УДК 004.65

МЕТОДИКА ТЕСТИРОВАНИЯ РЕЗУЛЬТАТОВ ВЕРТИКАЛЬНОЙ КЛАСТЕРИЗАЦИИ ОТНОШЕНИЙ РЕЛЯЦИОННЫХ БАЗ ДАННЫХ

М.В. ГРАНКОВ, А.И. ЖУКОВ
(Донской государственный технический университет)

Рассмотрена методика тестирования результатов структурной оптимизации отношений реляционных баз данных, основанная на нивелировании влияния кэш-системы и доказана возможность ее практической реализации за счет использования трасс с равномерным распределением объектов.
***Ключевые слова**: методы структурной оптимизации, вертикальная кластеризация, HBVP, декомпозиция отношений.*

Введение. В современных информационных системах (ИС) базы данных (БД) являются одним из ключевых компонентов, поэтому повышение эффективности их использования в средних и крупных проектах является важнейшим фактором, влияющим на производительность ИС в целом.

Наиболее известными классами методов повышения эффективности ИС, использующих реляционные БД (РБД) являются *методы кэширования информации* и *методы структурной оптимизации*. Методы первого класса заключаются в комбинировании двух видов памяти (основной и кэш-памяти) и повышении скорости доступа к информации за счет сохранения в кэш-памяти наиболее востребованных объектов ИС. Методы второго класса основаны на различных вариантах декомпозиции отношений РБД.

Методы данных классов аддитивны в том смысле, что использование методов структурной оптимизации совместно с методами кэширования позволяет повысить эффективность последних и наоборот. Объектом исследования эффективности методов структурной оптимизации являются системы управления базами данных (СУБД), как правило, реализующие некоторую модель повышения эффективности доступа к информации на базе собственной кэш-системы, полное исключение которой из схемы функционирования СУБД представляется затруднительным, а в большинстве случаев невозможным. Поэтому для проведения теоретических и экспериментальных исследований методов второго класса необходимо нивелировать влияние методов первого класса.

Одним из методов структурной оптимизации является метод *вертикальной кластеризации* (*секционирования*) отношений РБД. На базе этого метода в ДГТУ аспирантом кафедры «ПОВТ и АС» Нго Т.Х. был разработан эвристический алгоритм вертикальной кластеризации HBVP [1], который заключается в получении декомпозиции исходного отношения, приводящего к повышению вероятности кэш-попадания при заданном распределении запросов к БД в независимости от эффективности используемого алгоритма кэширования. При обосновании данного метода была выдвинута гипотеза о том, что при практических и теоретических исследованиях методов структурной оптимизации необходимо использовать поток запросов с равномерным распределением объектов ИС [1]. Целью настоящей статьи является теоретическое доказательство данной гипотезы.

Постановка задачи. Рассмотрим модель информационной системы для проведения исследований методов структурной оптимизации. Пусть данная ИС реализует в своем составе некоторый алгоритм замещения объектов в кэш-памяти, определим ее основные понятия:

– *объект информационной системы* (объект трассы, объект системы кэширования) – минимальная единица информации, сохраняемая в кэше (в нашем случае, кортеж). Допустим также, что каждый объект имеет идентификатор, уникальным образом определяющий его на множестве всех объектов ИС;

– *трасса* – это последовательность обращений к объектам информационной системы, соответствующая некоторому потоку запросов к БД. Трасса формируется на основании пользовательских запросов, каждый из которых может подразумевать запрос в источнике данных (база

данных или файловое хранилище) некоторого числа объектов. Таким образом, трасса может быть представлена как последовательность идентификаторов объектов ИС;

– *дистанция* – участок трассы для объекта *a*, который начинается и заканчивается обращением к объекту *a* и внутри себя не содержит обращений к этому объекту.

Необходимо доказать, что использование трасс с равномерным распределением объектов позволяет нивелировать влияние кэш-системы на эффективность информационной системы в целом, таким образом, объективно оценить эффективность проведения структурной оптимизации.

Доказательство. Величина временного интервала между двумя соседними вызовами объектов в исследованиях методов структурной оптимизации не играет роли и обычно принимается равной 1 [2,3]. Таким образом, позиция объекта в трассе может быть интерпретирована как момент времени, в который данный объект был запрошен пользователем ИС (рис.1).

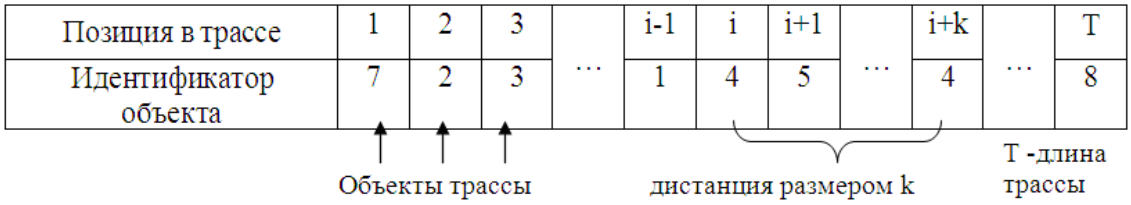

Рис.1. Схема трассы потока объектов кэш-системы

Будем считать, что понятию *«объект ИС»* в реляционных системах соответствует понятие *«кортеж»*. Рассмотрим отношение, состоящее из N кортежей и только те отношения, в которых $N>>1$.

Пусть вероятность появления объекта в трассе в некоторый момент времени *i* не зависит ни от объекта, ни от позиции в трассе и равна:

$$p = 1/N \tag{1}$$

Вероятность того, что объект не появится в любой позиции трассы в момент времени i, выражается соотношением:

$$q = 1 - p = 1 - 1/N = (N-1)/N \tag{2}$$

Обозначим ξ – дискретную случайную величину, равную дистанции для некоторого объекта и изменяющуюся в диапазоне (1, ∞). Пусть в момент времени *i* в трассе появляется объект *a*. Тогда с вероятностью $(N-1)/N^2$ он может появиться в (i+1)-ой позиции, с вероятностью $1/N\left((N-1)/N\right)^2$ – в (i+2)-ой позиции и в (i+k-1)-ой позиции с вероятностью:

$$p_{i+k-1} = 1/N\left((N-1)/N\right)^{k-1}, \tag{3}$$

где $i = 1,2,\ldots$

Введем в рассмотрение E_K:

$$E_k = \sum_{l=1}^{k} 1/N\left((N-1)/N\right)^{l-1} \cdot l \tag{4}$$

Выполнив преобразования в соответствии с (2), получаем:

$$E_k = 1/N \sum_{l=1}^{k} q^{l-1} \cdot l \tag{5}$$

Тогда математическое ожидание случайной величины ξ:

$$E(\xi) = \lim_{k\to\infty} E_k \tag{6}$$

Введем дополнительное обозначение для суммы: $S_k = \sum_{l=1}^{k} q^{l-1} \cdot l$ и рассчитаем несколько первых значений для определения закономерности: $S_1 = 1$, $S_2 = 2q$, $S_3 = 3q^2$. Тогда, очевидно: $S_k = S_1 + S_2 + S_3 + \ldots + S_m$, при $m=k$. Представим полученные значения в виде квадратной матрицы, в которой на каждой j-ой строке расположим составные части j-ого значения для S_j, $j = \overline{1,m}$. При этом, $\tilde{S}_j$ – сумма элементов в j-ом столбце:

$$\begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ \\ S_m \\ \\ \end{array} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ q & q & 0 & 0 & 0 & 0 \\ q^2 & q^2 & q^2 & 0 & 0 & 0 \\ q^3 & q^3 & q^3 & q^3 & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ q^{m-1} & q^{m-1} & q^{m-1} & q^{m-1} & \ldots & q^{m-1} \end{pmatrix}. \qquad (7)$$
$$\begin{array}{cccccc} \tilde{S}_1 & \tilde{S}_2 & \tilde{S}_3 & \tilde{S}_4 & \ldots & \tilde{S}_m \end{array}$$

Очевидно, что $S_k = \sum_{j=1}^{m} S_j = \sum_{j=1}^{m} \tilde{S}_j$, кроме того, $S_2 = S_1 - 1$, $S_3 = S_2 - q$, $S_4 = S_3 - q^2$, из чего следует, что j-ая сумма по столбцам есть разность двух геометрических прогрессий:

$$\tilde{S}_j = \sum_{t=1}^{m} q^{t-1} - \sum_{t=1}^{j} q^{t-1} \qquad (8)$$

Для нахождения $\tilde{S}_j$ из (8) воспользуемся формулой геометрической прогрессии:

$$S_k = \sum_{j=1}^{m} \tilde{S}_j = \sum_{j=1}^{m} \left(\frac{1-q^m}{1-q} - \frac{1-q^{j-1}}{1-q} \right) = \frac{1}{1-q} \left(m\left(1-q^m\right) - m + \sum_{j=1}^{m} q^{j-1} \right) = \frac{1}{1-q} \left(-mq^m + \frac{1-q^m}{1-q} \right). \qquad (9)$$

Случайная величина ξ – целая, положительная и теоретически неограничена, поэтому ее математическое ожидание можно вычислить по формуле:

$$E\left(\xi\right) = \lim_{k\to\infty} E_k(\xi) = \frac{1}{N} \lim_{k\to\infty} \sum_{l=1}^{k} q^{l-1} \cdot l \qquad (10)$$

Учитывая выражение, полученное для S_k, а также подставив значения для q, предельное значение для математического ожидания появления каждого объекта из рассматриваемого множества мощности N ($N >> 1$) на дистанции неограниченной длины, равно:

$$\begin{aligned} E(\xi) &= \frac{1}{N} \lim_{m\to\infty} S_k = \frac{1}{N} \lim_{m\to\infty} \left(\frac{1}{1-q} \left(-mq^m + \frac{1-q^m}{1-q} \right) \right) = \frac{1}{N} \cdot \frac{1}{\left(1-q\right)^2} \cdot \lim_{m\to\infty} \left(1 - q^m - \frac{mq^m}{1-q} \right) = \\ &= \frac{1}{N} \cdot \frac{1}{\left(1-q\right)^2} \cdot \lim_{m\to\infty} \left(1 - \frac{q^m - q^{m+1} + mq^m}{1-q} \right) = \frac{1}{N} \cdot \frac{1}{\left(1-q\right)^2} \cdot \lim_{m\to\infty} \left(1 - \frac{q^m}{1-q} - \frac{q^{m+1}}{1-q} + \frac{mq^m}{1-q} \right) \end{aligned} \qquad (11)$$

Так как q<1, а также в связи с тем, что показательная функция растет на бесконечности быстрее любой полиномиальной, получаем:

$$E(\xi) = \frac{1}{N} \cdot \frac{1}{\left(1-q\right)^2} \qquad (12)$$

Подставим значение для q:

$$E(\xi) = \frac{1}{N} \cdot \frac{1}{\left(1-\left(1-N\right)\right)^2} = N \qquad (13)$$

Таким образом, если вероятность появления каждого объекта в трассе является величиной постоянной и зависит только от мощности начального множества объектов, то математическое ожидание дистанции каждого объекта трассы равно количеству объектов и не зависит от других параметров системы.

Теорема А0 Ахо доказывает [4], что оптимальной стратегией вытеснения объектов из кэш-памяти является утилизация объектов с наибольшим математическим ожиданием дистанции появления в трассе. Также доказано, что этот алгоритм уступает по эффективности только оптимальному алгоритму Биледи, для которого будущая трасса должна быть известна, что практически нереализуемо [2]. Однако, очевидно, что при равенстве математического ожидания дистанции для всех объектов трассы, оптимальный алгоритм А0 неэффективен, а значит, любой другой алгоритм кэширования, кроме алгоритма Биледи, имеет эффективность меньше эффективности алгоритма А0.

Заключение. В работе доказано, что объективная оценка эффективности алгоритмов структурной оптимизации в теоретических и экспериментальных исследованиях может быть получена на трассах с равномерным распределением объектов.

Библиографический список

1. Нго Тхань Хунг. Метод вертикальной кластеризации отношений реляционных баз данных / Тхань Хунг Нго // Вестн. Донск. гос. техн. ун-та. – 2008. – №4.

2. Аль-Згуль Мосаб Басам. Гибридные алгоритмы в системах кэширования объектов / Мосаб Басам Аль-Згуль // Вестн. Донск. гос. техн. ун-та. – 2008. – №4.

3. Жуков А.И. Математическая модель метода бигибридизации алгоритмов кэширования / А.И. Жуков, Мосаб Басам Аль-Згуль // «В мире научных открытий». – №4(10). – Ч.13. – Красноярск, 2010.

4. Aho A.V., Denning P.J., Ulman J.D., Principles of optimal page replacement, J. ACM, vol. 18, no. 1, 1971.

Материал поступил в редакцию 06.06.2011.

References

1. Ngo Txan` Xung. Metod vertikal`noj klasterizacii otnoshenij relyacionny`x baz danny`x / Txan` Xung Ngo // Vestn. Donsk. gos. texn. un-ta. – 2008. – #4. – In Russian.

2. Al`-Zgul` Mosab Basam. Gibridny`e algoritmy` v sistemax ke`shirovaniya ob``ektov / Mosab Basam Al`-Zgul` // Vestn. Donsk. gos. texn. un-ta. – 2008. – #4. – In Russian.

3. Zhukov A.I. Matematicheskaya model` metoda bigibridizacii algoritmov ke`shirovaniya / A.I. Zhukov, Mosab Basam Al`-Zgul` // «V mire nauchny`x otkry`tij». – #4(10). – Ch.13. – Krasnoyarsk, 2010. – In Russian.

4. Aho A.V., Denning P.J., Ulman J.D., Principles of optimal page replacement, J. ACM, vol. 18, no. 1, 1971.

RESULTS TESTING TECHNIQUE OF VERTICAL CLUSTERING RELATIONAL DATABASE

M.V. GRANKOV, A.I. ZHUKOV
(Don State Technical University)

The results testing technique of the structural optimization of the relational databases founded on the effect leveling of the cache-system is considered. Its feasibility through the paths with object flat sharing is proved.
Keywords: *structural optimization methods, vertical clustering, HBVP, relation decomposition.*