INFORMATION TECHNOLOGY DEPARTMENT

MINI PROJECT SYNOPSIS

(KCS554)

A COMPARATIVE STUDY OF STRING PATTERN MATCHING ALGORITHMS
(1901640130034)

Supervisor: Write your supervisor name here

PSIT – PRANVEER SINGH INSTITUTE OF TECHNOLOGY, KANPUR

Vision & Mission Statements of the Department

Vision of the Department:

To be recognized as an impeccable department, that escalates the scholars to innovative IT engineers with apt professionalism and capabilities to adapt to the ever-changing IT industry.

Mission of the Department:

M1: Impart quality education in information technology of global standards by providing in depth knowledge of core subjects as well as extensive practical exposure of modern technologies to align our students toward successful career.

M2: Promote continuous improvement in teaching learning process, foster innovation, and research for attaining academic excellence by students and faculty.

M3: Promote collaboration with industry to bridge the gap between academic and industrial application in emerging IT Technologies.

M4: Inculcate value based socially committed professionalism for the cause of overall development of students and society.

PEOs, POs, PSOs of the Department

Program Educational Objectives (PEOs):

1. The graduate will excel in IT with strong foundation that prepares them in use of recent tools and techniques and be employable /pursue higher studies.
2. The graduate will possess in depth knowledge of IT technologies, concepts of programming languages, software development process, computer networking and communication technology.
3. The graduate will possess excellent communication skills, ethical thinking, leadership qualities and self/lifelong learning attitude.
4. The graduate will possess broad based knowledge, logical reasoning, and innovative approaches to solve real life problems.

Program Outcomes (POs):

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
7. **Environment and sustainability**: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes (PSOs):

1. Use algorithms, data structures/management, software design, concepts of programming languages and computer organization and architecture.
2. Understand the processes that support the delivery and management of information systems within a specific application environment.

Course Outcomes:

On successful completion of this course, students will be able to:

1. Students will be able to apply engineering knowledge to identify real world problems for sustainable development and indulge them in lifelong learning.
2. Students will be able to observe the complex engineering problems and demonstrate the solution for the same.
3. Students will be able to select and apply appropriate tools or technologies for solving real world engineering problems.
4. Students will be able to express their proposed solution by presenting and defending through reports and presentations.

Table of Content

Declaration

I hereby declare that the project entitled – "A Comparative Study Of String Pattern Matching Algorithm", which is being submitted as Mini Project in 5th semester of Information Technology to Pranveer Singh Institute of Technology is an authentic record of our genuine work done under guidance of Prof. Ashish Chakawarti, Department of Information Technology, Pranveer Singh Institute of technology, Kanpur.

Name :- Khushi Gupta

Roll No. :- 1901640130034

Date :-

Signature:-

Acknowledgement

It is my privilege to express our sincerest regards to our project guide. Mr. Ashish Chakawarti, for his valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project. We deeply express our sincere thanks to our Head of Department Mr. Piyush Bhushan Singh for encouraging and allowing us to present the project on the topic “A Comparative Study Of String Pattern Matching Algorithm” at our department premise for the partial fulfilment of the requirement leading to the award of B-Tech degree. We take the opportunity to thank all our lecturers who have directly or indirectly helped our project. We pay our respect and love to our parent and all other family member and friends for their love and encouragement through our career. In last I would like to thank Pranveer Singh Institute of Technology, for providing us such an opportunity to learn from these experiences.

Thank you All.

Date :-

Name :- Khushi Gupta

Roll No :- 190164013003

Signature :-

PSIT
Kanpur

Certificate

This to certify that the project entitled “A Comparative Study Of String Pattern Matching Algorithm” is submitted to department of Information Technology (IT), Pranveer Singh Institute of Technology, Kanpur is the record of bonfire project work carried out under the supervision of Mr. Ashish Chakawarti, for the fulfilment of the requirements of bachelor of technology in department of information Technology of the institute.

Date :-

_____ _____

(Mr. Ashish Chakawarti) (Mr. Piyush Bhushan Singh)

(Supervisor) (Head of the Department)

CHAPTER 1

Introduction

Data Searching is a process which involves matching the data in a systematic order to make it easier to find, work and analyze. In computer science, string searching algorithms are also called string matching algorithms. String matching or searching algorithms search the pattern or alphabets from the array of elements. It is most required when you have data in bulk and we have to find a particular item amongst hundreds, thousands or more items.

Pattern Searching is a process of checking and finding a pattern from a string. Although there are huge numbers of searching algorithms available, but here our work intends to show an overview of comparison between three different types of searching algorithms. We have tried to cover some technical aspects of these three searching algorithms. This research provides a detailed study of how all the three algorithms work & give their performance analysis with respect to time complexity.

This is the way of making our study, research process easy and fast. Our project deals with comparative study about few Pattern Matching Algorithms. These algorithms check the possibility of presence of a sequence of characters from a particular string. If the sequence of characters in found in the string, pattern matching is performed. We work on the time consumed and the space occupied on searching a particular pattern from the data and searches the best result from the entire set of characters as fast and in the most convenient manner.

So our study shows the comparison of the different pattern matching algorithms.

Keywords: Data, Matching, Searching, Memory and Time Complexity, Hash codes, C.

PSIT Kanpur

Problem Statement

There is numerous data which is analyzed and worked upon in every sector in the industry. Every time an item is to be searched an algorithm needs to be implemented to get the best results. A simple example is Google, its uses an algorithm to search for our searched item throughout its database matching the items and then showing us the output. It matches each word with the data and shows us any relevant data to it.

So this process needs to be as fast as possible for efficient searching.

• String matching is needed to search and retrieve items and important data from bulk of information which takes a lot of time if done manually.

• Working with large data and complex categories matching and searching, it becomes difficult and complex.

• Searching data becomes very time consuming and there is a need for better searching methods to make this task fast.

• Amongst all algorithms it's important to compare and understand which algorithm is most suitable for the work.

• Best algorithm is one which takes the least time and space and gives the best output.

• String matching and searching can also be used for plagiarism check to match and compare the strings of Document 1 and Document 2.

When working with data in researches and studies, searching is a common method which helps to pick, update, delete and use data in an easier manner. Searching makes the study easy to understand and work on instead of going through all of it just to find one pattern or text. Working with patterns and matching of data also requires sorting. There are many ways to search data with different time and space complexities. Bases on these complexities and the following test cases we can decide the best algorithm for the data.

Problem Motivation

We use searching so frequently while working on word or maybe power point or any other software. It's simply searching, but this simple search can be done in such effective way. Reading out about so many different algorithms only for searching makes us more eager to go in depth of it. These algorithms propose completely new ideas about how a simple search can be made. This gives us motivation to try and compare these algorithms and try building one of our own.

Objectives :

This comparative study will help student to get the perfect idea of the algorithms they should use for different data projects. This study gives a clear view of the different functions of the pre existing algorithms with new updates. We also aim to lay a comparison between our generated algorithm and the previously compared algorithms.

Sub objective -

• This will help students to understand the searching algorithms and their best case scenarios.

CHAPTER 2

Literature Review

Knuth, Morris and Pratt discovered first linear time string-matching algorithm by following a tight analysis of the naïve algorithm. Knuth-Morris-Pratt algorithm keeps the information that naïve approach wasted gathered during the scan of the text. The algorithm was conceived by James H. Morris and independently discovered by Donald Knuth “a few weeks later” from automata theory. Morris and Vaughan Pratt published a technical report in 1970.

The Rabin Karp algorithm performs the matching by using hash function which was created by Richard M. Karp and Michael O. Rabin (1987). The Rabin–Karp algorithm is inferior for single pattern searching to Knuth–Morris–Pratt algorithm, Boyer–Moore string search algorithm. The hash function used here basically converts every string value into numeric value.

Until now Boyer-Moore is considered as the most efficient algorithm for pattern matching. It was developed by Robert S. Boyer and J Strother Moore in 1977. This algorithm preprocesses the pattern that has to be searched in the string. There are two rules followed here. Two rules that are followed here are the good suffix rule and the bad character rule.

CHAPTER 3

Module-1 Rabin-Karp Algorithm

Rabin Karp Algorithm works very similar to the brute force approach or the naive pattern matching algorithm, in which we traverse step by step while matching each character. Very similar to this is Rabin-Karp algorithm in which we once match the hash value of the substring if the hash value is matched then only we start with the matching of the characters individually. It algorithm uses hash functions and the rolling hash technique. A hash function is a function that maps the data of arbitrary size to fixed size values returning a values called hash values, hash codes, digests or hashes.

In this algorithm there is only one comparison taking place per text subsequence. The character searching and matching is done only when the hash values for the sequences match.

Steps for Rabin-Karp algorithm:

1. Find the hash function for the pattern for item.

2. Find the hash for sub patterns or string of the item (pattern length) .

3. Match and search the substring/pattern of text and its pattern from the block of data. Search the item if the hash of the sub pattern is equal to the pattern's hash.

Module-2 Knuth Morris Pratt

Knuth Morris Pratt Algorithm is based on the concept of generating a suffix-prefix table also known as the Pi table or the lps table. The pi table is generated using the substring (this is the pre-processing part). The way to generate the substring is the main part in this algorithm. The main concept behind this approach is we save the pattern. As soon as we detect the mismatch while searching for the pattern we already know a part of the pattern in the next window.

Module-3 Boyer-Moore

Boyer-Moore Algorithm unlike the other two algorithm starts matching from the last character of the pattern. In this algorithm we create two approaches , the bad character heuristic and the good suffix heuristic. In this approach we shift the character which is initially in the mismatch state and bring it to a position where a match is found. Upon comparing the last character of the pattern if the match is not found the entire pattern is shifted by the length of the pattern.