УДК 551.50

МНОГОМЕРНАЯ МОДЕЛЬ МЕТЕОРОЛОГИЧЕСКИХ ДАННЫХ ДЛЯ АНАЛИТИЧЕСКОЙ ОБРАБОТКИ

В.С. МУХА, А.Н. КОЗЯЧИЙ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки,6, Минск, 220013, Беларусь

ОАО "ИнтэксСофт"
Кобяка, 8/4, офис 7, Гродно, Беларусь

Поступила в редакцию 25 апреля 2009

Приводятся теоретические положения, лежащие в основе многомерной модели данных, терминология OLAP-систем согласуется с терминологией теории многомерных матриц, разрабатывается структура многомерной модели метеорологических данных и программное средство, реализующее эту модель.

Ключевые слова: OLAP, многомерная модель данных, многомерные матрицы, метеорологические данные.

Введение

В настоящее время серьезное внимание уделяется концепциям построения и средствам реализации информационных систем, ориентированных на аналитическую обработку данных [1]. Аналитическая обработка предполагает быстрый доступ к большим объемам данных и выполнение большого объема вычислений, причем используемые данные чаще всего должны быть упорядочены во времени. Это накладывает определенные ограничения на способы их хранения, т.е. на базы данных, наиболее распространенными из которых в настоящее время являются реляционные. Несмотря на то что существование реляционных баз данных, а также их использование, в том числе и для аналитической обработки данных, не подвергается сомнению, осуществляется поиск других подходов к организации баз данных. Одним из них является многомерный подход. Впервые он был представлен для широкого обсуждения основоположником теории реляционных баз данных Э. Коддом [2], где была определена также категория OLAP-систем (Online Analytical Processing). Из 12 правил оценки OLAP-систем, представленных Э. Коддом в этой работе, важнейшим является многомерность модели данных.

Теоретические основы многомерной модели данных

В основе многомерной модели данных лежит понятие многомерной матрицы [3, 4]. Многомерной (p -мерной) $(n_1 \times n_2 \times ... \times n_p)$-матрицей называется система элементов $a_{i_1 i_2 ... i_p}$, $i_\alpha = \overline{1, n_\alpha}$, $\alpha = \overline{1, p}$, расположенных в точках p-мерного пространства, определяемого осями координат $i_1, i_2, ..., i_p$. Обозначается p-мерная матрица как

$$A = (a_{i_1 i_2 ... i_p}),\ i_\alpha = \overline{1, n_\alpha},\ \alpha = \overline{1, p}. \quad (1)$$

Переменные $i_1, i_2, ..., i_p$ многомерной матрицы называются индексами. Каждый индекс i_α многомерной матрицы принимает значения от 1 до n_α. В случае различных чисел $n_1, n_2, ..., n_p$ матрица называется гиперпрямоугольной. Если $n_1 = n_2 = ... = n_p = n$, то матрица называется гиперквадратной или p-мерной матрицей n-го порядка.

Совокупность элементов матрицы A (1) с фиксированными значениями $i^*_{\alpha_1}, i^*_{\alpha_2}, ..., i^*_{\alpha_m}$ индексов $i_{\alpha_1}, i_{\alpha_2}, ..., i_{\alpha_m}$ $(1 \le m \le p-1,\ 1 \le \alpha_1 < ... < \alpha_m \le p)$ называется m-кратным сечением матрицы A (1) ориентации $(i_{\alpha_1}, i_{\alpha_2}, ..., i_{\alpha_m})$. Это сечение представляет собой $(p-m)$-мерную матрицу. Например, сечение двухкратной ориентации (i_α, i_β) матрицы A (1) представляет собой $(p-2)$-мерную матрицу

$$B = (b_{i_1,...i_{\alpha-1},i_{\alpha+1},...,i_{\beta-1},i_{\beta+1},...,i_p}) = (a_{i_1,...i_{\alpha-1},i^*_\alpha,i_{\alpha+1},...,i_{\beta-1},i^*_\beta,i_{\beta+1},...,i_p}).$$

M-кратное сечение матрицы A (1) ориентации $(i_{\alpha_1}, i_{\alpha_2}, ..., i_{\alpha_m})$ со значениями индексов $i^*_{\alpha_1}, i^*_{\alpha_2}, ..., i^*_{\alpha_m}$ будем обозначать в виде $B = (A)_{i^*_{\alpha_1} i^*_{\alpha_2} ... i^*_{\alpha_m}}$. Множество всех возможных m-кратных сечений матрицы A (1) ориентации $i_{\alpha_1}, i_{\alpha_2}, ..., i_{\alpha_m}$ совпадает, очевидно, с исходной матрицей A, т.е. матрицу A можно представить с помощью ее сечений в виде $A = \{(A)_{i_{\alpha_1} i_{\alpha_2} ... i_{\alpha_m}}\}$.

Совокупность p индексов $(i_1, i_2, ..., i_p)$ p-мерной матрицы будем называть мультииндексом или p-мультииндексом и обозначать $i_{(p)} = (i_1, ..., i_p)$. С применением мультииндексов многомерную матрицу (1) можно записать в виде

$$A = (a_{i_{(p)}}),\ i_{(p)} = (i_1, i_2, ..., i_p).$$

Если каждому индексу $i_1, i_2, ..., i_p$ p-мерной матрицы A (1) присвоить определенное значение, то мы получим упорядоченную последовательность значений $(i^*_1, i^*_2, ..., i^*_p)$ (кортеж), которую будем называть значением мультииндекса $i_{(p)} = (i_1, ..., i_p)$ и обозначать $i^*_{(p)} = (i^*_1, i^*_2, ..., i^*_p)$. Значения $i^*_{(p)}$ мультииндекса $i_{(p)}$ можно упорядочить каким-либо образом и поставить в соответствие множеству целых чисел. Эти числа будем называть скалярными значениями мультииндекса.

Во многих случаях полезно определять структуру многомерной матрицы. Для этого мультииндекс $i_{(p)} = (i_1, i_2, ..., i_p)$ разбивают на составляющие мультииндексы l, s, c следующим образом:

$$i_{(p)} = (i_1, i_2, ..., i_p) = (l, s, c),$$

где

$$l = (l_1, l_2, ..., l_\kappa),\ s = (s_1, s_2, ..., s_\lambda),\ c = (c_1, c_2, ..., c_\mu),$$

причем $\kappa + \lambda + \mu = p$, и каждое из чисел κ, λ, μ может быть равным нулю. Матрица A (1) с такой структурой индексов записывается в виде

$$A = A_{(\kappa,\lambda,\mu)} = (a_{l,s,c}). \qquad (2)$$

Пусть A — матрица структуры (2), и $l^{(1)},l^{(2)},...,l^{(n^{(\kappa)})}$, $s^{(1)},s^{(2)},...,s^{(n^{(\lambda)})}$, $c^{(1)},c^{(2)},...,c^{(n^{(\mu)})}$ — упорядоченные значения мультииндексов l,s,c этой матрицы. Клеточно-диагональная матрица

$$\tilde{A}_{(\kappa,\lambda,\mu)} = \operatorname{diag}\ A^{(1)}_{(\kappa,0,\mu)}, A^{(2)}_{(\kappa,0,\mu)}, \ldots, A^{(n^{(\lambda)})}_{(\kappa,0,\mu)}\ ,$$

составленная из элементов матрицы A, где диагональные клетки $A^{(h)}_{(\kappa,0,\mu)}$, $h=1,2,...,n^{(\lambda)}$, — двухмерные $(n^{(\kappa)}\times n^{(\mu)})$-матрицы,

$$A^{(h)}_{(\kappa,0,\mu)} = (a_{\tilde{l},s^{(h)},\tilde{c}}),\ \tilde{l} = l^{(1)},l^{(2)},...,l^{(n^{(\kappa)})},\ \tilde{c} = c^{(1)},c^{(2)},...,c^{(n^{(\mu)})},$$

называется (κ,λ,μ)-ассоциированной с матрицей A. Ассоциированная матрица $\tilde{A}_{(\kappa,\lambda,\mu)}$ вполне представляет исходную многомерную матрицу $A = A_{(\kappa,\lambda,\mu)}$, так как содержит все ее элементы. Ассоциированные матрицы можно использовать для графического изображения многомерных матриц.

Матрица $A^T = (a^T_{i_1,i_2,...i_p})$, $i_\alpha = \overline{1,n_\alpha}$, $\alpha = \overline{1,p}$, элементы которой связаны с элементами матрицы A (1) соотношениями

$$a^T_{i_1,i_2,...i_p} = a_{i_{\alpha_1},i_{\alpha_2},...i_{\alpha_p}}, \qquad (3)$$

где $i_{\alpha_1},i_{\alpha_2},...,i_{\alpha_p}$ — какая-нибудь перестановка индексов $i_1,i_2,...,i_p$, называется транспонированной относительно матрицы A соответственно подстановке $T = \begin{pmatrix} i_1,..., i_p \\ i_{\alpha_1},...,i_{\alpha_p} \end{pmatrix}$ и обозначается как A^T или $A^{\begin{pmatrix} i_1,i_2,..., i_p \\ i_{\alpha_1} i_{\alpha_2},...i_{\alpha_p} \end{pmatrix}}$.

Организация многомерной модели данных

В связи с тем что терминология теории многомерных матриц является достаточно устоявшейся и общепринятой, желательно, чтобы терминология OLAP-систем и многомерных баз данных согласовывалась с теорией многомерных матриц.

Под многомерной моделью данных в OLAP-системе будем понимать организацию данных в виде многомерной матрицы, или, иначе, в виде гиперпрямоугольника. Случай гиперквадрата (гиперкуба по терминологии работы [1]) является скорее исключением, чем правилом. Для организации гиперпрямоугольника данных необходимо определить оси координат (индексы) *p*-мерного пространства и данные, располагаемые в узлах сетки гиперпрямоугольника. Значения индексов будем называть также метками на осях координат. Значению $i^*_{(p)} = (i^*_1,i^*_2,...,i^*_p)$ мультииндекса будет соответствовать упорядоченный по осям набор меток. Оси выбираются таким образом, чтобы значение мультииндекса однозначно определяло данные. Данные чаще всего представляет собой строки, аналогичные строкам реляционной таблицы данных, так что мультииндекс является составным ключом этой таблицы. Выбор осей осуществляется для каждой подсистемы OLAP-системы на интуитивном уровне и не представляет трудностей. При описанном способе построения модели данных не нужно выполнять нормализацию таблиц данных, что существенно упрощает процесс проектирования многомерной базы данных по сравнению с реляционной.

При организации осей их наименования можно использовать в качестве начальных значений индексов (начальных меток осей). В этом случае значение $i^*_{(p)}$=(1,1,…,1) мультииндекса (начало координат) будет указывать на наименования осей, и физически многомерная модель

данных будет состоять их двух файлов, один из которых определяет оси с их наименованиями, а другой — данные, соответствующие этим осям.

OLAP-системы, использующие многомерную модель данных, получили название MOLAP-систем, в отличие от ROLAP-систем, использующих реляционную модель данных.

Кроме задач аналитической обработки данных (расчет статистических характеристик, прогнозирование и др.) в MOLAP необходимо решать вопросы манипулирования данными: ввод или загрузка данных, отображение данных (формирование текстовых отчетов), удаление данных. В свою очередь эти операции связаны с такими операциями, как формирование сечений и вращение.

В MOLAP-системе можно организовать ввод данных в интерактивном режиме. Однако в системах, предназначенных для обработки больших объемов данных, это вряд ли целесообразно. Чаще всего данные хранятся в реляционных базах данных, электронных таблицах или в текстовых файлах. В этом случае следует организовать автоматический импорт данных из этих источников. Поскольку эти источники представляют собой двухмерные массивы данных, которые могут быть помещены в определенные двухмерные сечения гиперпрямоугольника, то необходимо формировать сечения в соответствии с определением сечения, приведенным выше, т.е. путем выбора определенных меток осей.

Формирование текстовых отчетов состоит в отображении на бумагу (на плоскость) информации, содержащейся в гиперпрямоугольнике. Это можно сделать путем формирования двухмерной матрицы $\tilde{A}_{(\kappa,0,\mu)}$, $(\kappa,0,\mu)$-ассоциированной с имеющейся p-мерной матрицей $A_{(\kappa,0,\mu)}$. Целесообразно, например, использовать матрицу $\tilde{A}_{(p-1,0,1)}$, $(p-1,0,1)$-ассоциированную с p-мерной матрицей $A_{(p-1,0,1)}$.

Большое разнообразие отчетов можно получить путем вращения данных. Операция вращения представляет собой не что иное, как операцию транспонирования многомерной матрицы в соответствии с некоторой подстановкой, и определяется формулой (3). Важно иметь в виду, что при выполнении операции транспонирования имеющегося гиперпрямоугольника данных необходимо также соответствующим образом переформировать имеющиеся данные. Отчет формируется после выполнения операции вращения.

Многомерная модель метеорологических данных

Приведенные рекомендации были положены в основу создания MOLAP-системы для аналитической обработки метеорологических данных.

В настоящее время метеорологические данные доступны в виде отдельных таблиц с данными, как это имеет место, например, в http://gismeteo.ru или http://pogoda.by. Имеющиеся таблицы с данными относятся к определенной метеостанции. Это затрудняет выполнение параллельной обработки данных, т.е. одновременной обработки данных ряда метеостанций. Кроме того, данные обычно относятся к фиксированному интервалу времени (году, месяцу). Это затрудняет обработку данных за произвольные интервалы времени. Указанных недостатков лишена многомерная модель метеорологических данных.

Структура файла осей многомерной модели метеорологических данных представлена в табл. 1.

Таблица 1. **Структура файла осей многомерной модели метеорологических данных**

Наименование оси	Метеостанция	Год	Месяц	День	Час
Номер оси	Ось 1	Ось 2	Ось 3	Ось 4	Ось 5
Позиция в файле	1	2	3	4	5

Таким образом, в качестве осей многомерной модели метеорологических данных нами выбраны метеостанция, год, месяц, день, час, так как они однозначно определяют данные. Файл осей данных выглядит следующим образом:

Метеостанция,26850,…,26666,

Год,1998,1999,…,2009,

Месяц,1,2,3,4,5,6,7,8,9,10,11,12,
День,1,2,3,…31,
Час,0,3,6,9,12,15,18,21,

Метеостанции определяются принятым во Всемирной метеорологической организации (ВМО) кодом. Так, 26850 — это Минск, 26666 — Витебск. Метеорологические данные собираются с периодичностью в 3 ч, начиная с 0 ч по Гринвичу, что отражено в оси "Час".

Структура файла данных многомерной модели метеорологических данных (многомерной базы данных) представлена в табл. 2.

Таблица 2. **Структура файла данных многомерной модели метеорологических данных**

Данные	Позиция в файле
Скалярное значение мультииндекса	1
Температура воздуха	2
Атмосферное давление на уровне моря	3
Направление ветра	4
Скорость ветра	5
Относительная влажность воздуха	6
Количество облаков	7
Атмосферное давление на уровне станции	8
Величина барической тенденции	9
Атмосферное давление на уровне моря	10
Величина барической тенденции	11

Строки файла данных (на примере двух строк) выглядят следующим образом:

1,0,0,0,0,0,0,0,0

17266,18.5,1016.7,40,4,40.3,9999,9999,0.5

Первая строка со скалярным значением мультииндекса 1 (ему соответствует векторное значение мультииндекса $i^*_{(p)}=(1,1,\ldots,1)$) соответствует началу координат и указывает на наименования осей. По определению эта строка не содержит данных, поэтому они обозначены как нулевые. Значение 9999 означает, что данные отсутствуют. В частности, во второй строке отсутствуют данные о количестве облаков и атмосферном давлении на уровне моря.

Описанная многомерная модель метеорологических данных была реализована в виде программного средства. Импорт данных в многомерную модель организован из текстовых файлов, структура которых приведена в табл. 3.

Таблица 3. **Структура файла данных для импорта в многомерную базу данных**

Данные	Позиция в файле
Номер (код) метеостанции	1
Дата наблюдения	2
Срок наблюдения	3
Форма (количество) облаков	4
Направление ветра	5
Скорость ветра	6
Температура	7
Относительная влажность	8
Атмосферное давление на уровне станции	9
Атмосферное давление на уровне моря	10
Величина барической тенденции	11

Были реализованы также такие процедуры, как изменение длин осей, удаление данных, формирование текстовых отчетов, вращение, просмотр осей. Программное средство подтвердила свою работоспособность.

MULTIDIMENSIONAL MODEL OF METEOROLOGICAL DATA FOR ANALYTICAL PROCESSING

V.S. MUKHA, A.N. KOZYACHIY

Abstract

The theoretical bases of multidimensional data model are given. OLAP-systems terminology with the terminology of multidimensional matrix theory is coordinated. The structure of multidimensional model of meteorological data is designed.

Литература

1. *Сахаров А.А.* // Системы управления базами данных. 1996. № 3. С. 44–59.
2. *Codd E.F., Codd S.B., Salley C.T.* Providing OLAP to User-Analyst: An IT Mandate – E.F. Codd Associates. 1993.
3. *Муха В.С.* Анализ многомерных данных. Минск, 2004.
4. *Соколов Н.П.* Введение в теорию многомерных матриц. Киев, 1971.