УДК:004.655

Д.Б. БУЙ, Ю.А. БОГАТЫРЕВА

Киевский национальный университет имени Тараса Шевченко, Украина

ТЕОРИЯ МУЛЬТИМНОЖЕСТВ: БИБЛИОГРАФИЯ, ПРИМЕНЕНИЕ В ТЕОРИИ ТАБЛИЧНЫХ БАЗ ДАННЫХ

В статье рассматриваются элементы теории мультимножеств. Дано формальное определение мультимножеству, его характеристической функции, а также операциям объединения и пересечения мультимножеств. Строится решетка мультимножеств. Приводятся результаты о структуре семейства мультимножеств. Раскрывается библиография по мультимножествам. Также рассматривается применения мультимножеств в табличных базах данных. Даны формальные определения операциям проекции, селекции и декартового соединения.

Ключевые слова: *мультимножество, базы данных, полная решетка, объединение мультимножеств, пересечение мультимножеств.*

Введение

Содержательно говоря, мультимножества – это совокупности с повторениями.

Понятие “мультимножество” (multiset, bag) было впервые предложено Н.Г. де Брейном в частной корреспонденции с Д. Кнутом. В 70-х годах этот термин широко распространился и сейчас является стандартным термином [1].

Для представления и исследования объектов, особенностью которых является множественность и повторяемость данных, в качестве модели удобно использовать мультимножество. Поэтому не удивительно, что мультимножества используются в самых различных областях и сферах.

1. Основы теории мультимножеств

Введем формальное определение мультимножества.

Мультимножество α с основой U – это функция вида $\alpha : U \to N^+$, где U – некоторое множество (в классическом канторовском понимании), а $N^+ = \{1, 2, ...\}$ – множество натуральных чисел без нуля [1, 2].

Пусть задано мультимножество α с основой $U_\alpha = \mathrm{dom}\,\alpha$. Здесь $\mathrm{dom}\,\alpha$ – множество первых компонент пар, которые составляют функцию, т.е. область определения мультимножества как функции.

Характеристической функцией мультимножества α называется функция вида $\chi_\alpha : D \to N$, значение которой задается следующей кусочной схемой

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

для всех $d \in D$, где D – универсум элементов основ мультимножеств [1, 2].

Очевидно, что по характеристической функции соответствующее мультимножество восстанавливается однозначно.

Введем бинарное отношение включения на мультимножествах.

Мультимножество β включается в мультимножество α (обозначается $\beta \preceq \alpha$), если для их характеристических функций выполняется утверждение:

$$\chi_\beta(d) \le \chi_\alpha(d), \quad \forall d \in D.$$

Непосредственно проверяется, что отношение включения является частичным порядком.

Дадим определение операциям объединения и пересечения мультимножеств.

Операция $\bigcup_{All}$ мультимножествам α и β сопоставляет мультимножество $\alpha \bigcup_{All} \beta$, значение характеристической функции которого на произвольном аргументе d задается выражением $\max(\chi_\alpha(d), \chi_\beta(d))$.

Операция $\bigcap_{All}$ мультимножествам α и β сопоставляет мультимножество $\alpha \bigcap_{All} \beta$, значение характеристической функции которого на произвольном аргументе d задается выражением

$$\min(\chi_\alpha(d), \chi_\beta(d)).$$

Операции объединения и пересечения мультимножеств обладают нижеперечисленными стандартными свойствами.

Лемма 1 (о идемпотентности, коммутативности и ассоциативности операций объединения и пересечения мультимножеств). Операции $\bigcup_{All}$ и $\bigcap_{All}$ идемпотентны (т.е. $\alpha \bigcup_{All} \alpha = \alpha$, $\alpha \bigcap_{All} \alpha = \alpha$), коммутативны и ассоциативны.

Доказательство вытекает из того, что теоретико-числовые операции $\max$, $\min$ имеют те же самые свойства.

Таким образом, можно рассматривать две коммутативные идемпотентные полугруппы $\langle \mathbf{A} , \bigcup_{All} \rangle$ и $\langle \mathbf{A} , \bigcap_{All} \rangle$, где A – семейство мультимножеств соответствующего универсума D.

Используя результат теории решеток [3, § 8, с. 151, теорема 1], можно полугруппу по объединению превратить в верхнюю полурешетку, а полугруппу по пересечению – в нижнюю. Частичные порядки верхней полурешетки и нижней полурешетки задаются соответственно следующими определениями:

$$\alpha \underline{\leq} \beta \overset{def}{\Leftrightarrow} \alpha \bigcup_{All} \beta = \beta \,,\ \alpha \underline{\trianglelefteq} \beta \overset{def}{\Leftrightarrow} \alpha \bigcap_{All} \beta = \alpha \,,$$

причем

$$\sup\nolimits_{\underline{\leq}} \{\alpha,\beta\} = \alpha \bigcup_{All} \beta \,,\ \inf\nolimits_{\underline{\trianglelefteq}} \{\alpha,\beta\} = \alpha \bigcap_{All} \beta \,.$$

Непосредственно проверятся, что эти порядки совпадают с порядком включения мультимножеств $\preceq$.

Таким образом, семейство мультимножеств A с частичным порядком $\preceq$ является одновременно и верхней, и нижней полурешеткой, т.е. решеткой.

Приведем более сильные результаты о структуре семейства мультимножеств A.

Семейство мультимножеств A имеет наименьший элемент (пустое мультимножество – $\varnothing$). Его характеристической функцией является константная функция всюду равная нулю.

Кроме того, любое его подмножество, ограниченное сверху, имеет точную верхнюю грань.

Таким образом, следуя терминологии работ [4, 5], семейство мультимножеств A с частичным порядком включения является одновременно условно полным множеством и полной полурешеткой (complete semilattice).

Кроме того, можно показать, что пополнение указанного семейства мультимножеств наибольшим элементом превращает его в полную решетку.

2. Обзор современной библиографии по мультимножествам

Обзор современной литературы по мультимножествам показал, что исследования, посвященные этой тематике, можно условно разделить на два вида: работы по построению теории мультимножеств, а также работы, связанные с вопросами применения мультимножеств в различных прикладных областях.

Теорию мультимножеств рассматривали в своих работах Дж. Альберт (J. Albert) [24], В. Близард (W. Blizard) [26], А. Сиропоулос (A. Syropoulos) [13], А.Б. Петровський [1], а также В.Н. Редько, Ю.И. Брона, Д.Б. Буй, С.А. Поляков [2].

В. Близард в своей работе [26] представил развернутый обзор развития теории мультимножеств, различные определения мультимножества, а также некоторые специфические применения мультимножеств. В статье в хронологическом порядке излагаются основные идеи и достижения, полученные исследователями в этом направлении. Отмечено при этом, что у различных авторов понятие мультимножества возникает под разными названиями.

Одним из первых, кто обратил внимание на то, что существует необходимость рассмотрения мультимножества как отдельного математического объекта, был Д. Кнут. В своей книге [19] он дает содержательное определение мультимножеству и определяет операции объединения, пересечения и сложение мультимножеств.

Дж. Альберт в своей статье [24] не только дает определения мультимножеству и операций над ними, но и представляет некоторые результаты, относящиеся к алгебраическим свойствам мультимножеств.

В статье [13] А. Сиропоулос систематизирует все то, что имеет отношение к мультимножествам: дает определение мультимножеству и операций над ними, описывает так называемые гибридные множества, мультимножества в теории категорий, нечеткие и частично-упорядоченные мультимножества.

А.Б. Петровский в своей небольшой монографии [1] вводит основные определения, относящиеся к теории мультимножеств: мультимножества, его характеристической функции, операций над мультимножествами. Кроме того, он рассматривает некоторые свойства основных операций над мультимножествами, методы графического представления и приводит краткий обзор применений мультимножеств в различных областях. Однако следует заметить, что изложенная теория требует уточнений и дополнений, а также устранения очевидных некорректностей.

Автор продолжает свое исследования в книге [25], в которой рассматриваются метрические пространства множеств и мультимножеств, устанавливаются основные свойства мер множеств и мультимножеств, описываются новые типы пространств измеримых множеств и измеримых мультимножеств, а также новые виды метрик.

В работе [12] рассматриваются различные представления мультимножеств: в мультипликативной форме, в линейной форме, в виде последовательности, как семейство множеств, в виде числовой функции. Определяются операции над мультимножествами, а также дается краткий обзор применений мультимножеств в математике, компьютерных науках и других областях.

Мультимножества так же используются в декларативных языках программирования. Дж. Ллойд (J.W. Lloyd) в статьях [9], [10] предлагает новый способ поддержки мультимножеств в декларативном языке программирования Escher.

Он вводит стандартное определение мультимножества, а потом определяет мультимножество соответствующими средствами языка. Автор также реализует операции над мультимножествами (сложение, объединение, пересечение, разность) и ряд вспомогательных функций.

В декларативном языке ограничений OCL такой структурный тип, как мультимножество (BAG), определен явно. Этот тип является одним из разновидностей коллекций и имеет набор соответствующих операций [20].

Одним из возможных применений теории мультимножеств является представление и кодирование информации в терминах данной теории. Этому вопросу посвящена работа [29]. Информационный ресурс в данном случае рассматривается как ресурс, порождающий мультимножественные сообщения (т. е. сообщения, представляющие собой мультимножества символов). Исследуется норма энтропии мультимножественного информационного ресурса.

Д. Кнут в своей работе [4] использует мультимножества в контекстно-свободных мультиязыках. Он определяет мультиязык как мультимножество строк и строит контекстно-свободный мультиязык. Автор обращает внимание на то, что замена множества строк на мультимножество строк, является более естественным с точки зрения программирования.

Мультимножества применяют при определении основных понятий сетей Петри [15, 27] и в задачах распознавания символов [28].

Кроме собственно компьютерных наук мультимножества используются в математике (в частности, в λ-исчислении [14]), физике, философии, логике, лингвистике [12, 26], а также в новой области знаний – так называемых вычислениях на ДНК [21].

Анализ литературы свидетельствует о достаточно широком применении мультимножеств при решении практических задач, что, в свою очередь, вызывает необходимость в дальнейшем расширении и уточнении соответствующих аспектов теории мультимножеств.

3. Применение мультимножеств в табличных базах данных

Одним из наиболее естественных применений мультимножеств является их применение в табличных базах данных.

Определение таблицы (модели данных в табличных базах данных) в терминах мультимножества приведено в [2], а также в книге Г. Гарсия-Молина и др. [18].

Мультимножество рассматривается как совокупность кортежей, с возможными повторениями. Над мультимножествами вводится как основные (объединение, пересечение, разность, произведение, соединение), так и дополнительные (агрегирование, группирование, сортировка) операции.

Вопрос расширения возможностей баз данных (далее БД) за счет использования мультимножеств освещен в работе Ж. Ламперти (G. Lamperti) и др. [6]. Авторы отмечают, что современные коммерческие реляционные системы БД позволяют проводить мультимножественно-ориентированные манипуляции над таблицами, даже если они основаны на формальной модели, которая является множественно-ориентированной.

В статье даются определения операциям проекции, селекции, произведения, соединения (natural и theta-), переименования таблиц как мультимножества строк, а также определения аналогов теоретико-множественных операций (объединения, пересечения, разности).

Приведем определения некоторых операций.

Операция проекции таблицы, представляющей собой мультимножество строк:

$$\pi_Y(m) \overset{\text{def}}{=} \left[t_{(Y)} \mid t \in m \right],$$

где m – мультимножество строк над схемой M, состоящей из множества атрибутов X, причем $Y \subseteq X$.

Здесь и далее используются обозначения из [6]: в частности, $t_{(Y)}$ – ограничение кортежа t по множеству атрибутов Y, запись $t \in m$ следует понимать как принадлежность кортежа t основе мультимножества m; $\mathrm{Occ}(t, m)$ – количество дубликатов (экземпляров) кортежа t в мультимножестве m.

В отличие от операции проекции множеств срок, дубликаты строк, появившиеся после выполнения операции, не удаляются. Количество дубликатов каждого кортежа определяется следующей формулой:

$$\mathrm{Occ}(t, \pi_Y(m)) \overset{\text{def}}{=} \sum_{t' \in m, t'_{(Y)}=t} \mathrm{Occ}(t', m).$$

Определим операцию селекции мультимножеств. Операция селекции на мультимножестве строк m просто применяет предикат селекции к каждому кортежу и возвращает как результат такие кортежи $t \in m$, на которых предикат принимает значение истины:

$$\delta_\rho(m) \overset{\text{def}}{=} \left[t \in m \mid \rho(t) \right].$$

В зависимости от значения предиката на кортеже t все дубликаты этого кортежа или входят в результирующую таблицу или нет:

$$\mathrm{Occ}(t, \delta_\rho(m)) \overset{\text{def}}{=} \begin{cases} \mathrm{Occ}(t, m), \text{ если } \rho(t), \\ 0, \text{ иначе.} \end{cases}.$$

Рассмотрим еще операцию декартового соединения мультимножеств.

Пускай имеется два мультимножества m_1 и m_2, определенные на схемах X и Y соответственно, причем $X \cap Y = \varnothing$. Произведение $m_1 \otimes m_2$ – это мультимножество строк схемы $X \cup Y$, состоящие из всех кортежей, образовавшихся в результате конкатенации кортежей из m_1 и m_2:

$$m_1 \otimes m_2 \overset{\text{def}}{=} \left[t \text{ над } X \cup Y \;\middle|\; \exists t_1 \exists t_2 \begin{pmatrix} t_1 \in m_1, t_2 \in m_2, \\ t_{(X)} = t_1, t_{(Y)} = t_2 \end{pmatrix} \right].$$

Содержательно говоря, каждый кортеж m_1 объединяется с каждым кортежем m_2, независимо от того – дубликат это или нет; количество дубликатов находится так:

$$\mathrm{Occ}(t_1 \cup t_2, m_1 \times m_2) \overset{\text{def}}{=} \mathrm{Occ}(t_1, m_1) \cdot \mathrm{Occ}(t_2, m_2).$$

Отметим, что рассмотрение таблиц (и операций над ними) в виде мультимножеств строк было предложено в работах [2, 17], т.е. гораздо раньше, чем это было сделано в работе [6]

Мультимножества также достаточно широко используются в SQL-подобных языках [23]. Начиная со стандарта SQL:2003, в язык был введен конструктор типа MULTISET. Значения мультимножеств задаются путем использования специальной конструкции значений-мультимножеств (`multiset value constructor`).

Кроме того, для мультимножеств обеспечиваются операции объединения (`multiset union`), пересечения (`multiset intersect`) и разности (`multiset except`). Также введены новые агрегатные функции (`collect`, `fusion`, `intersect`). Естественно, что введение конструктора типа мультимножества открывает новые возможности для применения языка SQL.

Л. Либкин (L. Libkin) и Л. Вонг (L. Wong) рассматривали теоретические вопросы баз данных, основой которых выступают мультимножества. В работе [7] они строят язык запросов для мультимножеств BQL (Bag Query Language) и исследуют связь между полученным языком и так называемой вложенной реляционной алгеброй (nested relation algebra). Работа [8] посвящена изучению выразительной силы языка запросов для мультимножеств, а также обсуждению проблемы использования конструкций типа “структурная рекурсия” и “ограниченный цикл” для мультимножеств, множеств и списков.

Авторы К. Росс (K. Ross) и Ю. Стоянович (J. Stoyanovich) в своей работе [11] вводят так называемую симметрическую связь между k-арными сущностями БД как мультимножество мощности k, где k – натуральное число. Мультимножества, ограниченные по мощности (cardinality-bounded multiset), естественным образом возникают при решение реальных задач.

В статье приводятся аргументы о необходимости поддержки базами данных мультимножеств, ограниченных по мощности, и предлагаются методы, реализующие это. Авторы также описывают синтаксис расширения SQL, что позволит формулировать запросы над такими симметрическими связями.

Заключение

В статье изложены основы теории мультимножеств. Дано формальное определение мультимножеству, его характеристической функции, а также операциям объединения и пересечения мультимножеств. Построена решетка мультимножеств. Также приведены некоторые более сильные результаты о структуре семейства мультимножеств.

Во второй части работы приведен обзор библиографии по мультимножествам. Отдельно рассмотрено применения мультимножеств в табличных базах данных. Даны формальные определения операциям проекции, селекции и декартового соединения.

Литература

1. Петровський А.Б. Основные понятия теории мультимножеств / А.Б. Петровський. – М.: Едиториал УРСС, 2002. – 80 с.

2. Редько В.Н. Реляційні бази даних: табличні алгебри та SQL-подібні мови / В.Н. Редько, Ю.Й. Брона, Д.Б. Буй, С.А. Поляков. – К.: Академперіодика, 2001. – 198 с.

3. Скорняков Л.А. Элементы алгебры / Л.А. Скорняков. – М.: Наука, 1986. – 240 с.

4. Davey B.A. Introduction to Lattice and Order / B.A. Davey, H.A. Priestly. – Cambridge: Cambridge University Press, 1990. – 248 p.

5. Биркгоф Г. Теория решеток / Г. Биркгоф. – М.: Наука, 1984. – 564 с.

6. 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.

7. Libkin L. Query Language for Bags and Aggregates Function / L. Libkin, L. Wong // Journal of Computer and System Sciences. – 1997. – Vol. 55, No. 1. – P. 241-272.

8. Libkin L. Some Properties of Query Language for Bags / L. Libkin, L. Wong // Proceedings of 4th International Workshop on Database Programming Languages. – New York, 1993. – P. 97-114.

9. Lloyd J. Programming with Multisets / J.W. Lloyd // Department of Computer Science University of Bristol, 1998.

10. Lloyd J. Programming with Sets and Multisets / J.W. Lloyd // Department of Computer Science University of Bristol, 1998.

11. Ross K. Symmetric relations and cardinality-bounded multisets in database systems / K.A. Ross, J. Stoyanovich // Very Large Database Endowment: international conference, August 31-September 03, 2004, Toronto, Canada: proceedings. – 2004. – Vol. 30. – P. 912-923.

12. Singh D. An Overview of the Applications of Multisets / D. Singh, A.M. Ibrahim, T. Yohanna, J.N. Singh // Novi Sad Journal of Mathematics. – 2007. – Vol. 37, No. 2. – P. 73-92.

13. Syropoulos A. Mathematics of Multisets / Apostolos Syropoulos // Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View, number 2235 in Lecture Notes in Computing Since. – Berlin: Springer-Verlag, 2001. – P. 347-358.

14. Барендрегт Х. Ламбда-исчисление. Его синтаксис и семантика: [пер. с англ.] / Х. Барендрегт. – М.: Мир, 1985. – 606 с.

15. Башкин В.А. Подобие обобщенных ресурсов в сетях Петри [Электронный ресурс] / В.А. Башкин, И.А. Ломазова. – Режим доступа: http://lvk.cs.msu.su/files/mco2005/bashkin.pdf.

16. Knuth D.E. Context-Free Multilanguages / D.E. Knuth // Theoretical Studies in Computer Science. – Academic Press, 1992. – P. 1-13.

17. Буй Д.Б. Композиційна семантика SQL-подібних мов: мультимножини, рядки, впорядковані таблиці / Д.Б. Буй, С.А. Поляков // Вісник Київського університету. Сер.: фіз.-мат. науки. – 1999. – Вип. 2. – С. 183-194.

18. Гарсиа-Молина Г. Системы баз данных: [полный курс.: пер. с англ.] / Г. Гарсиа-Молина, Дж. Ульман, Дж. Уидом. – М.: Вильямс, 2004. – 1088 с.

19. Кнут Д. Искусство программирования: [2 том, 3-е изд.: пер. с англ.] / Д. Кнут. – М.: Вильямс, 2000. – 832 с.

20. Кузнецов С.Д. Концептуальное проектирование реляционных баз данных с использованием языка UML [Электронный ресурс] / С.Д. Кузнецов. – Режим доступа: ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic9/Doc13.doc.

21. Малинецкий Г.Г. Вычисления на ДНК. Эксперименты. Модели. Алгоритмы. Инструментальные средства [Электронный ресурс] / Г.Г. Малинецкий, С.А. Науменко. – Режим доступа: http://www.keldysh.ru/papers/2005/prep57/ prep2005_57.html.

22. Мальцев А.И. Алгебраические системы / А.И. Мальцев. – М.: Наука, 1970. – 392 с.

23. Наиболее интересные новшества в стандарте SQL:2003 [Электронный ресурс] / Режим доступа: http://www.nestor.minsk.by/sr/2004/03/40331.html.

24. Albert J. Algebraic properties of bag data types / J. Albert // Seventeenth International Conference on Very Large Data Bases. – Barcelona, Spain, 1991. – P. 211–219.

25. Петровський А.Б. Пространства множеств и мультимножест / А.Б. Петровський. – М.: Едиториал УРСС, 2003. – 248 с.

26. Blizard W.D. The Development of Multiset Theory / W.D. Blizard // Notre Dame Journal of Formal Logic. – 1989. – Vol. 30, No. 1. – P. 36-66.

27. Сети Петри [Электронный ресурс] / Режим доступа: http://www.iacp.dvo.ru/lab_11/otchet/ot2000/pn3.html#top.

28. Славин О.А. Использование мультимножеств в распознавании символов [Электронный ресурс] / О.А. Славин. – Режим доступа: ftp://ftp.dol.ru/pub/users/cgntv/download/sbornic/sbornic9/Doc13.doc.

29. Bonchis C. Information Theory over Multiset / C. Bonchis, C. Izbasa, G. Ciobanu // Research Institute "re-Austria", Institute of Computer Science, 2005.

30. Скорняков Л.А. Элементы теории структур / Л.А. Скорняков. – М.: Наука, 1982. – 160 с.

Поступила в редакцию 2.03.2010

Рецензент: д-р техн. наук, проф. Г.Н. Жолткевич, Национальный университет им. В.Н. Каразина, Харьков, Украина.

ТЕОРІЯ МУЛЬТИМНОЖИН: БІБЛІОГРАФІЯ, ЗАСТОСУВАННЯ В ТЕОРІЇ ТАБЛИЧНИХ БАЗ ДАНИХ

Д.Б. Буй, Ю.О. Богатирьова

У статті розглядаються елементи теорії мультимножин. Надано формальне визначення мультимножини, її характеристичній функції, а також операцій об'єднання і перетину мультимножин. Будуються грати мультимножин. Наводяться результати про структуру сімейства мультимножин. Розкривається бібліографія по мультимножинам. Також розглядається застосування мультимножини в табличних базах даних. Дані формальні визначення операціям проекції, селекції і декартового з'єднання.

Ключові слова: мультимножина, бази даних, повна решітка, об’єднання мультимножин, перетин мультимножин.

MULTISET’S THEORY: BIBLIOGRAPHY, APPLICATION IN TABLE DATA BASES THEORY

D.B. Buy, J.O. Bogatyreva

The elements of multisets theory are considered. Formal definition of multisets, its characteristic function and operations of association and crossing of multisets is given. The lattice of multisets is constructed. The results on multisets family structure are described. The analysis of bibliography on multisets is conducted. Application multisets in relation data bases are examined.. Formal definitions for projection, selection and cartesian connection operations are proposed.

Keywords: multiset, data bases, complete lattice, multiset union, multiset intersection.

Буй Дмитрий Борисович – д-р физ.-мат. наук, ст. научн. сотр. НИС «Проблем программирования», заместитель декана по научной работе факультета кибернетики, Киевский национальный университет им. Тараса Шевченко, Киев, Украина, e-mail: buy@unicyb.kiev.ua.

Богатырева Юлия Александровна – аспирантка кафедры кибернетики, национальный университет им. Тараса Шевченко, Киев, Украина, e-mail: j_bogatyreva@ukr.net.