{p}(\boldsymbol{u}_1, ..., \boldsymbol{u}_m)$ – атом.

Побудуємо формули із атомів, використовуючи логічні зв'язки $\neg$, $\wedge$, $\vee$, квантори $\exists, \forall$ та дужки $(,)$.

f1. Будь-який атом є формулою.

f2. Якщо $\boldsymbol{P}$ і $\boldsymbol{Q}$ – формули, тоді вирази $(\neg \boldsymbol{P}), (\boldsymbol{P} \wedge \boldsymbol{Q})$, $(\boldsymbol{P} \vee \boldsymbol{Q})$ також є формулами.

f3. Нехай $\boldsymbol{x}$ – змінний рядок, $\boldsymbol{P}$ – формула, $R \subseteq \boldsymbol{A}$ – схема, тоді вирази $\exists \boldsymbol{x}(R)\boldsymbol{P}$ і $\forall \boldsymbol{x}(R)\boldsymbol{P}$ також є формулами.

f4. Якщо $\boldsymbol{P}$ – формула, тоді вираз $(\boldsymbol{P})$ – також формула.

f5. Інших формул, крім зазначених в пунктах f1-f4 немає.

Будемо використовувати $\boldsymbol{P}$, $\boldsymbol{Q}$ та $\boldsymbol{G}$ як синтаксичні змінні, областю зміни яких є формули.

Визначимо клас дозволених формул числення рядків для мультимножинної табличної алгебри, використовуючи поняття вільних та зв'язаних рядків, схеми $scheme(\boldsymbol{x},\boldsymbol{P})$ змінного рядка $\boldsymbol{x}$ та множини атрибутів $attr(\boldsymbol{x},\boldsymbol{P})$, з якими рядок $\boldsymbol{x}$ зустрічається у формулах.

Вирази $scheme(\boldsymbol{x},\boldsymbol{P})$ та $attr(\boldsymbol{x},\boldsymbol{P})$ визначені, якщо рядок $\boldsymbol{x}$ має хоча б одне вільне входження у формулу $\boldsymbol{P}$. Крім того має місце включення $attr(\boldsymbol{x},\boldsymbol{P}) \subseteq scheme(\boldsymbol{x},\boldsymbol{P})$, якщо дані вирази визначені.

Виділимо клас, так званих, дозволених формул числення рядків мультимножинної табличної алгебри.

Почнемо з термів:

1) якщо $\boldsymbol{u} = \boldsymbol{d}$, то $attr(\boldsymbol{x},\boldsymbol{u}) = \varnothing$;
2) якщо $\boldsymbol{u} = \boldsymbol{x}(\mathcal{A})$, то $attr(\boldsymbol{x},\boldsymbol{u}) = \{\mathcal{A}\}$, а $attr(\boldsymbol{x},\boldsymbol{y}(\mathcal{A})) = \varnothing$, якщо $\boldsymbol{x} \neq \boldsymbol{y}$;
3) якщо $\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{u}_1,...,\boldsymbol{u}_n)$ де $\boldsymbol{u}_i$ – терми, то $attr(\boldsymbol{x},\boldsymbol{u}) = \bigcup\limits_{i=1}^{n} attr(\boldsymbol{x},\boldsymbol{u}_i)$.

Іншими словами, $attr(\boldsymbol{x},\boldsymbol{u})$ – це множина атрибутів, які повинна мати схема рядка $\boldsymbol{x}$.

Нехай $\boldsymbol{P}$ – атомарна формула, тоді

a1') якщо $\boldsymbol{P} = \alpha_R(\boldsymbol{x})$, то єдине входження змінного рядка $\boldsymbol{x}$ є вільним у формулі $\boldsymbol{P}$ і $scheme(\boldsymbol{x},\boldsymbol{P}) = attr(\boldsymbol{x},\boldsymbol{P}) = R$;

a1'') аналогічно, якщо $\boldsymbol{P} = X_R(\boldsymbol{x})$, то єдине входження змінного рядка $\boldsymbol{x}$ є вільним у формулі $\boldsymbol{P}$ і $scheme(\boldsymbol{x},\boldsymbol{P}) = attr(\boldsymbol{x},\boldsymbol{P}) = R$;

a2) якщо $\boldsymbol{P} = \boldsymbol{p}(\boldsymbol{u}_1,...,\boldsymbol{u}_m)$, де $\boldsymbol{u}_i$ – терми, $\boldsymbol{x}_1,...,\boldsymbol{x}_k$ – усі змінні цих термів, тоді входження цих змінних рядків є вільними у формулі $\boldsymbol{P}$, при цьому схема $scheme(\boldsymbol{x}_i,\boldsymbol{P})$ не визначена, а $attr(\boldsymbol{x}_i,\boldsymbol{P}) = \bigcup\limits_{j=1}^{m} attr(\boldsymbol{x}_i,\boldsymbol{u}_j)$, $i = 1,...,k$.

Атоми завжди дозволені. Побудову дозволених формул числення рядків для мультимножинної табличної алгебри проведемо індукцією за довжиною формули. Нехай $\boldsymbol{G}$ і $\boldsymbol{Q}$ дозволені формули.

f1. Якщо $\boldsymbol{P} = \neg\boldsymbol{G}$, то формула $\boldsymbol{P}$ – дозволена, а входження змінних у формулі $\boldsymbol{P}$ будуть вільними або зв'язаними залежно від того, вільні або зв'язані входження цих змінних у формулі $\boldsymbol{G}$. Якщо рядок $\boldsymbol{x}$ входить у формулу $\boldsymbol{G}$ вільно, то мають місце рівності $scheme(\boldsymbol{x},\boldsymbol{P}) \simeq scheme(\boldsymbol{x},\boldsymbol{G})$ і $attr(\boldsymbol{x},\boldsymbol{P}) = attr(\boldsymbol{x},\boldsymbol{G})$, де $\simeq$ – узагальнена рівність, яка означає, що обидві частини рівності або невизначені, або визначені та мають рівні значення.

f2. Якщо $\boldsymbol{P} = \boldsymbol{G} \wedge \boldsymbol{Q}$ або $\boldsymbol{P} = \boldsymbol{G} \vee \boldsymbol{Q}$, то входження змінних у формулі $\boldsymbol{P}$ будуть вільними або зв'язаними, залежно від того, вільні або зв'язані входження цих змінних у підформулах $\boldsymbol{G}$ або $\boldsymbol{Q}$. Припустимо, що змінний рядок $\boldsymbol{x}$ входить у підформули $\boldsymbol{G}$ та/або $\boldsymbol{Q}$ вільно. Для схеми та множини атрибутів, з якими рядок $\boldsymbol{x}$ зустрічається у формулах маємо три випадки:

a. Схеми формул $scheme(\boldsymbol{x},\boldsymbol{G})$ та $scheme(\boldsymbol{x},\boldsymbol{Q})$ визначені. Формула $\boldsymbol{P}$ дозволена, якщо виконується рівність $scheme(\boldsymbol{x},\boldsymbol{G}) = scheme(\boldsymbol{x},\boldsymbol{Q})$. Покладемо за означенням $scheme(\boldsymbol{x},\boldsymbol{P}) = scheme(\boldsymbol{x},\boldsymbol{G})$.

b. Схема визначена тільки для однієї з підформул. Припустимо, що схема $scheme(\boldsymbol{x},\boldsymbol{G})$ визначена, а схема $scheme(\boldsymbol{x},\boldsymbol{Q})$ – невизначена. Формула $\boldsymbol{P}$ дозволена, якщо виконується включення $attr(\boldsymbol{x},\boldsymbol{Q}) \subseteq scheme(\boldsymbol{x},\boldsymbol{G})$. Покладемо за означенням $scheme(\boldsymbol{x},\boldsymbol{P}) = scheme(\boldsymbol{x},\boldsymbol{G})$.

c. Схема невизначена для обох підформул. Формула $\boldsymbol{P}$ є дозволеною, але $scheme(\boldsymbol{x},\boldsymbol{P})$ теж невизначена.

Для будь-якого з розглянутих випадків a - c $attr(\boldsymbol{x},\boldsymbol{P}) = attr(\boldsymbol{x},\boldsymbol{G}) \cup attr(\boldsymbol{x},\boldsymbol{Q})$.

f3. Якщо $\boldsymbol{P} = \exists\boldsymbol{x}(R)\boldsymbol{G}$ і змінний рядок $\boldsymbol{x}$ входить у підформулу $\boldsymbol{G}$ вільно, то формула $\boldsymbol{P}$ – дозволена. При визначеності схеми $scheme(\boldsymbol{x},\boldsymbol{G})$ та за умови виконання включення $attr(\boldsymbol{x},\boldsymbol{G})\subseteq R$ повинна виконуватися рівність $scheme(\boldsymbol{x},\boldsymbol{G})=R$. Оскільки змінна $\boldsymbol{x}$ не входить вільно у формулу $\boldsymbol{P}$, то $scheme(\boldsymbol{x},\boldsymbol{P})$ і $attr(\boldsymbol{x},\boldsymbol{P})$ невизначені. Якщо $\boldsymbol{y}\neq\boldsymbol{x}$, то будь-яке входження змінної $\boldsymbol{y}$ в $\boldsymbol{P}$ вільне або зв'язане, залежно від того, вільне або зв'язане входження $\boldsymbol{y}$ в $\boldsymbol{G}$. Якщо $\boldsymbol{y}$ входить у $\boldsymbol{P}$ вільно, то $scheme(\boldsymbol{y},\boldsymbol{P})\simeq scheme(\boldsymbol{y},\boldsymbol{G})$ і $attr(\boldsymbol{y},\boldsymbol{P})=attr(\boldsymbol{y},\boldsymbol{G})$.

f4. Якщо $\boldsymbol{P}=\forall\boldsymbol{x}(R)\boldsymbol{G}$ і змінний рядок $\boldsymbol{x}$ входить у підформулу $\boldsymbol{G}$ вільно, то формула $\boldsymbol{P}$ – дозволена, а всі визначення і обмеження аналогічні випадку f3.

f5. Якщо $\boldsymbol{P}=(\boldsymbol{G})$, то формула $\boldsymbol{P}$ – дозволена, а вільні та зв'язані входження змінних, схема і множина атрибутів, з якими змінний рядок зустрічається у формулах, залишаються такими як і для підформули $\boldsymbol{G}$.

Іншими словами, рівність $attr(\boldsymbol{x},\boldsymbol{P})=R$, означає, що для конкретної інтерпретації формули $\boldsymbol{P}$, коли змінна $\boldsymbol{x}$ набуває значення у вигляді рядка s схеми R', повинне виконуватися включення $R\subseteq R'$.

Дозволеність формули забезпечує її коректну інтерпретацію при розгляді виразів числення рядків для мультимножинної табличної алгебри.

Вирази числення рядків для мультимножинної табличної алгебри мають вигляд $\{\boldsymbol{x}^n(R)\,|\,\boldsymbol{P}(\boldsymbol{x})\}$, де

1. формула $\boldsymbol{P}$ – дозволена;
2. змінна $\boldsymbol{x}$ – єдина змінна, яка входить у формулу $\boldsymbol{P}$ вільно;
3. якщо $scheme(\boldsymbol{x},\boldsymbol{P})$ визначена, то $scheme(\boldsymbol{x},\boldsymbol{P})=R$, інакше, $attr(\boldsymbol{x},\boldsymbol{P})\subseteq R$
4. n – кількість дублікатів рядка $\boldsymbol{x}$.

Варто підкреслити, що результатом виконання запиту, визначеного виразом $\{\boldsymbol{x}^n(R)\,|\,\boldsymbol{P}(\boldsymbol{x})\}$, є мультимножина рядків, які описує вираз $\boldsymbol{P}(\boldsymbol{x})$.

Нехай формула $\boldsymbol{P}(\boldsymbol{x})$ – дозволена, $R\subseteq\boldsymbol{A}$. Підставивши конкретний рядок s схеми R замість $\boldsymbol{x}$ у формулу $\boldsymbol{P}$ отримаємо формулу $\boldsymbol{P}(s/\boldsymbol{x})$, значення якої визначається шляхом модифікації кожного атома з $\boldsymbol{P}$ за наступними правилами:

a1') нехай рядок $\boldsymbol{x}$ підформули $\alpha_{R'}(\boldsymbol{x})$ вільний у формулі $\boldsymbol{P}$. За означенням дозволеної формули маємо включення $R'\subseteq R$. Атом $\alpha_{R'}(\boldsymbol{x})$ істинний у разі підстановки конкретного рядка s замість змінної $\boldsymbol{x}$, якщо $s\,|\,R'\in t$, інакше атом $\alpha_{R'}(\boldsymbol{x})$ хибний

a1'') нехай рядок $\boldsymbol{x}$ підформули $X_{R'}(\boldsymbol{x})$ вільний у формулі $\boldsymbol{P}$. Аналогічно випадку a1'), має місце включення $R'\subseteq R$. Атом $X_{R'}(\boldsymbol{x})$ істинний у разі підстановки конкретного рядка s замість змінної $\boldsymbol{x}$, якщо $s\,|\,R'\in\bar{X}$, де $\bar{X}\in P(S(R'))$ – значення змінної таблиці;

a2) нехай рядок $\boldsymbol{x}$ підформули $\boldsymbol{p}(\boldsymbol{u}_1,...,\boldsymbol{u}_m)$ вільний у формулі $\boldsymbol{P}$, тоді замінимо $\boldsymbol{x}(A_i)$ на $d_i\in\boldsymbol{D}$, де $\langle A_i,d_i\rangle\in s$ (d_i значення атрибута A_i в рядку s) при підстановці конкретного рядка s замість змінної $\boldsymbol{x}$. Атом $\boldsymbol{p}(\boldsymbol{u}_1,...,\boldsymbol{u}_m)$ буде істинний, якщо предикат $\boldsymbol{p}$ – істинний на відповідних значеннях, інакше атом буде хибний.

Інтерпретацією формули є множина значень істинності всіх атомів цієї формули. Припустимо, що формула $\boldsymbol{P}$ – дозволена без вільних змінних. Визначимо інтерпретацію формули $\boldsymbol{P}$ для кожного випадку.

f1. Якщо $\boldsymbol{P}=\neg\boldsymbol{G}$, то в $\boldsymbol{G}$ не повинно бути вільних змінних. $\boldsymbol{P}$ істинна формула, коли підформула $\boldsymbol{G}$ хибна, і хибна, коли підформула $\boldsymbol{G}$ істинна.

f2. Якщо $\boldsymbol{P}=\boldsymbol{G}\wedge\boldsymbol{Q}$ або $\boldsymbol{P}=\boldsymbol{G}\vee\boldsymbol{Q}$, то в підформулах $\boldsymbol{G}$ та $\boldsymbol{Q}$ не повинно бути вільних змінних. При $\boldsymbol{P}=\boldsymbol{G}\wedge\boldsymbol{Q}$, формула $\boldsymbol{P}$ істинна тоді, коли підформули $\boldsymbol{G}$ та $\boldsymbol{Q}$ одночасно істинні, та хибна у всіх інших випадках. При $\boldsymbol{P}=\boldsymbol{G}\vee\boldsymbol{Q}$, формула $\boldsymbol{P}$ хибна тоді, коли підформули $\boldsymbol{G}$ і $\boldsymbol{Q}$ одночасно хибні, та істинна у всіх інших випадках.

f3. Якщо $\boldsymbol{P} = \exists \boldsymbol{x}(R)\boldsymbol{G}$, то $\boldsymbol{x}$ – єдина змінна, яка входить у підформулу $\boldsymbol{G}$ вільно. Формула $\boldsymbol{P}$ істинна, якщо знайдеться принаймні один рядок $s \in S(R)$ такий, що формула $\boldsymbol{G}(s/\boldsymbol{x})$ отримана при підстановці s замість $\boldsymbol{x}$ – істинна, інакше $\boldsymbol{P}$ – хибна.

f4. Якщо $\boldsymbol{P} = \forall \boldsymbol{x}(R)\boldsymbol{G}$, то $\boldsymbol{x}$ – єдина змінна, яка входить у підформулу $\boldsymbol{G}$ вільно. Формула $\boldsymbol{P}$ істинна, якщо для кожного рядка $s \in S(R)$ формула $\boldsymbol{G}(s/\boldsymbol{x})$, отримана в результаті підстановки, буде істинна, інакше формула $\boldsymbol{P}$ хибна.

f5. Якщо $\boldsymbol{P} = (\boldsymbol{G})$, то формула $\boldsymbol{P}$ істинна, коли підформула $\boldsymbol{G}$ істинна і хибна, коли підформула $\boldsymbol{G}$ хибна.

Нехай $E = \{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$ – вираз числення рядків для мультимножинної табличної алгебри. Значенням виразу E назвемо таблицю $\langle \varphi, R \rangle$ мультимножинної табличної алгебри, яка складається з рядків $s \in S(R)$ таких, що формула $\boldsymbol{P}(s/\boldsymbol{x})$ істинна, а кількість дублікатів рядка s в таблиці $\langle \varphi, R \rangle$ визначається так:

1) нехай $\boldsymbol{P} = \alpha_R(\boldsymbol{x})$, тоді $n = Occ(s,\varphi) = Occ(s,\alpha)$;

2) нехай $\boldsymbol{P} = X_R(\boldsymbol{x})$, тоді $n = Occ(s,\varphi) = Occ(s,X)$;

3) нехай $\boldsymbol{P} = \boldsymbol{p}(\boldsymbol{u}_1,...,\boldsymbol{u}_m)$, де $\boldsymbol{u}_j$ – терми, $j = \overline{1,m}$, $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_k$ – всі змінні цих термів, тоді маємо

$$n = Occ(s,\varphi) = \sum_{\substack{s' \in \Theta(\psi), s' \mid R' = s \\ \boldsymbol{p}(u_1,\ldots,u_m) = true}} Occ(s',\psi),$$

де $\langle \psi, R \rangle$ – таблиця до якої будується запит, $R' = \bigcup_{j=1}^{m} attr(\boldsymbol{x}_i, \boldsymbol{u}_j)$, $i = \overline{1,k}$;

4) нехай $\boldsymbol{P} = \neg \boldsymbol{G}$ та формула $\boldsymbol{G}$ породжує m дублікатів рядка s, тоді $n = Occ(s,\varphi) = Occ(s, C(\psi)) \dot{-} m$, де $C(\psi)$ – мультимножина таблиці $C(\langle \psi, R \rangle)$, яка є насиченням таблиці $\langle \psi, R \rangle$, що є значенням виразу $\{\boldsymbol{y}(R) \mid \boldsymbol{G}(\boldsymbol{y})\}$, а зрізана різниця $\dot{-}$ розуміється в звичайному сенсі $x \dot{-} y = \begin{cases} x - y, \text{якщо } x \ge y \\ 0, \text{якщо } x < y \end{cases}$;

5) нехай $\boldsymbol{P} = \boldsymbol{G} \wedge \boldsymbol{Q}$ та формула $\boldsymbol{G}$ породжує k дублікатів рядка s, а формула $\boldsymbol{Q}$ породжує m дублікатів рядка s, тоді $n = Occ(s,\varphi) = \min(k,m)$;

6) нехай $\boldsymbol{P} = \boldsymbol{G} \vee \boldsymbol{Q}$ та формула $\boldsymbol{G}$ породжує k дублікатів рядка s, а формула $\boldsymbol{Q}$ породжує m дублікатів рядка s, тоді $n = Occ(s,\varphi) = k + m$;

7) нехай $\boldsymbol{P} = \exists \boldsymbol{x}(R)\boldsymbol{G}$ та формула $\boldsymbol{G}$ породжує k дублікатів рядка s, тоді $n = Occ(s,\varphi) = k$;

8) нехай $\boldsymbol{P} = \forall \boldsymbol{x}(R)\boldsymbol{G}$ та формула $\boldsymbol{G}$ породжує k дублікатів рядка s, тоді $n = Occ(s,\varphi) = k$;

9) нехай $\boldsymbol{P} = (\boldsymbol{G})$ та формула $\boldsymbol{G}$ породжує k дублікатів рядка s, тоді $n = Occ(s,\varphi) = k$.

Твердження 2. Якщо F – вираз мультимножинної табличної алгебри, то можна ефективно побудувати еквівалентний йому вираз E числення рядків для мультимножинної табличної алгебри.

Доведення. Згідно Твердження 1 у процесі доведення досить розглянути вирази мультимножинної табличної алгебри, які міститимуть тільки операції об'єднання, різниці, селекції, проєкції, з'єднання та перейменування. Доведемо дане твердження методом математичної індукції за числом операцій у виразі F.

База індукції. У цьому випадку вираз F не містить операцій. Це або стала, або змінна таблиці.

Нехай $F = \langle \alpha, R \rangle$ – стала таблиця, де α – мультимножина рядків схеми R. Покладемо $E = \{\boldsymbol{x}^n(R) \mid \alpha_R(\boldsymbol{x})\}$.

Нехай $F = \langle X, R \rangle$ – змінна таблиця, тоді $E = \{\boldsymbol{x}^n(R) \mid X_R(\boldsymbol{x})\}$. □

Крок індукції. Припустимо, що твердження виконується для всіх виразів мультимножинної табличної алгебри, які містять менше i операцій. Розглянемо вираз F, який містить i операцій.

Випадок 1 (об'єднання). $F = F_1 \bigcup_{All}^{R} F_2$, де вирази F_1 і F_2 мають менше i операцій. Тоді існують вирази числення рядків $\{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$ і $\{\boldsymbol{x}^m(R) \mid \boldsymbol{Q}(\boldsymbol{x})\}$, які еквівалентні F_1 і F_2 відповідно. Значеннями цих виразів є таблиці, в які рядок $\boldsymbol{x}$ входить n та m разів відповідно. Покладемо E рівним $\left\{\boldsymbol{x}^k(R) \mid \boldsymbol{P}(\boldsymbol{x}) \vee \boldsymbol{Q}(\boldsymbol{x})\right\}$, де кількість дублікатів рядка $\boldsymbol{x}$ у вихідній таблиці дорівнює $k = n + m$. □

Випадок 2 (різниця). $F = F_1 \setminus_{All}^{R} F_2$, де вирази F_1 і F_2 мають менше i операцій. Нехай $\{\boldsymbol{x}^n(R_1) \mid \boldsymbol{P}(\boldsymbol{x})\}$ і $\{\boldsymbol{x}^m(R) \mid \boldsymbol{Q}(\boldsymbol{x})\}$ – вирази числення рядків, еквівалентні F_1 і F_2 відповідно. Значеннями цих виразів є таблиці, в які рядок $\boldsymbol{x}$ входить n та m разів. Покладемо E рівним $\left\{\boldsymbol{x}^k(R) \mid \boldsymbol{P}(\boldsymbol{x}) \wedge \neg \boldsymbol{Q}(\boldsymbol{x})\right\}$, причому кількість дублікатів рядка $\boldsymbol{x}$ у вихідній таблиці дорівнює $k = n \dot{-} m$. □

Випадок 3 (селекція). $F = \sigma_{\tilde{p},R}\left(F_1\right)$, де вираз F_1 має менше i операцій. Тоді існує вираз $\{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$, еквівалентний F_1. Покладемо $E = \left\{\boldsymbol{x}^k(R) \mid \boldsymbol{P}(\boldsymbol{x}) \wedge \boldsymbol{p}(\boldsymbol{x}(A_1),...,\boldsymbol{x}(A_m))\right\}$, де $R = \{A_1,...,A_m\}$ схема таблиці, що є значенням виразу F_1, а кількість дублікатів рядка $\boldsymbol{x}$ у вихідній таблиці не змінюється, тому $k = n$. Предикат-параметр селекції заданий наступним чином $\tilde{p}(s) = T \Leftrightarrow \boldsymbol{p}(s(A_1),...,s(A_m)) = T$, $s \in S(R)$, де $\boldsymbol{p}$ – m-арний предикатний символ. □

Випадок 4 (проєкція). $F = \pi_{X,R}(F_1)$, де вираз F_1 має менше i операцій. Тоді існує вираз числення рядків $\{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$, еквівалентний F_1. Значенням цього виразу є таблиця, до якої рядок $\boldsymbol{x}$ входить n разів. Покладемо E рівним $\{\boldsymbol{y}^k(X \bigcap R) \mid \exists \boldsymbol{x}(R)(\boldsymbol{P}(\boldsymbol{x}) \wedge \bigwedge_{A \in X \bigcap R} \boldsymbol{y}(A) = \boldsymbol{x}(A))\}$, де кількість дублікатів рядка $\boldsymbol{y}$ у вихідній таблиці дорівнює $k = \sum_{\bigwedge_{A \in X \bigcap R} \boldsymbol{y}(A) = \boldsymbol{x}(A)} n$. □

Випадок 5 (з'єднання). $F = F_1 \underset{R_1,R_2}{\otimes} F_2$. Тоді існують вирази числення рядків $\{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$ і $\{\boldsymbol{x}^m(R) \mid \boldsymbol{Q}(\boldsymbol{x})\}$, які еквівалентні F_1 і F_2 відповідно. Значеннями цих виразів є таблиці, до якиїх рядок $\boldsymbol{x}$ входить n разів, а рядок $\boldsymbol{y}$ – m разів відповідно. Покладемо вираз E рівним $\{\boldsymbol{z}^k(R_1 \bigcup R_2) \mid \exists \boldsymbol{x}(R_1) \exists \boldsymbol{y}(R_2)(\boldsymbol{P}(\boldsymbol{x}) \wedge \boldsymbol{Q}(\boldsymbol{y}) \wedge$, $\wedge \bigwedge_{A \in R_1} \boldsymbol{z}(A) = \boldsymbol{x}(A) \wedge \bigwedge_{A \in R_2} \boldsymbol{z}(A) = \boldsymbol{y}(A))\}$ причому кількість дублікатів рядка $\boldsymbol{z}$ у вихідній таблиці дорівнює $k = n \times m$. □

Випадок 6 (перейменування). $F = Rt_{\xi,R}(F_1)$, де $\xi : \boldsymbol{A} \tilde{\to} \boldsymbol{A}$ – ін'єктивна функція, що здійснює перейменування атрибутів. Нехай $\{\boldsymbol{x}^n(R) \mid \boldsymbol{P}(\boldsymbol{x})\}$ – вираз числення рядків, еквівалентний F_1. Значенням цього виразу є таблиця, до якої рядок $\boldsymbol{x}$ входить n разів. Покладемо E рівним $\{\boldsymbol{y}^k(R_2) \mid \exists \boldsymbol{x}(R)(\boldsymbol{P}(\boldsymbol{x}) \wedge \bigwedge_{C \in R \setminus dom\xi} \boldsymbol{y}(C) = \boldsymbol{x}(C) \wedge$ $\wedge \bigwedge_{A \in R \bigcap dom\xi} \boldsymbol{x}(A) = \boldsymbol{y}(\xi(A)))\}$, де $R_2 = R \setminus dom\xi \bigcup \xi[R]$, а кількість дублікатів рядка $\boldsymbol{y}$ у вихідній таблиці дорівнює n. □ □

3. Зв'язок з мовою запитів SQL

Фундаментальним об'єктом даних у мові SQL є не класичне відношення Е. Кодда, а скоріше таблиця. Причому таблиці SQL містять, власне кажучи, не множини, а мультимножини рядків, тобто допускаються повторення елементів. Основні оператори SQL не є реляційними операторами у сенсі цього терміну, а є аналогами реляційних операторів, які призначені для роботи з мультимножинами. При створенні нової таблиці за допомогою запиту, система SQL, як правило, не видаляє дублікати рядків, а повертає результат, в якому один і той же рядок може зустрічатися декілька разів. Так, щоб виключити появу дублікатів, за оператором SELECT потрібно поставити ключове слово DISTINCT.

Продемонструємо доцільність побудови числення рядків для мультимножинної табличної алгебри на наступному прикладі.

Розглянемо таблицю <*Scores*, *R*>, де схема R ={№, Name, Topic 1, Topic 2, Topic 3, Quiz} (див. Табл. 1).

Таблиця 1

Таблиця <*Scores*, *R*>

№	Name	Topic 1	Topic 2	Topic 3	Quiz
1.	Баков А.	5	15	14	16
2.	Бойко І.	6	14	15	16
3.	Борода К.	7	17	20	20
4.	Геранов О.	9	20	19	20
5.	Кайдан Ю.	5	18	15	18
6.	Кузенко Є.	6	19	13	20
7.	Кулак П.	4	8	9	16

Відповідь на питання «Які бали отримали за тест (Quiz) перші п'ять студентів?» на мові SQL матиме вигляд:

```
SELECT Quiz
FROM Scores
LIMIT 5;
```

Таблиця-результат <Quiz, R_1>, де схема R_1={Quiz}, міститиме значення-дублікати (див. Табл. 2).

Таблиця 2

Таблиця <Quiz, R_1>

Quiz
16
16
20
20
18

Реалізувати цей запит в термінах класичного числення рядків неможливо, оскільки результатом його виконання буде множина рядків, а не мультимножина (як очікується), тобто дублікати не будуть враховуватися в результат.

У численні рядків для мультимножинної табличної алгебри вираз еквівалентний даному запиту матиме вигляд: $\{x^n(\text{Quiz}) \mid \exists y(R)(Scores(y) \wedge (y(\text{№}) = 1 \vee y(\text{№}) = 2 \vee y(\text{№}) = 3 \vee y(\text{№}) = 4 \vee y(\text{№}) = 5) \wedge \wedge x(\text{Quiz}) = y(\text{Quiz}))\}$.

Результат, який він описує, аналогічний результату отриманому при виконанні відповідного запиту на мові SQL.

Отже, як видно з прикладу, побудоване числення рядків для мультимножинної табличної алгебри дозволяє адекватно формалізувати мови запитів, зокрема SQL, врахувавши мультимножинну семантику закладену в їхній основі.

Висновки

У статті запропоновано числення рядків для мультимножинної табличної алгебри. Визначено алфавіт, синтаксис термів, атомів і формул числення рядків. Використовуючи концепцію вільних і зв'язаних рядків, поняття схеми та множини атрибутів, з якими рядок зустрічається у формулах, введено клас дозволених формул. Показано, що запропоноване числення рядків є не менш виразним, ніж мультимножинна таблична алгебра. На прикладі продемонстровано доцільність побудови числення рядків для мультимножинної табличної алгебри, враховуючи те, що мови запитів орієнтовані на роботу з базами даних передбачають повторюваність елементів у таблиці.

Проблематика наступних досліджень полягає у встановленні відповідного дуального результату.

Література

1. Codd E.F. Relational Completeness of Data Base Sublanguages. *Data Base Systems*. New York: Prentice-Hall, 1972. P. 65-98
2. Lacroix M., Pirotte A. Domain-oriented Relational Languages. Proc. 3rd Int. Conf. on Very Large Data Bases. Tokyo, October, 1977. P. 370-378.
3. Буй Д. Б., Глушко І. М. Числення на розширення сигнатур табличних алгебр: монографія. Ніжин: НДУ ім. М. Гоголя, 2016. 151 с.
4. 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.
5. Богатирьова Ю. О. Теорія мультимножин та її застосування: дис. Канд. фіз.-мат. наук: 01.05.03. К., 2011. 113с.

References

1. E.F. Codd Relational Completeness of Data Base Sublanguages, in: Data Base Systems (1972) 65-98.
2. Lacroix M., Pirotte A. Domain-oriented Relational Languages, in: Proceedings of. 3rd Int. Conf. on Very Large Data Bases., 1977, pp. 370-378.
3. V.N. Redko, et al. Relational Databases: Table Algebras and SQL-like Language. Kyiv: Publishing house Academperiodica, 2001. [in Ukrainian]
4. D.B Buy, I.M. Glushko, Calculi and extensions of table algebras signature. Nizhyn: NDU im. M. Gogol, 2016. [in Ukrainian]
5. J.A. Bogatyreva Multisets theory and its applications. Ph.D. thesis, Kyiv National Taras Shevchenko University, 2011. [in Ukrainian]

Одержано: 12.02.2024
Внутрішня рецензія отримана: 19.02.2024
Зовнішня рецензія отримана: 08.03.2024

Про авторів:
Лисенко Ірина Миколаївна,
к.ф.-м.н., доцент.
http://orcid.org/0000-0003-2549-5356.

Місце роботи автора:
Ніжинський державний
університет імені Миколи Гоголя,
E-mail: iryna.glushko@ndu.edu.ua
Сайт: http://www.ndu.edu.ua/