УДК 519.2 : 681.3

ПОСТРОЕНИЕ ОПТИМАЛЬНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ СОЕДИНЕНИЙ ОТНОШЕНИЙ В ЗАПРОСАХ РЕЛЯЦИОННОЙ БАЗЫ ДАННЫХ

Л.А. ПОНОМАРЕНКО, С.С. ТАНЯНСКИЙ, В.А. ФИЛАТОВ

Рассматриваются принципы обработки реляционных запросов и способы их реализации. В качестве основной операции, влияющей на эффективность выполнения запроса, выделена операция соединения отношений. Предложен метод минимизации числа операций при поиске последовательности соединения. Доказана оптимальность такого метода.

ВВЕДЕНИЕ

Цель выполнения процедуры обработки запросов в реляционной базе данных — преобразование запроса, реализованного на языке высокого уровня, например SQL, в корректную и эффективную последовательность действий, представленную на языке низкого уровня в терминах операций реляционной алгебры [1]. При этом важнейшим аспектом обработки запроса является его оптимизация. В результате преобразования исходного запроса может быть получено множество его эквивалентных вариантов.

Основная проблема оптимизации — это оценка сложности выполнения вычислительных операций с большим количеством отношений. Как правило, на практике чаще всего выбирается стратегия ближайших оптимальных решений. Большинство таких стратегий предусматривает эквивалентные преобразования алгебраических выражений, на которых строится запрос.

Языки манипулирования данными высокого уровня позволяют формировать сложные запросы для баз данных с большим количеством отношений. Скорость обработки таких запросов может быть в значительной степени увеличена, если перед их выполнением модифицировать план выполнения алгебраических операций. Цель таких модификаций — получить эквивалентное выражение, но требующее меньше времени и памяти для его выполнения.

Одним из важных свойств, влияющих на эффективность запроса, является последовательность вычисления соединений в цепочке операций реляционной алгебры. Соединение — это определяющая операция, так как время ее выполнения пропорционально времени выполнения операции объедине-

ния отношений. В качестве критерия сложности структуры запроса можно выбрать оценку количества промежуточных операций соединения и порядок выполнения коммутирующих и ассоциативных операций.

При подробном рассмотрении операции соединения можно отметить важность способа ее вычисления. Выбранный способ может определяться либо тем, как хранится отношение, либо наличием связанных дополнительных структур. При этом процедура вычисления может проводиться различными методами, например, последовательным или индексным доступом, с учетом или без учета дубликатов и т.п.

Для организации эффективной реализации запроса необходимо рассмотреть технику управления файлами, взаимосвязь между обращениями к внешней памяти и операциями реляционной алгебры, а также некоторые технические характеристики вычислительной техники.

Рассматриваемые в статье способы реализации запроса используют свойства операций алгебры и теории множеств. Дальнейшие выкладки будут основываться на свойствах операции соединения и правилах, позволяющих осуществлять алгебраические преобразования [2].

АРХИТЕКТУРА ПЕРЕБОРА

Алгоритм перебора должен выбрать оптимальный план выполнения запроса на основе исследования пространства поиска. Алгоритмы перебора многих существующих систем рассчитаны на выбор только оптимального порядка линейных соединений.

На практике желательно иметь такой алгоритм перебора, который мог бы легко приспосабливаться к изменениям пространства поиска (добавление новых преобразований, новых физических операций, например, новых реализаций соединений) и к изменениям методов оценки стоимости. Современные архитектуры оптимизации построены на основе этой парадигмы и называются расширяемыми оптимизаторами [3].

Построение расширяемого оптимизатора — сложная задача, так как необходимы не только улучшенный алгоритм перебора, но и обеспечение инфраструктуры для развития техники оптимизации. Однако общность архитектуры должна быть сбалансирована с потребностью эффективного перебора. Таким образом, одной из важных составляющих частей оптимизатора является наличие эффективного метода перебора, определяющего последовательность выполнения операций запроса.

МИНИМИЗАЦИЯ СУММЫ ПРОМЕЖУТОЧНЫХ РЕЗУЛЬТАТОВ СОЕДИНЕНИЙ ОТНОШЕНИЙ БАЗЫ ДАННЫХ

Одним из критериев повышения эффективности выполнения запроса является уменьшение числа кортежей в отношениях при многократном соединении. При условии, что операция соединения предполагает обращение к каждому кортежу соединяемых отношений, возникает задача отыскания такой последовательности, которая гарантирует наименьшее суммарное число обращений к кортежам при последовательном соединении.

Пусть база данных R содержит шесть отношений $r(R) = \{r_1(R), r_2(R), r_3(R), r_4(R), r_5(R), r_6(R)\}$, и пусть каждое отношение содержит соответственно $r(R) = \{100,\ 50,\ 80,\ 5,\ 20,\ 10\}$ кортежей.

Тогда при последовательном соединении, когда выполняется условие

$$R = R_1 \triangleright\triangleleft R_2 \triangleright\triangleleft R_3 \triangleright\triangleleft R_4 \triangleright\triangleleft R_5 \triangleright\triangleleft R_6\ ,$$

суммарное количество просмотренных кортежей будет соответствовать выражению $r = r_1 + r_2 + r_3 + r_4 + r_5$, где $r_1 = 100 * 50$; $r_2 = r_1 * 80$; $r_3 = r_2 * 5$; $r_4 = r_3 * 20$; $r_5 = r_4 * 10$, т.е. $r(R) = 442\,405\,000$.

Используя свойства ассоциативности и коммутативности операции соединения и изменяя общую последовательность соединения, можно уменьшить (увеличить) значение $r(R)$. Изменим последовательность операций. Поменяем местами R_1 и R_6, тогда отношение R' будет представлено выражением

$$R' = R_6 \triangleright\triangleleft R_2 \triangleright\triangleleft R_3 \triangleright\triangleleft R_4 \triangleright\triangleleft R_5 \triangleright\triangleleft R_1$$

и, соответственно, $r(R') = 404240500$. Таким образом, для вычисления отношения R при такой последовательности потребуется на 38164500 дисковых операций меньше. Учитывая, что количество перестановок конечно и соответствует $n!$ (для данного примера $6!$), всегда можно получить наименьшую сумму последовательного произведения.

Пусть $X = \{x_1,\ x_2, \ldots, x_n\} \in N$, где N — множество натуральных чисел, и пусть результаты операций произведения образуют множество $P = \{p_1,\ p_2, \ldots, p_m\}$, где $p_1 = x_1 * x_2,\ p_2 = p_1 * x_3, \ldots, p_m = p_{m-1} * x_n$. Необходимо найти такую последовательность $p_1,\ p_2, \ldots, p_m$, при которой $\sum_{i=1}^{m} p_i \to \min$.

Теорема *о наименьшей сумме промежуточных результатов последовательного произведения*. Значение суммы промежуточных результатов p_i, $(i = 1, \ldots, m)$, последовательного произведения элементов множества $X = \{x_1,\ x_2, \ldots, x_n\}$ будет наименьшим, если значения $x_1,\ x_2, \ldots, x_n$ упорядочены по возрастанию, т. е. $x_1 < x_2 < \ldots < x_n$.

Доказательство. Упорядочим значения $x_1,\ x_2, \ldots, x_n$ по возрастанию: $x_1 < x_2 < \ldots < x_n$. Запишем общий вид суммы промежуточных результатов последовательных умножений элементов множества X при следующем упорядочивании:

$$x_1 * x_2 + \ldots + x_1 * x_2 * \ldots * x_{i-1}\, x_i + x_1 * x_2 * \ldots * x_{i-1}\, x_i\, x_{i+1} + \ldots$$

$$\ldots + x_1 x_2 * \ldots * x_i * \ldots * x_{j-1} + x_1 x_2 * \ldots * x_i * \ldots * x_j + \ldots + x_1 x_2 * \ldots * x_i * \ldots * x_j * \ldots * x_n$$

и обозначим ее S_1.

Поменяем местами произвольные элементы x_i и x_j из множества X и найдем сумму S_2 промежуточных результатов последовательных умножений при такой перестановке:

$$x_1 * x_2 + \ldots + x_1 * x_2 * \ldots * x_{i-1}\, x_j + x_1 * x_2 * \ldots * x_{i-1}\, x_j\, x_{i+1} + \ldots$$

$$\ldots + x_1 x_2 * \ldots * x_j * \ldots * x_{j-1} + x_1 x_2 * \ldots * x_j * \ldots * x_i + \ldots + x_1 x_2 * \ldots * x_i * \ldots * x_j * \ldots * x_n .$$

Вычислим разность $\Delta = S_2 - S_1$. С учетом коммутативности операции умножения получаем

$$\Delta = (x_1 * x_2 * \ldots * x_{i-1}\, x_j + x_1 * x_2 * \ldots * x_{i-1}\, x_j\, x_{i+1} + \ldots + x_1 x_2 * \ldots * x_j * \ldots * x_{j-1}) -$$

$$-(x_1 * x_2 * \ldots * x_{i-1}\, x_i + x_1 * x_2 * \ldots * x_{i-1}\, x_i\, x_{i+1} + \ldots + x_1 x_2 * \ldots * x_i * \ldots * x_{j-1}) =$$

$$= (x_j - x_i)(x_1 * x_2 * \ldots * x_{i-1} + x_1 * x_2 * \ldots * x_{i-1}\, x_j\, x_{i+1} + \ldots + x_1 x_2 * \ldots * x_{j-1}) > 0 .$$

Таким образом, при последовательном перемножении некоторых значений $x_1, x_2, \ldots, x_n \in X$ сумма промежуточных результатов будет наименьшей, если рассматриваемые значения упорядочены по возрастанию $x_1 < x_2 < \ldots < x_n$. ■

Из теоремы о наименьшей сумме промежуточных результатов последовательного произведения следует, что при изменении порядка перемножения сумма результатов также изменится.

Поскольку операция соединения является коммутативной, последовательность ее выполнения необязательно должна быть линейной. В частности, запрос для БД R может быть алгебраически представлен как $R = (R_1 \triangleright\!\triangleleft R_2) \triangleright\!\triangleleft (R_3 \triangleright\!\triangleleft R_4) \triangleright\!\triangleleft (R_5 \triangleright\!\triangleleft R_6)$.

При такой последовательности значение $r(R)$ будет равно $400\,000\,000$, что значительно меньше предыдущих значений, полученных последовательным выполнением операции соединения. В дальнейшем такую последовательность будем называть парной.

За конечное число шагов можно найти такую последовательность пар произведений, при которой сумма промежуточных результатов будет минимальной. Отличие парного соединения от последовательного заключается в том, что общая сумма промежуточных результатов достигается несколькими независимыми шагами. Сначала суммируются произведения произвольных пар, после чего результаты перемножаются парами в произвольном порядке, пока не будут перемножены все возможные варианты.

Теоретически не исключен вариант смешанного произведения, когда часть операций выполняется линейно, а часть попарно [1]. Формулу такого произведения можно представить как последовательность $R = (R_1 \triangleright\!\triangleleft R_2) \triangleright\!\triangleleft (R_3 \triangleright\!\triangleleft R_4) \triangleright\!\triangleleft (R_5 \triangleright\!\triangleleft R_6)$.

Сравним результаты соединения, полученные различными способами. Для линейной, упорядоченной по значениям последовательности

$R = R_4 \rhd\lhd R_6 \rhd\lhd R_5 \rhd\lhd R_2 \rhd\lhd R_3 \rhd\lhd R_1$ результат будет соответствовать значению $r(R) = 400000000$. Для произвольного парного соединения вида $R = (R_1 \rhd\lhd R_4) \rhd\lhd (R_2 \rhd\lhd R_6) \rhd\lhd (R_3 \rhd\lhd R_5)$ результат будет $r(R) = 40000000$. При произвольном смешанном способе соединения $R = (R_1 \rhd\lhd R_2) \rhd\lhd \rhd\lhd (R_3 \rhd\lhd R_4) \rhd\lhd (R_5 \rhd\lhd R_6)$ результат будет $r(R) = 20000000$.

Анализируя результаты, полученные различными способами, в контексте рассматриваемой задачи необходимо сформулировать и решить задачу поиска минимальной суммы промежуточных результатов на каждом шаге выбора пары соединения. В общем случае задачу можно представить как подбор таких значений из некоторого множества целых чисел, произведение которых даст минимальную сумму промежуточных результатов.

Будем утверждать, что минимальная сумма произвольных пар произведений достигается при перемножении наименьших и наибольших значений заданного множества. Для доказательства этого факта рассмотрим теорему.

Теорема *о наименьшей сумме промежуточных результатов парного произведения.* Пусть задано множество $X = \{x_1,\ x_2, \ldots, x_n\} \in N$, тогда минимум суммы парных произведений $\sum_{i,\,j=1}^{n} \left(x_i\, x_j\right) \to \min$ достигается при $x_i = \min\{x_1, x_2, \ldots, x_n\}$ и $x_j = \max\{x_1, x_2, \ldots, x_n\}$.

Доказательство. Основываясь на том, что количество возможных пар множества *X* конечно и зависит от перестановок элементов *X*(их количество равно *n*!) покажем, что всегда можно найти такие пары элементов, которые определят наименьшую сумму при перемножении.

Пусть $x_i, x_j, x_k, x_l \in X$ и пусть $x_i < x_j < x_k < x_l$. Рассмотрим произведение пар $p_1 = x_i * x_k$, $p_2 = x_j * x_l$, $p_3 = x_i * x_l$ и $p_4 = x_j * x_k$. Покажем, что $(p_1 + p_2) > (p_3 + p_4)$.

Представим комбинацию парных произведений элементов x_i, x_j, x_k, x_l как произведение разностей $(x_j - x_i) * (x_l - x_k)$. Очевидно, что $(x_j - x_i) > 0$ и $(x_l - x_k) > 0$. Таким образом $(x_j - x_i) * (x_l - x_k) > 0$. Раскрыв скобки, получим неравенство $x_j x_l - x_i x_l - x_j x_k + x_i x_k > 0$. Сгруппировав отрицательные и положительные пары произведений, получим $x_j x_l + x_i x_k > x_i x_l + x_j x_k$, т. е. $(p_2 + p_1) > (p_3 + p_4)$.

Применяя это свойство ко всем элементам множества X, всегда можно получить наименьшую сумму промежуточных результатов, в среднем за $2n!$ возможных перестановок. Таким образом, начиная с любой пары за конечное число шагов можно найти последовательность, соответствующую минимальной сумме произведений элементов множества X. ■

Из теоремы о минимальной сумме парных произведений следует, что для упорядоченной последовательности элементов алгоритм составления пар произведений должен выбирать крайние элементы (минимальный и максимальный) и перемещаться к середине множества, пока не станут равны индексы сдвига справа и слева.

ОБОБЩЕННАЯ ЗАДАЧА ПОСТРОЕНИЯ ПЛАНА ВЫПОЛНЕНИЯ ЗАПРОСА

Во многих системах последовательность операций соединения синтаксически ограничена. Как было показано выше, парная последовательность соединений требует вычислительных затрат на создание промежуточных отношений. Хотя такая последовательность приводит к более эффективному плану выполнения запроса, она значительно увеличивает вычислительные затраты на перебор пространства поиска. С другой стороны, более существенна не стоимость генерации синтаксических порядков соединений, а процедура выбора физических операций и оценка каждого возможного плана. Для каждой таблицы статистическая информация содержит число кортежей в структуре потока данных и соответствующие требования к буферам оперативной памяти. Важно различать статистические свойства выполнения запроса и вычислительные затраты, связанные с построением плана [4].

ЗАКЛЮЧЕНИЕ

Полученные в статье результаты позволяют формализовать алгоритм выполнения операций соединения отношений в запросе и при этом минимизировать число кортежей на промежуточных итерациях. Общая задача построения плана выполнения запроса сводится к оценке вычислительных затрат по заданным критериям. Разработанные преобразования запросов позволяют в значительной степени увеличить скорость их обработки путем модификации плана выполнения алгебраических операций.

ЛИТЕРАТУРА

1. *Коннолли Т., Бегг К., Страчан А.* Базы данных: проектирование, реализация и сопровождение. Теория и практика. 2-е изд. / Пер. с англ.: Учеб. пособие. — М.: Издательский дом «Вильямс», 2000. — 1120 с.
2. *Ульман Дж.* Основы систем баз данных / Пер. с англ. М.Р. Когаловского и В.В. Когутовского. Под ред. М.Р. Когаловского. — М.: Финансы и статистика, 1983. — 334 с.
3. *Чаудхари С.* Методы оптимизации запросов в реляционных системах // Системы управления базами данных. — 1998. — № 3.— С. 22–27.
4. *Пономаренко Л.А., Філатов В.О.* Програмні агентні технології в адмініструванні баз даних // Вісник Київського торговельно-економічного університету. Вип. 3. — 2001. — С. 68–73.

Поступила 05.03.2003