

УДК 514.743

## ПОСТРОЕНИЕ МОДЕЛИ ДАННЫХ С ПРИМЕНЕНИЕМ ТЕНЗОРНОЙ МЕТОДОЛОГИИ

Попова Н. А.

*ФГБОУ ВПО «Пензенский государственный университет», Пенза, Россия (440026, г. Пенза, ул. Красная, 40), e-mail: popov.maxim@bk.ru*

Рассматриваются вопросы построения модели данных информационно-аналитической системы (ИАС) с применением тензорной методологии, развитой в работах Г. Крона. Определены предпосылки использования тензорного исчисления для формализованного описания моделей данных. Предложено задание и преобразование базисов многомерных предметных пространств в виде индексных объектов со скользящими индексами на верхнем (системном) уровне модельного представления. Для детального описания всех объектов многомерного предметного пространства (сущностей) вводится набор фиксированных индексов. В этом случае любой объект модели можно представить в многомерном пространстве в виде индексного объекта первой валентности, содержащего набор фиксированных индексов и один скользящий индекс. Определены инварианты в пространствах данных, что позволило производить анализ структур данных. Решена задача оценки объемов хранимых данных при переходе от одного базиса пространства к другому.

Ключевые слова: модель данных, тензорное исчисление, индексные объекты, базис пространства, инварианты.

## BUILDING A MODEL DATA PRIENENIEM TENSOR METHODOLOGIES

Popova N. A.

*Penza State University, Penza, Russia (440026, Penza, Krasnaya street, 40), e-mail: popov.maxim @ bk.ru*

The problems of building the data model of information-analytical system (IAS) with the use of tensor methodology developed in the works of Crohn's. Preconditions for the use of tensor calculus formalized description of the data model. Proposed assignment and conversion of the bases of multidimensional spaces subject as index moving objects with indices on the top (the system) level model representation. For a detailed description of all the objects of the multidimensional space of subject (the entity), introduces a set of fixed indices. In this case, any object model can be represented in a multidimensional space as an index of the first object valence, containing a set of fixed index and a sliding index. Invariants are defined spaces in the data allowing to analyze data structures. We solve the problem of estimating the amount of data stored in the transition from one type of space to another.

Keywords: data model, tensor calculus, index objects, a basis of invariants.

Информационно-аналитические системы (ИАС) предназначены для решения задач принятия решений при управлении на основе данных, получаемых в процессе производственной, образовательной или коммерческой деятельности. Комплекс ИАС предприятия или организации может затрагивать всю управленческую вертикаль: корпоративную отчетность, финансово-экономическое и стратегическое планирование. При этом ИАС являются надстройкой над уже функционирующими информационными приложениями и не требуют их замены [2]. В ИАС аккумулируются данные по различным видам деятельности, включая состояния производственной или финансовой деятельности. Как правило, ИАС подразделений предполагают более углубленную детализацию и сложную аналитическую обработку данных.

В процессе проектирования любая ИАС может рассматриваться как четверка  $SI = \langle U, Q, V, f \rangle$ , где  $U$  – множество объектов,  $Q$  – множество свойств (атрибутов),

$V = \bigcup_{q \in Q} V_q$  – множество всех возможных значений свойств, а  $f : U \times Q \rightarrow V$  – информационная функция [5]. При этом разработка модельного представления проблемно-ориентированной ИАС может выполняться как с использованием исчисления предикатов, так и тензорного исчисления [1, 4].

Тензорное исчисление эффективно в различных областях деятельности: гидромеханика, электротехника, системы массового обслуживания и телекоммуникационные системы. Поскольку большинство тензоров описываются с помощью многомерных массивов [1], такое же представление имеют таблицы баз и хранилищ данных, было решено использовать аппарат тензорной алгебры для формализованного описания многомерного пространства данных ИАС. Существует ряд предпосылок для применения тензорного исчисления при формализованном описании моделей данных:

- классы реальных объектов предметной области ИАС представляются в виде отдельных сущностей (таблиц);
- формализованное описание объектов осуществляется с учетом заданной системы измерений предметной области (координат).
- преобразование объекта при переходе из одной системы координат в другую выполняется с учетом выполнения требований аксиом Армстронга.
- наличие в модельном представлении данных инвариантов – параметров, значения которых не меняются при переходе из одной системы координат в другую.

Одной из основных частей ИАС является средства хранения данных, состоящие из транзакционных баз данных (БД) и хранилищ данных (ХД). Предположим, что модель фрагмента БД приведена на рисунке 1. В соответствии с этой моделью сущность «Запись» отражает значения показателей деятельности аспирантов за определенный период времени.

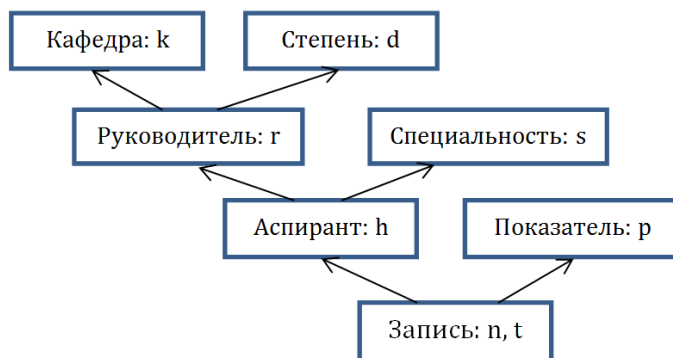


Рисунок 1. Модель данных «Снежинка»

На основе тензорной методологии, развитой в работах Г. Крона, любую структуру многомерного пространства данных или базис пространства можно задать в виде индексного объекта второй валентности [3]. Базис пространства модели данных «Снежинка», представленной на рисунке 1, можно задать в виде квадратной матрицы, описывающей

индексный объект со скользящими индексами  $\alpha, \beta$ , на верхнем (системном) уровне модельного представления:

$$Q_{\alpha, \beta} = \begin{array}{c} \begin{array}{c} \beta \\ \rightarrow \\ \hline \end{array} \\ \begin{array}{c} \alpha \\ \downarrow \\ \hline \end{array} \end{array} \begin{array}{cccccc} h & p & r & s & k & d & (n, t) \\ \left[ \begin{array}{cccccc} 1 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \begin{array}{l} h \\ p \\ r \\ s \\ k \\ d \\ (n, t) \end{array} \end{array}$$

Индексный объект  $Q_{\alpha, \beta}$  представляет собой выбранный базис предметного пространства данных «Снежинка» с размерностью равной семи, определяемой количеством измерений. Для детального описания всех объектов рассматриваемого многомерного предметного пространства (сущностей) вводим набор фиксированных индексов:  $(h, p, r, s, k, d, (n, t))$ . В этом случае любой объект модели можно представить в многомерном пространстве в виде индексного объекта первой валентности, содержащего набор фиксированных индексов и один скользящий индекс. Наличие первичного ключа задается единицей со знаком «+», наличие внешнего ключа – единицей со знаком «-». Например, класс «Аспирант» модели данных можно представить как индексный объект:

$$Q_{h, \beta} = [1 \ 0 \ -1 \ -1 \ 0 \ 0 \ 0]. \quad (1)$$

Поскольку в выражении (1) скользящие индексы  $p, k, d, (n, t)$  имеют нулевое значение, то возможно использование сокращенной записи индексного объекта  $Q_h^{r, s}$ . В этой записи нижний фиксированный индекс соответствует первичному ключу, верхние фиксированные индексы соответствуют внешним ключам.

С использованием рассмотренных представлений индексных объектов может быть решена задача преобразования базисов пространств. Зададим базис многомерного пространства, определяемого моделью данных «Звезда», приведенной на рисунке 2.

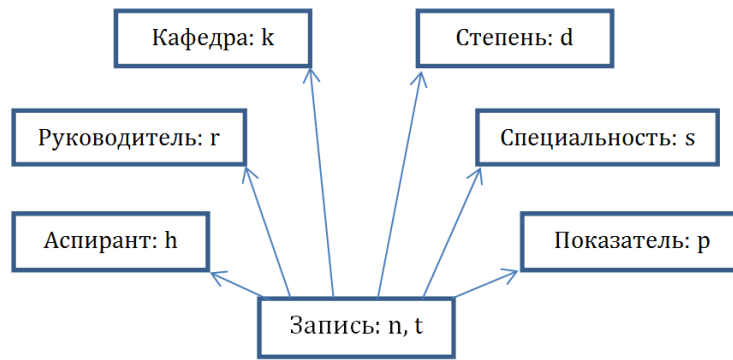


Рисунок 2 – Модель данных «Звезда»

В соответствии с рисунком 2 базис многомерного пространства имеет вид:

$$Q_{\gamma,\beta}^* = \begin{bmatrix} h & p & r & s & k & d & (n,t) \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & -1 & -1 & -1 & -1 & -1 & 1 \end{bmatrix} \begin{matrix} h \\ p \\ r \\ s \\ k \\ d \\ (n,t) \end{matrix}$$

Для заданных базисов  $Q_{\alpha,\beta}$ ,  $Q_{\gamma,\beta}^*$  многомерных пространств можно задать преобразования следующего вида:

$$Q_{\gamma,\beta}^* = \sum_{\alpha} t_{\gamma}^{\alpha} \cdot Q_{\alpha,\beta} \quad \text{или} \quad Q_{\gamma,\beta}^* = T_{\gamma}^{\alpha} \cdot Q_{\alpha,\beta},$$

где  $T_{\gamma}^{\alpha}$  – тензор преобразования базы пространства данных «Снежинка» в базу пространства «Звезда» с валентность равной двум.

Применение индексных объектов позволяет производить анализ структур данных. Для этого необходимо определить инварианты в пространствах данных:

- объем данных для описания одного и того же набора измерений в различных предметных пространствах фиксировано;
- объем данных, хранимых в таблицах мер различных предметных пространств, без учета структуры функциональных зависимостей неизменно.

Определение инвариантов позволяет осуществлять расчет объема хранимой информации при любой структуре и выполнять преобразование одной структуры данных в другую без потери данных. Расчет объема хранимой информации производится в два этапа.

На первом этапе выполняется анализ структуры модели данных, содержащей сущности без указания внешних зависимостей (рисунок 3).



Рисунок 3. Модель данных  $Q_{\alpha, \gamma}^*$  без внешних связей

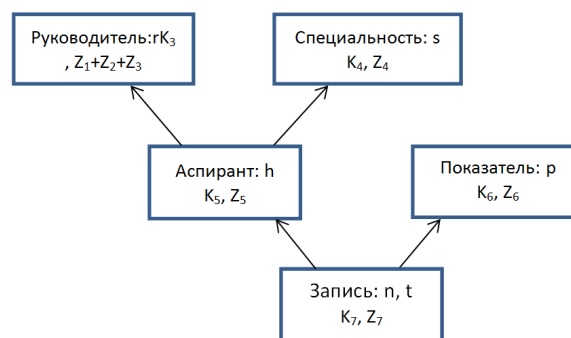
В модели на рисунке 3 использованы следующие обозначения:  $K_i, i = \overline{1,7}$  – количество записей в сущностях,  $Z_i, i = \overline{1,7}$  – размерность экземпляра сущности в атрибутах. В соответствии с базисом пространства, в котором представлена модель  $Q_{\alpha, \gamma}^*$ , расчет хранимой информации с учетом количества записей и размерности экземпляра сущности может быть выполнен с использованием выражения:

$$\begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \\ G_6 \\ G_7 \end{bmatrix} = \begin{bmatrix} K_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & K_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & K_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & K_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & K_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & K_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & K_7 \end{bmatrix} \cdot \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \end{bmatrix}.$$

При этом, объем хранимой информации, измеряемый в атрибутах, будет рассчитываться по формуле:

$$Q = \sum_{i=1}^7 G_i. \tag{2}$$

На втором этапе осуществляется анализа структуры данных при выявленных функциональных зависимостях  $F : r \rightarrow k, k \rightarrow d, h \rightarrow r, h \rightarrow s, (n, t) \rightarrow h, (n, t) \rightarrow p$  и ограничении уровня иерархии измерений, равном 2. В этом случае модель структуры данных имеет вид, приведенный на рисунке 4.



#### Рисунок 4. Модель с ограничением уровня иерархии измерений

Как видно из рисунка, базис пространства был преобразован. В модели, убрав один уровень иерархии, а функциональные зависимости, заданные в модели данных «Снежинка» (рисунок 1), сохранены. В этом случае базис пространства имеет вид:

$$Q_{\alpha,\beta}^{***} = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 1 \end{bmatrix}$$

На втором этапе также определяется объем данных, необходимый для представления внешних ключей. Предположим, что для рассматриваемой модели объемы данных для представления внешних ключей имеют следующие значения:

$$C_1 = 0, C_2 = 0, C_3 = 2, C_4 = 0, C_5 = 2.$$

Таким образом, объем хранимой информации в каждой сущности модели с учетом внешних связей можно рассчитать по следующей формуле:

$$\begin{bmatrix} G_1 \\ G_2 \\ G_3 \\ G_4 \\ G_5 \end{bmatrix} = \begin{bmatrix} K_3 & K_3 & K_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & K_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & K_5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & K_6 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & K_7 \end{bmatrix} \cdot \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \\ Z_6 \\ Z_7 \end{bmatrix} + \begin{bmatrix} K_3 & 0 & 0 & 0 & 0 \\ 0 & K_4 & 0 & 0 & 0 \\ 0 & 0 & K_5 & 0 & 0 \\ 0 & 0 & 0 & K_6 & 0 \\ 0 & 0 & 0 & 0 & K_7 \end{bmatrix} \cdot \begin{bmatrix} C_1 \\ C_2 \\ C_3 \\ C_4 \\ C_5 \end{bmatrix}$$

Объем данных, хранимых во всей БД, рассчитывается по формуле (2).

Рассмотрим пример расчета объема данных при следующих начальных данных:

$$K_1 = 5, K_2 = 2, K_3 = 20, K_4 = 8, K_5 = 100, K_6 = 11, K_7 = 3 \cdot K_5 \cdot K_6$$

$$Z_1 = 10, Z_2 = 12, Z_3 = 11, Z_4 = 15, Z_5 = 12, Z_6 = 10, Z_7 = 9.$$

Объем хранимых данных при структуре, представленной на рисунке 4, равен:

$$Q = 3.859 \times 10^4,$$

а при структуре со всеми уровнями иерархии измерений (рисунок 1):

$$Q' = 3.188 \times 10^4.$$

Следовательно, если убрать один уровень иерархии измерений в структуре данных, то объем хранимых данных увеличится на 21 %. Такие вычисления целесообразно производить при выборе оптимальной структуры хранилища или базы данных, подборе технологической платформы реализации выбранной структуры, поскольку некоторые аналитические

платформы (например, Deductor Studio компании BaseGroup Labs) поддерживают только два уровня иерархии.

Таким образом, введение понятия базиса многомерного пространства данных, задание базисов пространства и классов объектов в виде индексных объектов, определение инвариантов реляционных моделей данных существенно расширяют применение тензорной методологии при проектировании баз и хранилищ данных ИАС. Дальнейшее развитие тензорной методологии применительно к информационным системам связано с решением задачи анализа процессов сбора, предварительной обработки, оперативного и интеллектуального анализа данных в ИАС.

### Список литературы

1. Бакулева М. А. Тензорная модель работы реляционной СУБД // Информационные технологии в проектировании. Межвуз. сб. научн. трудов. Рязань: РГРТА, 2004. С. 39–43.
2. Белов В. С. Информационно-аналитические системы. Основы проектирования и применения: учебное пособие, руководство, практикум / Московский государственный университет экономики, статистики и информатики. – М., 2004. – 116 с.
3. Крон Г. Тензорный анализ сетей: Пер. с англ. / Под ред. Л. Т. Кузина, П. Г. Кузнецова. М.: Сов. Радио, 1978. – 720 с.
4. Кулик Б. А. Обобщенный подход к моделированию и анализу интеллектуальных систем на основе алгебры кортежей // Труды VI Межд. конф. «Идентификация систем и задачи управления» SICPRO. М., 2007. С. 679–715.
5. Рутковский Л. Методы и технологии искусственного интеллекта / Пер. с польск. И. Д. Рудинского. – М.: Горячая линия –Телеком, 2010. – 520 с.

### Рецензенты:

Зинкин С. А. д.т.н., профессор, профессор кафедры «Вычислительная техника» Пензенского государственного университета, г. Пенза.

Макарычев П. П. д.т.н., профессор, зав. кафедрой «Математическое обеспечение и применение ЭВМ» Пензенского государственного университета, г. Пенза.