ИНФОРМАТИКА И ПРОГРАММИРОВАНИЕ

УДК 004.652

О.А. Бистерфельд

ОЦЕНКА РЕСУРСОВ ПАМЯТИ, НЕОБХОДИМЫХ ДЛЯ РЕАЛИЗАЦИИ КАТЕГОРИРОВАННЫХ ОТНОШЕНИЙ В РЕЛЯЦИОННЫХ БАЗАХ ДАННЫХ

Предложена модель базы категорированных данных, учитывающая объем и особенности распределения данных по категориям. Выведены формулы зависимостей требуемых ресурсов памяти от параметров структуры категорированных данных для наиболее распространенных вариантов представления категорированных отношений. Аналитические выражения метода уточнены с помощью имитационной программы. Получаемые оценки используются при решении задачи выбора вариантов представления категорированных данных.

Ключевые слова: *базы данных*, *отношения категоризации*, *объем памяти.*

1. Способы представления категорированных отношений

В базах данных (БД) информационных систем значительная часть данных связана отношениями категоризации.

Наиболее развиты способы реализации категорированных отношений в СУБД Oracle. Варианты СУБД Oracle охватывают практически все возможные варианты других СУБД, поэтому они приняты за основу при разработке аналитического метода оценки объемов памяти, необходимых для реализации категорированных отношений в реляционных БД. В многотомном описании методологии проектирования Oracle [1, с. 71–76.] приведены только словесные описания вариантов реализации с кратким перечнем достоинств и недостатков каждого.

На рис. 1 показана модель отношения категоризации по нотации Баркера («супертип» – «тип»).

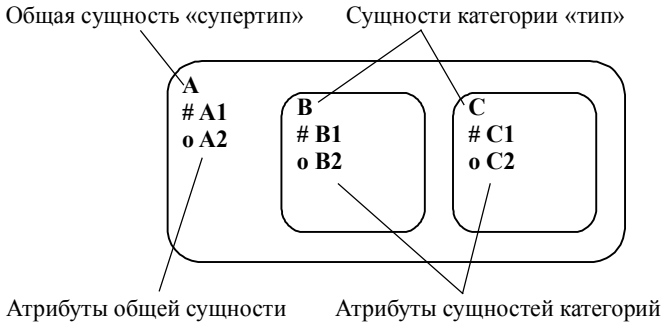

Рис. 1. Базовая ER-диаграмма

Категорированные данные могут быть представлены (рис. 2):
- в одной таблице (вариант *а*);
- в нескольких (по количеству категорий) таблицах (вариант *б*);
- в нескольких (по количеству категорий + одна) таблицах (варианты *в*, *г*).

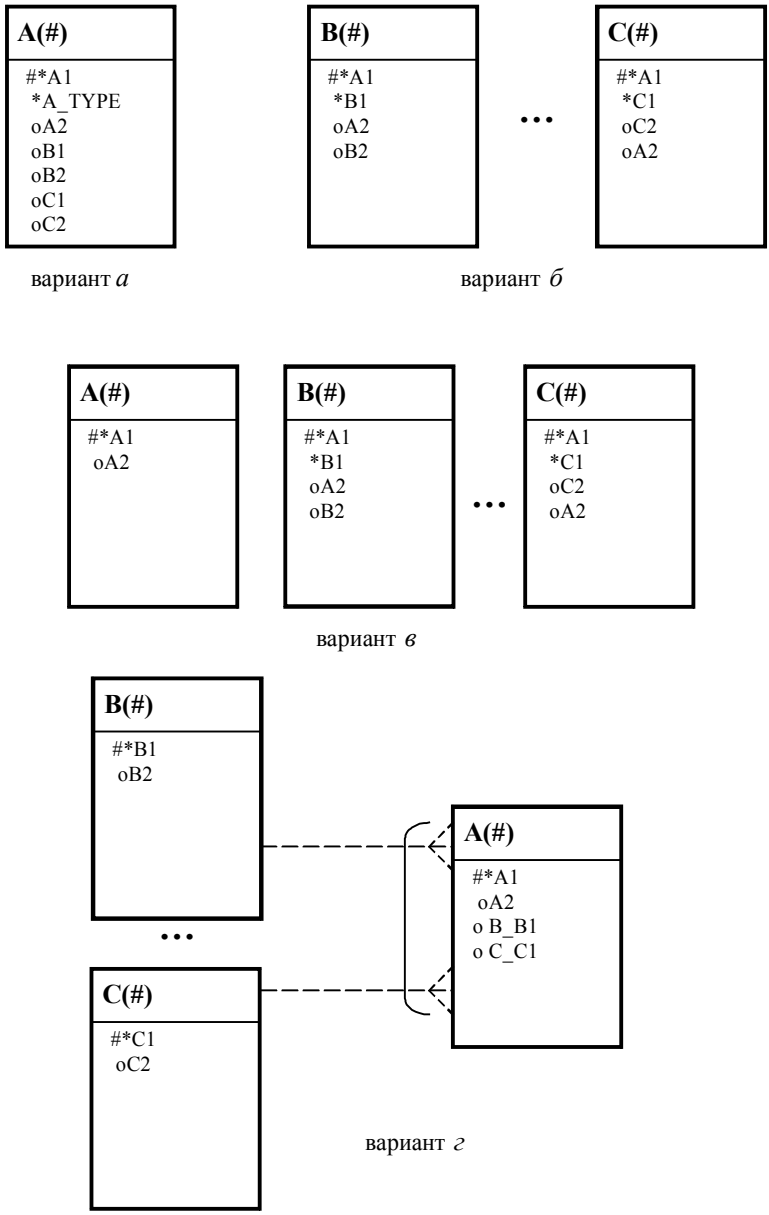

Рис. 2. Варианты реализации категорированных отношений

2. Аналитический метод оценки различий в требуемых ресурсах памяти БД вариантов представления категорированных отношений

Введем обозначения: N – число записей в БД; $N_{\max}$ – максимально возможное число записей в БД; $K_{\text{у.и}}$ – размер уникального идентификатора; $A_{\text{о.а}}$ – размер общих атрибутов; $K_{\text{к}}$ – число категорий; A_i – размер частных атрибутов (для категории i); N_i – число записей категории i; $N_{i\max}$ – максимально возможное число записей категории i; $K_{\text{у.и}i}$ – размер уникального идентификатора для категории i. Должны выполняться следующие неравенства:

$$K_{\text{у.и}} >= \log_2 N_{\max}\ ;\ K_{\text{у.и}i} >= \log_2 N_{i\max}\ , \tag{1}$$

однако на практике для уникальных идентификаторов записей используют десятичные числа, поэтому имеем

$$K_{\text{у.и}} >= \lg N_{\max}\ ;\ K_{\text{у.и}i} >= \lg N_{i\max}\ . \tag{2}$$

Кроме того,

$$\sum_{i=1}^{k_k} N_i = N\ . \tag{3}$$

Пределы суммирования во всех формулах данной статьи одинаковы, поэтому ниже они не указываются (вместо $\sum_{i=1}^{k_k}$ используется Σ).

В некоторых упрощениях (при несущественном влиянии на результат) принимается также, что $K_{\text{у.и}i} = K_{\text{у.и}}$.

Объем памяти по варианту *а* – O_a (без учета несущественного по объему атрибута TYPE):

$$\text{O}_a = N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, A_i). \tag{4}$$

Объем по варианту *б* – ($\text{O}_б$):

$$\text{O}_б = \Sigma\, N_i(K_{\text{у.и}} + A_{\text{о.а}} + A_i). \tag{5}$$

Объем по варианту *в* – ($\text{O}_в$):

$$\text{O}_в = N(K_{\text{у.и}} + A_{\text{о.а}}) + \Sigma\, N_i(K_{\text{у.и}} + A_{\text{о.а}} + A_i). \tag{6}$$

Объем по варианту *г* – ($\text{O}_г$):

$$\text{O}_г = N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, K_{\text{у.и}i}) + \Sigma\, N_i(K_{\text{у.и}i} + A_i). \tag{7}$$

Для анализа эффективности различных вариантов представления категорированных отношений в реляционных БД представляют интерес:

- относительные оценки: $(\text{O}_a - \text{O}_б)/\text{O}_a$; $(\text{O}_a - \text{O}_в)/\text{O}_a$; $(\text{O}_a - \text{O}_г)/\text{O}_a$;
- абсолютные оценки: O_a; $\text{O}_a - \text{O}_б$; $\text{O}_a - \text{O}_в$; $\text{O}_a - \text{O}_г$.

С учетом формул (1) – (7):

$$(\text{O}_a - \text{O}_б)/\text{O}_a = \{N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma A_i) - \Sigma N_i(K_{\text{у.и}} + A_{\text{о.а}} + A_i)\}/N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma A_i); \tag{8}$$

$$(\text{O}_a - \text{O}_в)/\text{O}_a = \{N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, A_i) - [N(K_{\text{у.и}} + A_{\text{о.а}}) + \Sigma\, N_i(K_{\text{у.и}} + A_{\text{о.а}} + A_i)]\}/$$
$$/N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, A_i); \tag{9}$$

$$(\text{O}_a - \text{O}_г)/\text{O}_a = \{N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, A_i) - [N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, K_{\text{у.и}i}) + \Sigma\, N_i(K_{\text{у.и}} + A_i)]\}/$$
$$/N(K_{\text{у.и}} + A_{\text{о.а}} + \Sigma\, A_i). \tag{10}$$

После преобразований, с учетом формул (2), (3) и некоторых приближений:

$$(\text{O}_a - \text{O}_б)/\text{O}_a = A_{\text{о.а}}[\Sigma A_i/A_{\text{о.а}} - \Sigma\,(N_i/N)(A_i/A_{\text{о.а}})]/[A_{\text{о.а}}(1 + \Sigma A_i/A_{\text{о.а}}) + \lg N_{\max}]; \tag{11}$$

$$\text{O}_a - \text{O}_б = NA_{\text{о.а}}[\Sigma A_i/A_{\text{о.а}} - \Sigma(N_i/N)(A_i/A_{\text{о.а}})]; \tag{12}$$

$$O_a - O_в)/O_a = \{A_{o.a}[\Sigma A_i/A_{o.a} - 1 - \Sigma(N_i/N)(A_i/A_{o.a})] - \lg N_{max}\}/$$
$$/[A_{o.a}(1+ \Sigma A_i/A_{o.a}) + \lg N_{max}]; \quad (13)$$

$$O_a - O_в = N\{A_{o.a}[\Sigma A_i/A_{o.a} - 1 - \Sigma(N_i/N)(A_i/A_{o.a})] - \lg N_{max}\}; \quad (14)$$

$$(O_a - O_г)/O_a = \{A_{o.a}[\Sigma A_i/A_{o.a} - \Sigma(N_i/N)(A_i/A_{o.a})] - \Sigma \lg N_{imax} - \lg N_{max}\}/$$
$$/[A_{o.a}(1+\Sigma A_i/A_{o.a}) + \lg N_{max}]; \quad (15)$$

$$(O_a - O_г) = N\{A_{o.a}[\Sigma A_i/A_{o.a} - \Sigma (N_i/N)(A_i/A_{o.a})] - \Sigma \lg N_{imax} - \lg N_{max}\}; \quad (16)$$

$$O_a = N[A_{o.a}(1 + \Sigma A_i/A_{o.a}) + \lg N_{max}]; \quad (17)$$

$$O_б = N[A_{o.a}\{1 + \Sigma (N_i/N)(A_i/A_{o.a})\} + \lg N_{max}]; \quad (18)$$

$$O_в = N[A_{o.a}\{2 + \Sigma (N_i/N)(A_i/A_{o.a})\} + 2\lg N_{max}]; \quad (19)$$

$$O_г = N[A_{o.a}\{1 + \Sigma (N_i/N)(A_i/A_{o.a})\} + \Sigma \lg N_{imax} + 2\lg N_{max}]. \quad (20)$$

Выражения (11) – (16) могут быть использованы для проведения исследований и анализа эффективности различных вариантов представления категорированных отношений в реляционных БД [2, с. 86–87].

3. Модель базы категорированных данных

Следует обратить внимание на набор компонентов выражений (11) – (20):

$$\{N; N_{max}; N_{imax}; A_{o.a}; \Sigma A_i/A_{o.a}; \Sigma (N_i/N)(A_i/A_{o.a})\}.$$

Эти компоненты достаточно адекватно представляют реальную БД и особенности структуры данных.

Введем обозначения: $a_i = A_i/A_{o.a}$; $n_i = N_i/N$ и преобразуем формулы (11) – (20):

$$(O_a - O_б)/ O_a = A_{o.a}(\Sigma a_i - \Sigma n_i a_i)/[A_{o.a}(1 + \Sigma a_i) + \lg N_{max}]; \quad (21)$$

$$O_a - O_б = NA_{o.a}(\Sigma a_i - \Sigma n_i a_i); \quad (22)$$

$$(O_a - O_в)/O_a = [A_{o.a}(\Sigma a_i - 1 - \Sigma n_i a_i) - \lg N_{max}]/[A_{o.a}(1+ \Sigma a_i) + \lg N_{max}]; \quad (23)$$

$$O_a - O_в = N[A_{o.a}(\Sigma a_i - 1 - \Sigma n_i a_i) - \lg N_{max}]; \quad (24)$$

$$(O_a - O_г)/O_a = [A_{o.a}(\Sigma a_i - \Sigma n_i a_i) - \Sigma \lg N_{imax} - \lg N_{max}]/[A_{o.a}(1+ \Sigma a_i + \lg N_{max})]; \quad (25)$$

$$(O_a - O_г) = N[A_{o.a}(\Sigma a_i - \Sigma n_i a_i) - \Sigma \lg N_{imax} - \lg N_{max}]; \quad (26)$$

$$O_a = N[A_{o.a}(1 + \Sigma a_i) + \lg N_{max}]; \quad (27)$$

$$O_б = N[A_{o.a}(1 + \Sigma n_i a_i) + \lg N_{max}]; \quad (28)$$

$$O_в = N[A_{o.a}(2 + \Sigma n_i a_i) + 2\lg N_{max}]; \quad (29)$$

$$O_г = N[A_{o.a}(1 + \Sigma n_i a_i) + \Sigma \lg N_{imax} + 2\lg N_{max}]. \quad (30)$$

Формулы (21) – (30) положены в основу аналитического метода оценки требуемых объемов памяти по вариантам представления категорированных отношений.

Целесообразно использовать набор $\{ N; N_{max}; |N_{imax}|; A_{o.a}; |a_i|; |n_i| \}$ как модель базы категорированных данных для исследований зависимостей объемов памяти, необходимых для реализации категорированных отношений.

Замечание. Если Σa_i имеет простой «физический» смысл (эта сумма показывает соотношение между суммарными размерами общих и частных атрибутов), то с компонентом $\Sigma n_i a_i$ несколько сложнее. Диапазон значений $\Sigma n_i a_i$: $- 0 < \Sigma n_i a_i < \Sigma a_i$. Если значения $\Sigma n_i a_i$ в левой части диапазона (от 0,5 Σa_i к 0), то в БД незначительна доля записей, относящихся к категориям, имеющим существенные размеры част-

ных атрибутов, и, наоборот, существенна доля записей, относящихся к категориям, имеющим незначительные размеры частных атрибутов. Если значения $\Sigma\, n_i\, a_i$ в правой части диапазона (от 0,5 Σa_i к Σa_i), то в БД значительна доля записей, относящихся к категориям, имеющим существенные размеры частных атрибутов, и, наоборот, несущественна доля записей, относящихся к категориям, имеющим незначительные размеры частных атрибутов.

4. Зависимости затрат памяти от параметров модели базы категорированных данных

Некоторые зависимости затрат памяти, полученные по выражениям (21)-(30), приведены на диаграммах (рис. 3, 4).

Зависимость затрат памяти от структуры категорированных записей ($\Sigma a_i n_i$)
($N = 500\,000$; $\Sigma a_i = 10$; $\lg N_{\max} = 6$; $A_{\text{о.а}} = 512$ симв.)

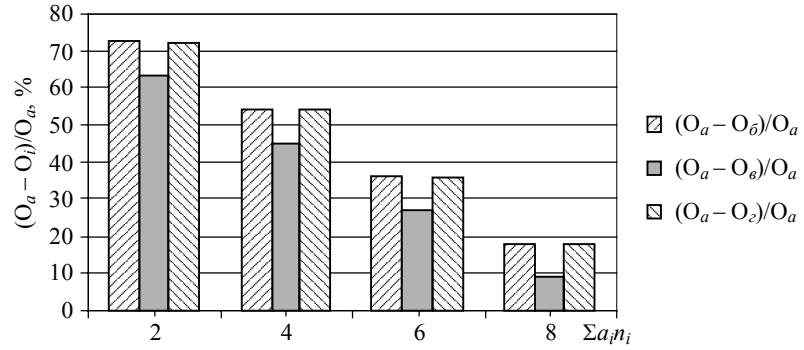

Рис. 3. Уменьшение (экономия) ресурсов памяти при переходе от варианта представления категорированных данных *а* к вариантам *б*, *в*, *г* в зависимости от структуры категорированных данных ($\Sigma a_i n_i$) в БД

Зависимость затрат памяти от соотношения общих и частных атрибутов (Σa_i)
($\Sigma n_i a_i = 0{,}5\Sigma a_i$; $N = 500\,000$; $\lg N_{\max} = 6$; $A_{\text{о.а}} = 512$ симв.)

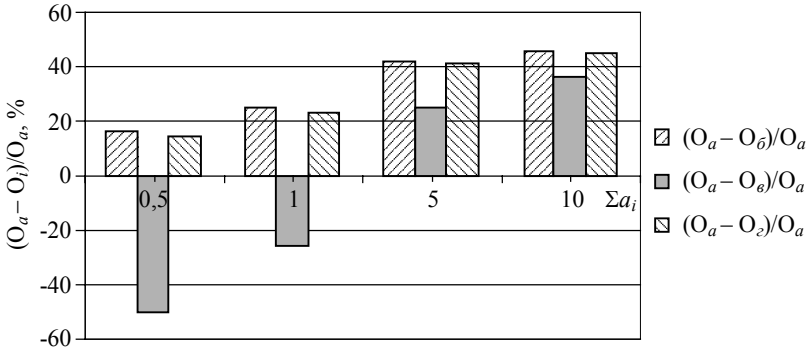

Рис. 4. Уменьшение (экономия) ресурсов памяти при переходе от варианта представления категорированных данных *а* к вариантам *б*, *в* и *г* в зависимости от соотношения общих и частных атрибутов (Σa_i)

Полученные с помощью аналитических выражений данные показывают потенциальные возможности минимизации затрат ресурсов памяти с помощью выбора вариантов представления категорированных данных. В предельных случаях ресурсы памяти, необходимые для реализации категорированных данных, могут быть сокращены на 60 – 70 %.

5. Имитационная программа для уточнения аналитических выражений метода оценки требуемых ресурсов памяти

В аналитических выражениях метода оценки объемов памяти (разделы 2, 3) единицей измерения объема памяти является символ данных, заносимых в базу данных – $S_д$. Такой выбор удобен для определения параметров модели категорированных данных по анализу информационных моделей. В то же время при оценке затрат ресурсов памяти общепринятыми являются единицы измерения: бит, кбит, Мбит и Гбит.

Помимо этого, в ЭВМ и в СУБД используются методы сжатия данных. При записи таких типов данных, как числовые, несмотря на то, что отдельным атрибутам (в СУБД) резервируется определенные максимально возможные значения, нулевые значения слева от значащих символов числа не записываются и не сохраняются в памяти ЭВМ. Как правило, не заносятся (и не требуют ресурсов памяти) пустые значения атрибутов.

Современные СУБД обладают развитыми средствами повышения производительности баз данных. Ряд методов повышения производительности достигают за счет избыточного хранения данных (например, методы индексации записей в таблицах БД).

Для практического применения аналитических выражений (21) – (30) необходимо определить коэффициент перевода единиц измерения, учета сжатия и избыточности данных – k_{ec} размерностью бит/символ данных. При этом оценки (в единицах измерения – бит) необходимых затрат ресурсов памяти для реализации различных вариантов представления категорированных отношений:

$$O_a^{\text{бит}} = k_{ec}^{\ a}\, O_a;\ O_б^{\text{бит}} = k_{ec}^{\ б}\, O_б;\ O_в^{\text{бит}} = k_{ec}^{\ в}\, O_в;\ O_г^{\text{бит}} = k_{ec}^{\ г}\, O_г.$$

Для перевода единиц измерения необходимо учитывать особенности кодирования символов данных — $S_д$. Каждый символ данных представляется обычно десятичным или шестнадцатеричным кодом, содержащим несколько символов (символов кода – $S_к$). Наиболее употребительным является стандарт ASCII (American Standard Code for Information Interchange, стандарт ANSI – американского национального института стандартов). Часто используемая для передачи данных американская версия семибитовой кодировки символов кода ($S_к$) утверждена ISO. Восьмой бит символа кода ASCII обычно является битом контроля четности (паритета). Таким образом, одним из компонентов k_{ec} является коэффициент ($n_{S_к}$) двоичного представления символа кода $S_к$. Для ASCII $n_{S_к} = 8$ (бит/символ); $n_{S_к} = 1$ (байт/символ); $n_{S_к} = 1/1024$ (кбайт/символ) и т.д.

На ПК используется так называемый расширенный код ASCII, в котором первые 128 комбинаций совпадают со стандартным, а остальные используются для представления национальных алфавитов, псевдографики и специальных знаков. Кодировка символов данных отличается количеством знаков кода ($N_{S_к}$). Например, цифры и большая часть латинских символов кодируются двумя десятичными знаками ($N_{S_к} = 2$), а символы кириллицы — тремя ($N_{S_к} = 3$). Необходимый для

представления данных объем памяти зависит от набора символов, необходимого для определенного типа данных, заносимых в базу данных. В первом приближении может быть принято, что набор символов данных носит случайный характер, и этот набор может быть представлен математическим ожиданием значения числа символов кода – $\overline{N_{S_к}}$. По стандарту ASCII для большинства наборов символов значение $\overline{N_{S_к}}$ будет находиться в диапазоне от 2 до 3.

Таким же образом может быть признана случайной величина коэффициента учета сжатия ($k_{сж}$) данных при записи в ЭВМ (при записи данных сжатие проводится только по отдельным атрибутам, и степень сжатия различна). Фактор сжатия данных должен приводить к уменьшению $k_{ес}$. При сжатии данных, например в 1,5 – 2 раза, общая оценка $k_{ес}$ должна находиться в диапазоне от 1 до 2 (при измерении в единицах байт/символ).

Избыточность данных, прежде всего при использовании индексирования записей, также может быть учтена коэффициентом ($k_{и.д}$). При индексировании записей СУБД может создавать дополнительные колонки, с помощью которых упорядочиваются записи. В связи с этим введение $k_{и.д}$ адекватно отражает изменения в требуемых ресурсах памяти.

С учетом представления коэффициента сжатия средним значением общий вид $k_{ес}$:

$$k_{ес} = k_{сж}\, k_{и.д}\, n_{Sк}\, N_{Sк}.$$

Значения $n_{S_к}$ и $k_{и.д}$ могут быть определены при достаточно тщательном анализе конкретных вариантов реализаций фрагментов БД с категорированными данными. Аналитически определить $k_{сж}$ и $N_{S_к}$ гораздо сложнее. Предлагается использовать для оценки $k_{ес}$ имитационную программу без детального анализа составляющих $k_{ес}$. Программа представляет собой тестовую базу данных со структурой, соответствующей исследуемой структуре категорированных данных (соотношение общих и частных атрибутов, в соответствии с моделью категорированных данных, раздел 3) и вариантам представления категорированных отношений.

Шаги процедуры определения значения $k_{ес}$:

- внесение набора числа тестовых записей (ряд значений N) и фиксация затрат (в битах) ресурсов памяти – $O_a^{имит}(N)$; $O_б^{имит}(N)$; $O_в^{имит}(N)$; $O_г^{имит}(N)$;
- расчет затрат ресурсов по выражениям (27) – (30) в символах ($S_д$) для каждого значения из ряда N – $O_a(N)$; $O_б(N)$; $O_в(N)$; $O_г(N)$;
- расчет $k_{ес}$ для каждого N_i:

$$k_{ес}^{а}(N) = O_a^{имит}(N) / O_a(N);\ k_{ес}^{б}(N) = O_б^{имит}(N) / O_б(N);$$

$$k_{ес}^{в}(N) = O_в^{имит}(N) / O_в(N);\ k_{ес}^{г}(N) = O_г^{имит}(N) / O_г(N);$$

- определение среднего значения $k_{ес}^{а}$; $k_{ес}^{б}$; $k_{ес}^{в}$; $k_{ес}^{г}$;
- построение зависимостей $O_a^{бит}\, O_б^{бит}\, O_в^{бит}\, O_г^{бит}$.

Использование оценки на основе среднего значения ($k_{ес}$), получаемого по данным имитационной программы, позволяет получить аналитические выражения, отличающиеся от данных имитационной программы на 3 – 5 %, что вполне допустимо для использования их для оценок затрат ресурсов при проектировании реальных базах данных.

Величина $k_{ес}(N)$ при различных значениях N и вариантах структуры категорированных данных находилась в пределах от 0,8 до 3,2 (оценка диапазона при анализе факторов, влияющих на $k_{ес}(N)$, – от 1 до 2). Расширение диапазона влево

можно объяснить тем, что при малом числе записей более существенно влияет фактор сжатия данных при записи в БД (например, для значений уникальных атрибутов достаточно одного-двух символов вместо $\lg N_{\max}$). Расширение диапазона вправо до значения $k_{\text{ec}}(N) = 3{,}2$ также объяснимо. Часть тестовых данных использовалась на текстовых фрагментах, содержащих только символы кириллицы, которые представляются тремя символами кода по ASCII. Кроме того, дополнительные затраты ресурсов памяти связаны с индексацией первичных ключей.

Заключение

Оценить различия в объеме памяти БД вариантов представления категорированных отношений возможно аналитическими расчетами. Аналитические выражения определяют зависимости требуемых объемов памяти от параметров структуры категорированных данных для наиболее распространенных на практике вариантов представления категорированных отношений.

В аналитической модели должны быть учтены особенности структуры категорированных данных. Модель особенностей категорированных данных используется в аналитических выражениях метода оценки различий в объеме памяти БД вариантов представления категорированных отношений. Представление моделью реальных баз категорированных данных распространяет конкретные аналитические выражения метода на группу баз категорированных данных и сокращает затраты на моделирование.

Уточнить аналитические выражения метода оценки требуемых ресурсов памяти можно с помощью имитационной программы. Программа обеспечивает проверку достоверности аналитических выражений метода. Данные, получаемые имитационной программой, используются для определения коэффициента учета сжатия данных в СУБД, особенностей кодирования символов и дополнительных затрат ресурсов памяти при использовании в СУБД, например избыточных методов повышения производительности.

ЛИТЕРАТУРА

1. *CDM* – метод разработки информационных систем фирмы Oracle // Oracle Magazine: russian edition. 1997. № 2.
2. *Бистерфельд О.А. Сидоров М.В.*, *Таганов Р.А.* Исследование зависимости затрат памяти на представление категорированных отношений в реляционных базах данных // Новые информационные технологии в научных исследованиях и в образовании: тез. докл. 4-й Всероссийской научно-технической конференции. Рязань, 1999.

Бистерфельд Ольга Александровна
Рязанский государственный университет имени С.А. Есенина
E-mail: o.bisterfeld@rsu.edu.ru

Поступила в редакцию 18 июля 2011 г.