

**Московский авиационный институт
(национальный исследовательский университет)**

**Факультет информационных технологий и прикладной
математики**

Кафедра вычислительной математики и программирования

Лабораторная работа №5 по курсу «Дискретный анализ»

Студент: А. А. Боглаев
Преподаватель: А. А. Кухтичев
Группа: М8О-306Б-22
Дата:
Оценка:
Подпись:

Москва, 2024

Лабораторная работа №5

Задача: Необходимо реализовать алгоритм Укконена построения суффиксного дерева за линейное время. Построить суффиксное дерево для входной строки для решения своего варианта задания.

Вариант: Найти образец в тексте используя статистику совпадений.

Алфавит строк: Строчные буквы латинского алфавита (от a до z).

1 Описание

Требуется реализовать алгоритм Укконена для построения суффиксного дерева за линейное время.

Согласно [1]: «Алгоритм Укконена строит неявное суффиксное дерево τ_i для каждого префикса $S[1..i]$ строки S , начиная с τ_1 и увеличивая i на единицу, пока не будет построено τ_n . Настоящее суффиксное дерево для S получается из τ_n , и вся работа требует времени $O(n)$.»

Начинают знакомство с алгоритмом Укконена с наивной реализацией, имеющей сложность $O(n^3)$. А после пытаются ускорить алгоритм, чтобы сложность стала $O(n)$.

В чем же проблема алгоритма за $O(n^3)$?

Вместо того, чтобы вставить все суффиксы строки, происходит вставка суффиксов всех префиксов строки. Для поиска и вставки элемента каждый раз осуществляем проход из корня. Алгоритм не эффективен по памяти, так как на ребре хранятся символы строки.

Для ускорения алгоритма будем строить все суффиксы строки, чтобы избежать большого числа повторений, как это было в алгоритме за $O(n^3)$.

Допустим, в суффиксном дереве есть строка $x\alpha$, где x - первый символ строки, а α - оставшаяся подстрока (возможно пустая). Значит в суффиксном дереве есть и строка α . Пусть для строки $x\alpha$ есть внутренняя вершина v , тогда для строки α существует внутренняя вершина p . Значит из v в p можно постоить «путь», который называется суффиксной ссылкой. В программной реализации суффиксная ссылка будет являться указателем из v в p .

Суффиксные ссылки позволяют перемещаться по дереву быстрее, так как нам известно, какие символы будут находится на ребре до внутренней вершины, значит можно не делать проход из корня (можно не делать повторное сравнение). Чтобы построить суффиксную ссылку, нужно хранить указатель на последнюю внутреннюю вершину, и, когда мы создадим новую внутреннюю вершину, необходимо связать ее с последней.

Но это еще не все, для ускорения алгоритма, надо избегать повторных сравнений.

Чтобы перейти по суффиксной ссылке, нужно перейти по ребру до ближайшей внутренней вершины, а после пройти такое же количество символов, после перемещения по ссылке. Избежать лишних сравнений помогут «прыжки по счётчику». Будем считать, сколько символов на ребре было пройдено, до перехода по суффиксной ссылке, и столько же символов «перепрыгнем» после перехода.

Теперь сложность алгоритма будет составлять $O(n^2)$. Причиной всему память, которая затрачивается на хранение символов на ребре. На данный момент на ребре хранится подстрока. Значит символы идут подряд. Тогда можно хранить только индекс начала и конца подстроки на каждом ребре. При добавлении нового символа на

ребро будем увеличивать индекс конца подстроки для всего дерева. Такое «продление» суффиксов в среднем работает за $O(1)$.

Если использовать все вышеописанные эвристики, то алгоритм Укконена будет иметь сложность $O(n)$.

2 Исходный код

Для реализации алгоритма Укконена опишем необходимо написать класс для суффиксного дерева *TSuffixTree*. Дерево состоит из узлов, значит напишем класс для узла *TNode*. Также у дерева есть параметры, которые мы должны хранить для правильного построения. Напишем структуру для хранения параметров *TreeData*.

```
1  #include <iostream>
2  #include <string>
3  #include <vector>
4  #include <unordered_map>
5
6
7  const char SENTINEL = '$';
8  const int TERMPOS = -1;
9  const int TERMVAL = 0;
10 const bool ISLEAF = true;
11
12
13 class TSuffixTree {
14     private:
15         class TNode {
16             public:
17                 int begin; // start curve
18                 int *end; // end curve, ptr, for change
19                 TNode *suffix_link; // suffix link
20                 bool is_leaf; // node status
21                 std::unordered_map<char, TNode *> child; // child
22
23                 TNode(int start, int *finish, TNode* s_link, bool leaf); // node
                     constructor
24         };
25
26         struct TreeData {
27             TNode *current_node;
28             int current_index;
29             int jump_counter;
30             int plannedSuffixs;
31             TNode *last_inner_node;
32         };
33
34         TNode *root;
35         std::string str;
36         int suffixTreeEnd;
37         TreeData params;
38
39         void CreateTree();
40         void AddSuffix(int position);
41 }
```

```

42     void DestroyTree(TNode *node);
43     int CurveLength(TNode *node);
44     void SplitCurve(TNode *node, int position);
45
46 public:
47     TSuffixTree(std::string &input_str);
48     ~TSuffixTree();
49     void MatchStatistic(std::vector<int> &ms, const std::string &str);
50 };

```

Стоит отметить, что строение узла суффиксного дерева, параметры дерева и ряд методов (разделение ребра, подсчет длины подстроки на ребре) доступны только внутри класса. Вне класса пользователю доступен только конструктор с деструктором и метод для подсчета статистики совпадений.

Для начала определим, что информация о ребре находится в узле, в который оно приходит.

Также заметим, что в классе, описывающем узел, есть статус узла, он принимает два значения (*true*, *false*). Так мы сможем отличить внутренний узел от листового. Конечно, вместо булевых значений можно использовать целочисленные и внутреннему узлу присвоить значение -1 , а листовому номер суффикста в строке.

Корень является внутренней вершиной, его статус будет равен *false*. У корня есть только исходящие ребра, значит позиции начала и конца подстроки будет равно -1 .

Опишем методы для класса суффиксного дерева. Так как полный листинг методов займет много места, опишем их в таблице ниже.

suffix_tree.hpp	
void CreateTree();	Создание и построение дерева
void AddSuffix(int position)	Добавление нового суффикса в дерево
void DestroyTree(TNode *node)	Удаление дерева
int CurveLength(TNode *node)	Подсчет длины подстроки на ребре
void SplitNode(TNode *node, int position)	Разделение ребра на два
TSuffixTree(std::string &input_str)	Конструктор для суффиксного дерева
TSuffixTree()	Деструктор дерева
void MatchStatistic(std::vector<int> &ms, const std::string &str)	Подсчет статистики совпадений

При создании дерева в конструкторе будет вызван метод *CreateTree*. Но перед этим, чтобы не копировать входную строку в переменную класса *str*, переместим ее. Так будет затрачено меньше памяти на построение суффиксного дерева.

Конец подстроки на ребре удобно сделать указателем, так можно удобно добавить символ на каждое ребро.

Метод разделения ребра *SplitCurve* создает внутреннюю вершину и «прикрепляет» к ней два листа, старый и новый. Также не забываем прикрепить суффиксную ссылку, если она есть.

Основной метод для конструирования дерева это *AddSuffix*. Тут надо воссоздать все эвристики, которые были описаны выше.

В этой функции ведется подсчет суффиксов, которые нужно создать. В качестве такого счетчика выступает переменная *plannedSuffixes* из структуры *TreeData*. В обычном случае эта переменная сначала увеличивается на 1, а в конце метода уменьшается. Но, если символ, который мы вставляем уже есть, то происходит выход из функции и значение счетчика не уменьшается. Это нужно для последующего прохода по остальным суффиксам и деления ребер.

Основная логика прописана в цикле. Тут мы пытаемся найти новый символ в дереве. Если его нет, то создаем новый лист. Но если символ есть (точнее такой путь), то нужно проверить, есть ли он на ребре, если да, то по правилу 3 в [1], ничего не делаем. Чтобы сделать эффективную реализацию, будем увеличивать счетчик совпадений *jump_counter*.

Если такой путь не нашли, то считаем длину ребра и проверяем, что больше: счетчик прыжков или длина ребра. Если счетчик стал больше, то нужно его уменьшить на длину ребра, а текущую позицию увеличить. И, соответственно, перейти на другое ребро.

После всех преобразований остается разделить ребра, чтобы сохранялось свойство *compactTrie*.

Далее важно понять, в каком узле мы находимся, чтобы подготовиться к новой итерации. Если мы в корне, то увеличим текущую позицию и уменьшим счетчик совпадений. Важно не забыть, что количество суффиксов стало меньше, значит уменьшим значение счетчика суффиксов.

Если же остановились не в корневом узле, то нужно перейти по суффиксной ссылке. Как сказано в [1]: «Определим $ms[i]$ как длину наибольшей подстроки, начинающейся с позиции i , которая совпадает где-то (но мы не знаем, где) с подстрокой. Эти значения называются статистикой совпадений.»

В данной реализации был создан вектор *ms* длины текста, куда будет записываться статистика совпадений. Статистика совпадений это тот же поиск по суффиксному дереву, только с некоторыми улучшениями (ускорениями). Ведь не просто так было построено суффиксное дерево при помощи алгоритма Укконена. Для статистики совпадений будем задействовать суффиксные ссылки и идею с счетчиком совпадений и прыжками.

Для начала будем искать последовательность совпадающих символов. Как только это процесс прервется (встретим различные символы), сохраним длину такой последовательности в вектор *ms*. И продолжим поиск.

Чтобы поиск был эффективным, перейдем по суффиксной ссылке, ведь как мы знаем, символы до внутренней вершины у нас совпадают. Более того, мы знаем, сколько символов совпадает после внутренней вершины, ведь мы только что из сравнили на предыдущем шаге. Тогда в следующую позицию текста можно записать значение из предыдущей ячейки меньшее на 1. При этом проверив, нет ли еще совпадений.

Метод подсчета статистики использует вышеописанные преимущества. Используется три цикла. Берется цикл по тексту. Далее создается переменная $j = i$. Главное условие, чтобы j не вышло за пределы текста.

Внутри цикла делаем поиск символа. Если символ найден, то проходим по ребру иравниваем символы.

Если все символы совпали, то переходим на следующее ребро, иначе нужно перейти по суффиксной ссылке и проверить другое ребро. Так будут сохранены все совпадающие подстроки паттерна и текста.

Если символ не найден, то выходим из цикла.

3 Консоль

```
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ ./lab5
aba
qababaz
2
4
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ ./lab5
baobab
aobbaobabababba
4
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ cat <tests/01.t
fdvmldnvlswa
agtkgdifdvmlndnvlswattgtidmnai
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ ./lab5 <tests/01.t
8
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ cat <tests/05.t
hnvlovfjgz
kvzubkvfswhgkxqtxwvtpqzjmlkhnvlovfjgzyq
alex@wega:~/Рабочий стол/вуз/2course/da_labs/Lab_05$ ./lab5 <tests/05.t
28
```

4 Тест производительности

Реализованный алгоритм Z-функции сравним с наивным алгоритмом Z-функции.

Будем тестировать на последовательности из: 10^3 , 10^4 , 10^5 , 10^6 .

100, 50 1000, 500 10000, 5000 100000, 50000

```
[info] [2024-10-08 21:10:57] Running tests/01.t
Ukkonen+MS: 0.009 ms
std::search: 0.002 ms
[info] [2024-10-08 21:15:32] Running tests/01.t
Ukkonen+MS: 0.136 ms
std::search: 0.008 ms
[info] [2024-10-08 21:16:04] Running tests/01.t
Ukkonen+MS: 1.366 ms
std::search: 0.103 ms
[info] [2024-10-08 21:16:42] Running tests/01.t
Ukkonen+MS: 11.739 ms
std::search: 0.723 ms
[info] [2024-10-08 21:17:46] Running tests/01.t
Ukkonen+MS: 60.507 ms
std::search: 8.749 ms
```

5 Выводы

Выполнив пятую лабораторную работу по курсу «Дискретный анализ», я изуил и реализовал алгоритм Укконена для построения суффиксного дерева за линейное время. Ознакомился с приложениями суффиксного дерева, а точнее со статистикой совпадений.

Статистика совпадений позволяет искать совпадающие подстроки в паттерне и тексте, при этом суффиксное дерево требует меньше памяти, потому что строиться по паттерну. Статистика совпадений активно используется в базах данных для ускоренного поиска.

Суффиксные деревья лучше всего подходят для поиска нескольких шаблонов в одном тексте, так как остальные алгоритмы (Кнута-Морриса-Пратта, Бойера-Мура) не смогут также эффективно справиться с этой задачей. Суффиксные деревья применяются в химии и биологии для работы с ДНК.

Список литературы

[1] Гасфилд Дэн

Строки, деревья и последовательности в алгоритмах: Информатика и вычислительная биология. Пер. с англ. И. В. Романовского. — СПб.: Невский Диалект; БХВ-Петербург, 2003. — 654 с: ил. ISBN 5-7940-0103-8 ("Невский Диалект"), ISBN 5-94157-321-9 ("БХВ-Петербург")

[2] *Алгоритм Укконена — Викиконспекты.*

URL: https://neerc.ifmo.ru/wiki/index.php?title=Алгоритм_Укконена (дата обращения: 05.10.2024). *Алгоритм Укконена — MAXimal.*

URL: <http://www.e-maxx-ru.1gb.ru/algo/ukkonen> (дата обращения: 05.10.2024).