МАТЕМАТИЧЕСКАЯ МОДЕЛЬ МНОГОМЕРНОГО ПРЕДСТАВЛЕНИЯ ДАННЫХ

Иванов Е.С.

Российский университет дружбы народов,

Rumbiddiebooo@gmail.com

В работе дано описание математической модели многомерного представления данных, в рамках которой рассмотрены основные функции преобразования куба данных.

Ключевые слова: многомерный куб, куб данных, аналитическое измерение, показатель, факт, ячейка, уровень детализации.

Введение

В процессе анализа сложных явлений возникает потребность в описании и преобразовании огромных объемов информации. При этом появляется необходимость в создании формальной математической модели, на которой основывались бы концептуальная модель и реализация системы. В рамках многомерной модели данных может быть реализован удобный и наглядный способ анализа данных в компьютерных системах, соответствующий интуитивному представлению человека.

Математическая многомерная модель

Источником первичной информации для анализа является факт, представленный в рамках многомерного подхода. Свойства факта выражаются набором величин – показателей,– которые рассматриваются как величины, зависящие от условий, при которых он произошел (например время и место). Характеристики, выражающие эти условия, называются измерениями, и задаются множеством значений измерения. Таким образом показатели можно рассматривать как функции нескольких переменных. Для того, чтобы отделить разные факты друг от друга, можно разбить эти множества на непересекающиеся подмножества. Таким образом, измерение D представляет собой множество, на котором возможно построить разбиение. Разбиением D' множества D назовем его представление в виде объединения непересекающихся его подмножеств.

$$D' = \{\overline{D_i} \subseteq D\}, \overline{D_i} \cap \overline{D_j} = \varnothing (i \neq j), D = \cup \overline{D_i} \quad (1)$$

Разбиения являются уровнями детелизации, если для них можно задать иерархию. Наличие уровней детализации позволеют переходить при анализе фактов от детального рассмотрения к более общему, позволяя получить получить оценку явления в целом и сделать выводы, необходимые для принятия решения. Если заданы два уровня детализаций D^1, D^2 одного измерения D, то множество D^3 всех попарных пересечений их элементов тоже можно представить как уровень детелизации. Этот подход позволяет сформировать более сложные группировки элементов:

$$D^1 = \{d^1_i\}, D^2 = \{d^2_j\}, D^3 = \{d^1_i \cap d^2_j\} \cdot \quad (2)$$

Существует два особых уровня детализации, для которых введены следующие обозначения:

$$D^A = \{d_i\}, \; D^E = \cup d_i \;, \quad (3)$$

где D^A – „атомарный уровень” – на нем заданы значения показателей, а D^E – уровень „все” – одноэлементный уровень, представляющий измерение вцелом.

Прямое произведение конечного набора измерений задает пространство, называемое многомерным кубом C факта F:

$$C^F = D_1 \times \ldots \times D_n \cdot \quad (4)$$

Показатель m факта – это функция, заданная на атомарном уровне детализации всех измерений куба:

$$m : D_1^A \times \ldots \times D_n^A \to X \ . \qquad (5)$$

Для остальных уровней детализации значения показателей последовательно вычисляются с помощью функции агрегации. Функция агрегации f_a – это любая функция вида

$$f_a : X \times X \to X \ , \qquad (6)$$

образующая полугруппу с множеством определения показателя.

Факт можно определить как множество значений показателей M на кубе C^F:

$$C_M^F = \{m_i(\bar{d})\} = \{m_1(\bar{d}),\ldots,m_k(\bar{d})\} \ . \qquad (7)$$

Множество значений показателей в заданной точкте пространства факта называется ячейкой:

$$c_M^F = \{m_i(\bar{d}_0)\} = \{m_1(d_1^0,\ldots,d_n^0),\ldots,m_k(d_1^0,\ldots,d_n^0)\} \ . \qquad (8)$$

Таким образом, факт состоит из ячеек, в каждой из которых находятся конкретные значения показателей факта.

Операции над многомерным кубом данных

Инструментарием для осуществления анализа данных куба может служить небольшой, но эффективный набор операций. Они позволяют рассматривать факты в разных масштабах, с разных точек наблюдения, выделять интересующие показатели из полного набора, рассматривать только часть всего аналитического пространства.

Проекция π – задает подмножество множества показателей М факта:

$$\pi(C_M^F) = C_{\bar{M}}^F, \bar{M} \subseteq M \ . \qquad (9)$$

Данная операция позволяет избавиться от ненужных показателей рассматриваемого факта.

Кросс-детализация δ – задает новый факт на уже имеющемся кубе:

$$\delta(C_{M_1}^F) = C_{M_2}^F \ . \qquad (10)$$

Операция сопоставляет два факта между собой. Необходимым для выполнения кросс-детализации является условие того, чтобы показатели обоих фактов определяли одинаковые пространства.

Кубический срез σ – задает подмножество пространства факта:

$$\sigma(C_M^F) = \bar{C}_M^F, \bar{C}_M^F \subseteq C_M^F, \bar{C} = \bar{D}_1 \times \ldots \times \bar{D}_n \ . \qquad (11)$$

Данная операция позволет выделить для рассмотрения и получения оценок только часть факта. Результатом операции является новый куб с меньшим количеством элементов, заданных для одного либо нескольких измерений.

Смена измерений γ – изменяет уровень детализации куба следующим образом:

$$\gamma(D_1 \times \ldots \times D_n, I_D) = D'_1 \times \ldots \times D'_n,$$

$$D'_i = \begin{cases} D_i^E, i \notin I_D \\ D_i, i \in I_D .AND. D_i \neq D_i^E \\ D_i^A, i \in I_D .AND. D_i = D_i^E \end{cases} \qquad (12)$$

Здесь I_D – множество индексов измерений, к которым нужно перейти. Нужно отметить, что хотя количество измерений не изменяется, переход к уровню D^E означает, что показатели больше не зависят от данного измерения, так как на нем задан только один элемент.

Пусть задан факт вместе со всеми показателями всех его ячеек. Пусть, также, заданы функции агрегации для всех показателей:

$$\bar{f}_a^M = (f_a^{m_1},\ldots,f_a^{m_k}) \ . \qquad (13)$$

Укрупнение ρ – изменяет уровень детализации измерений куба и задает показтели на нем:

$$(C_M^{\rho}, \underline{F}^M f_a) = \{ f_{a\ d_1 \in d'_1}^{m_i} \dots f_{a\ d_n \in d'_n}^{m_i}\ m_i(d_1, \dots, d_n) \} = \{ m_i(d'_1, \dots, d'_n) \}, \quad (14)$$

где D'_i – уровни детализации, к которым нужно перейти. Операция укрупнения увеличивает масштаб измерений куба. При этом значения показателей новых ячеек являются обобщением значений показателей нижнего уровня. По смыслу укрупнение – это анализ данных. Меняя функции агрегации можно получить разные оценки факта.

Нужно отметить, что описанные операции составляют минимальный и полный набор функций над кубом. То есть, все функции, которые можно определить, будут являться либо производными от данных функций, либо их композициями.

Выводы

В данной работе были предложены основные понятия математической модели многомерного представления данных. Также был описан необходимый минимальный набор операций над кубом, при помощи которого можно анализировать данные, представленные в виде многомерного куба.

Литература

1. *Висков А.В.* - Модель многомерного представления данных и методы ее анализа, 2010.

MATHEMATICAL MULTIDIMENSIONAL DATA MODEL

Ivanov E.S.

Peoples' Friendship University of Russia,

Rumbiddiebooo@gmail.com

The paper proposes a formal mathematical definition of the multidimensional data model basic terms and functions.

Key words: multidimensional data model, data cube, dimension, measure, fact, cell, level.