УДК 681.3.06

Т.П. Колесник

Харьковский национальный университет внутренних дел, Харьков

ОБ ОДНОМ МЕТОДЕ МИНИМИЗАЦИИ РЕЗУЛЬТАТОВ ОПЕРАЦИИ СОЕДИНЕНИЯ

В статье рассматриваются принципы обработки реляционных запросов и способы их реализации. В качестве основной операции, влияющей на эффективность выполнения запроса, выделена операция соединения. Предложены варианты последовательности соединения отношений, минимизирующих количество кортежей в промежуточных результатах, доказана оптимальность таких последовательностей.

Ключевые слова: *запрос, соединения, модификация, кортежи, минимизация, база данных, алгоритм перебора, сложность выполнения.*

Введение

Постановка проблемы. Цель выполнения процедуры обработки запросов является преобразования этого запроса, записанного на языке высокого уровня, например SQL, в корректную и эффективную последовательность действий, представленную на языке низкого уровня, реализующем операции реляционной алгебры. При этом важнейшим аспектом обработки запроса является его оптимизация. В результате преобразования исходного запроса может быть получено множество его эквивалентных вариантов.

Основная проблема связана со сложностью выполнения вычислительных операций с большим количеством отношений. Как правило, на практике чаще всего выбирается стратегия ближайшая из возможных оптимальных решений. Большинство таких стратегий предусматривает эквивалентные преобразования алгебраических выражений, на которых строится запрос.

Анализ литературных источников показал, что: 1) языки манипулирования данными высокого уровня позволяют формулировать сложные запросы для большого количества отношений, выполнение которых требует много времени; 2) скорость обработки таких запросов может быть в значительной степени увеличена, если перед его выполнением модифицировать план выполнения алгебраических операций; 3) цель таких модификаций – получить эквивалентное выражение, требующее меньше времени и памяти для его выполнения [1 – 5].

Целью настоящей публикации является рассмотрение принципов обработки реляционных запросов и способов их реализации. В качестве основной операции, влияющей на эффективность выполнения запроса, выделена операция соединения. Предложены варианты последовательности соединения отношений, минимизирующих количество кортежей в промежуточных результатах. Доказана оптимальность таких последовательностей.

Изложение основного материала

Общие принципы обработки запросов. Одним из важных свойств, влияющих на эффективность запроса, является порядок вычисления соединений в цепочке операций. Соединение – это решающая операция, так как требуемое время пропорционально произведению размеров соединяемых отношений. В качестве критериев улучшения структуры запроса можно выбрать размер промежуточных результатов 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 храниться отношение, либо имеющимися дополнительными структурами. При этом само вычисление может проводиться различными методами, например, последовательный или индексный просмотр, с учетом или без учета дубликатов и т.п.

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

Рассматриваемые в статье способы реализации запроса использует свойства операций алгебры и теории множеств.

Дальнейшие выкладки будут основываться на свойствах операции соединения и правилах, позволяющих осуществлять некоторые алгебраические преобразования [2].

Архитектура перебора. На практике желательно иметь такой алгоритм перебора, который мог бы легко приспосабливаться к изменениям про-

странства поиска по причине добавления новых преобразований, добавления 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 – это трудная задача, поскольку необходимо не только наличие улучшенного алгоритма перебора, но и обеспечение инфраструктуры для развития техники оптимизации [4]. 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 поиска такой последовательности, которая гарантирует наименьшее суммарное число обращений к кортежам при последовательном соединении.

Пусть, например, база данных R содержит шесть отношений

$$r(R) = \{r_1(R_1), r_2(R_2), r_3(R_3), r_4(R_4), r_5(R_5), r_6(R_6)\},$$

и пусть каждое отношение содержит

$$r(R) = \{100, 50, 80, 5, 20, 10\}$$

кортежей соответственно.

Тогда, при последовательном соединении

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

суммарное число просмотренных кортежей будет соответствовать выражению

$$r = r_1 + r_2 + r_3 + r_4 + r_5,$$

где $r_1 = 100 * 50$, $r_2 = r_1 * 80$,

$r_3 = r_2 * 5$, $r_4 = r_3 * 20$, $r_5 = r_4 * 10$,

то есть $r(R) = 442405000$.

Используя свойства ассоциативности и коммутативности операции соединения, изменяя общую последовательность соединения, можно уменьшить (увеличить) значение $r(R)$.

Изменим последовательность операций. Поменяем местами R_1 и R_6, тогда отношение R' будет получено 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 $X = \{x_1, x_2, ..., x_n\} \in N$, где N – множество натуральных чисел, и пусть результаты операций произведения образуют множество

$$P = \{p_1, p_2, ..., p_m\},$$

где $p_1 = x_1 * x_2$, $p_2 = p_1 * x_3$, …, $p_m = p_{m-1} * x_n$.

Необходимо найти такую последовательность

$$p_1, p_2, ..., p_m,$$

чтобы

$$\sum_{i=1}^{m} p_i \to \min.$$

Теорема (о наименьшей сумме промежуточных результатов последовательного произведения).

Значение суммы промежуточных результатов p_i, $(i = \overline{1,m})$ последовательного произведения элементов множества $X = \{x_1, x_2, ..., x_n\}$ будет наименьшим, если значения $x_1, x_2, ..., x_n$ упорядочены по возрастанию, то есть $x_1 < x_2 < ... < x_n$.

Доказательство. Пусть $x_i, x_j \in X$, $i \neq j$, $i, j = \overline{1,n}$, и пусть $p = x_i * x_j$, очевидно, что произведение с очередным элементом множества X имеет вид $p_i = p_{i-1} * x_{i+2}$ и будет наименьшим если $x_{i+2} = \min_k X$, $k = \overline{1, n-(i+1)}$. С другой стороны значение p будет наименьшим, если начальные значения x_i, x_j будут наименьшими, то есть $x_i, x_j = \min_k X$, $k = \overline{1,n}$.

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

Из теоремы о наименьшей сумме промежуточных результатов последовательного произведения следует, что если изменить порядок перемножения, то сумма также изменится.

Поскольку операция соединения является коммутативной, последовательность ее выполнения не обязательно 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 400000000, что значительно меньше предыдущих значений, полученных последовательным выполнением операции соединения. В дальнейшем такую последовательность будем называть попарной.

Таким образом, за конечное число шагов можно найти такую последовательность пар произведений, при которой сумма промежуточных результатов будет минимальной.

Отличие попарного соединения от последовательного заключается в том, что общая сумма промежуточных результатов получается за несколько независимых шагов. Сначала суммируются произведения произвольных пар, после чего результаты также перемножаются парами в произвольном порядке, пока не будет перемножены все возможные значения.

Теоретически не исключен вариант смешанного произведения, когда часть операций выполняется линейно, а часть попарно [1]. Формулу такого произведения можно изобразить, например, как последовательность

$$R = (R_1 \triangleright\triangleleft R_2) \triangleright\triangleleft (R_3 \triangleright\triangleleft R_4 \triangleright\triangleleft R_5 \triangleright\triangleleft R_6).$$

Используя числовые значения приведенного примера, сравним результаты соединения различными способами. Как было отмечено выше, для линейной упорядоченной по значениям последовательности

$$R = R_4 \triangleright\triangleleft R_6 \triangleright\triangleleft R5 \triangleright\triangleleft R_2 \triangleright\triangleleft R_3 \triangleright\triangleleft R_1$$

результат будет соответствовать значению $r(R) = 400000000$, для произвольного попарного соединения, например вида

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

результат будет также $r(R) = 40000000$, при произвольном смешанном способе соединения, например,

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

результат будет соответствовать

$$r(R) = 2000000000.$$

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 дает 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 попарного произведения). Пусть задано множество $X = \{x_1, x_2, ..., x_n\} \in N$. Минимум суммы попарных произведений $\sum_{i,j=1}^{n} (x_i x_j) \to \min$ достигается при $x_i = \min\{x_1, x_2, ..., x_n\}$ и $x_j = \max\{x_1, x_2, ..., x_n\}$.

Доказательство. Основываясь на том, что количество возможных пар множества X конечно и зависит от перестановок элементов X, которое соответствует $n!$, 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

$$p_1 = x_i * x_k,\ p_2 = x_j * x_l,$$

$$p_3 = x_i * x_l \text{ и } p_4 = x_j * x_k.$$

Покажем, что $(p_1 + p_2) < (p_3 + p_4)$.

Представим элементы x_i, x_j, x_k, x_l как произведение разности $(x_j - x_i) * (x_l - x_k)$. 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

$$x_j x_l - x_j x_k - x_i x_l - x_i x_l + x_i x_k > 0,$$

сгруппировав отрицательные и положительные пары произведений, получим

$$x_j x_l + x_i x_k > x_j x_k + x_i x_l,$$

то есть $(p_2 + p_1) > (p_4 + p_3)$.

Применяя это свойство для всех элементов 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.

Исходя из теоремы о минимальности суммы попарных произведений, следует, что если элементы упорядочены, то алгоритм составления пар произведений должен выбрать крайние элементы (то есть минимальный и максимальный) и сдвигать счетчик

к середине множества, пока не будут равны индексы сдвига справа и слева.

Обобщенная задача построения плана выполнения запроса. Во многих системах последовательность операций соединения синтаксически ограничивается для ограничения размеров пространства поиска. Как было показано выше, попарная последовательность соединений требуют материализации промежуточных отношений. И хотя такая последовательность приводит к более дешевому плану выполнения запроса, она значительно увеличивают расходы на перебор пространства поиска. С другой стороны, наиболее велика не стоимость генерации синтаксических порядков соединений, а выбор физических операций и оценка стоимости каждого возможного плана.

Для каждой таблицы необходимая статистическая информация включает число кортежей в потоке данных, поскольку этот параметр определяет стоимость сканирования данных, соединений и соответствующие требования к памяти. Важно различать статистические свойства выполнения запроса и стоимостью плана. Статистическая информация – это логическое свойство, а стоимость плана – это свойство физическое [4].

Выводы

Полученные в статье результаты определяют статистические свойства, строго задавая последовательность выполнения операций, при этом минимизируя число кортежей в промежуточных результатах, и общая задача построения плана выполнения запроса сводится к оценке стоимости по заданным критериям.

С другой стороны, оптимизация означает намного больше, чем преобразования и эквивалентность запросов. Разработка эффективных преобразований запросов является трудной задачей. Несмотря на многие годы работы, существенные проблемы остаются открытыми.

Одним из важных направлений является то, в котором допускается генерация полных планов при условии доступности информации времени выполнения. Кроме того, открытой остается проблема учета других важных ресурсов (загруженность центрального процессора, время доступа к страницам памяти, методы доступам к данным и т.п.) при определении планов выполнения запроса. Кроме того, при использовании БД в контекстах мультимедиа и Web, появились интересные задачи в связи с нечеткими (неточными) запросами [3].

Список литературы

1. Коннолли Т. Базы данных: проектирование, реализация и сопровождение. Терия и практика, [Текст]: Уч. пос. / Т. Коннолли, К. Бегг, А. Страчан; пер. с англ. 2-е изд. – М.: Издательский дом "Вильямс", 2000. – 1120 с.: ил.

2. Ульман Дж. Основы систем баз данных [Текст] / Дж. Ульман; пер. с англ. М.Р. Когаловского и В.В. Когутовского; под ред. М.Р. Когаловского. – М.: Финансы и статистика, 1983. – 334 с., ил.

3. Чаудхари С. Методы оптимизации запросов в реляционных системах [Текст] / С. Чаудхари // Системы управления базами данных. – 1998. – № 3(98). – С. 22-36.

4. Кузнецов С. Методы оптимизации выполнения запросов в реляционных СУБД [Электронный ресурс] / С. Кузнецов // Центр информационных технологий. – Режим доступа к ресурсу: http://www.citforum.ru/database/articles/art_26.shtml.htm.

5. Руденко Д.А. Модификация ограничений на ведение данных для обеспечения целостности крупномасштабных информационных систем [Текст] / Д.А. Руденко, С.С. Танянский, В.В. Тулупов // Вестник НТУ "ХПИ" Сб. науч. трудов. Тематический выпуск: Информатика и моделирование. – Х.: НТУ "ХПИ", 2006. – № 23. – С. 137-144.

Поступила в редколлегию 27.01.2011

Рецензент: д-р техн. наук, проф. С.Г. Удовенко, Харьковский национальный университет радиоэлектроники, Харьков.

ПРО ОДИН МЕТОД МІНІМІЗАЦІЇ РЕЗУЛЬТАТІВ ОПЕРАЦІЇ З'ЄДНАННЯ

Т.П. Колісник

У статті розглядаються принципи обробки реляційних запитів і способи їхньої реалізації. Як основну операцію, що впливає на ефективність виконання запиту виділено операцію з'єднання.

***Ключові слова**: запит, з'єднання, модифікація, кортежі, мінімізація, база даних, алгоритм перебору, складність виконання.*

ONE OF THE METHODS OF CONNECTION OPERATION RESULTS MINIMIZATION

T.P. Kolesnik

The article describes the principles of relational query processing and the means of their performance. The connection operation was determined as a basic operation influencing on the query effectiveness.

***Keywords:** query, connection, modification, tuples, minimization, database, search algorithm, performance complexity.*