**ОТЧЕТ**

о выполнении долгосрочного домашнего задания

по дисциплине:

«Технологии и методы программирования»

на тему: «Поиск подстроки в строке. Алгоритм Рабина-Карпа»

Выполнил:

Лимасов С.В.\_\_\_\_\_\_\_\_

(подпись)

Проверил:

Кузнецов А.В.\_\_\_\_\_\_\_

(подпись)

Дата: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

2019 г.

Оглавление

[1. Описание задания 3](#__RefHeading___Toc993_284433866)

[2. Математическая модель, теоретическое описание построения и вычислительная сложность 3](#__RefHeading___Toc995_284433866)

[3. Тестирование алгоритма 4](#__RefHeading___Toc997_284433866)

[Заключение 6](#__RefHeading___Toc999_284433866)

[Приложение Листинг файла RKMatcher.h 7](#__RefHeading___Toc1001_284433866)

## Описание задания

Задача заключается в реализации алгоритма поиска в строке. А так же в оценке сложности операции поиска подстроки в алгоритме Рабина-Карпа.

Алгоритм Рабина — Карпа— это алгоритм [поиска строки](https://ru.wikipedia.org/wiki/Поиск_подстроки), который ищет шаблон, то есть подстроку, в тексте, используя [хеширование](https://ru.wikipedia.org/wiki/Хеширование). Он был разработан в [1987 году](https://ru.wikipedia.org/wiki/1987_год) [Майклом Рабином](https://ru.wikipedia.org/wiki/Рабин,_Майкл_Озер) и [Ричардом Карпом](https://ru.wikipedia.org/wiki/Карп,_Ричард).

## Математическая модель, теоретическое описание построения и вычислительная сложность

Алгоритм редко используется для поиска одиночного шаблона, но имеет значительную теоретическую важность и очень эффективен в поиске совпадений множественных шаблонов одинаковой длины. Для текста длины n и шаблона длины m его среднее и лучшее время исполнения равно [O](https://ru.wikipedia.org/wiki/O_нотация)(n) при правильном выборе хеш-функции (смотрите ниже), но в худшем случае он имеет эффективность O(nm), что является одной из причин того, почему он не слишком широко используется. Для приложений, в которых допустимы ложные срабатывания при поиске, то есть, когда некоторые из найденных вхождений шаблона на самом деле могут не соответствовать шаблону, алгоритм Рабина — Карпа работает за гарантированное время O(n) и при подходящем выборе рандомизированной хеш-функциивероятность ошибки можно сделать очень малой. Также алгоритм имеет уникальную особенность находить любую из заданных k строк одинаковой длины в среднем (при правильном выборе хеш-функции) за время O(n) независимо от размера k.

Ключами к производительности алгоритма Рабина — Карпа являются низкая вероятность коллизий и эффективное вычисление хеш-значения последовательных подстрок текста.

Алгоритм начинается с подсчета ***h******a******s******h******(******s******[******0..******m******−******1******]******) и*** ***h******a******s******h******(******p******[******0..******m******−******1******]******)***, а также с подсчета ***p******m*,** для ускорения ответов на запрос.

Для ***i******∈******[******0..******n******−******m******]*** вычисляется ***h******a******s******h******(******s******[******i******.******.******i******+******m******−******1******]******)*** и сравнивается с ***h******a******s******h******(******p******[******0..******m******−******1******]******).*** Если они оказались равны, то образец ***p***, скорее всего, содержится в строке ***s*** начиная с позиции *i****,*** хотя возможны и ложные срабатывания алгоритма. Если требуется свести такие срабатывания к минимуму или исключить вовсе, то применяют сравнение некоторых символов из этих строк, которые выбраны случайным образом, или применяют явное сравнение строк, как в [наивном алгоритме поиска подстроки в строке](https://neerc.ifmo.ru/wiki/index.php?title=Наивный_алгоритм_поиска_подстроки_в_строке). В первом случае проверка произойдет быстрее, но вероятность ложного срабатывания, хоть и небольшая, останется. Во втором случае проверка займет время, равное длине образца, но полностью исключит возможность ложного срабатывания.

Если требуется найти индексы вхождения нескольких образцов, или сравнить две строки − выгоднее будет предпосчитать все степени ***p***, а также хеши всех префиксов строки *s*.

## Тестирование алгоритма

Для проверки алгоритма поиска подстроки проведем эксперименты по поиску подстроки в строке, состоящей из разного числа элементов.

Проведем пять экспериментов по поиску подстроки в строке из n элементов (см. таблицу 2.1).

Таблица 1 – Измерение времени поиска элемента в дереве из n элементов.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Кол-во элементов в строке | Кол-во элементов в подстроке | Измеренное время заполнения (с) | | | | |
| 1 эксперимент | 2 эксперимент | 3 эксперимент | 4 эксперимент | 5 эксперимент |
| 100 | 50 | 0,000043 | 0,063 | 0,066 | 0,064 | 0,064 |
| 1000 | 50 | 0,000181 | 0,093 | 0,093 | 0,094 | 0,095 |
| 10000 | 50 | 0,001872 | 0,141 | 0,143 | 0,141 | 0,142 |
| 100000 | 50 | 0,01152 | 0,182 | 0,183 | 0,185 | 0,183 |
| 200000 | 50 | 0,022387 | 0,217 | 0,218 | 0,215 | 0,021249 |
| 1000000 | 50 | 0,101774 | 0,103076 | 0,101097 | 0,102595 | 0,102142 |
| 2000000 | 50 | 0,207967 | 0,207243 | 0,213205 | 0,207106 | 0,206697 |
| 3000000 | 50 | 0,311964 | 0,310005 | 0,310383 | 0,310334 | 0,309606 |
| 4000000 | 50 | 0,410614 | 0,421347 | 0,407398 | 0,408144 | 0,407834 |
| 5000000 | 50 | 0,522727 | 0,519605 | 0,519497 | 0,517419 | 0,517249 |
| 10000000 | 50 | 1,04774 | 1,03517 | 1,03641 | 1,0418 | 1,03425 |

Результаты практических опытов наглядно представлены на графике (см. рисунок 2).

Рисунок 2 – График измерения времени поиска элемента.

График зависимости возрастает как O(n), что совпадает с теорией.

## Заключение

С точки зрения построения алгоритма, операции вставки тесно связана с поиском элемента в дереве. График функции выше показывает, что алгоритм работает с вычислительной сложность O(n), что совпадает с теоретическими значениями.

Выводы: в результате проделанной работы я получил навыки работы со средой CLion, повторил и подтвердил пройденный мною лекционный материал, освоил методы оценки сложности алгоритмов, приобрел опыт в анализе оценки времени работы алгоритмов.

# Приложение Листинг файла RKMatcher.h

//

// Created by apbfor on 26.03.19.

//

#include <iostream>

#include <string>

#include "cstdio"

#include <ctime>

#include <math.h>

#include <vector>

using namespace std;

#ifndef TEST\_EMPTY\_DDZ\_H

#define TEST\_EMPTY\_DDZ\_H

#endif //TEST\_EMPTY\_DDZ\_H

/\*

\* Поиск подстроки в строке.

\* Простейший алгоритм. Алгоритм Рабина-Карпа.

\* Сравнение времени выполнения реализаций указанных алгоритмов.

\*/

#define PRIME\_NUMBER 11

#define ALPHABET\_ENG 26

#define START\_ASKII\_ENG 97

using namespace std;

int tonum(char c){

return int(c-START\_ASKII\_ENG);

}

int my\_pow(int d,int m,int q){

d=d%q;

int k=(m-1)%(q-1);

if(int(pow(d,k))%q>=0)

return int(pow(d, k))%q;

return int(pow(d, k))%q+q;

}

vector<int> rabin\_karp\_matcher(string T,string P,int d=ALPHABET\_ENG,int q=PRIME\_NUMBER)//d-мощность алфавита q-произвольное простое число

{

vector<int> vec;

if ((T.size()==0)||(P.size()==0))

return vec;

int n =T.size();

int m =P.size();

int h =my\_pow(d,m,q);

int p=0;

vector<int> t;

t.push\_back(0);

for(int i=0;i<m;i++){

p=(d\*p+tonum(P[i]))%q;

t[0]=(d\*t[0]+tonum(T[i]))%q;

}

for(int s=0;s<n-m+1;++s){

if(p==t[s])

if (P==T.substr(s,m)){

vec.push\_back(s);

}

if (s<n-m)

if(((d\*(t[s]-tonum(T[s])\*h)+tonum(T[s+m]))%q)<0)

t.push\_back(((d\*(t[s]-tonum(T[s])\*h)+tonum(T[s+m]))%q)+q);

else t.push\_back(((d\*(t[s]-tonum(T[s])\*h)+tonum(T[s+m]))%q));

}

return vec;

}