УДК 656.13

МНОГОМЕРНАЯ МОДЕЛЬ ПРЕДСТАВЛЕНИЯ ДАННЫХ МОНИТОРИНГА ПЕРЕВОЗОК

С.В. Пронин, доцент, к.т.н., ХНАДУ

Аннотация: *В статье рассматривается подход к моделированию многомерных моделей данных для их применения в системах анализа и представления информации на автомобильном транспорте.*

Ключевые слова*: многомерная модель данных, агрегация, база данных, гиперкуб*

БАГАТОМІРНА МОДЕЛЬ ПРЕДСТАВЛЕННЯ ДАНИХ МОНІТОРИНГУ ПЕРЕВЕЗЕНЬ

С.В. Пронін, доцент, к.т.н., ХНАДУ

Анотація: *У статті розглядається підхід до моделювання багатомірних моделей даних для їхнього застосування в системах аналізу та представлення інформації на автомобільному транспорті.*

Ключові слова*: багатомірна модель даних, агрегація, база даних, гіперкуб*

MULTIDIMENSIONAL MODEL PROVIDE MONITORING TRANSPORT

S. Pronin, assistant professor, cand. eng. sc., KhNAHU

Abstract*: The article discusses the approach to modeling multi-dimensional data models for application in systems analysis and report on road transport.*

Keywords*: multidimensional data model, aggregation, database, hypercube*

Введение

Современный этап развития баз данных характеризуется активным исследованием различных подходов основанных на использовании как реляционных так и нереляционных баз данных возникающих при создании систем мониторинга построенных на основе информационных технологий и веб-приложений.

Также современные веб-приложения развиваются в направлении реализации функций аналитики (Web OLAP), предоставляющих пользователям гибкий, простой и удобный доступ к гиперкубам данных с разной степенью детализации.

Традиционные, ориентированные на OLAP, хранилища данных (Data Warehouse) имеют специальную организацию данных в виде многомерной модели (Multidimensional Model) или особой реляционной модели данных (структуры типа «звезда» или «снежинка»).

Анализ публикаций

Основным требованием к системам OLAP является скорость выполнения запросов, так как анализ должен проходить в интерактивном режиме. Предложенные в литературе алгоритмы OLAP основаны на дисковых структурах данных или структурах данных в оперативной памяти. Дисковые

структуры данных являются медленными или вынуждены хранить практически полностью агрегированные кубы для достижения скорости, что приводит к большим расходам памяти. Структуры в оперативной памяти могут обрабатывать лишь небольшие объёмы данных. Если объёмы данных очень велики, то предагрегация может значительно ускорить выполнение запросов. Также агрегирование может применяться для ответа на запросы пользователя с одновременным требованием просмотра многих агрегатных данных (например, при отображении сводной таблицы).

Первые алгоритмы агрегирования куба основываются на существенном использовании диска и являются достаточно медленными. Алгоритмы MemoryCube и BUC компактно используют оперативную память для проведения вычислений, но их планы выполнения являются неоптимальными.

Задача агрегирования куба состоит в вычислении всех агрегатных данных по заданным исходным данным, что соответствует одновременному выполнению агрегатных запросов для всех подмножеств множества измерений, или подкубов.

Алгоритмы выполнения отдельного агрегатного запроса делятся на две группы: основанные на сортировке и основанные на хешировании. Алгоритмы агрегирования куба являются обобщениями этих алгоритмов с применением ряда оптимизаций: разделение сортировок и разбиений между подкубами, одновременное вычисление нескольких подкубов, вычисление по родительскому подкубу вместо исходного отношения. В настоящее время для решения вышеперечисленных задач разработаны такие алгоритмы как PipeSort, PipeHash, Overlap, ArrayCube, PartitionedCube, MemoryCube, BUC и другие [1].

Алгоритм агрегации данных в многомерной модели

Сформируем многомерную модель данных по результатам мониторинга пассажирского транспорта на основе многомерной базы данных которая предоставляет информацию по показателям работы маршрутов городского пассажирского транспорта.

Измерения сыграют роль индексов, используемых для идентификации значений показателей, которые находятся в ячейках гиперкуба. Комбинация членов разных измерений сыграют роль координат, которые определяют значение определенного показателя. Поскольку для куба может быть определено несколько показателей, то комбинация членов все измерения будет определять несколько ячеек со значениями каждого из показателей.

Иерархии в измерениях необходимые для возможности агрегации и детализации значений показателей соответственно иерархической структуре.

В этом случае можно применить сбалансированную иерархию [2]. Степень агрегации такого куба можно определить следующим выражением:

$$a = \frac{a}{a^*}, \qquad (1)$$

где a – реальное количество агрегированных значений показателей, a* – максимально возможное количество агрегатных значений исходных данных куба.

Обозначим R, D, P – множества членов соответствующих измерений «респондент», «время», «параметры движения». Обозначим также количество членов в каждом из измерений: nr = |P|, nd = |D|, np = |P|. Члены этих измерений будем обозначать соответственно mr, md, mp.

Для получения агрегированных значений в разрезе маршрутов и времени мы должны просуммировать первоначальные значения показателей по всем моделям для каждой комбинации (md, mp) маршрутов и времени. Нетрудно понять, что количество агрегированных таким образом значений равняется ndnp.

По аналогии, получим число агрегатов для всех комбинаций (mr, mp) при суммировании показателей по всем членам измерения “маршруты”. Оно равняется nrnp.

Количество агрегатов для всех комбинаций

(mr, md) при агрегации по временному измерению равняется ndnr.

Теперь необходимо получить число агрегатов в разрезе членов одного из измерений. Очевидно, что количество таких агрегатов равняется числу членом соответствующего измерения nd, nr и np.

Учитывая значение полного агрегата, который определяет в нашем случае суммарный объем продаж по всем транспортным средствам, маршрутам и всему временному периоду, получим суммарное количество всех агрегатов:

$$a^* = n_r n_d + n_d n_p + n_r n_p + n_r + n_d + n_p + 1, \quad (2)$$

Для дальнейших соображений введем новые обозначения. Допустим, имеем m измерений с ni количество членов в i-ом измерении, где i = 1..m. Приведем в порядок некоторым образом существующие измерения и сопоставим каждому измерения порядковый номер i соответственно указанной сортировке. Обозначим множество таких порядковых номеров I, так что m = | I |.

Для трехмерного случая с учетом того, что i(R)=1, i(D)=2, i(P)=3 – порядковые номера измерений, получим:

$$a^* = a_{011} + a_{101} + a_{110} + a_{001} + a_{010} + a_{100} + a_{000}. \quad (3)$$

Легко понять, что количество агрегатов, полученных агрегированием по некоторым измерениям, равняется произведению числа членов всех других измерений, т. е.

$$a_{l1..li..lm} = \prod_{i \in I \setminus I^0} n_i. \quad (4)$$

Очевидно, что количество всяких агрегатов равняется сумме al1…li…lm при всяческих вариантов последовательности l1,…,li,…,lm, кроме множества исходных данных. Нетрудно понять, что таких вариантов всего 2m -1.

С другой стороны при агрегации исходных данных по всяческим k измерениях, где k E {1,…,m},мы получаем некоторую совокупность множества агрегации с тем самым уровнем детализации, равным l = m - k. Обозначим эту совокупность Al, с количеством агрегатов al. Для расчета количества множества агрегации Al1…li…lm в Am-k нам необходимо в индексе сосчитать количество вариантов размещения k нулей по m позициях. Это случай соединения k элементов по m, равный Ckm.

Формула расчета полного числа агрегатов с учетом (4) может быть представленная так

$$a^* = \sum_{k=1}^{m} \sum_{j=1}^{C_m^k} \prod_{i=1}^{m-k} n_i. \quad (5)$$

Кроме числа исходных данных, получим:

$$a^* = \prod_{i=1}^{m} (n_i + 1) - \prod_{i=1}^{m} n_i. \quad (6)$$

Процедура формирования агрегатов

Расчет множества агрегатов Al1…li…lm может быть описанный следующим выражением:

$$A_{l1..li..lm} = \left\{ A_{i, Al1..li-1..lm} * i \right\}, i = 1...n. \quad (7)$$

Так как затраты на каждую из этих операции агрегации разные, то из всех возможных альтернатив агрегации выбираем ту, которая требует наименьших вычислительных расходов.

При распределение агрегатов между множествами одного уровня детализации рассмотрим совокупность множества агрегации Al, что отвечают уровню детализации l (l=1...1... …1*). Допустим, на этот момент все множества большей степени детализации, чем l, сформированное и общее их количество равняется a’. Обозначим множество всех агрегатов уровня l как al. В зависимости от заданной для текущего куба степени агрегации возможные следующие варианты дальнейшего обращения.

Оцениваемое суммарное количество агрегатов, которые отвечают уровню детализации более или равной l, не превышает разрешенного количества агрегатов a, обусловленного заданной степенью детализации, т. е. a’+ ai<= a. В этом случае полностью формируются все множества агрегатов Al1…li…lm, соответствующему уровню детализации l.

Оцениваемое суммарное количество агрегатов, которые отвечают уровню детализации более или равной l, превышает разрешенное количество агрегатов a, обусловленное заданной степенью детализации, т.е. a'+ al> a. Поскольку нет оснований для определения приоритетов и преимуществ среди агрегатов l-го уровня, то остаточное количество агрегатов равноценно распределяем по этому множеству. Равноценность припускает распределение остаточного количества агрегатов сред множеств l-го уровня детализации в прямой пропорции максимально возможному количеству их элементов al1…li…lm.

Вывод

В статье был рассмотрен и предложен достаточно универсальный подход для моделирования многомерных моделей данных для их применения в системах анализа и представления информации на автомобильном транспорте.

Литература

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009. — 512 с.
2. Хрусталёв Е.М. Агрегация данных в OLAP-кубах [Электронный ресурс]: Режим доступа: URL: http://www.olap.ru/home/mut.asp

Рецензент А.В. Бажинов профессор, д.т.н., ХНАДУ

Статья поступила в редакцию 14.05.15 г.
.