Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 9, №3 (2017) http://naukovedenie.ru/vol9-3.php

URL статьи: http://naukovedenie.ru/PDF/77TVN317.pdf

Статья опубликована 24.06.2017

Ссылка для цитирования этой статьи:
Гришин Д.С. Способ подборки данных в хранилищах текстовых данных на основе продукционного подхода и моделирование его работы // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 9, №3 (2017) http://naukovedenie.ru/PDF/77TVN317.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 681.324

Гришин Дмитрий Сергеевич[1]
ФГБОУ ВО «Юго-Западный государственный университет», Россия, Курск
Аспирант
E-mail: Grish1nds@yandex.ru
РИНЦ: http://elibrary.ru/author_profile.asp?id=789742

Способ подборки данных в хранилищах текстовых данных на основе продукционного подхода и моделирование его работы

Аннотация. В работе рассматривается задача подборки текстовых данных в хранилищах. Данная задача чрезвычайно важна поскольку одномерные формы представления текстовых данных, их размер, многообразие связей, вариативность границ отдельных информационных единиц и другие свойства, приводят к экспоненциальным и другим времяемким пропорциям выполнения подборки данных. Для решения данной задачи в работе предлагается использовать способ подборки на основе продукционного подхода, который осуществляет подборку данных в модифицированном позиционном представлении текста, обеспечивающем направленность поиска, содержащем позиции вхождения в текст всех его собственных подстрок, до задаваемой константы, что позволяет исключить ряд неперспективных операций и значительно сократить временные затраты подборки данных в текстовых хранилищах. В целях определения рационального значения константы для модифицированного позиционного представления текста в работе производится моделирование работы способа подборки в данном представлении при различных значениях этой константы на двух типах данных. Для сравнения и анализа характеристик предложенного способа подборки со способом подборки Бойера-Мура и способом подборки в суффиксном дереве производится моделирование работы этих способов также на двух типах входных данных. На основе результатов моделирования производится оценка целесообразности использования способа подборки в модифицированном позиционном представлении текста.

Ключевые слова: продукционные системы; продукционные вычисления; обработка запросов; хранилища данных; обработка символьной информации; текстовые данные; поиск подстроки

[1] 305000, г. Курск, ул. Почтовая, д. 2, кв. 45

Введение

Современный этап развития поисковых систем, систем принятия решения и анализа текстовых данных в хранилищах характеризуется обработкой нечисловых данных большого размера в виде естественных или искусственных текстов [1, 4]. Среди способов обработки таких данных значительное место занимает схема принятия решение «условие → действие» [5, 7]. Система формальных правил (продукционная система) как формализация данной схемы, а также аппаратные и программные средства, реализации базовых операций составляют основу процессов обработки запросов и администрирования данных [6, 7].

Основу продукционных вычислений определяет система двух базовых крупноблочных операций:

- операция подборки и ее вариации (подборка пересечений, дополнений или объединений);
- операция модификации данных в минимальной подобранной позиции вхождения.

Система данных операций означает, что выходные данные операции подборки безусловно являются определяющей частью входных данных для операции модификации. Общеизвестно, что операция подборки, имеет самостоятельное значение в технических решениях информационно-поисковых систем [4, 9], но она так же применяется как самостоятельная операция в экспертных системах, машинах баз данных, аппаратных средствах поддержки естественно-языковых систем [8, 10], специализированных устройствах систем цифровой связи. Вместе с тем, организация вычислений на основе продукционной логики предусматривает модификацию данных в определенных позициях с изменением размера обрабатываемого текста. Такую модификацию реализует операция замены. Многообразие способов и структур данных для реализации операции подборки служит доминирующим фактором в выборе способов и структур данных для операции модификации. Считая логически связанными эти две операции среди структур данных для организации этих операций можно выделить три типа:

- структуры данных, основанные на принципе последовательного доступа к элементам (список, дерево, очередь, множество и др.);
- структуры данных, основанные на принципе произвольного доступа к элементам (массив);
- гибридные структуры данных, основанные на принципе комбинирования последовательно и произвольного доступа к данным (хеш-карты, хеш-деревья).

Для реализации продукционных операций в хранилищах, целесообразно использовать структуры гибридного типа, которые имеют дополнительную информацию о внутренней организации текста, что позволяет сократить временные затраты как подборки, так и модификации данных [2, 3, 4]. При этом временные затраты на построение гибридной структуры не сопоставимы с «временем жизни» текстовых данных [1, 2, 3].

В качестве гибридной структуры для решения данной задачи предлагается использовать модифицированное позиционное представление текста (МППТ), подборка в котором эффективна для неизменяемых текстовых данных [3]. МППТ содержит все собственные подстроки текста размером до задаваемого пользователем размера $\mathcal{L}$, с соответствующим набором позиций вхождений этих подстрок. **Пример МППТ «abcabaabca» при $\mathcal{L}$ = 3 изображен на рисунке 1.**

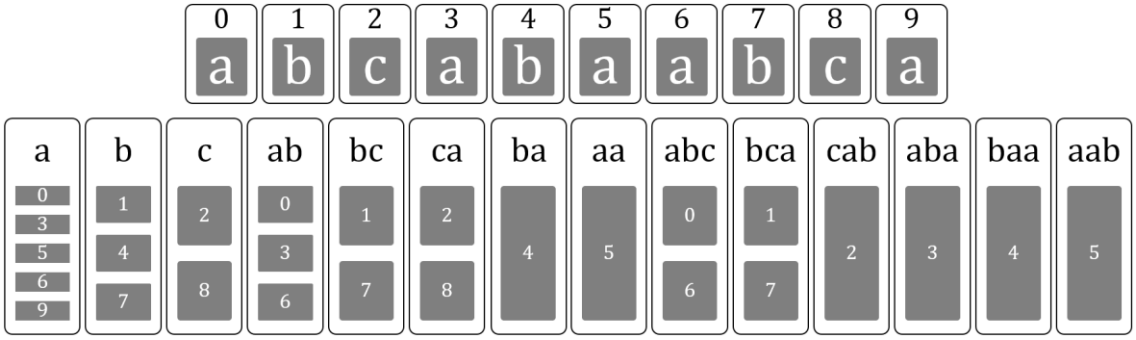

Рисунок 1. *Модифицированное позиционное представление текста «abcabaabca» при $\mathcal{L} = 3$ (составлен автором)*

Способ подборки данных в МППТ состоит из трех шагов [3]:

1. Построение модифицированного позиционного представления.
2. Вычисление набора позиций возможных вхождений.
3. Подборка на основе вычисленных позиций.

Однако в зависимости от значения $\mathcal{L}$ временные затраты подборки и предобработки входного текста для способа подборки в МППТ могут значительно изменяться [3], поэтому в первую очередь необходимо выяснить какое значение $\mathcal{L}$ будет оптимальным для реализации продукционных вычислений в МППТ.

Моделирование

В качестве ЭВМ для моделирования продукционных вычислений в МППТ использовался персональный компьютер, имеющий следующую конфигурацию:

- процессор: Intel Core i7-3770K 4.5 ГГц;
- оперативная память: 24 Гб;
- операционная система: Windows 8;
- тип системы: 64-разрядная операционная система.

В качестве значений $\mathcal{L}$ для моделирования использовались следующие значения: 1, 2, 3, 4, 5, 8, 16, 32, 64, 128. Данное моделирование производилось для двух типов входных данных, для которых были установлены следующие параметры:

1. Случайные данные.
   - размер алфавита: от 2 до 30 (28 итераций);
   - размер текста: от 150 до 100000 (100 итераций);
   - размер подстроки: от 1 до 128 (127 итераций).
2. Естественные данные.
   - размер текста: от 150 до 100000 (100 итераций);
   - размер подстроки: от 1 до 128 (127 итераций).

В результате данного моделирования получены средние временные затраты подборки в МППТ для каждого значения $\mathcal{L}$ (таблица 1).

Таблица 1

Средние временные затраты (составлена автором)

$\mathcal{L}$	Время подборки, мс		Время предобработки, мс	
	Случайные	Естественные	Случайные	Естественные
1	0.001759	0.032493	5.710337	5.500272
2	0.001086	0.010239	13.14473	12.18415
3	0.001005	0.005002	21.59398	21.59398
4	0.000923	0.002875	27.60416	26.22924
5	0.000912	0.001825	41.54946	36.14888
6	0.000929	0.001608	52.48224	52.37958
7	0.000931	0.001562	63.94176	71.17359
8	0.000934	0.001535	86.21683	82.07583
16	0.000981	0.001506	169.3975	163.1859
32	0.001033	0.001563	440.7578	433.7774
65	0.001084	0.001621	785.7504	771.4676
128	0.001279	0.00182	1479.913	1451.068

По результатам моделирования работы способа подборки в МППТ построены графики зависимости временных затрат подборки (рисунок 2а) и предобработки входного текста (рисунок 2б) относительно константы $\mathcal{L}$. Для определения рационального значения константы $\mathcal{L}$ графики были усреднены по типу данных и объединены по временным затратам, от минимальных до максимальных (рисунок 2в).

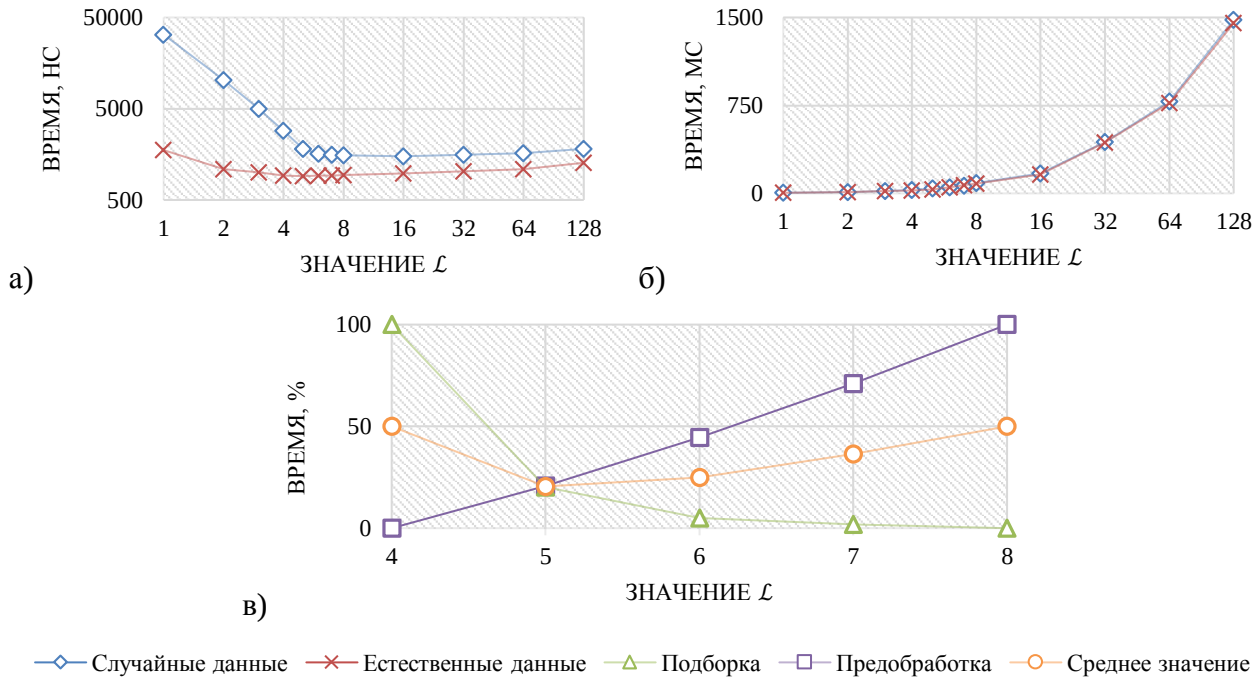

Рисунок 2. *Графики зависимости временных затрат способа подборки в МППТ относительно константы $\mathcal{L}$ (составлен автором)*

Результаты данного моделирования показали, что значение $\mathcal{L}$ равное 5, обеспечивает эмпирически оптимальное соотношение временных затрат подборки данных и предобработки входного текста.

Сравнение временных характеристик подборки в МППТ производилось со способом подборки в суффиксном дереве (МСД), который является одним из самых быстрых способов

подборки, используемых для подборки в неизменяемых текстовых данных, а также со способом подборки Бойера-Мура (БМ), который является одним из самых быстрых среди способов общего назначения.

Способ подборки Бойера-Мура модифицирован для подборки всех вхождений в текст чтобы соответствовать способу подборки в МППТ. Суффиксное дерево модифицировано, таким образом, чтобы все позиции, хранящиеся в листьях для всех поддеревьев в дереве, хранились в корневой вершине данного поддерева, строится на основе алгоритма Эско Укконена.

Моделирование работы этих способов производилось также для двух типов входных данных, для которых были установлены следующие параметры:

1. Случайные данные.
   - размер алфавита: от 2 до 30 (28 итераций);
   - размер текста: от 50 до 500000 (100 итераций);
   - размер подстроки: от 1 до 20 (19 итераций).
2. Естественные данные.
   - размер текста: от 50 до 500000 (100 итераций);
   - размер подстроки: от 1 до 20 (300 итераций).

В результате данного моделирования получены средние временные затраты подборки данных для каждого способа таблица 2.

Таблица 2

Средние временные затраты (составлена автором)

Способ	Время подборки, нс		Время предобработки, мс	
	Случайные	Естественные	Случайные	Естественные
БМ	268676	196839	0	0
в МСД	168.2	179.1	700	697
в МППТ	151.4	148.2	121	67

Вычислим по полученным средним временным затратам подборки процент сокращения временных затрат способом подборки в МППТ относительно способа подборки в МСД. Сокращение временных затрат подборки вычисляется по формуле:

$$\frac{t_{\text{в МСД}} - t_{\text{в МППТ}}}{t_{\text{в МСД}}} \cdot 100\%,$$

где: $t_{\text{в МСД}}$ - средние временные затраты подборки данных в МСД;

$t_{\text{в МППТ}}$ - средние временные затраты подборки данных в МППТ.

Таким образом способ подборки в МППТ позволяет сократить временные затраты относительно способа подборки в МСД для случайных данных на

$$\frac{168.2 - 151.4}{168.2} \cdot 100\% \approx 10\%,$$

а для естественных данных на

$$\frac{179.1 - 148.2}{179.1} \cdot 100\% \approx 17\%.$$

Однако для предлагаемого способа необходимо МППТ, построение которого занимает довольно много времени, в то время как способ БМ производит подборку на исходном тексте

т.е. общие временные затраты работы с текстом при единичной подборки у способа БМ будут намного ниже, чем у предлагаемого. Однако, после того как МППТ построено, с каждой подборкой общие временные затраты работы с текстом у способа подборки в МППТ будут уменьшаться относительно временных затрат способа БМ [3].

Произведем расчет необходимого значения используя средние временные затраты подборки и предобработки входного текста. Количество необходимых операций подборки q, вычисляется при помощи неравенства:

$$(t_{\text{БМ}} \cdot q) > t' + (t \cdot q),$$

где: $t_{\text{БМ}}$ - средние временные затраты подборки данных способом Бойера-Мура;

t' - средние временные затраты предобработки входного текста способом для которого вычисляется количество операций подборки;

t - средние временные затраты подборки данных способом для которого вычисляется количество операций подборки.

Формула вычисления количества q операций подборки данных имеет вид:

$$q = \left\lceil \frac{t'}{t_{\text{БМ}} - t} \right\rceil.$$

Результаты вычисления приведены в таблице 3.

Таблица 3

Количество операций подборки, необходимое чтобы способы с предобработкой входного текста стали эффективней способа подборки Бойера-Мура (составлена автором)

Способ	Количество операций	
	Искусственные	Естественные
в МСД	2608	3545
в МППТ	451	341

Графики зависимости суммарных временных затрат работы с текстом относительно количества операций подборки для случайных данных приведены на рисунке 3а и на рисунке 3б для естественных данных.

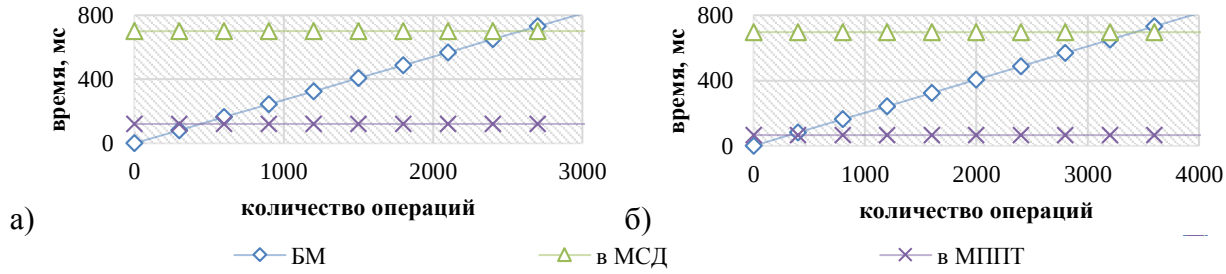

Рисунок 3. *Графики зависимости суммарных временных затрат работы с текстом относительно количества операций подборки (составлен автором)*

Заключение

Результаты моделирования показали, что способ подборки в МППТ, позволяет выполнять гибкую настройку соотношения временных затрат подборки к временным затратам его построения, установкой подходящего значения максимального размера подстрок, хранящихся в данном представлении, что позволяет расширить область его применения и использовать априорные знания о входных данных на естественном языке для сокращения

временных затрат подборки. Если такая информация отсутствует, то задаваемый пользователем размер подстроки рекомендуется выбирать равным 5, что обеспечивает эмпирически оптимальное соотношение временных затрат подборки данных и предобработки входного текста.

Проведенный анализ временных затрат подборки в МППТ при рекомендуемом значении $\mathcal{L}$ равном 5, позволяет сократить временные затраты относительно одного из самых быстрых способов, используемых для подборки в неизменяемых текстовых данных (способа подборки на основе суффиксного дерева) на 10% для случайных данных и на 17% для естественных данных.

Количество операций подборки, которое необходимо способу подборки в МППТ, чтобы общие временные затраты работы с текстом данного способа стали меньше, чем общие временные затраты работы с текстом одного из самых быстрых способов подборки данных общего назначения (способа подборки Бойера-Мура), для случайных данных равно 451 в то время как для естественных данных это количество равно 341.

ЛИТЕРАТУРА

1. Батьков, В.О. Анализ проблем современных хранилищ данных / В.О. Батьков // Труды Международного симпозиума «Надежность и качество». - 2013. - Том 1 - С. 259 - 261.
2. Гришин, Д.С. Алгоритм поиска подстроки на основе позиционного представления текста для архивно-текстовых данных / Д.С. Гришин, Е.А. Титенко, Н.А. Милостная [и др.] // Известия Юго-Западного государственного университета. - 2015. - № 3(16). - С. 43 - 48.
3. Гришин, Д.С. Алгоритм построения структуры, представляющей строку в виде хеш-таблицы, состоящей из хешей подстрок данной строки и алгоритм поиска в ней / Д.С. Гришин, Е.А. Титенко // Известия Юго-Западного государственного университета. - 2015. - № 6(63). - С. 62 - 69.
4. Гришин, Д.С. Ассоциативное матричное устройство для обработки строковых данных в хранилищах текстовой информации / Д.С. Гришин, Е.А. Титенко // Информационные системы и технологии. - 2017. - № 3(101). - С. 72 - 81.
5. Евсюков, В.С. Архитектуры и аппаратные решения обработки символьной информации / В.С. Евсюков, Е.А. Титенко // Инфокоммуникационные системы. - 2009. - № 3. - С. 77 - 80.
6. Зерин, И.С. Метод, алгоритм и техническое решение параллельного поиска и подстановки на ассоциативной памяти / И.С. Зерин, О.И. Атакищев, Е.А. Титенко [и др.] // В мире научных открытий - Красноярск: Научно-инновационный центр, 2012 - № 1.1. - С. 166 - 180.
7. Титенко, Е.А. Структурно лингвистический подход определения продукционных исчислительных систем для задания недетерминированных вычислительных процессов / Е.А. Титенко, М.В. Шиленков // Инфокоммуникационные системы. - 2009 - № 3. - С. 77 - 80.
8. Титенко, Е.А. Устройство и алгоритм ассоциативного поиска вхождений / Е.А. Титенко, В.С. Евсюков, Е.А. Семенихин // Известия Курского государственного технического университета. - 2009. - № 2(27). - С. 56 - 62.
9. Breslauer, D. Simple real-time constant-space string matching / D. Breslauer, R. Grossi, F. Mignosi // Theoretical Computer Science. - 2013. - Т. 483. - С. 2 - 9.
10. Liu, J. A Parallel Algorithm of Multiple String Matching Based on Set-Partition in Multi-core Architecture / J. Liu, F. Li, G. Sun // International Journal of Security and Its Applications. - 2016. - Т. 10. - № 4. - С. 267 - 278.

Grishin Dmitriy Sergeevich
Southwest state university, Russia, Kursk
E-mail: Grish1nds@yandex.ru

Data Warehouse string matching algorithm based on the production system processing. Modeling of the algorithm

Abstract. The paper deals with the task of Data Warehouse string matching. This task is very important because structure of Data Warehouse text data is difficult and very big. It makes string matching algorithms too slow. To solve this problem, it is proposed to use special text data structure and string matching algorithm in the structure. The algorithm let to search Data Warehouse text data very fast because this one based on the production system processing and the special text data structure contains metadata to make searching faster. This paper contains modeling of the algorithm to get optimal way to use special text data structure for Data Warehouse string matching and also to compare the algorithm with suffix tree algorithm and Boyer-Moore algorithm. The modelling results are analyzed to estimate performance of the algorithm for task of Data Warehouse string matching.

Keywords: production system; production processing; query processing; Data Warehouse; symbol information processing; text data; string matching