ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

УДК 681.3
DOI 10.21685/2072-3059-2016-3-1

П. П. Макарычев, Д. В. Артамонов

МОДЕЛЬНЫЕ ПРЕДСТАВЛЕНИЯ ДАННЫХ НА ОСНОВЕ ПРЯМОГО ТЕНЗОРНОГО ИСЧИСЛЕНИЯ

Аннотация.

Актуальность и цели. В информационных системах анализ данных, как правило, выполняется с точки зрения множественности измерений. Концептуально модельные представления строятся на основе таких понятий, как объект, класс, отношение. Для формализованного описания данных используются матричное исчисление, алгебра кортежей, тензорное исчисление.

Материалы и методы. Данные могут быть представлены в виде обобщенной тензорной модели, которая может интерпретироваться в различные предметные области.

Результаты. Разработана модель представления классов, объектов и отношений в предметном пространстве информационной системы. Модель отличается прямой тензорной формой записи отношений, арифметических и логических операций. Классы (объекты) предметного пространства задаются в виде диад, характеризующих предметное пространство. Предметные пространства «Звезда», «Снежинка» и «Созвездие» задаются в виде тензоров второго порядка – неупорядоченной совокупности диад, ассоциированных с классами. Разработано тензорное представление запросов к реляционным структурам данных, арифметических и логических операций. Рассмотрены процедуры обработки и анализа данных. Приведен пример реализации кластерного анализа данных.

Выводы. Предложено тензорное модельное представление реляционных структур данных. Представление обеспечивает формализованное описание простых и сложных запросов к базам данных, арифметических и логических операций, процедур обработки и анализа данных.

Ключевые слова: модель данных, тензорное исчисление, диада, инвариант, тензорси соотношения, тензорси предиката.

P. P. Makarychev, D. V. Artamonov

MODEL-BASED PRESENTATIONS OF DATA ON THE BASIS OF DIRECT TENSOR CALCULUS

Abstract.

Background. In information systems, data analyses are usually performed from the point of view of the plurality of measurements. Conceptual model representations are based on concepts such as an object, a class, a relation. For a formalized description of data one uses matrix calculus, algebra of tuples and tensor calculus.

Materials and methods. Data can be presented in the form of a generalized tensor model, which may be interpreted in different subject areas.

Results. The authors developed a model of representation of classes, objects and relationships of the subject space of information systems. The model has a direct tensor notation of relations, arithmetic and logical operations. Classes (objects) of the subject space are defined in the form of dyads that characterize the object space. The subject spaces "Star", "Snowflake" and "Constellation" are specified in the form of tensors of the second order – disordered aggregate dyads associated with the classes. The researchers developed a tensor representation of queries in relational data structures, arithmetic and logical operations, reviewed procedures for data processing and analysis, and described an example of cluster analysis implementation.

Conclusions. The authors have suggested the tensor model representation of relational data structures. The representation provides formalized description of simple and complicated queries to databases, arithmetic and logical operations, procedures of data processing and analysis.

Key words: data model, tensor calculus, dyad, invariant, tensors ratio, tensors predicate.

Введение

Как известно, тензорное исчисление представляет собой математическое средство, с помощью которого формулируются инвариантные соотношения между величинами изучаемых объектов. В теории информационных систем признание принципа инвариантности означает, что объекты существуют независимо от субъективных систем координат (измерений) предметного пространства, заданных наблюдателем [1, 2]. В современном тензорном исчислении используются три формы записи соотношений: координатная (арифметическая, индексная), матричная и безындексная (прямая). При координатной форме тензор представляется в виде набора чисел $T^{\alpha...\beta}_{\gamma...\delta}$, в котором индексы $\alpha...\delta$ пробегают значения от 1 до $n \geq 1$. Эта форма очень практична и широко распространена [2]. При бескоординатной форме описания тензора индексы не пробегают целочисленные значения, а рассматриваются как метки, несущие информацию о типе тензора и возможных операциях над тензором [3]. Матричное представление предполагает соблюдение соответствия между размерностью матрицы и рангом тензора, между строками, столбцами и индексами [4].

Все три формы практически равнозначны. Однако матрицы сами по себе не соответствуют структурам данных информационных систем, представленным в виде сущностей реляционных баз данных. Безындексная форма концептуально более предпочтительна в том смысле, что тензоры естественно рассматривать не как наборы чисел, а как некоторые линейные отображения. Безындексная форма представления тензоров обеспечивает несложное преобразование к индексной и матричной формам представления. В данной работе приводится обоснование применимости безындексной формы записи тензорных модельных представлений данных, алгебраических и логических операций.

Тензорное представление предметного пространства

На рис. 1 приведена структура обобщенной реляционной модели данных «Снежинка» (хранилища данных), содержащая измерения и факты [5].

В качестве измерений модель содержит сущности (классы): «Кафедра», «Факультет», «Научный Руководитель», «Научная Специальность», «Аспирант» и «Показатель». Сущность «Запись» содержит факты о деятельности аспирантов. Каждый экземпляр сущности соответствует записи оценки деятельности аспиранта по одному из показателей. Факт фиксируется с указанием номера записи n и времени выполнения записи t. Линии связи между сущностями отражают функциональные зависимости между ключевыми атрибутами сущностей $(n,t) \to h, p;\ \ h \to r, s;\ r \to k, d$.

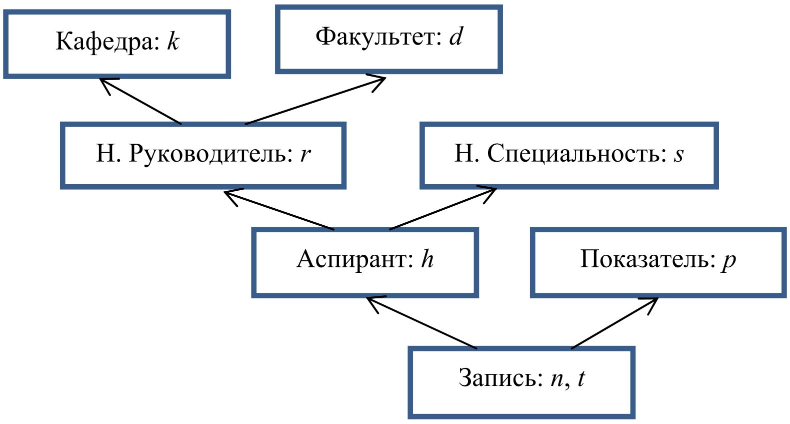

Рис. 1. Структура модели данных «Снежинка»

Для построения тензорного модельного представления расположим и зафиксируем классы предметного пространства в следующей последовательности: $A = (h, p, r, s, k, d, (n,t)) = (h, p, r, s, k, d, z)$.

Предположим, что $\mathbf{h}, \mathbf{p}, \mathbf{r}, ..., \mathbf{z}$ и $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_7$ являются тензорами первого ранга (векторами), характеризующими классы и отношения классов предметного пространства информационной системы. Размерности векторов равны. Векторы-столбцы $\mathbf{h}, \mathbf{p}, \mathbf{r}, ..., \mathbf{z}$ пространства T_1 характеризуют функциональные отношения между классами, а векторы-строки $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, ..., \mathbf{e}_7$ пространства T_2 – отношение заданного порядка в последовательности A. В этом случае для сущностей измерений $\mathbf{h}, \mathbf{p}, \mathbf{r}, ..., \mathbf{d}$ и сущности фактов $\mathbf{z}$ рассматриваемого предметного пространства тензоры первого ранга имеют вид (рис. 1):

$$\mathbf{h}^{\mathrm{T}} = (m_{11}\ \ 0\ \ m_{31}\ \ m_{41}\ \ 0\ \ 0\ \ 0);\ \mathbf{e}_1 = (n_1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0);$$

$$\mathbf{p}^{\mathrm{T}} = (0\ \ m_{22}\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0);\ \mathbf{e}_2 = (0\ \ n_2\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0);$$

$$\mathbf{r}^{\mathrm{T}} = (0\ \ 0\ \ m_{33}\ \ 0\ \ m_{53}\ \ m_{63}\ \ 0);\ \mathbf{e}_3 = (0\ \ 0\ \ n_3\ \ 0\ \ 0\ \ 0\ \ 0);$$

$$\mathbf{s}^{\mathrm{T}} = (0\ \ 0\ \ 0\ \ m_{44}\ \ 0\ \ 0\ \ 0);\ \mathbf{e}_4 = (0\ \ 0\ \ 0\ \ n_4\ \ 0\ \ 0\ \ 0);$$

$$\mathbf{k}^{\mathrm{T}} = (0\ \ 0\ \ 0\ \ 0\ \ m_{55}\ \ 0\ \ 0),\ \ \mathbf{e}_5 = (0\ \ 0\ \ 0\ \ 0\ \ n_5\ \ 0\ \ 0);$$

$$\mathbf{d}^{\mathrm{T}} = (0\ \ 0\ \ 0\ \ 0\ \ 0\ \ m_{66}\ \ 0),\ \ \mathbf{e}_6 = (0\ \ 0\ \ 0\ \ 0\ \ 0\ \ n_6\ \ 0);$$

$$\mathbf{z}^{\mathrm{T}} = (m_{17} \quad m_{27} \quad 0 \quad 0 \quad 0 \quad 0 \quad m_{77})\,;\ \mathbf{e}_7 = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad n_7)\,. \qquad (1)$$

Значение каждого элемента $m_{ii}, i = 1,2,...,7,$ вектора пространства T_1 равно количеству байт, необходимых для регистрации атрибутов экземпляра сущности (объекта класса). Значения элементов $m_{ij}, i \neq j,$ $i = 1,2,...,7,$ $j = 1,2,...,7,$ векторов пространства T_1 равны количеству байт, необходимых для регистрации внешних ключей сущностей модели данных. Векторы пространства T_2 содержат элементы n_j, $j = 1,2,...,7$, значения которых определяются количеством экземпляров сущностей (объектов класса). Каждую упорядоченную пару векторов ($\mathbf{h} \otimes \mathbf{e}_1, \mathbf{p} \otimes \mathbf{e}_2, ..., \mathbf{z} \otimes \mathbf{e}_7$) из пространств T_1, T_2 можно рассматривать как единое целое (диаду векторов), представляющее собой элемент множества прямого декартового произведения двух векторных пространств [1]. Например, для сущности «Научный Руководитель» диада имеет следующий вид:

$$\mathbf{r} \otimes \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ m_{33} \\ 0 \\ m_{53} \\ m_{63} \\ 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & n_3 & 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_{33}n_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_{53}n_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & m_{63}n_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

где $m_{33}n_3$ – количество байт в записи данных о научном руководителе, $m_{53}n_3$, $m_{63}n_3$ – количество байт, содержащихся в записях внешних ключей.

При этом предметное пространство может быть определено как тензор второго ранга, заданный суммой диад:

$$\mathbf{V} = \mathbf{h} \otimes \mathbf{e}_1 + \mathbf{p} \otimes \mathbf{e}_2 + \mathbf{r} \otimes \mathbf{e}_3 + ... + \mathbf{z} \otimes \mathbf{e}_7\,. \qquad (2)$$

Таким образом, тензор $\mathbf{V}$ является конечной неупорядоченной совокупностью упорядоченных пар тензоров первого ранга. След каждой диады в (2) равен скалярному произведению тензоров первого ранга:

$$\mathrm{tr}(\mathbf{a} \otimes \mathbf{e}_i) = \mathbf{a} \cdot \mathbf{e}_i,\ \mathbf{a} \in \{\mathbf{h}, \mathbf{p}, \mathbf{r}, ..., \mathbf{z}\},\ i = 1,2,...,7,$$

составляющих диаду.

Для модели данных, представленной на рис. 1, тензор второго ранга, характеризующий предметное пространство в матричной форме, имеет вид

$$\mathbf{V} = \begin{bmatrix} m_{11}n_1 & 0 & 0 & 0 & 0 & 0 & m_{17}n_7 \\ 0 & m_{22}n_2 & 0 & 0 & 0 & 0 & m_{27}n_7 \\ m_{31}n_1 & 0 & m_{33}n_3 & 0 & 0 & 0 & 0 \\ m_{41}n_1 & 0 & 0 & m_{44}n_4 & 0 & 0 & 0 \\ 0 & 0 & m_{53}n_3 & 0 & m_{55}n_{55} & 0 & 0 \\ 0 & 0 & m_{63}n_3 & 0 & 0 & m_{66}n_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & m_{77}n_7 \end{bmatrix}.$$

Таким образом, след тензора второго ранга определяется формулой

$$\mathrm{tr}(\mathbf{V}) = \mathrm{tr}(\mathbf{h}\cdot\mathbf{e}_1 + \mathbf{p}\cdot\mathbf{e}_2 + \mathbf{r}\cdot\mathbf{e}_3 + ... + \mathbf{z}\cdot\mathbf{e}_7) = \sum\nolimits_i^7 m_{ii}n_i.$$

Эта важная характеристика тензора второго ранга является одним из инвариантов и обладает следующим свойством: $\mathrm{tr}(\mathbf{V}) = \mathrm{tr}(\mathbf{V}^{\mathrm{T}})$.

Пусть определен второй вариант предметного пространства типа «Звезда» с тем же набором объектов (сущностей). Данное пространство имеет такую же размерность, но число уровней иерархии равно двум [5]. Для этого пространства тензор имеет вид

$$\mathbf{W} = \begin{bmatrix} m_{11}n_1 & 0 & 0 & 0 & 0 & 0 & m_{17}n_7 \\ 0 & m_{22}n_2 & 0 & 0 & 0 & 0 & m_{27}n_7 \\ m_{31}n_1 & 0 & m_{33}n_3 & 0 & 0 & 0 & 0 \\ m_{41}n_1 & 0 & 0 & m_{44}n_4 & 0 & 0 & 0 \\ m_{51}n_1 & 0 & 0 & 0 & m_{55}n_5 & 0 & 0 \\ m_{61}n_1 & 0 & 0 & 0 & 0 & m_{66}n_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & m_{77}n_7 \end{bmatrix}.$$

Предположим, что для тензора второго ранга $\mathbf{V}$ определитель $\det\mathbf{V} \neq 0$ и имеется обратный тензор $\mathbf{V}^{-1}$ такой, что $\mathbf{V}^{-1}\cdot\mathbf{V} = \mathbf{V}\cdot\mathbf{V}^{-1} = \mathbf{E}$. В этом случае, используя скалярное произведение двух тензоров, можно определить тензор преобразования предметных пространств. Например:

$$\mathbf{N} = \mathbf{W}\cdot\mathbf{V}^{-1}.$$

При этом скалярное произведение тензора второго ранга $\mathbf{N}$ на тензор первого ранга представляет собой линейное преобразование объектов одного пространства в объекты другого пространства. Это линейное преобразование осуществляется следующим образом:

$$\mathbf{b} = \mathbf{N}\cdot\mathbf{a}, \tag{3}$$

где $\mathbf{a} \in \{\mathbf{h}, \mathbf{p}, \mathbf{r}, ..., \mathbf{z}\}$, $\mathbf{b} \in \left\{\mathbf{h}^*, \mathbf{p}^*, \mathbf{r}^*, ..., \mathbf{z}^*\right\}$.

Рассмотренный выше подход к заданию тензорного модельного представления данных можно использовать при анализе различных информационных и технических систем. Введенные в рассмотрение тензоры второго ранга $\mathbf{V}, \mathbf{W}$ характеризуют не только структуру предметного пространства информационной системы. Одновременно тензоры $\mathbf{V}, \mathbf{W}$ содержат компоненты количественных характеристик: общее число экземпляров в одной сущности (количество строк в таблице), объем хранимых данных (атрибутов, байт, бит) в одном экземпляре сущности (строке таблицы). Например, пусть диада $\mathbf{r}\otimes\mathbf{e}_3$ образована из тензоров следующего вида:

$$\mathbf{r}^{\mathrm{T}} = (0 \quad 0 \quad m_{33} \quad 0 \quad m_{53} \quad m_{63} \quad 0) \quad \mathbf{e}_3 = (0 \quad 0 \quad n_3 \quad 0 \quad 0 \quad 0 \quad 0),$$

где $m_{33} = 233$, $m_{53} = 10$, $m_{63} = 10$, $n_3 = 17$.

Очевидно, что след диады имеет вид

$$tr(\mathbf{r} \otimes \mathbf{e}_3) = m_{3,3}n_3 = 3961 \text{ байт}$$

и общий объем данных, соответствующий в структуре данных сущности «Научный Руководитель» (см. рис. 1), равен

$$Q_r = \sum_{j=1}^{7} (r \otimes e_3)_{j,3} = m_{3,3}n_3 + m_{5,3}n_3 + m_{6,3}n_3 = 4301 \text{ байт.}$$

Аналогично можно определить и объемы данных для остальных шести диад (объектов) модели данных «Снежинка». В результате получим общий объем данных в структуре предметного пространства «Снежинка»:

$$Q_V = \sum_{i}^{7}\sum_{j}^{7} V_{i,j} = \sum_{i}^{7}\sum_{j}^{7} m_{i,j}n_j = 5.521 \cdot 10^4 \text{ байт.} \tag{4}$$

При этом объем хранимых данных в структуре предметного пространства «Звезда» составит

$$Q_W = \sum_{i}^{7}\sum_{j=1}^{7} W_{i,j} = \sum_{i}^{7}\sum_{j=1}^{7} m_{i,j}n_j = 5.663 \cdot 10^4 \text{ байт}. \tag{5}$$

Из выражений (4), (5) видно, что объем данных, хранимых в структуре «Звезда», увеличился.

Тензорное модельное представление объектов

Каждую сущность предметного пространства можно представить в виде суммы диад. Например, для сущности «Аспирант» сумма диад

$$\mathrm{q}_h^{r,s} = h \otimes e_1 + r \otimes e_3 + s \otimes e_4,$$

где q – символ сущности; h, r, s – метки, отражающие элементы диад, значения которых могут быть отличны от нуля.

Таким образом, предметное пространство, соответствующее модели данных «Снежинка», как тензор второго ранга $\mathbf{V}$, может быть задано неупорядоченной последовательностью сумм диад:

$$\mathbf{V} = \mathbf{q}_h^{r,s} + \mathbf{q}_p + \mathbf{q}_r^{k,d} + \mathbf{q}_s + \mathbf{q}_k + \mathbf{q}_d + \mathbf{q}_z^{a,p}. \tag{6}$$

В выражении (4) верхние и нижние символы в неупорядоченной последовательности диад не являются индексами. Это метки, которые определяют структуры диад и возможные действия с диадами. Нижние метки соответствуют первичным ключам сущностей, верхние – внешним ключам сущностей. В результате выполнения операции преобразования $\mathbf{N}$ линейного предметного пространства (6) имеем

$$\mathbf{W} = \mathbf{NV} = \mathbf{q}_h^{r,s,k,d} + \mathbf{q}_p + \mathbf{q}_r + \mathbf{q}_s + \mathbf{q}_k + \mathbf{q}_d + \mathbf{q}_z^{h,p}. \tag{7}$$

Из выражения (7) следует, что линейное преобразование $\mathbf{N}$ связано с уменьшением уровней иерархии за счет операции денормализации применительно к сущности «Аспирант», представленной в выражении (7) суммой из 5 диад $q_h^{r,s,k,d}$ (рис. 2).

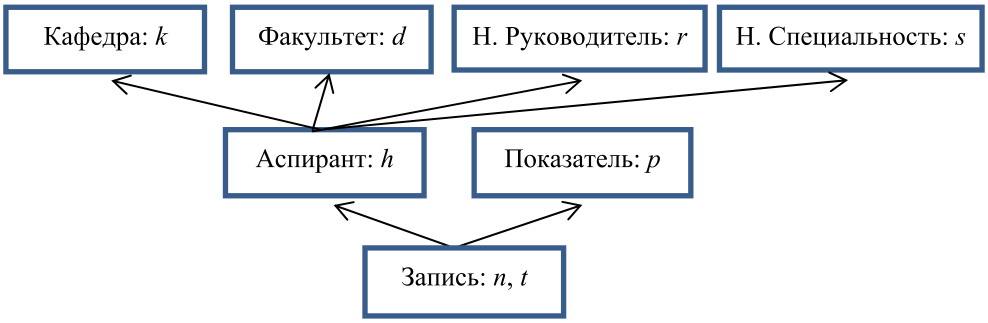

Рис. 2. Структура модели данных

Рассмотрим применение тензорного исчисления при разработке модельного представления и анализа данных на уровне сущностей. Если воспользоваться понятием тензорси (3), то класс объектов «Аспирант» можно представить в виде функциональных графиков. При этом представлении классов (сущностей) определяется закон формирования функциональных графиков $\langle h,r\rangle$ и $\langle h,s\rangle$ из множеств H,R,S, где H – множество аспирантов (экземпляров сущности), R – множество руководителей, S – множество научных специальностей. Согласно этому подходу графики q_h^r, q_h^s, соответствующие суммам диад, не содержат пар с одинаковыми нижними и различными верхними метками. В этом представлении сумму диад $q_h^{r,s}$ можно рассматривать как объединение двух функциональных графиков $q_h^{r,s} = q_h^r \cup q_h^s$, сумму диад $q_a^{r,s,k,d}$ – четырем функциональным графикам.

При проектировании модели данных содержание каждой диады должно быть раскрыто как можно более полно. Например, для сущности «Кафедра» из рассматриваемой модели данных диада $\mathbf{q}_k = \mathbf{k} \otimes \mathbf{e}_1$ может быть суммирована с диадами, отражающими функциональную зависимость неключевых атрибутов от первичного ключа $\langle k,a_1\rangle,\langle k,a_2\rangle,\ldots,\langle k,a_m\rangle$:

$$\mathbf{q}_k^{a_1,a_2,\ldots,a_m} = \mathbf{k} \otimes \mathbf{e}_1 + \mathbf{a}_1 \otimes \mathbf{e}_2 + \mathbf{a}_2 \otimes \mathbf{e}_3 + \ldots + \mathbf{a}_m \otimes \mathbf{e}_{m+1},$$

где $\mathbf{k}$ – вектор первичного ключа; $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ – векторы неключевых атрибутов сущности.

Следует отметить, что математический объект $\mathbf{q}_k^{a_1,a_2,\ldots,a_m}$ в алгебре кортежей рассматривается как C-кортеж, представляющий собой множество элементарных кортежей $[k,a_1,a_2,\ldots,a_m]$ одной и той же размерности или многоместное отношение [5]. На основе элементарных кортежей и C-кортежей в алгебре кортежей вводятся и другие структуры данных и операции: операций алгебры множеств, операции с атрибутами, включая перестановку атрибутов, добавление фиктивного атрибута и элиминацию атрибута. Аналогичные операции могут быть осуществлены и использованием тензорного представления структур данных.

Таким образом, сущность в модели данных можно рассматривать как сумму диад или кортеж функциональных графиков:

$$q_r^{k,d}(a_r^1, a_r^2, ..., a_r^m) = q_r^{k,d} \cup q_r^{a_1, a_2, ..., a_m} = q_r^{k,d,a_1.a_2,...,a_m}.$$

В выражениях (6), (7) возможные типы атрибутов: Int (Целое число), Real (Действительное число), Date (формат MM/DD/YY), Text (Строка), List (Список значений, разделенных запятой), Command (Команда – выполняемая строка).

Над диадами можно осуществлять преобразования, описываемые тензорси соотношений и тензорси предикатов [3]. Например, операцию сложения значений атрибутов $b = a_1 + a_2$ можно представить как тензорси $g1_{a_1,a_2}^b = \left(+_{a_1,a_2}^b\right)$, а операцию умножения значений атрибутов $b = a_1 \cdot a_2$ – тензорси $g2_{a_1,a_2}^b = \left(\times_{a_1,a_2}^b\right)$. При этом запись операции сложения атрибутов диады с использованием суффиксной формы будет выглядеть следующим образом:

$$q_h^{r,s} q_h^{a_1,a_2,a_3,a_4} = q_h^{r,s} q_h^{a_1,a_2,a_3} g1_{a_1,a_2}^{a_4} = q_h^{r,s} q_h^{a_1,a_2,a_3}\left(+_{a_1,a_2}^{a_4}\right).$$

Аналогично можно выполнять запись и логических операций типа $\vee, \wedge$. Например: $\left(\vee_{a_i,a_j}^{a_k}\right)$, $\left(\wedge_{a_i,a_j}^{a_k}\right)$, $i, j, k \in \{1, 2, ..., m\}$ и т.д.

В работе [5] введен тензорси предикатов, обозначающий множество действий над множеством пар атрибутов $\langle a, b \rangle$. Результаты действий принимают значения из множества $\Re = \{0,1\}$. Следуя работе [5], будем задавать тензорси предикатов в виде символа γ, снабженного метками и обозначающего множество вычислений над множествами пар атрибутов. Например, запись $\gamma_{a,b}^{\Re}$ эквивалента записи $(\le_{a,b}^{\Re})$. На месте символа «$\le$» в тензорси предикатов могут быть символы «<», «=» и др. Эти вычисления характерны для двухместных предикатов, имеющих значение «истина» или «ложь».

С использованием принятых соглашений простой запрос на формирование списка научных руководителей, имеющих научную степень «кандидат технических наук, «*ктн*», можно записать в следующем виде:

$$q_h^{r,s} q_h^{a_1,a_2,"ктн",a_4} = q_r^{k,s} q_h^{a_1,a_2,a_3,a_4}\left(=_{a_3,"ктн"}^{\Re}\right),$$

где суффикс, отражающий формирование факта наличия в составе научных руководителей лиц со степенью «кандидат технических наук».

Область применения

Эффективная работа с реляционными базами данных осуществляется на основе запросов. Для описания запросов, как правило, используют язык QBE (Query By Example, язык запросов по образцу) или язык SQL (Structured Query Language, структурированный язык запросов). Однако данные языки предназначены в основном для управления данными конкретной базы данных. Вместе с тем на начальном этапе проектирования базы данных целесообразно проведение анализа объектов и операторов, обеспечивающих доступ к данным пользователя формализованными методами. Один из эффективных подходов к этому анализу осуществляется на основе тензорного исчисления в форме безындексной записи соотношений.

Рассмотрим возможности прямой формы записи тензоров на примере кластерного анализа неиерархическим методом. Предположим, что имеется множество объектов $A: q_k^{a_i}$, $k = 1,2,...,m$, $i = 1,2,...,p$, и множество центров кластеров $B: q_l^{a_i}$, $l = 1,2,...,n$.

Первый шаг. Для фиксированного $k^* \in \{1,2,...,m\}$ формируем тензор разности между атрибутами центров кластеров и атрибутами выбранного объекта (функциональный график):

$$q_l^{b_i} = q_{k^*}^{a_i} - q_l^{a_i}, \; l = 1,2,...,n, \; i = 1,2,...,p.$$

Второй шаг. Вычисляем свертку тензоров и формируем тензор $q_l^{a_i,c}$:

$$q_l^c = q_l^{b_i} \circ q_l^{b^i}, \; q_l^{a_i,c} = q_l^{a_i} \cup q_l^c, \; l = 1,2,...,n.$$

Третий шаг. Определяем расстояния d от объекта $q_{k^*}^{a_i}$ до центров всех кластеров $q_l^{a_i}$, $l = 1,2,...,n$, и формируем тензор $q_l^{a_i,c,d}$. При этом

$$q_l^{c,d} = q_l^c \left(sqrt\,_c^d \right), \; l = 1,2,...,n,$$

где $sqrt\,_c^d$ – функция извлечения корня квадратного ($d = \sqrt{c}$).

Четвертый шаг. Определяем минимальное расстояние от объекта до центра кластера $q_l^{a_i,c,d}$ из множества B:

$$q_{k^*}^{a_i,e} = q_l^d \left(\min\nolimits_{D(d)}^e \right),$$

где $D(d)$ – домен значений атрибута d в функциональном графике q_l^d.

Пятый шаг. Размещаем выбранный объект в кластере l. Другими словами, определяем в функциональном графике q_l^d значение метки l, для которой $d = e$:

$$q_{k^*}^{l,a_i} = q_{k^*}^{e,a_i} q_l^{c,d} \left(match\,_{(e,q_l^d)}^{\,l} \right),$$

где $\mathrm{match}_{(e,d)}^{l}$ – функция определения значения метки центра кластера l, до которого расстояние от объекта $q_{k^*}^{e,a_i}$ минимально, $e = match(e, q_l^d)$.

Рассмотрим реализацию кластерного анализа в среде математического пакета Mathcad. Поскольку пакет не поддерживает тензорного представления, то воспользуемся матричной формой. Пусть A – множество объектов $q_k^{a_i}$, $k = 1,2,...,N$, $i = 1,2,...,6$; B – множество заданных центров кластеров $q_l^{a_i}$, $l = 1,2,3$, $i = 1,2,...,6$. Множества A, B заданы таблично. Ввод выполнен с применением операции импорта из базы данных. Листинг (рис. 3) содержит множество B и фрагмент множества A.

A :=

	1	2	3	4	5	6
1	4.9	5.1	4.3	6.2	5.1	3.1
2	6.8	7.2	6.2	8.3	7.1	5.2
3	5.2	5.2	4.1	6.1	5.2	3
4	6.7	7.2	7.1	8.1	7.3	7.1
5	4.9	5.2	4.2	6.2	5.1	...

B :=

	1	2	3	4	5	6
1	6.7	6.7	6.6	6.6	6.6	6.5
2	5	5	5	5	5	5
3	7	7	7	7	7	...

Рис. 3. Импорт данных

Решение задачи размещения объектов в кластерах (листинг) приведено на рис. 4. Программа разработана с использованием встроенных функций пакета Mathcad.

Результаты кластерного анализа приведены в листинге вывода данных (рис. 5). В этом листинге столбцы с номерами 1–6 содержат значения атрибутов, столбец 7 – номера кластеров.

Приведенный пример кластерного анализа является не единственным. Тензорное представление данных может быть использовано в различных методах интеллектуального анализа данных и методах искусственного интеллекта.

Заключение

Рассмотренные выше модельные представления данных позволяют сделать следующие выводы.

1. При проектировании реляционных баз данных может эффективно использоваться безындексная форма записи тензорных модельных представлений предметных пространств, классов и объектов.

2. С использованием безындексной формы записи тензоров возможна запись арифметических и логических операций над тензорами первого ранга, представляющими объекты предметного пространства информационной системы.

```
Pr(A,B) := | "определение количества объектов и центров"
           | m ← rows(A)
           | n ← rows(B)
           | "определение количества параметров объектов"
           | p ← cols(B)
           | for k ∈ 1..m
           |   | for q ∈ 1..n − 1
           |   |   | "выделение объекта и центра кластера"
           |   |   | U ← submatrix(A,k,k,1,p)
           |   |   | W ← submatrix(B,q,q,1,p)
           |   |   | "вычисление расстояний объект-центры"
           |   |   | V ← U − W
           |   |   | S_q ← √(V·V^T)
           |   | "вычисление минимального расстояния"
           |   | D_k ← min(S)
           |   | "размещение объекта в кластере"
           |   | A_k,p+1 ← |match(D_k,S)|
           | A
```

Рис. 4. Размещение объектов в кластерах

$Pr(A,B) =$

	1	2	3	4	5	6	7
1	4.9	5.1	4.3	6.2	5.1	3.1	2
2	6.8	7.2	6.2	8.3	7.1	5.2	1
3	5.2	5.2	4.1	6.1	5.2	3	2
4	6.7	7.2	7.1	8.1	7.3	7.1	3
5	4.9	5.2	4.2	6.2	5.1	3.1	2
6	4.9	5.1	4.3	6.3	5.2	3.2	2
7	6.8	7.3	6.5	8.3	7.1	5.2	1
8	6.5	7.1	6.2	8.2	7.2	5.1	1
9	5.2	5.2	4.3	6.1	5.3	4.9	2
10	7.2	6.7	6.5	7.3	7.5	6.7	3
11	7.8	8.1	8.2	7.9	7.5	7.7	3
12	4.7	4.1	3.7	4.7	4.4	3.9	2
13	8.1	7.7	6.7	6.6	8.4	8.1	3
14	6.6	6.3	5.9	6.7	5.7	6.7	1
15	4.4	4.7	5.3	5.5	5.8	4.9	2
16	5.1	5.5	5.4	5.1	5.2	5.3	2

Рис. 5. Вывод результатов

3. Безындексная форма записи тензоров обеспечивает формализованную запись двухместных предикатов над множеством пар атрибутов объектов предметного пространства.

4. На основе безындексной формы тензорных модельные представлений данных можно анализировать параметры структур данных, реализуемость простых и сложных запросов к реляционным базам данных.

Список литературы

1. **Арменский, А. Е.** Тензорные методы построения информационных систем / А. Е. Арменский. – М. : Наука, 1989. – 152 с.
2. **Крон, Г.** Тензорный анализ сетей : пер. с англ. / Г. Крон ; под ред. Л. Т. Кузина, П. Г. Кузнецова. – М. : Сов. радио, 1978. – 720 с.
3. **Чижухин, Г. Н.** Тензорная методология в дискретной системотехнике / Г. Н. Чижухин, Ю. Г. Бочкарева. – Пенза : Изд-во ПГУ, 2006. – 184 с.
4. **Кулик, Б. А.** Алгебраический подход к интеллектуальной обработке данных и знаний / Б. А. Кулик, А. А. Зуенко, А. Я. Фридман. – СПб. : Изд-во Политехн. ун-та, 2010. – 235 с.
5. **Макарычев, П. П.** Построение моделей классов и объектов с применением тензорной методологии / П. П. Макарычев, Н. А. Попова // Университетское образование : сб. ст. XVII Международная науч.-метод. конф. – Пенза : Изд-во ПГУ, 2013. – С. 457–458.

References

1. Armenskiy A. E. *Tenzornye metody postroeniya informatsionnykh sistem* [Tensor methods of information system building]. Moscow: Nauka, 1989, 152 p.
2. Kron G. *Tenzornyy analiz setey: per. s angl.* [Tensor analysis: translation from English]. Moscow: Sov. radio, 1978, 720 p.
3. Chizhukhin G. N., Bochkareva Yu. G. *Tenzornaya metodologiya v diskretnoy sistemotekhnike* [Tensor methodology in discrete systems engineering]. Penza: Izd-vo PGU, 2006, 184 p.
4. Kulik B. A., Zuenko A. A., Fridman A. Ya. *Algebraicheskiy podkhod k intellektual'noy obrabotke dannykh i znaniy* [Algebraic approach to intelligent treatment of data and knowledge]. Saint-Petersburg: Izd-vo Politekhn. un-ta, 2010, 235 p.
5. Makarychev P. P., Popova N. A. *Universitetskoe obrazovanie: sb. st. XVII Mezhdunarodnaya nauch.-metod. konf.* [University education: proceedings of XVII International scientific and methodological conference]. Penza: Izd-vo PGU, 2013, pp. 457–458.

Макарычев Петр Петрович
доктор технических наук, профессор, заведующий кафедрой математического обеспечения и применения ЭВМ, Пензенский государственный университет (Россия, г. Пенза, ул. Красная, 40)

Makarychev Petr Petrovich
Doctor of engineering sciences, professor, head of sub–department of computer application and software, Penza State University (40 Krasnaya street, Penza, Russia)

E-mail: makpp@yandex.ru

Артамонов Дмитрий Владимирович
доктор технических наук, профессор, кафедра автономных информационных и управляющих систем, Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40)

Artamonov Dmitriy Vladimirovich
Doctor of engineering sciences, professor, sub-department of autonomous information and control systems, Penza State University
(40 Krasnaya street, Penza, Russia)

E-mail: aius@pnzgu.ru

УДК 681.3

Макарычев, П. П.

Модельные представления данных на основе прямого тензорного исчисления / П. П. Макарычев, Д. В. Артамонов // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2016. – № 3 (39). – С. 3–15. DOI 10.21685/2072-3059-2016-3-1