является то, насколько система сама может определить качество распознавания и исправлять ошибки. Например, в качестве критерия распознавания можно использовать минимум разности между полученными и эталонными значениями. Перспективным является применение технологии адаптивного распознавания, когда формирование эталонов происходит в процессе распознавания и система в случае ошибки проводит "дораспознавание". Также, несмотря на ресурсоемкость, надежной является система голосования машин, при условии, что параллельно работающие машины распознавания основаны на различных алгоритмах и обучающих выборках.

Список литературы: 1. *Ford R.* Identifications and Monitoring // Modern Railways, 1998, N 592. C. 25-27. 2. *Rahn W.-H.* Signal und Draht // Modern Railways, 1998. N 9. C. 5-8. 3. *Бондарев В., Трестер Г., Чернега В.* Цифровая обработка сигналов: методы и средства. Севастополь: СевГТУ, 1999. 397 с. 4. *Ян Д.Е., Анисимович К.В., Шамис А.Л.* Новая технология распознавания символов. М.: Препринт. 1995. 150с. 5. *Аркадьев А.Г., Браверман Э.М.* Обучение машины классификации объектов. М.:Наука, 1971, 192с. 6. *Васильев В.И.* Распознающие системы. К.: Наук. думка, 1969, 292с.

Поступила в редколлегию 14.05.01

Горелова Елена Викторовна, магистр кафедры программного обеспечения ЭВМ ХТУРЭ. Научные интересы: программирование систем управление и контроля, изучение операционных систем реального времени. Адрес: Украина, 61166, Харьков, пр. Ленина, 14.

Белоус Наталья Валентиновна, канд. техн. наук, доцент кафедры программного обеспечения ЭВМ ХТУРЭ. Научные интересы: математическое моделирование сложных объектов и систем, теоретические основы дискретной математики. Адрес: Украина, 61166, Харьков, пр. Ленина, 14, (0572) 40-98-21, e-mail: lgorelik@ukrpost.net.

УДК 681.519

А.В. КЛИМЕНКО

АВТОМАТИЗАЦИЯ ПРОЦЕДУРЫ ФОРМИРОВАНИЯ МНОГОМЕРНЫХ БАЗ ДАННЫХ

Рассматривается вопрос конвертирования реляционных баз данных (БД) в многомерные для информационных систем, использующих многомерную модель организации данных. Предлагается алгоритм формирования многомерных БД, позволяющий повысить скорость обработки информации.

В настоящее время все больше организаций убеждаются в том, что без наличия своевременной и объективной информации о состоянии рынка, прогнозирования его перспектив, постоянной оценки эффективности функционирования собственных структур и анализа взаимоотношений с бизнес партнерами и конкурентами их дальнейшее развитие практически невозможно. Поэтому не удивительно то внимание, которое сегодня уделяется средствам реализации и концепциям построения информационных систем, ориентированных на аналитическую обработку данных. И в первую очередь это касается систем управления базами данных, основанными на многомерном подходе.

Многомерная модель организации данных не зависит от физической природы их хранения. Строительными блоками реляционной модели являются сущности, каждая из которых впоследствии представляет собой отдельную таблицу. Строительными блоками многомерной модели являются таблицы фактов (иногда называемые таблицами показателей) и измерений, организованные в специальные структуры данных (схемы). Последние включают характерные для организационных объектов понятия: подразделения, временные периоды, объемы показателей и т. д. Таблица фактов содержит в численном выражении то, что в результате необходимо выяснить.

Основные из них – схемы звездочки, наиболее популярные во многомерном моделировании, и снежинки.

Таблицы измерений представляют собой более сложную структуру, чем таблицы фактов. В них можно выделить следующие основные понятия: элементы, иерархию и атрибуты.

Элементы отражают иерархию измерения. Например, дни, недели и месяцы составляют иерархию "Время". Элементы более низкого уровня могут быть свернуты в элементы более высокого. В одном измерении может быть несколько иерархий. Скажем, дни свернуты в недели и месяцы, а недели в месяц свернуть нельзя, поскольку он не делится без остатка на недели. Поэтому в одном измерении имеются две ветки иерархии. Значение иерархий заключается в том, что их можно применять для детализации и обобщения анализа. Кроме того, элементы можно использовать для фильтрации данных.

Следующей популярной схемой представления данных (после схемы звездочки) является схема снежинки, та же схема звездочки, но с нормализованными таблицами измерений. В ней атрибуты элементов выносятся в отдельные таблицы. Измерения с теми же названиями сохраняются, но служат для связи с таблицами фактов и атрибутов (по одной на каждый элемент измерения). Скорость выполнения запросов несколько уменьшается, но только в том случае, когда необходимы таблицы фактов. А если речь идет о манипуляциях с таблицами измерений, скорость обработки, наоборот, резко увеличивается из-за уменьшения количества обрабатываемых строк и значительно сокращается размер памяти, требуемый для хранения. К недостаткам следует отнести усложнение схемы данных.

И, наконец, в качестве компромисса используется схема неполной снежинки. В ней нормализации подвергаются лишь отдельные измерения. Это дает возможность повысить производительность при некоторой экономии дискового пространства.

Многомерную модель организации данных можно реализовать с помощью процедуры формирования многомерных БД. В этом представлении данные хранятся как совокупность логически упорядоченных массивов, отражающих их многомерную природу. Многомерное представление данных позволяет уменьшить объем занимаемой памяти за счет устранения избыточности хранимых данных и увеличить скорость обработки и анализа данных путем более простой и эффективной системы индексации. Поскольку отдельный элемент хранения в многомерной БД крупнее, чем в реляционной, то индекс, соответствующий ему, имеет меньший размер.

Многомерная модель данных представляется в виде множества классов (орт), которым соответствует точка в пространстве. Каждый элемент пространства имеет два состояния – активное или пассивное. Активное состояние он принимает в случае, если значения его координат определены на множестве активных классов (значения орт, которые должны быть обязательно определены). В противном случае он принимает пассивное состояние. При этом в памяти хранятся только активные элементы, что значительно уменьшает объем занимаемой памяти. Запросы на выборку данных в многомерной модели осуществляются с помощью сечений по соответствующим ортам, а результатом являются активные элементы на пересечении всех сечений. При этом значения активных элементов определяются по их координатам.

Если необходимо использовать несколько многомерных БД, то связь между ними устанавливается по соответствующим классам, а результат выбора данных в одной из них будет условием поиска (сечением) в других связанных с данным классом многомерных базах данных.

Таким образом, многомерная модель данных позволяет уменьшить объем занимаемой памяти, быстро производить выбор и анализ данных, а также обеспечивает целостность всей системы.

Для создания многомерной модели данных разработан алгоритм представления многомерной базы данных (МБД), который основывается на разбиении разреженной многомерной БД на несколько более плотных МБД.

Алгоритм разбиения разреженной многомерной БД на множество плотных МБД.

Дано:

Многомерная разреженная БД A, характеризующаяся множеством измерений:

$$A = \{a_1, a_2, ..., a_j, ..., a_n\},$$

где n – количество измерений МБД или количество характеристик организационной системы.

Каждое измерение a_j имеет свою размерность p_j, определяемую множеством значений a_j.

К МБД A при функционировании информационных систем осуществляется множество различных запросов на выбор информации E, которые назовем структурными элементами:

$$E = \{e_1, e_2, ..., e_i, ..., e_m\},$$

здесь m – количество структурных элементов.

Каждый элемент e_i характеризуется набором измерений, к которым он обращается при реализации запроса

$$e_1 = \{a'_{11}, a'_{12}, ..., a'_{1i_1}, ..., a'_{1k_1}\},$$

$$e_2 = \{a'_{21}, a'_{22}, ..., a'_{2i_2}, ..., a'_{2k_2}\},$$

$$\cdot$$

$$e_m = \{a'_{m1}, a'_{m2}, ..., a'_{1i_m}, ..., a'_{mk_m}\},$$

где $a'_{11}, a'_{12}, ...$ – элементы принадлежащие множеству A, т.е. $\bigcup a'_{ij} \subseteq A$.

Необходимо МБД A разбить на несколько МБД меньшей размерности, таким образом, чтобы при реализации запросов E образовывалось наименьшее количество связей между ними, т.е.

$$A \to \{A'_1, A'_2, ..., A'_{m_n}\},$$

m′ – количество МБД.

Решение задачи достигается путем объединения элементов множества E, т.е.

$$\bigcup_{i=1}^{m} e_i : \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} \to \min,$$

где $x_{ij} = \begin{cases} 1, \text{если } a'_{ij} = a_j, \\ 0, \text{в другом случае}. \end{cases}$ и при ограничениях

$$\sum_{i=1}^{s}\left(\prod_{j=1}^{n} p_j x_{ij}\right) < M^* : x_{ij} \# 0;$$

Здесь $\sum_{i=1}^{s}\left(\prod_{j=1}^{n} p_j x_{ij}\right)$ – размер сформированных МБД; M^* – максимальный размер формируемых МБД, определяемый как

$$M^* < \sum_{i=1}^{m}\left(\prod_{j=1}^{n} p_j x_{ij}\right) : x_{ij} \# 0.$$

Алгоритм разбиения разреженной МБД на множество плотных следующий.

1 шаг: Каждому структурному элементу e_i задается отдельная МБД A'_i. При этом количество МБД $m' = m$.

2 шаг: Формируются матрица $X = |x_{ji}|$ принадлежности измерений МБД к структурным элементам и матрица $D = |d_{ij}|$ связей между структурными элементами:

$$x_{ji} = \begin{cases} 1, \text{если } a'_{ji} = a_i, \\ 0, \text{в другом случае;} \end{cases} \quad d_{ij} = \sum_{k=1}^{m} (x_{ik} \wedge x_{jk}).$$

3 шаг: Рассчитывается матрица емкости $L = \{l_k\}, k = [1, m']$, определяющая размер каждой МБД A'_p:

$$l_k = \sum_{i=1}^{n} p_i x_{ki}.$$

4 шаг: Рассчитывается матрица весовых коэффициентов $V = v_{ij}$, $i = [1, n]$, $j = [1, m']$, характеризующая связь между структурными элементами:

$$v_{ij} = \frac{d_{ij}}{d_{ii} + d_{jj}}.$$

5 шаг. Определяется максимальное значение матрицы V, т.е.

$$\max |V| = v_{i'j'},$$

где i' и j' – индексы максимального элемента матрицы V.

6 шаг: Происходит сравнение общего объема МБД A′ с заданным значением,

$$\sum_{k=1}^{m'} l_k \le M.$$

7 шаг: Если условие 6 шага истинно, то происходит объединение 2-х структурных элементов или 2-х МБД в одну, т.е.

$$x_{i'j} \, Y \, x_{j'j}, j = [1, m] \text{ или } A'_i \, Y \, A'_j, \; m' = m' - 1$$

и происходит переход на шаг 2.

8 шаг. Если условие 6 шага ложно, то осуществляется поиск следующего максимального элемента и происходит переход на шаг 6. Поиск продолжается пока не будет найден максимальный элемент или не будут перебраны все элементы.

9 шаг. Формируются МБД A′ согласно матрице X, т.е.

$$A'_k = \{a_1 x_{1k}, a_2 x_{2k}, ..., a_j x_{jk}, ..., a_n x_{nk}\}, \; k = [1, m'].$$

Предлагаемый алгоритм разбиения разреженной МБД на множество плотных МБД реализован автором при разработке ряда информационных систем.

Литература: 1. *Warehouse* Cornerstones. BYTE, jan.,1997. P.85-90. 2. Groupware enablers and business solutions / *Morton Marjorie S., Holman Richard A., Bess David A.* // IEEE Int. Conf. Syst., Man. and Cybern. "Emergent Innov. Inf. Transfer Process and Decis. Mak.", Chicago, I11., Oct. 18-21, 1992. Vol. 2. Piscataway (N.J.), 1992. P. 943-948. 3. .*Linthicum D.S.* Power Tools for Date Drilling. BYTE, jan., 1997. P.143-144. 4. *Deje-sus E.X.* Dimensions of Data. BYTE,1995, apr. p.139-143. 5. *Информационные* системы для обеспечения деятельности кафедр высших учебных заведений / Гайдаров К.А.; Изв.вузов. Геод. и аэрофотосъемка. 1992. N2. P.166-176.

Поступила в редколлегию

Клименко Александр Васильевич, канд. техн. наук, старший преподаватель кафедры информационных управляющих систем ХТУРЭ. Научные интересы: управление распределенными информационными системами. Увлечения и хобби: путешествия. Адрес: Украина, 61166, Харьков, пр.Ленина 14, тел. 409-451.