УДК 20.23.17

ОСНОВАННЫЕ НА ВЕЙВЛЕТАХ ГИСТОГРАММЫ ДЛЯ ОЦЕНКИ СЕЛЕКТИВНОСТИ ЗАПРОСОВ

А.М.ЛОГВИНОВ, Ю.Е. ПОЛЕНОВА, Г.А. ТРАВИН, М.Г. ТРАВИН

Белгородский государственный университет

e-mail: travin@bsu.edu.ru

В статье предлагается метод, основанный на кратномасштабном вейвлет-преобразовании для построения гистограмм на основных распределениях данных применительно к базам данных, статистике и моделированию. Гистограммы, основанные на совокупных значениях данных, дают очень хорошее приближение с ограниченным используемым объемом памяти. Предлагаются быстрые алгоритмы построения гистограмм и их использование для оценки селективности в режиме прямого доступа.

Ключевые слова: база данных, вейвлет, гистограмма, запрос, селективность

Введение

Некоторые важные компоненты систем управления базами данных (СУБД) требуют точной оценки селективности конкретного запроса. Например, оптимизаторы запросов используют это для оценки точности конкретного запроса при оценке затрат различных вариантов его выполнения и выбора наиболее удачного варианта. При этом множество рассматриваемых предикатов представляют запросами выбора, в частности предикатов диапазона вида $a < X < b$, где X — неотрицательный атрибут области отношений R, а a и b — константы. Множество равных предикатов — подмножество предикатов диапазона, соответствующих $a = b$. Множество односторонних предикатов диапазона — особый случай предикатов диапазона, в котором $a = -\infty$ или $b = \infty$.

Для оптимизации запросов требуемые данные представляют в виде данных различных гистограмм [1], по которым делается оценка селективности запросов. Математическая модель постановки задачи состоит в следующем.

Область определения $D = \{0,1,2,\ldots,N-1\}$ атрибута X есть множество всех значений X. Множество значений $V \subseteq D$ включает n определенных значений X, фактически представленных в отношении R. Пусть $v_1 < v_2 < \ldots v_n$ — n значений V. Разброс s_i от v_i определяется, как $s_i = v_{i+1} - v_i$ (мы принимаем $s_0 = v_1$, а $s_n = 1$). Частота f_i от v_i представляет собой число записей, в которых X принимает значение v_i. Совокупная частота c_i от v_i — число записей $t \in R$ с $t.R \le v_i$, т.е. $c_i = \sum_{j=1}^{i} f_i$. Представление данных для X состоит из множества пар $T = \{(v_1, f_1),(v_2, f_2),\ldots,(v_n, f_n)\}$. Обобщенное представление данных для X состоит из множества пар $T^C = \{(v_1, c_1),(v_2, c_2),\ldots,(v_n, c_n)\}$. Расширенное обобщенное представление данных для X, обозначаемое T^{C+}, образуется из множества T^C путем дополнения его до размера области D присвоением нулевого значения всем частотам в $D - V$.

Для оценки селективности может использоваться случайное осуществление выборки. Самый простой способ использовать случайное осуществление выборки для оценки селективности во время автономной фазы состоит во взятии случайной выборки определенного размера (в зависимости от ограничения размера каталога) по отношению. Когда запрос представлен в фазе оперативного режима, запрос сравнивается с выборкой, и селективность оценивается явным способом: Если размер результата за-

проса при использовании выборки размера *t* является s, селективность оценивается как *sM/t,* где *M* — размер отношения.

Наилучшие показатели получены на основе применения гистограмм MaxDiff(V,A) [1]. В то же время дальнейшее совершенствование гистограмм в оценке селективности запросов представляет собой **актуальную задачу**. **Цель работы** – усовершенствование гистограмм в оценке селективности запросов при ограниченном объеме памяти базы данных, а также разработка быстрых алгоритмов построения гистограмм и их использование для оценки селективности в режиме прямого доступа. Поставленная цель достигнута за счет использования предложенного метода, основанного на кратномасштабном вейвлет-преобразовании для построения гистограмм на основных распределениях данных применительно к базам данных, статистике и моделированию.

Предлагаемый метод, основанный на вейвлетах

Вейвлеты представляют собой математический инструмент для иерархического разложения функций. Вейвлеты выражают функцию в элементах грубой полной формы и деталях, которые располагаются иерархически от грубых к более детальным. Независимо от того, является ли интересующая функция изображением, кривой или поверхностью, вейвлеты предлагают изящную технику для представления различных уровней детализации функции пространственно эффективным способом.

Верхний уровень предлагаемого алгоритма построения основанных на вейвлетах гистограмм работает следующим образом:

1. На шаге предварительной обработки формируется расширенное обобщенное распределение данных T^{C+} атрибута *X* путем преобразования предварительно вычисленного представления *T* из исходных данных или из случайной выборки данных.

2. Вычисляется вейвлет-разложение T^{C+}, что дает ряд *N* коэффициентов вейвлетов. В своих экспериментах при разложении мы использовали вейвлеты Хаара и линейные сплайны.

3. Сохраняется только *m* наиболее значащих коэффициентов вейвлетов для некоторого *m*, соответствующего желаемому использованию памяти. Выбор числа *m* сохраняемых коэффициентов зависит от специфичности используемого метода пороговой обработки.

Пороговая обработка

Учитывая ограниченность памяти для гистограммы, мы можем хранить только определенное число коэффициентов вейвлетов *N*. Пусть *m* обозначает число коэффициентов вейвлетов, для которых имеется пространство для хранения; остальные коэффициенты будут неявно установлены в 0. Обычно мы имеем $m \ll N$. Цель пороговой обработки заключается в определении «лучших» *m* сохраняемых коэффициентов для минимизации ошибки приближения.

Пусть *m* обозначает число коэффициентов вейвлетов, для которых имеется пространство для хранения; остальные коэффициенты будут неявно установлены в 0.

Мы можем измерить ошибку приближения гистограммами несколькими способами. Пусть S_i — реальный размер запроса q_i, и пусть S_j' — предполагаемый размер запроса. Мы использовали следующие три различных меры для ошибки e_i запроса q_i:

1. Абсолютная ошибка запроса:
$$e_i^{\text{abs}} = \left|S_i - S_i'\right|.$$
2. Относительная ошибка запроса:
$$e_i^{\text{rel}} = \frac{e_i^{\text{abs}}}{S_i} = \frac{\left|S_i - S_i'\right|}{S_i}, \text{ где } S_i > 0.$$
3. Совместная ошибка запроса:
$$e_i^{\text{comb}} = \min\left\{\alpha e_i^{\text{abs}}, \beta e_i^{\text{rel}}\right\},$$

где α и β — положительные константы. Если $S_i > 0$, мы полагаем $e_i^{\text{comb}} = \alpha e_i^{\text{abs}}$.

Совместная ошибка отражает важность наличия, как хорошей относительной ошибки, так и хорошей абсолютной ошибки для каждой оценки. Например, для очень малых частот может быть достаточно, если абсолютная ошибка мала, даже если велика относительная ошибка, а для больших частот абсолютная ошибка, возможно, не является столь же значимой как относительная ошибка.

Как только мы определяем, какая из вышеупомянутых мер будет использоваться для представления ошибки определенных запросов, необходимо выбрать норму измерения ошибки совокупности запросов. Пусть $\mathbf{e} = (e_1, e_2, \ldots e_Q)$ — вектор ошибок последовательности Q запросов. Полагается, что используется одна из вышеупомянутых трех мер ошибки для каждой из ошибок конкретного запроса e_i. Например, для абсолютной ошибки, можно записать $\mathbf{e} = (e_1, e_2, \ldots e_Q) = \mathbf{e}^{\text{abs}} = (e_1^{\text{abs}}, e_2^{\text{abs}}, \ldots, e_Q^{\text{abs}})$. Мы определяем полную ошибку для Q запросов одной из следующих норм ошибки:

1. Средняя ошибка с одномерной нормой:

$$\|\mathbf{e}\|_1 = \frac{1}{Q}\sum_{i=1}^{Q} e_i \,.$$

2. Средняя ошибка с двумерной нормой:

$$\|\mathbf{e}\|_2 = \sqrt{\frac{1}{q}\sum_{1\le i\le Q} e_i^2} \,.$$

3. Средняя ошибка с бесконечномерной нормой:

$$\|\mathbf{e}\|_\infty = \max_{1\le i\le Q}\{e_i\} \,.$$

Эти меры ошибок представляют собой частные случаи средней ошибки с p-мерной нормой, $p > 0$:

$$\|\mathbf{e}\|_p = \left(\frac{1}{Q}\sum_{1\le i\le Q} e_i^p\right)^{1/p}, \; p > 0 \,.$$

Первый шаг пороговой обработки заключается во взвешивании коэффициентов определенным образом (соответствующим специфике используемого базиса, например, ортонормированного). В частности, для базиса Хаара нормализация проводится делением коэффициентов вейвлетов $\widehat{S}(2^j), \ldots \widehat{S}(2^{j+1}-1)$ на $\sqrt{2^j}$, где $0 \le j \le \log_2 N - 1$. Для любого данного частного взвешивания предлагаются следующие различные методы пороговой обработки:

1. Выбор m наибольших по абсолютному значению коэффициентов вейвлетов.
2. Выбор m коэффициентов вейвлетов поглощающим методом. Например, мы должны выбрать m наибольших по абсолютному значению коэффициентов вейвлетов, а затем повторно выполнить следующие два шага m раз:
   а) выбрать коэффициенты вейвлетов, включение которых приводит к наибольшему снижению ошибки;
   б) исключить коэффициенты вейвлетов, удаление которых приводит к наименьшему увеличению ошибки.

Другой подход состоит в неоднократном исполнении двух вышеупомянутых шагов до завершения цикла или до незначительного улучшения.

Возможны другие варианты поглощающего метода:

3. Начать с $m/2$ наибольших (по абсолютному значению) коэффициентов вейвлетов и производить выбор следующих $m/2$ коэффициентов методом поглощения.
4. Начать с наибольших $2m$ (по абсолютному значению) коэффициентов вейвлетов, уменьшая их количество до m методом поглощения.

После вышеупомянутого алгоритма мы получаем m коэффициентов вейвлетов Значения этих коэффициентов вместе с их положениями (индексы), сохраняются и служат гистограммой для восстановления приближенного распределения данных на

фазе прямого доступа (фазе запроса). Чтобы вычислить оценку числа кортежей, чье значение X находится в диапазоне $a < X < b$, мы восстанавливаем приближенные значения для b и $a-1$ в расширенном обобщенном распределении функции и затем вычитаем их.

Прямой метод выполнения каждой итерации метода поглощения требует $O(N^2)$ операций, таким образом, полное число операций составляет $O(mN^2)$. С поддержкой особой древовидной структуры динамического программирования мы можем значительно ускорить предварительную обработку.

На фазе запроса использовался запрос диапазона $a < X < b$. Восстанавливались аппроксимации обобщенных частот $a-1$ и b, обозначенные, как c'_a и c'_b, с использованием m коэффициентов вейвлетов. Ожидаемый размер запроса составляет $c'_b - c'_a$.

Время восстановления крайне важно для фазы оперативного режима. Следующий результат допускает быстрое восстановление.

Часто полезно представить гистограмму как явную кусочно гладкую функцию, а не как m коэффициентов вейвлетов. Для вейвлетов Хаара результирующая функция представляет собой ступенчатую функцию с не более $3m$ шагами в худшем случае, а для линейных сплайнов — линейную кусочную функцию с не более $5m$ изменениями наклона в худшем случае. В реальных данных можно ожидать, что число шагов или сегментов очень близко к m (во многих случаях оно точно m). Это свойство было подтверждено обширным набором экспериментов. Предыдущие методы представления гистограмм в виде кусочно гладкой функции требовали $O(N)$ времени, хотя некоторые исследователи подозревали, что возможна разработка алгоритмов, требующих времени $O(m\log_2 N)$. Мы разработали эффективную и практичную технику, используя очередь приоритетов, что предполагает существенное ускорение обработки запросов.

Результаты моделирования и экспериментальные данные

Далее приводится ряд экспериментов, которые проводились для сравнения работы предлагаемой методики, основанной на вейвлетах, с полученными результатами, основанными на случайной выборке и других методах [1, 3]. Используемые искусственные наборы данных взяты от предыдущих исследований формирования гистограмм по тесту TPC-D [4]. Они соответствуют исследованиям типичных данных, найденных в Интернете. Для простоты и облегчения повторяемости, мы использовали метод 1 для пороговой обработки во всех экспериментах с вейвлетами.

Приведем результаты сравнения эффективности гистограмм, основанных на вейвлетах, с гистограммами MaxDiff(V,A) и случайным осуществлением выборки. Для сравнения берутся характеристики гистограмм из предшествующих работ [1]. Так как в указанной работе делается заключение о наилучших показателях гистограмм MaxDiff(V,A), для сравнения были выбраны они.

В экспериментах мы использовали множество искусственных распределений данных, описанных подробно в [1]. Использованные распределения включают типы одномерных распределений по тесту TPC-D [4].

В экспериментах мы использовали восемь различных вариантов запросов:

A: $\{X \le b \mid b \in D\}$;

B: $\{X \le b \mid b \in V\}$;

C: $\{a \le X \le b \mid a, b \in D, a < b\}$;

D: $\{a \le X \le b \mid a, b \in V, a < b\}$

E: $\{a \le X \le b \mid a \in D, b = a + \Delta\}$, где Δ – положительная целая постоянная;

F: $\{a \le X \le b \mid a \in V, b = a + \Delta\}$, где Δ – положительная целая постоянная;

G: $\{X = b \mid b \in D\}$;

H: $\{X = b \mid b \in V\}$.

В проводимых экспериментах всем методам отводится один и тот же объем памяти. Объем памяти по умолчанию, использованный в экспериментах, составляет 42 четырехбайтовых числа для соответствия повторяемости экспериментов, проведенных в [1]. Ограничения объема памяти соответствуют практике в системах управления базами данных, чтобы выделять только незначительный вспомогательный объем памяти каждому отношению для оценки селективности [5]. Эти 42 числа соответствуют использованию 14 сегментов для гистограммы MaxDiff(V,A) и хранению $m = 21$ коэффициентов вейвлетов для основанных на вейвлетах гистограмм, а также поддержанию случайной выборки размера 42.

В этих экспериментах множество значений и набор частот распределены по закону Ципфа с параметром $z = 1.0$. Частоты случайным образом назначены элементам значений, заданы размер множества значений $n = 500$, размер области $N = 4096$ и размер отношения $M = 10^5$. В табл. 1-5 представлены результаты ошибки методов для запросов A, B, C, D и H, а на рис. 1 приведены графики качества аппроксимации обобщенного распределения данных различными методами.

Таблица 1

Ошибки различных методов для ряда запросов A

Норма ошибки	Линейные сплайны	Вейвлеты Хаара	MaxDiff(V,A)	Случайная выборка
$\left\|\mathbf{e}^{\text{rel}}\right\|_1$	0.6%	4.5%	8%	20%
$\left\|\mathbf{e}^{\text{abs}}\right\|_1 / M$	0.16%	0.8%	3%	8%
$\left\|\mathbf{e}^{\text{abs}}\right\|_2 / M$	0.26%	0.64%	3.2%	10%
$\left\|\mathbf{e}^{\text{abs}}\right\|_\infty / M$	1.5%	5.6%	11%	13%
$\left\|\mathbf{e}^{\text{comb}}\right\|_1$, $\alpha = 1$, $\beta = 100$	0.6	4.4	8	20
$\left\|\mathbf{e}^{\text{comb}}\right\|_1$, $\alpha = 1$, $\beta = 1000$	5	30	80	200
$\left\|\mathbf{e}^{\text{comb}}\right\|_2$, $\alpha = 1$, $\beta = 100$	5.1	70.4	12.8	19
$\left\|\mathbf{e}^{\text{comb}}\right\|_2$, $\alpha = 1$, $\beta = 1000$	19	224	192	243

Таблица 2

Ошибки различных методов для ряда запросов C

Норма ошибки	Линейные сплайны	Вейвлеты Хаара	MaxDiff(V,A)	Случайная выборка
$\left\|\mathbf{e}^{\text{abs}}\right\|_1 / M$	0.2%	1.1%	5%	3.5%
$\left\|\mathbf{e}^{\text{abs}}\right\|_2 / M$	0.035%	0.18%	0.71%	0.6%
$\left\|\mathbf{e}^{\text{abs}}\right\|_\infty / M$	2.4%	10%	20%	16%

Таблица 3

Ошибки различных методов для ряда запросов E при $\Delta = 10$

Норма ошибки	Линейные сплайны	Вейвлеты Хаара	MaxDiff(V,A)	Случайная выборка
$\left\|\mathbf{e}^{\text{abs}}\right\|_1 / M$	0.1%	0.42%	0.15%	0.35%
$\left\|\mathbf{e}^{\text{abs}}\right\|_2 / M$	0.19%	0.96%	0.26%	0.64%
$\left\|\mathbf{e}^{\text{abs}}\right\|_\infty / M$	1.5%	6%	3%	4.6%

Таблица 4

Ошибки различных методов для ряда запросов G

Норма ошибки	Линейные сплайны	Вейвлеты Хаара	MaxDiff(V,A)	Случайная выборка
$\left\|\mathbf{e}^{\text{abs}}\right\|_1 / M$	0.03%	0.04%	0.04%	0.04%
$\left\|\mathbf{e}^{\text{abs}}\right\|_2 / M$	0.077%	0.32%	0.096%	0.24%
$\left\|\mathbf{e}^{\text{abs}}\right\|_\infty / M$	1.6%	7%	2%	4.6%

Таблица 5

Ошибки различных методов для ряда запросов H

Норма ошибки	Линейные сплайны	Вейвлеты Хаара	MaxDiff(V,A)	Случайная выборка
$\left\|\mathbf{e}^{\text{abs}}\right\|_1 / M$	0.03%	0.42%%	0.2%	0.4%
$\left\|\mathbf{e}^{\text{abs}}\right\|_2 / M$	7.7%	16.1%	25.6%	38%
$\left\|\mathbf{e}^{\text{abs}}\right\|_\infty / M$	0.2%	7%	2%	5%

В дополнение к вышеупомянутым экспериментам мы также применяли метод MaxDiff(V,A), измененный путем хранения только двух чисел для каждого сегмента вместо трех (в частности не сохранялось число различных значений в каждом сегменте), что позволило получить в гистограмме 21 сегмент вместо 14. Точность оценки была улучшена. Преимуществу добавленных сегментов несколько противостояло менее точное моделирование в пределах каждого сегмента. Качественные результаты, однако, остаются теми же самыми: основанные на вейвлетах методы значительно более точны. Дальнейшие усовершенствования методов вейвлетов, конечно, возможны квантованием и кодированием энтропии, но они не рассматриваются в рамках этой статьи.

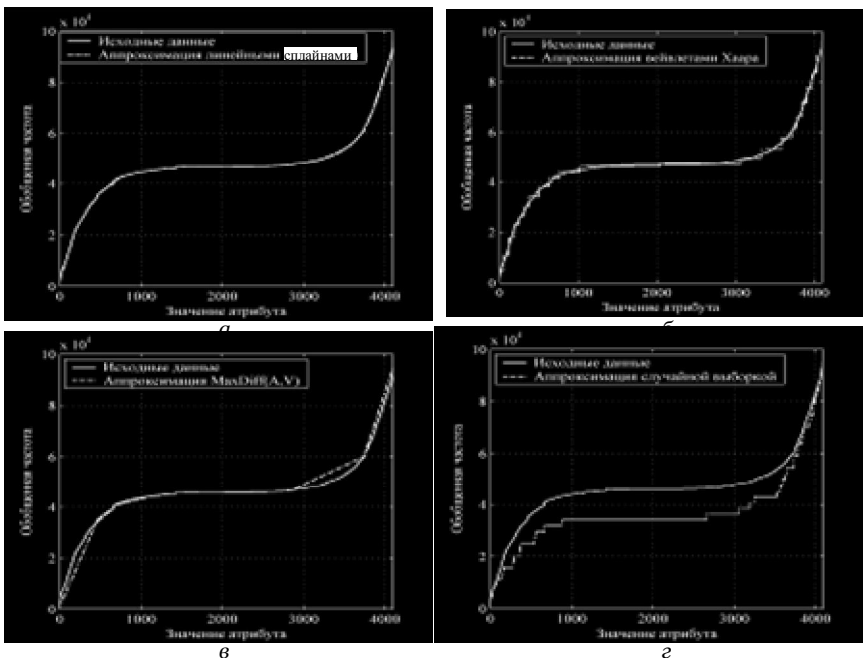

Рис. 1. Аппроксимация совокупного распределения данных различными методами: а) линейными сплайнами; б) вейвлетами Хаара; в) MaxDiff(V,A); г) осуществлением случайной выборки

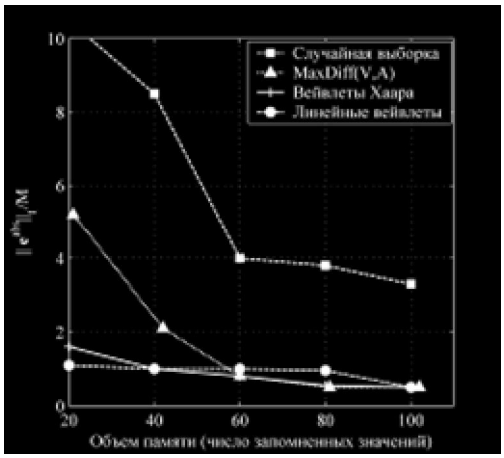

Рис. 2. Влияние объема памяти на различные гистограммы для набора запросов A

Выводы

В данной статье был предложен метод построения эффективных гистограмм с использованием вейвлет-разложения. Полученные гистограммы дают улучшение работы при оценке селективности по сравнению со случайным осуществлением выборки и предыдущими подходами.

Эксперименты показали, что применяя новый метод пороговой обработки в построении основанных на вейвлетах гистограмм, можно достичь намного большей точности даже для маломерных данных, которые рассматриваются в статье. Относительные ошибки могут быть существенно уменьшены (в три раза в типичных случаях), а абсолютные ошибки обычно сокращаются более чем на половину.

Основанные на вейвлетах гистограммы должны служить эффективной структурой накопления данных для оценки селективности в контексте итогового механизма оперативного режима [3]. Мы развиваем эффективные алгоритмы для поддержания основанных на вейвлетах гистограмм путем вставки и удаления данных в базовом отношении.

Литература

1. V. Poosala, Y. E. Ioannidis, P. J. Haas, and E. Shekita. Improved histograms for selectivity estimation of range predicates. In Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data, Montreal, Canada, May 1996.

2. Census Bureau Databases, http://www.census.gov/.

3. G. Piatetsky-Shapiro and C. Connell. Accurate estimation of the number of tuples satisfying a condition. In Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data, pages 256-276, 1984.

4. TCP benchmark D (decision support), 1995.

5. P. G. Selinger, M. M. Astrahan, D. D. Chamberlin, R. A. Lorie, and T. G. Price. Access path selection in a relational database management system. In Proceedings of the 1979 ACM SIGMOD International *Conference on Management of Data,* pages 23-34, 1979.

WAVELET-BASED HISTOGRAMS IN QUERY SELECTIVITY ESTIMATION

A.M. LOGVINOV , Yu.E. POLENOVA
G.A. TRAVIN , M.G. TRAVIN

Belgorod State University

e-mail: travin@bsu.edu.ru

In this paper, we present a technique based upon a multiresolution wavelet decomposition for building histograms on the underlying data distributions, with applications to databases, statistics, and simulation. Histograms built on the cumulative data values give very good approximations with limited space usage. We give fast algorithms for constructing histograms and using them in an on-line fashion for selectivity estimation.

Keywords: data base, histogram, query, selectivity, wavelet.