ных уравнений в частных производных в произвольных геометрических областях, составленной из непрограммируемых ячеек с фиксированными связями, возможно без добавления в состав каждой ячейки функционального блока граничных условий. Настройка на область решения достигается путем задания специальных коэффициентов в регистры ячейки, рассчитываемых из вида граничных условий. За счет этого аппаратные затраты на реализацию каждой ячейки снижаются и появляется возможность увеличения их количества в одной микросхеме.

СПИСОК ЛИТЕРАТУРЫ

1. Евреинов Э.В. Однородные вычислительные системы, структуры и среды. – М.: Радио и связь, 1981. – 208 с.
2. Каляев И.А., Левин И.И., Семерников Е.А., Шмойлов В.И. Реконфигурируемые мультиконвейерные вычислительные структуры. – Ростов на Дону: ЮНЦ РАН, 2008. – 393 с. URL: http://parallel.ru/FPGA/papers/rmvs.pdf (дата обращения: 31.03.2010).
3. Giefers H., Platzner M. A Many-Core Implementation Based on the Reconfigurable Mesh Model // IEEE Xplore DIGITAL LIBRARY. 2010. URL: http://ieeexplore.ieee.org/Xplore/defdeny.jsp?url=http://ieeexplore.ieee.org/stamp/stamp.jsp%3Ftp%3D%26arnumber%3D4380623&denyReason=-134&arnumber=4380623&productsMatched=null (дата обращения: 31.03.2010).
4. Ячейка однородной структуры для решения дифференциальных уравнений в частных производных: а.с. 783811 СССР. № 2727694/18-24; заявл. 21.02.1979; опубл. 30.11.1980, Бюл. № 44. – 2 с.
5. Ячейка однородной структуры для решения дифференциальных уравнений в частных производных: пат. 2359322 Рос. Федерация. № 2007141832/09; заявл. 12.11.07; опубл. 20.06.09, Бюл. № 17. – 6 с.
6. Каляев А.В. Теория цифровых интегрирующих машин и структур. – М.: Советское радио, 1970. – 472 с.
7. Лисейкин В.Д. Передовые технологии построения разностных сеток // РФФИ. 2010. URL: http://www.rfbr.ru/default.asp?doc_id=17662 (дата обращения: 31.03.2010).
8. Цифровые базовые матричные кристаллы. ОАО «Ангстрем» // 2010. URL: http://www.angstrem.ru/catalogue/element.php?IBLOCK_ID=2&SECTION_ID=5&ELEMENT_ID=120 (дата обращения: 31.03.2010).

Поступила 31.03.2010 г.

УДК 004.032.6;004.357

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ OLAP-КУБА В КОНТЕКСТЕ АГРЕГИРОВАНИЯ ПРОСТЫХ И ИЕРАРХИЧЕСКИХ ИЗМЕРЕНИЙ

В.П. Кулагин, В.Т. Матчин*

Государственный научно-исследовательский институт информационных технологий и телекоммуникаций, г. Москва
E-mail: kvp@informika.ru
*Московский государственный институт радиотехники, электроники и автоматики (технический университет)
E-mail: matchin@mirea.ru

Статья посвящена исследованию агрегации данных в многомерном OLAP-кубе в простом и иерархическом случае построения измерений. Получены формулы для расчета количества агрегатов и количества сочетаний агрегатов в простом и иерархическом случае построения измерений.

Ключевые слова:
Хранилище данных, база знаний, агрегирование данных, онтология, многомерный куб, иерархическое измерение.

Key words:
Databank, knowledge base, data aggregation, ontology, multidimensional cube, hierarchical measurement.

Известно, что основная цель управления знаниями – сделать знания доступными и повторно используемыми.

Чем больше накапливается информации, тем сложнее становится хранить ее на бумажных носителях или запоминать. И доступ к бумажным документам весьма ограничен. А если из организации уходит высококвалифицированный специалист, потеря ценных знаний и опыта зачастую оказывается невосполнимой. Поэтому целесообразным является осуществлять переход к использованию хранилищ данных, чтобы использовать накопленные знания.

Как правило, одним из первых инструментариев управления знаниями на начальном этапе внедрения корпоративных систем являются хранилища данных, которые работают по принципу центрального склада. Хранилища данных отличаются от традиционных баз данных тем, что они проектируются для поддержки процессов принятия решений, а не просто для эффективного сбора и обработки данных. Как правило, хранилище содержит

многолетние версии обычной базы данных, физически размещаемые в той же самой базе. Данные в хранилище не обновляются на основании отдельных запросов пользователей. Вместо этого вся база данных периодически обновляется целиком.

Если хранилища данных содержат в основном количественные данные, то хранилища знаний ориентированы на качественные данные. Хранилища знаний генерируют знания из широкого диапазона баз данных, хранилищ данных, рабочих процессов, статей, новостей, внешних баз, Web-страниц. Таким образом, хранилища знаний подобны виртуальным складам, где знания распределены по большому количеству серверов.

Базы знаний оптимальных решений наполняются в процессе использования различных тестов при поиске эффективных путей решения задач. После того, как получено наилучшее решение, доступ к ним может быть открыт для сотрудников организации.

Разведка знаний – быстро развивающееся направление, использующее методы искусственного интеллекта, математики и статистики для извлечения знаний из хранилищ данных. Г. Пятецки–Шапиро и В. Фролей определяют термин «разведка знаний» как «нетривиальное извлечение точной, ранее неизвестной и потенциально полезной информации из данных». Метод включает инструментарий и различные подходы к анализу как текста, так и цифровых данных.

Метод в его современном прочтении опирается на использование в моделировании OLAP-куба таких понятий, как онтология, показатель, измерение, количество сочетаний агрегатов и некоторых других терминов.

Онтология – это точное описание концептуализации. В системах управления знаниями используются онтологические спецификации, ссылающиеся на таксономию задач, которые определяют знание для системы (*Таксономия* – теория классификации и систематизации сложноорганизованных областей деятельности, обычно имеющих иерархическое строение. Прим. авт.). Онтология определяет словарь, совместно используемый в системе для упрощения коммуникации, общения, запоминания и представления. Онтология необходима для того, чтобы пользователь мог работать с базами данных оптимальных решений, относящихся к широкому кругу проблем, и легко распознавать, какое решение может ему подойти в конкретной ситуации. Так как предприятия часто вовлечены в различные виды деятельности, то для одной системы управления знаниями может потребоваться несколько онтологий. Удобнее всего разрабатывать свою собственную онтологию.

Немаловажным аспектом является поиск знаний, поскольку базы имеют огромные размеры. Большинство современных методов поиска включают инструментальные средства, средства интеллектуального поиска и визуальные модели.

Показатель – числовая величина, которая является предметом анализа и хранится в ячейках таблиц.

Измерение – множество объектов одного или нескольких типов, организованных в виде иерархической структуры и обеспечивающих информационный контекст числового показателя.

Член измерения – отдельная строка или столбец таблицы, содержащая показатели.

Количество сочетаний агрегатов

Рассмотрим отдельную таблицу, содержащую два измерения A и B. Таблица имеет размер $m \times n$ ячеек. Рассчитаем количество возможных агрегатных состояний для такой таблицы.

Общее количество сочетаний агрегатов для m измерений рассчитывается следующим образом:

$$A = \prod_{x=1}^{m-1} n_x \sum_{y=1}^{n_m} C_{n_m}^{y}, \qquad (*)$$

где $x=1,2,...,m-1$ – порядковый номер измерения, за исключением одного, по которому рассчитывается сумма; n_x – указывает количество членов в x-м измерении; y – количество элементов в сочетании.

В случае, если необходимо вычислить количество сочетаний агрегатов в случае исчезновения членов измерений или появления новых членов измерений в количестве l у измерения n_k, в формулу (*) необходимо внести следующие изменения:

$$A = (n_k \pm l) \prod_{x=1}^{m-2} n_x \sum_{y=1}^{n_m} C_{n_m}^{y},$$

где l – количество появляющихся или исчезающих членов измерений.

Если подобные изменения имеют хаотический характер, то лучше заменить знак произведения членов измерений на раскрытую формулу произведения всех членов:

$$A = (n_1 \pm l_1)(n_2 \pm l_2)...(n_i \pm l_i) \sum_{y=1}^{n_m} C_{n_m}^{y}.$$

Формула (*) и ее производные формулы верны при любых положительных целых n_i.

Количество агрегатов

В случае трех измерений n_1, n_2, n_3 количество агрегатов можно представить так:

$$n_1 = n_{010}\ n_2 = n_{001}\ n_3 = n_{100}.$$

Эти точки задают оси, а также в случае присутствия единиц в двоичной форме записи означают наличие данного измерения, нуля – отсутствие. При перемножении измерений получаются производные от них точки

$$A^{*} = n_1 * n_2 + n_2 * n_3 + n_1 * n_3,$$

$$n_1 * n_2 = n_{011}\ n_2 * n_3 = n_{101}\ n_1 * n_3 = n_{110},$$

или то же самое можно записать в форме двоичных индексов:

$$A^* = n_{011} + n_{101} + n_{110}.$$

Общее количество измерений – m. Для приведения к общему виду необходимо учесть, что агрегация осуществляется максимум по $m-1$ измерению. В общем случае можно проводить агрегацию по $m-n$ измерениям. Чтобы рассчитать количество множеств агрегации нужно посчитать количество сочетаний z нулей по m позициям, что дает соответствующее количество слагаемых.

Формула для подсчета полного количества агрегатов может быть представлена в следующем виде:

$$A^* = \sum_{i=1}^{\sum_{z=1}^{m-z\geq 2} C_m^z} n_{x_1 x_2 x_3 \ldots x_i \ldots x_m},$$

где $x_1x_2x_3...x_i...x_m$ – двоичный вектор, состоящий из m двоичных разрядов.

Ограничение $m-z\geq 2$ указывает на необходимость двух и более измерений для агрегации. Верхний предел суммы для четырех измерений будет выглядеть следующим образом:

$C_4^1+C_4^2$, для пяти: $C_5^1+C_5^2+C_5^3$.

Случай иерархических измерений

Иерархию можно представить как объединение членов измерений в одно множество.

В этом случае для каждого k-го измерения существует t_k уровней иерархии. Общее количество членов k-го измерения состоит из суммы всех членов этого измерения

$$n_k = \sum_{m=1}^{t_k} n_{km}.$$

Количество агрегатов

Для каждого измерения A_k необходимо выбрать элемент с максимальным индексом m. Для каждого A_k может существовать свое количество m. Произведения всех индексов дадут максимально возможное количество агрегатов

$$g = \prod_{l=1}^{k} m_l.$$

Общее количество всех агрегатов получается суммированием числа агрегатов матрицы обобщенных членов измерений, определяющей всевозможные состояния агрегации

$$A^* = (n_{11} + n_{12} + \ldots + n_{1m}) * (n_{21} + n_{22} + \ldots + n_{2m}) * \ldots$$
$$\ldots * (n_{k1} + n_{k2} + \ldots + n_{km}).$$

Количество сочетаний агрегатов (иерархические измерения)

В случае иерархических измерений применима формула для подсчета количества сочетаний агрегатов. Необходимо учитывать, что сочетания различных членов иерархических измерений могут проводиться в различных уровнях иерархии

$$A = \sum_U \left(\prod_{x=1}^{m-1} n_x \sum_{y=1}^{n_m} C_{n_m}^y \right),$$

$U=U1+U2+...$ – множество всех иерархических уровней по всем измерениям

Формула является производной по отношению к формуле (*).

Количество информации

Каждый член измерения многомерного куба l_i вносит в модель, описывающую OLAP-куб, дополнительную информацию о состоянии системы. Сумма всех членов измерений L будет представлять состояние системы

$$L = \sum_i l_i.$$

Тогда количество информации системы (по формуле Хартли):

$$I = \log_2 L^{\varphi},$$

где φ – коэффициент эмерджентности Хартли.

Учитывая, что возможны смешанные состояния, являющиеся одновременной реализацией состояний системы «из L по m», всего возможно C_L^m состояний системы, являющихся сочетаниями исходных состояний. Тогда формулу для количества информации системы можно представить в виде:

$$I = \log_2 \sum_{m=1}^{M} C_L^m, \text{ при } M \leq L.$$

При $M=1$ формула приобретает вид классической формулы Хартли. Остальные слагаемые при $M>1$ дают дополнительное количество информации за счет наличия внутренних взаимосвязей системы.

Формулу можно представить также в раскрытом виде:

$$I = \log_2 (C_L^1 + C_L^2 + \ldots + C_L^m).$$

Дополнительная информация является информацией о внутренних взаимосвязях системы, состоящей из ряда подсистем различных уровней сложности. При $M=L$:

$$\sum_{m=1}^{M} C_L^m = 2^L - 1.$$

Это выражение дает оценку максимального количества информации, которое может содержаться в системе с учетом взаимосвязей различных подсистем. Подставив в качестве подлогарифмического выражения значение 2^L-1 и учитывая, что $L\to\infty$, получим, что количество информации стремится к L:

$$I = \sum_{m=1}^{M} C_L^m =_{L\to\infty} \log_2 (2^L - 1) \to L.$$

Приравняем два выражения формулы Хартли:

$$I = \log_2 L^{\varphi} = \log_2 \sum_{m=1}^{M} C_L^m.$$

Отсюда найдем коэффициент эмерджентности Хартли φ:

$$\varphi = \frac{\log_2 \sum_{m=1}^{M} C_L^m}{\log_2 L},$$

который представляет собой относительное превышение количества информации о системе при учете системных эффектов над количеством информации без учета системности. Тем самым коэффициент отражает уровень системности объекта.

Применив полученное значение для коэффициента эмерджентности, получим:

$$I = \log_2 L^{\frac{\log_2 \sum_{m=1}^{M} C_L^m}{\log_2 L}}.$$

Учитывая, что $I_{L\to\infty} \to L$, получим:

$$I = \log_2 L^{\frac{L}{\log_2 L}} = L.$$

Следовательно, количество информации в OLAP-кубе равно количеству членов измерения.

Коэффициент эмерджентности Хартли отражает уровень системности объекта и изменяется от 1 (системность минимальна) до $\frac{L}{\log_2 L}$ (системность максимальна).

Выводы

Рассмотрен и предложен достаточно универсальный обновленный подход к моделированию OLAP-кубов, опирающийся на современные онтологические и системные представления в этой области научных знаний.

СПИСОК ЛИТЕРАТУРЫ

1. Корн Г., Корн Т. Справочник по математике для научных работников и инженеров. Определения, теоремы, формулы / под общей ред. И.Г. Арамановича. – М.: Наука, 1974. – 832 с.

2. Выгодский М.Я. Справочник по элементарной математике. – М.: Физматгиз, 1962. – 420 с.

Поступила 25.01.2010 г.

УДК 004.657

СЕМАНТИКО-ЭНТРОПИЙНОЕ РЕГУЛИРОВАНИЕ ИНФОРМАЦИОННОГО МОРФИЗМА РЕАЛИЗАЦИЙ xOLAP

А.А. Миронов, А.С. Сигов

Московский государственный институт радиотехники, электроники и автоматики (технический университет)
E-mail: mironov@mirea.ru

Анализ опыта создания и сопровождения хранилищ данных говорит о том, что именно в этой области IT индустрии наиболее резко ощущаются трудности, порожденные отсутствием устоявшейся семантической теории информационных процессов и систем. Статья нацелена на изучение моделей xOLAP, целевым образом ориентированных на семантические методы управления, затрагивает понятия семантических разрывов применительно к xOLAP, их семантико-энтропийных оценок и регулирования.

Ключевые слова:
Оперативная аналитическая обработка данных, семантический разрыв, энтропийное регулирование, информационный морфизм.

Key words:
On-line analytical processing, semantic break, entropy control, information morphism.

Разнообразие версий OLAP достаточно велико и расширяется. Модели OLAP обретают новые классификационные признаки, свойства, изменяющие их особенности, достоинства и недостатки, впрочем, оцениваемые в зависимости от специфики решаемых задач. Так, наряду с такими известными модификациями как ROLAP, MOLAP и HOLAP [1], в последние годы появились и находят широкое применение SOLAP (*Spatial On-Line Analytical Processing*) – пространственная аналитическая обработка, предназначенная для изучения пространственных данных, объединяющая понятия из существенно отличающихся друг от друга сфер знаний, а именно географических информационных систем и OLAP, разработанная для интерактивного и быстрого анализа больших объемов данных; R-ROLAP (*Real-time ROLAP*) – OLAP реального времени, в отличие от ROLAP в R-ROLAP для хранения агрегатов не создаются дополнительные реляционные таблицы, а агрегаты рассчитыва-