УДК 519.677

И.А. Кулик, А.И. Новгородцев, Е.М. Скордина

Сумский государственный университет, Сумы

БИНОМИАЛЬНАЯ МОДЕЛЬ ВЕКТОРНОГО ПРЕДСТАВЛЕНИЯ БАЗЫ ДАННЫХ С КОЛОНОЧНОЙ СТРУКТУРОЙ

В работе предлагается биномиальная модель векторного представления базы данных, которая позволяет выделить структурную составляющую информации, под которой понимается расположение элементов множества значений в соответствующих столбцах таблицы данных. Позиции, в которых размещаются определенные информационные элементы, кодируются с помощью двоичных векторов. Такой подход позволяет более эффективно использовать методы сжатия в базах данных, которые учитывают не только статистические, но и структурные свойства информации, например, биномиальные методы сжатия.

Ключевые слова: *биномиальная модель, источник информации Бернулли, база данных, колоночная структура, векторное кодирование, биномиальное нумерационное сжатие.*

Постановка задачи

Оперативность получения и изменения информации в базах данных (БД) обуславливается активным и широким использованием сетевых и интернет технологий, облачных вычислений. Одним из наиболее важных параметров функционирования БД является время доступа к запрашиваемой информации, от которого напрямую зависит время выполнения запроса к базе или хранилищу данных [1, 2].

Целью данной научной работы является уменьшение времени доступа к информации в БД при ограничениях на объем внешней памяти и пропускную способность каналов передачи.

Эффективным средством достижения цели работы является применение методов сжатия данных, которые учитывали бы не только статистические особенности информации, сжимаемой в БД, но и ее структурные свойства. Важным составным элементом процесса сжатия является моделирование информационного источника, т.е. в нашем случае моделирование БД, а именно табличных данных.

В работах [3, 4] рассматривается биномиальная модель двоичного источника информации Бернулли, которая вызывает особый интерес при разработке методов кодирования с целью сжатия или защиты данных. Одними из основных преимуществ данной модели являются существенное снижение объема статистических данных, используемых для эффективного кодирования источника информации, а также выделение в составе биномиальной модели комбинаторного источника, что позволяет учитывать структурные свойства кодируемых множеств.

Эффективность биномиальной модели двоичного источника информации покажем на примере разработки модели векторного представления БД с колоночной структурой [5, 6].

Предлагаемая модель является дальнейшим развитием и усовершенствованием модели векторного представления БД [6] за счет применения биномиальной модели источника информации и теоретико-множественного подхода к описанию табличных данных.

1. Исходные данные к построению модели

Пусть имеется конечное множество

$$D = \{d_1, d_2, ..., d_i, ..., d_m\}$$

сообщений и на его основе формируется упорядоченная n-выборка

$$S_\alpha = (s_1, s_2, ..., s_j, ..., s_n),$$

состоящая из элементов $s_j = d_i$. Совокупность всех n-выборок S_α представляет собой теоретико-множественное произведение вида

$$D^n = \underbrace{D \times D \times ... \times D}_{n}$$

и, соответственно,

$$S_\alpha \in D^n,\ \alpha = \overline{1, \mathrm{Card}\left(D^n\right)}.$$

Рассмотрим векторное кодирование [7, 8], которое ставит во взаимно однозначно соответствие некоторому элементу d_i из n-выборки S_α двоичный вектор $Y_{d_i} = y_1 y_2 ... y_j ... y_n$, $y_j \in \{0,1\}$, содержащий единицы в тех разрядах, номера которых соответствуют позициям d_i в n-выборке S_α:

$$y_j = \begin{cases} 1, & s_j = d_i \\ 0, & s_j \neq d_i \end{cases} \text{ или } y_{j=\mathrm{pos}(d_i)} = 1, \qquad (1)$$

где $\mathrm{pos}(d_i)$ – функция, формирующая номера элементов из S_α, значения которых равно d_i.

Таким образом, упорядоченная n-выборка $S_\alpha = (s_1, s_2, ..., s_j, ..., s_n)$ преобразуется в упорядо-

ченную m -выборку $V_\alpha = \left(Y_{d_1}, Y_{d_2}, ..., Y_{d_i}, ..., Y_{d_m}\right)$, где $Y_{d_i} \in V = \{0,1\}^n$.

Совокупность всех выборок V_α представляет собой m -ю декартову степень множества V :

$$V^m = \underbrace{V \times V \times ... \times V}_{m}$$

и, соответственно, $V_\alpha \in V^m$, $\alpha = \overline{1, \mathrm{Card}\left(V^m\right)}$ и $\mathrm{Card}\left(V^m\right) = \mathrm{Card}\left(D^n\right)$.

В результате на бинарном соответствии $Z \subseteq D^n \times V^m$, $(S_\alpha, V_\alpha) \in Z$, которое задает биективное отображение:

$$f_v : D^n \to V^m,$$

определяется функция

$$V_\alpha = f_v\left(S_\alpha\right)$$

на множестве D^n со значениями на множестве V^m, которая является основанием для получения биномиальной модели векторного представления БД с колоночной структурой.

2. Биномиальная модель векторного представления базы данных

Анализ различных типов БД показывает, что наиболее распространенными и востребованными являются реляционные БД [1, 2].

Реляционные БД можно представить в виде множества связанных отношений G (таблиц значений). Сжатие в БД можно рассматривать с точки зрения независимого сжатия каждой из таблиц G. Такой подход позволит проводить операции восстановления сжатых данных лишь в тех таблицах, значения которых необходимы для выполнения запроса к БД.

С целью уменьшения времени доступа к данным и в связи ограниченностью аппаратно-программных ресурсов СУБД отношение G будем разбивать на страницы, содержащие заданное число записей (строк).

В модели векторного представления БД отношение G будет рассматриваться как множество страниц, каждая из которых содержит n записей. В случае не кратности числа строк в таблице G значению n, последняя страница дополняется записями до числа n путем введения числа недостающих пустых строк.

Из двух подходов к размещению данных в реляционных БД – строчному или колоночному – отдадим предпочтение колоночному подходу [2, 7]. Колоночные структуры размещают информацию на физическом носителе в виде последовательности записей для каждой колонки по отдельности. В БД, ориентированных на чтение данных с высокой скоростью поиска и извлечения информации, наибольшим быстродействием обладают колоночные структуры.

Помимо высокого быстродействия, колоночные структуры являются наиболее перспективными с точки зрения сжатия информации, поскольку сжатие выполняется над периодически повторяющимися однотипными данными в рамках каждой из колонок таблицы.

В реляционной БД, использующей колоночное размещение информации, таблицу G' размерности $n \times q$, где n − число записей, q − число полей, представим как множество колонок (столбцов)

$$G' = \left\{S_{\alpha\beta}\right\},\ \alpha = \overline{1, \mathrm{Card}\left(D_\beta^n\right)},\ \beta = \overline{1, q}.$$

Колонка

$$S_{\alpha\beta} = \left(s_{1\beta}, s_{2\beta}, ..., s_{j\beta}, ..., s_{n\beta}\right)$$

есть упорядоченная n -выборка, последовательность из значений $s_{j\beta} = d_{i\beta}$, которые принадлежат множеству

$$D_\beta = \left\{d_{1\beta}, d_{2\beta}, ..., d_{i\beta}, ..., d_{m\beta}\right\}$$

допустимых значений поля β.

Выборки $S_{\alpha\beta}$ являются элементами n -местного отношения

$$W_\beta \subseteq D_\beta^n = D_\beta \times D_\beta \times ... \times D_\beta,\ S_{\alpha\beta} \in W_\beta.$$

В результате на бинарном соответствии

$$Z_\beta \in W_\beta \times V^m,\ \left(S_{\alpha\beta}, V_{\alpha\beta}\right) \in Z_\beta,$$

где $V_{\alpha\beta} \in V^m$, $V_{\alpha\beta} = \left(Y_{d_1\beta}, Y_{d_2\beta}, ..., Y_{d_i\beta}, ..., Y_{d_m\beta}\right)$ − упорядоченная m -выборка, компонентами которой являются двоичные векторы

$$Y_{d_i\beta} = y_1 y_2 ... y_j ... y_n,\ y_j \in \{0,1\},$$

формируемые согласно (1), получаем биективное отображение

$$f_{v\beta} : D_\beta^n \to V^m,$$

которое выражается функцией вида

$$V_{\alpha\beta} = f_{v\beta}\left(S_{\alpha\beta}\right).$$

Далее, применив биномиальную модель двоичного источника информации Бернулли [1, 2], генерирующего двоичные векторы $Y_{d_i\beta} = y_1 y_2 ... y_j ... y_n$, получаем биномиальную модель векторного представления БД с колоночным размещением информации.

Графическое изображение биномиальной модели векторного представления БД с колоночной структурой представлено на рис. 1.

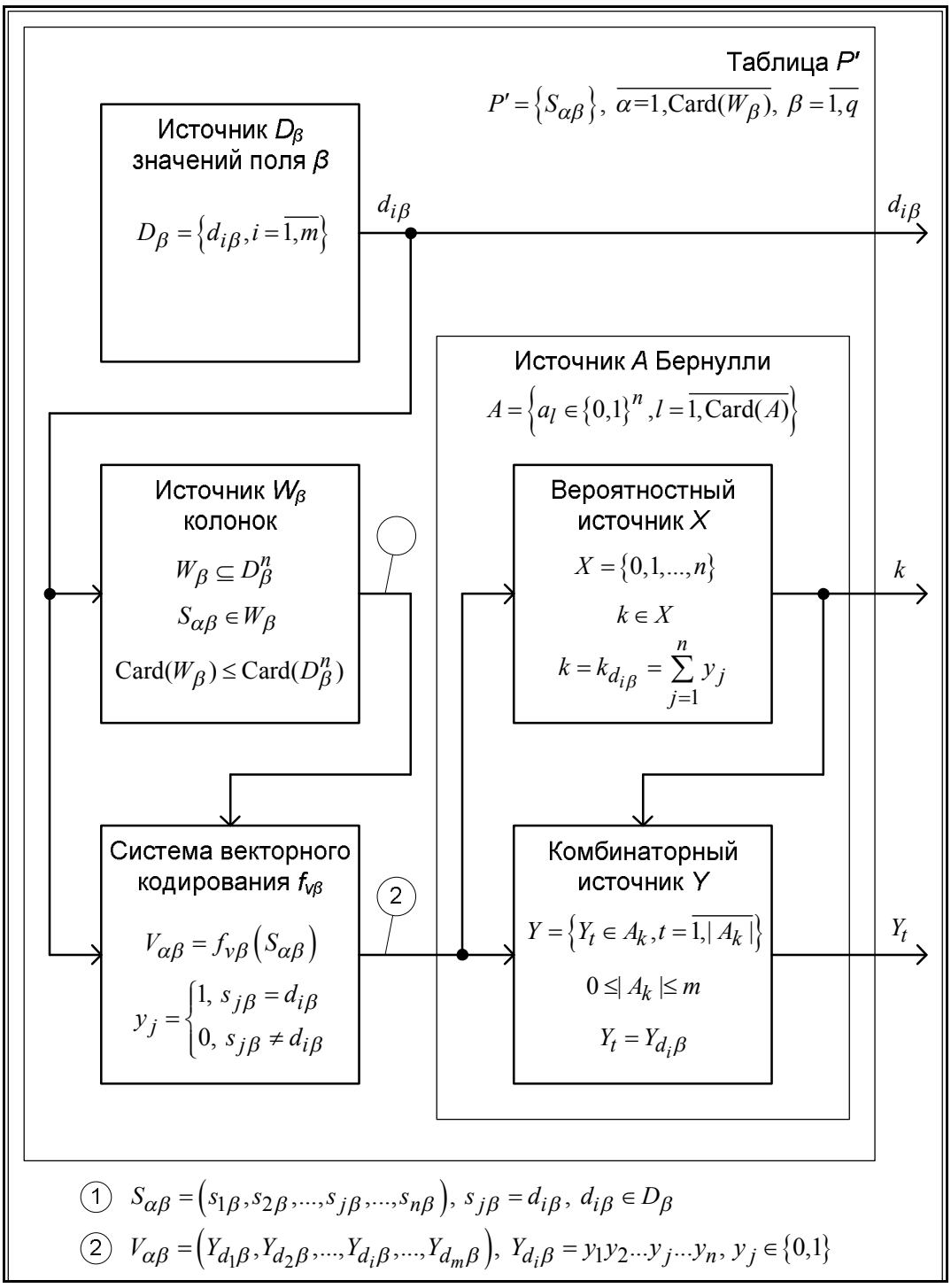

Рис. 1. Биномиальная модель векторного представления БД с колоночной структурой

Биномиальное моделирование БД с колоночным размещением информации на основе векторного представления значений столбцов (полей) состоит из следующих этапов.

Этап 1. Разбиение отношения G в реляционных БД на множество q -местных отношений P (таблиц значений в рамках одной страницы), каждое из которых состоит из n записей (строк):

$$P \subseteq D_1 \times D_2 \times ... \times D_\beta \times ... \times D_q,$$

$$\mathrm{Card}(P) = n,$$

где $D_\beta = \{d_{1\beta}, d_{2\beta}, ..., d_{i\beta}, ..., d_{m\beta}\}$ − множество допустимых значений поля β;

q − количество полей (колонок) отношения P .

Этап 2. Представление q -местного отношения P в виде множества P′ колонок $S_{\alpha\beta}$ (атрибутов):

$$P' = \{S_{\alpha\beta}\},\ S_{\alpha\beta} \in W_\beta,\ \alpha = \overline{1,\mathrm{Card}(W_\beta)},$$

где $S_{\alpha\beta} = (s_{1\beta}, s_{2\beta}, ..., s_{j\beta}, ..., s_{n\beta})$ − упорядоченная

n-выборка, $s_{j\beta} = d_{i\beta}$, $d_{i\beta} \in D_\beta$, $i = \overline{1,m}$, $j = \overline{1,n}$;

$W_\beta \subseteq D_\beta^n = D_\beta \times D_\beta \times ... \times D_\beta$ – n-местное отношение на множестве D_β^n, представляющего собой n-ю декартову степень D_β, $\mathrm{Card}\left(W_\beta\right) \le \mathrm{Card}\left(D_\beta^n\right)$.

Этап 3. Установление бинарного функционального соответствия

$Z_\beta \in W_\beta \times V^m$, $\left(S_{\alpha\beta}, V_{\alpha\beta}\right) \in Z_\beta$,

при котором согласно функции $V_{\alpha\beta} = f_{v\beta}\left(S_{\alpha\beta}\right)$ на основе системы равенств

$$y_j = \begin{cases} 1, & s_j = d_i; \\ 0, & s_j \ne d_i \end{cases}$$

для каждой колонки

$$S_{\alpha\beta} = \left(s_{1\beta}, s_{2\beta}, ..., s_{j\beta}, ..., s_{n\beta}\right)$$

определяется единственный образ

$$V_{\alpha\beta} = \left(Y_{d_1\beta}, Y_{d_2\beta}, ..., Y_{d_i\beta}, ..., Y_{d_m\beta}\right),$$

представляющий собой упорядоченную m-выборку, $V_{\alpha\beta} \in V^m$, $V = \{0,1\}^n$, компонентами которой являются двоичные векторы

$$Y_{d_i\beta} = y_1 y_2 ... y_j ... y_n,\ y_j \in \{0,1\}.$$

Этап 4. Вычисление для каждого двоичного вектора

$$Y_{d_i\beta} = y_1 y_2 ... y_j ... y_n$$

из образа

$$V_{\alpha\beta} = \left(Y_{d_1\beta}, Y_{d_2\beta}, ..., Y_{d_i\beta}, ..., Y_{d_m\beta}\right)$$

числа $k_{d_i\beta}$ единиц (числа повторений значения $d_{i\beta}$ в колонке $S_{\alpha\beta}$, прообразе $V_{\alpha\beta}$):

$$k_{d_i\beta} = \sum_{j=1}^{n} y_j$$

и формирование двоичной равновесной комбинации $Y_t = Y_{d_i\beta}$, содержащей число $k = k_{d_i\beta}$ единиц.

3. Оценка биномиальной модели векторного представления базы данных

Обоснование целесообразности применения биномиальной модели БД с колоночной структурой на основе векторного представления информации проведем с точки зрения уменьшения объема требуемой памяти для хранения данных, а значит и сокращения времени доступа к информации.

Введем некоторую функцию

$$y = \mathrm{size}\left(x\right), \qquad (2)$$

которая определяет размер аргумента x в битах, или, другими словами, количество y двоичных разрядов, необходимых для отображения аргумента x или хранения его в памяти.

Тогда объем ls памяти, занимаемой элементом $d_{i\beta}$ из n-выборки $S_{\alpha\beta} = \left(s_{1\beta}, s_{2\beta}, ..., s_{j\beta}, ..., s_{n\beta}\right)$, $s_{j\beta} = d_{i\beta}$, с учетом (2) составляет

$$ls\left(d_{i\beta}\right) = k_{d_i\beta} \cdot \mathrm{size}\left(d_{i\beta}\right). \qquad (3)$$

При векторном представлении объем lv памяти, занимаемой $d_{i\beta}$ из той же выборки $S_{\alpha\beta}$, определяется следующим образом:

$$lv\left(d_{i\beta}\right) = \mathrm{size}\left(d_{i\beta}\right) + n, \qquad (4)$$

где n – количество разрядов соответствующего двоичного вектора $Y_{d_i\beta}$.

Для неизменного значения $\mathrm{size}\left(d_{i\beta}\right)$ величина $ls\left(d_{i\beta}\right)$ лежит в диапазоне от 0 до $n \cdot \mathrm{size}\left(d_{i\beta}\right)$, а величина $lv\left(d_{i\beta}\right)$ является постоянной. При условии $\mathrm{size}\left(d_{i\beta}\right) > 1$ на некотором диапазоне значений $k_{d_i\beta}$ векторное представление будет сжимающим.

Из формул (3) и (4) можно определить граничное значение $k_{d_i\beta}$, при котором наблюдается уменьшение объема памяти при векторном представлении данных:

$$lv\left(d_{i\beta}\right) < ls\left(d_{i\beta}\right),$$

$$\mathrm{size}\left(d_{i\beta}\right) + n < k_{d_i\beta} \cdot \mathrm{size}\left(d_{i\beta}\right). \qquad (5)$$

Решение неравенства (5) дает следующие граничные значения числа $k_{d_i\beta}$ повторений элемента $d_{i\beta}$ в колонке $S_{\alpha\beta}$, при которых векторное представление будет сжимающим:

$$k_{d_i\beta} > 1 + \frac{n}{\mathrm{size}\left(d_i\beta\right)}. \qquad (6)$$

Для случая $n = 1024$ и $\mathrm{size}\left(d_{i\beta}\right) = 4$ векторное представление в БД с колоночной структурой будет иметь сжимающий эффект при $k_{d_i\beta}$, удовлетворяющем неравенству (6):

$$k_{d_i\beta} > 1 + \frac{1024}{4} = 257.$$

Обобщив (3) на все элементы $d_{i\beta}$ из $S_{\alpha\beta} = \left(s_{1\beta}, s_{2\beta}, ..., s_{j\beta}, ..., s_{n\beta}\right)$, $s_{j\beta} = d_{i\beta}$, $d_{i\beta} \in D_\beta$, $i = \overline{1,m}$, получим объем $Ls\left(S_{\alpha\beta}\right)$ информации, который занимает вся колонка $S_{\alpha\beta}$:

$$Ls\left(S_{\alpha\beta}\right) = \sum_{i=1}^{m} k_{d_i\beta} \cdot \mathrm{size}\left(d_i\beta\right). \qquad (7)$$

В то время как объем $\mathrm{Lv}\left(\mathrm{V}_{\alpha\beta}\right)$ информации, которым характеризуется векторное представление $\mathrm{V}_{\alpha\beta}$, определяется выражением:

$$\mathrm{Lv}\left(\mathrm{V}_{\alpha\beta}\right)=\sum_{i=1}^{m}\left(\mathrm{size}\left(\mathrm{d}_{i\beta}\right)+n\right)=m\cdot n+\sum_{i=1}^{m}\mathrm{size}\left(\mathrm{d}_{i\beta}\right). \quad (8)$$

При условии, что $\mathrm{size}\left(\mathrm{d}_{i\beta}\right)=\mathrm{size}\left(\mathrm{d}_{\beta}\right)$ для всех элементов $\mathrm{d}_{i\beta}\in\mathrm{D}_{\beta}$ множества D_{β} допустимых значений поля β, равенства (7) и (8) будут иметь вид соответственно:

$$\mathrm{Ls}\left(\mathrm{S}_{\alpha\beta}\right)=\mathrm{size}\left(\mathrm{d}_{\beta}\right)\sum_{i=1}^{m}k_{d_{i}\beta}=n\cdot\mathrm{size}\left(\mathrm{d}_{\beta}\right), \quad (9)$$

$$\mathrm{Lv}\left(\mathrm{V}_{\alpha\beta}\right)=\sum_{i=1}^{m}\left(\mathrm{size}\left(\mathrm{d}_{\beta}\right)+n\right)=m\cdot\left(n+\mathrm{size}\left(\mathrm{d}_{\beta}\right)\right). \quad (10)$$

Как следует из (9) и (10), степень сжатия при переходе к векторному представлению информации в БД с колоночной структурой зависит от соотношения параметров m, n и $\mathrm{size}\left(\mathrm{d}_{\beta}\right)$:

$$\mathrm{Lv}\left(\mathrm{V}_{\alpha\beta}\right)<\mathrm{Ls}\left(\mathrm{S}_{\alpha\beta}\right),$$

$$m\cdot\left(n+\mathrm{size}\left(\mathrm{d}_{\beta}\right)\right)<n\cdot\mathrm{size}\left(\mathrm{d}_{\beta}\right). \quad (11)$$

Анализ неравенства (11) показывает, что векторное представление $\mathrm{V}_{\alpha\beta}$ данных будет занимать меньший объем памяти, чем исходное отображение в виде колонки $\mathrm{S}_{\alpha\beta}$, при условии, если мощность $m=\mathrm{Card}\left(\mathrm{D}_{\beta}\right)$ множества допустимых значений $\mathrm{D}_{\beta}=\left\{\mathrm{d}_{1\beta},\mathrm{d}_{2\beta},...,\mathrm{d}_{i\beta},...,\mathrm{d}_{m\beta}\right\}$ поля β будет удовлетворять следующему неравенству:

$$m<\frac{n\cdot\mathrm{size}\left(\mathrm{d}_{\beta}\right)}{n+\mathrm{size}\left(\mathrm{d}_{\beta}\right)}. \quad (12)$$

С практической точки зрения больший интерес представляет решение неравенства (11) относительно числа записей n в таблице P:

$$n>\frac{m\cdot\mathrm{size}\left(\mathrm{d}_{\beta}\right)}{\mathrm{size}\left(\mathrm{d}_{\beta}\right)-m}, \quad (13)$$

которое позволяет рационально проводить разбиение отношения G в реляционной БД на множество q-местных отношений P (таблиц данных в рамках одной страницы), каждое из которых состоит из n записей (строк).

Как следует из (7), для хранения всего q-местного отношения P, состоящего из n записей (строк), необходимо выделить следующий объем памяти в битах:

$$\mathrm{Ls}\left(\mathrm{P}\right)=\sum_{\beta=1}^{q}\sum_{i=1}^{m}k_{d_{i\beta}}\cdot\mathrm{size}\left(\mathrm{d}_{i\beta}\right). \quad (14)$$

В свою очередь, основываясь на (8), отношение P, представленное в векторном виде, будет занимать объем памяти в битах:

$$\mathrm{Lv}\left(\mathrm{P}\right)=\sum_{\beta=1}^{q}\sum_{i=1}^{m}\left(\mathrm{size}\left(\mathrm{d}_{i\beta}\right)+n\right)=$$
$$=m\cdot n\cdot q+\sum_{\beta=1}^{q}\sum_{i=1}^{m}\mathrm{size}\left(\mathrm{d}_{i\beta}\right). \quad (15)$$

Граничное значение m в неравенстве (12) можно увеличить, если выполнить сжатие двоичных векторов $\mathrm{Y}_{d_{i}\beta}=y_1y_2...y_j...y_n$, являющихся компонентами образа $\mathrm{V}_{\alpha\beta}\in\mathrm{V}^{m}$, соответствующее прообразу $\mathrm{S}_{\alpha\beta}\in\mathrm{W}_{\beta}\subseteq\mathrm{D}_{\beta}^{n}$. Таким же способом можно уменьшить граничное значение числа n записей в выражении (13) при разбиении отношения G на составляющие его отношения P (или страничные таблицы P') в реляционной БД.

В качестве метода сжатия целесообразно использовать биномиальное нумерационное сжатие, построенное на основе двоичной (n,k)-биномиальной системы счисления [9, 10], используя числа k единиц, генерируемых источником X, и равновесные комбинации Y_t, порождаемых источником Y класса эквивалентности A_k (рис. 1).

Заключение

Основным назначением биномиальной модели векторного представления БД является выделение структурной составляющей информации, под которой понимается указание позиций элементов множества допустимых значений в соответствующих столбцах таблицы данных. Позиции, в которых располагаются определенные информационные элементы, кодируются с помощью двоичных векторов.

Двоичные вектора являются одним из видов индексирования данных. Они позволяют получать позиции, на которых размещены запрашиваемые данные, без выполнения операций просмотра и поиска этих данных.

Кроме того, использование логических операций над двоичными векторами в значительной степени упрощает выполнение сложных выборок из БД по нескольким параметрам.

Операции замены или изменения одного из значений атрибута также выполняются достаточно быстро. Добавление или удаление элемента в определенной позиции выполняется путем изменения бита в соответствующем разряде двоичного вектора на единичное или нулевое значение, что не изменяет

размеры двоичных векторов и множества значений атрибута.

Разделение элементов атрибута в биномиальной модели векторного представления БД с колоночной структурой предоставляет возможность выборочного извлечения из БД определенных данных. Особенно это актуально для сложных БД, имеющих расширенные области значений атрибутов. Выборочный доступ к данным позволяет в значительной степени уменьшить время поиска и извлечения информации с физического носителя.

Совокупность перечисленных преимуществ биномиальной модели векторного представления БД с колоночной структурой позволяет уменьшить время доступа к данным при выполнении наиболее часто используемых операций при работе с информацией в БД. Дополнительно стоит отметить тот факт, что представленную модель можно использовать как для кодирования отдельных атрибутов, отношений и БД целиком, так и для кодирования результирующих наборов, формируемых при выполнении запросов к БД.

Выполняя переход к биномиальной модели векторного представления информации при соблюдении неравенств (12) или (13), осуществляется сокращение структурной избыточности данных, содержащейся в колонках $S_{\alpha\beta}$ таблиц реляционной БД. Далее, избыточность, также имеющая структурный характер, вызывается уже представлением самих двоичных n-разрядных векторов $Y_{d_i\beta}$, а именно соотношением чисел k единиц и $n-k$ нулей.

Для устранения такого вида избыточности, а значит и дальнейшего увеличения эффективности применения биномиальной модели векторного представления информации в реляционной БД, предлагается использовать метод биномиального нумерационного сжатия двоичных последовательностей [9, 10]. Данный метод является методом сжатия без информационных потерь, обладает достаточно высоким коэффициентом сжатия при незначительных затратах времени на сжатие и восстановление данных.

Список литературы

1. Дейт К.Д. Введение в системы баз данных / К.Д. Дейт. – М.: Вильямс, 2001. – 1072 с.

2. Кириллов В.В. Введение в реляционные базы данных / В.В. Кириллов, Г.Ю. Громов. – СПб.: БХВ-Петербург, 2009. – 464 с.

3. Борисенко А.А. О разложении бернуллиевских источников / А.А. Борисенко // Вестник Сумского государственного университета. − 1995. − № 3. − С. 57-59.

4. Kulik I.A. A Binomial Model of Bernoulli Information Sources / I.A. Kulik, S. Kostel, E. Skordina // International Journal of Biomedical Soft Computing and Human Sciences. – 2011. − Vol.17, No.2. – P. 45-51.

5. Григорьев Ю.А. Модель обработки запросов в параллельной колоночной системе баз данных / Ю.А. Григорьев, Е.Ю. Ермаков // Информатика и системы управления. – Амурский гос. ун-т.− 2013. − № 1(31). – С. 3-15.

6. Кулик И.А. Модель векторного представления баз данных с колоночной структурой / И.А. Кулик, В.В. Гриненко, С.В. Костель, Е.М. Скордина // Вісник Сумського державного університету. Серія Технічні науки. − 2013. − №3. − С. 60-66.

7. Амелькин В.А. Методы нумерационного кодирования / В.А. Амелькин. – Новосибирск: Наука, 1986. – 159 с.

8. Орищенко В.И. Сжатие данных в системах сбора и передачи информации / В.И. Орищенко, В.Г. Санников, В.А. Свириденко. − М.: Радио и связь, 1985. − 184 с.

9. Кулик И.А. Повышение производительности СУБД на основе биномиального сжатия информации / И.А. Кулик, С.В. Костель // Системи обробки інформації. – Х.: ХУ ПС, 2014. – Вип. 2 (118), Т. 2. – С. 45-48.

10. Борисенко А.А. Биномиальное кодирование: монография / А.А. Борисенко, И.А. Кулик. − Сумы: СумГУ, 2010. – 206 с.

Поступила в редколлегию 21.03.2016

Рецензент: д-р техн. наук, проф. А.А. Борисенко, Сумский государственный университет, Сумы.

БІНОМІАЛЬНА МОДЕЛЬ ВЕКТОРНОГО ПРЕДСТАВЛЕННЯ БАЗИ ДАНИХ З КОЛОНОЧНОЮ СТРУКТУРОЮ

І.А. Кулик, А.І. Новгородцев, О.М. Скордіна

В роботі пропонується біноміальна модель векторного представлення бази даних, яка дозволяє виділити структурну складову інформації, за яку визначають розміщення елементів множини значень у відповідних стовпчиках таблиці даних. Позиції, на яких знаходяться певні інформаційні елементи, кодуються за допомогою двійкових векторів. Такий підхід дозволяє більш ефективно застосовувати методи стиснення інформації в базах даних, які враховують не тільки статистичні, але і структурні властивості інформації.

Ключові слова: *біноміальна модель, джерело інформації Бернуллі, база даних, колоночна структура, векторне кодування, біноміальне нумераційне стиснення.*

BINOMIAL MODEL OF VECTOR REPRESENTATION FOR DATABASE WITH COLUMNAR STRUCTURE

I.A. Kulyk, A.I. Novgorodsev, Ye.M. Skordina

Binomial model of vector representation for database, which allow us to separate out structural component of information. It means location of elements out of a set of values into corresponding columns of a data table. Positions, which contain the information elements, are coded with binary vectors. Such an approach permits us to use information compression methods in databases more effectively, which would take into account not only statistic properties of information, but also its structural features.

Keywords: *binomial model, Bernoulli's information source, database, column structure, vector coding, binomial numerical compression.*