TEXT SEARCHING ALGORITHM USING BOYER MOORE HORSPOOL (BMH) ALGORITHM

SITI NURAFIQAH BINTI JAAFAR

A thesis submitted in fulfillment of the

requirements for the award of the degree of Bachelor of

Computer Science (Graphics and Multimedia Technology)

Faculty of Computer Systems & Software Engineering

University Malaysia Pahang

DISEMBER 2015

ABSTRACT

The text searching is one of the famous technique uses to find data or text from the system for faster searching. Search algorithm is also known as a universal problem-solving mechanism. We study about the selected technique which is Boyer Moore Horspool algorithm that used for searching the occurrences of a pattern (keyword) in a random text in medical database. The matching between these two words which is pattern and text are analyze by display the accuracy percentage. The objective of this study is to develop a searching function using Boyer-Moore Horspool algorithm based on keyword in a medical database. Therefore, we intend to provide an automatic search routine in the medical first aid application and evaluate performance of the proposed searching algorithm through character match percentage. Based on the study, the average analysis result for this algorithm is 93% which is a strong match.

ABSTRAK

Pencarian teks adalah salah satu teknik yang terkenal digunakan untuk mencari data atau teks daripada sistem untuk carian yang lebih pantas. Algoritma carian juga dikenali sebagai mekanisme penyelesaian masalah yang universal. Kami belajar tentang teknik yang dipilih iaitu Boyer Moore Horspool algoritma yang mana digunakan untuk mencari kejadian sesuatu corak (kata kunci) dalam teks secara rawak di dalam pengkalan data perubatan. Persamaan perkataan bagi kedua-dua corak dan teks ini akan dianalisis untuk memaparkan peratusan ketepatannya. Objektif kajian ini adalah untuk membangunkan fungsi pencarian menggunakan Boyer-Moore Horspool algoritma berdasarkan kata kunci dalam pangkalan data perubatan. Oleh itu, kami berhasrat untuk menyediakan rutin carian automatik dalam permohonan pertolongan cemas perubatan dan menilai prestasi algoritma pencarian dicadangkan melalui peratusan persamaan perkataan. Berdasarkan kajian ini, purata bagi keputusan analisis untuk algoritma ini ialah 93% dimana ia adalah persamaan yang kuat.

TABLE OF CONTENT

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: METHODOLOGY

CHAPTER 4: IMPLEMENTATION

CHAPTER 5: TESTING AND RESULT DISCUSSION

LIST OF FIGURES

LIST OF TABLE

EQUATION

LIST OF ABBREVIATIONS

BM	-	Boyer-Moore algorithm
BMH	-	Boyer Moore Horspool algorithm
BF	-	Brute-Force algorithm
KMP	-	Knuth-Morris Pratt algorithm
RK	-	Rabin-Karp algorithm
T	-	Text
P	-	Pattern

CHAPTER 1

INTRODUCTION

1.0 INTRODUCTION

Search algorithm is known as a universal problem-solving mechanism. It is an algorithm for searching an item (pattern) with a collection of item (text). Normally, it is a list of data documents that match the predefined matching criteria which is the outcome of search forms. This algorithm is use to seeking for more valuable information from available dataset in medical database. For example, search algorithm is needed when to search for word or phrase in a document, search for a student's name in a student list and search for approximate file name in a directory. There are several types of text searching algorithm which is Boyer-Moore algorithm (BM), Boyer Moore Horspool algorithm (BMH), Brute-Force algorithm (BF), Knuth-Morris Pratt algorithm (KMP), and Rabin-Karp algorithm (RK).

This research is concern on pattern matching problems. It is used to check the similarities between pattern matching and the text from database. In many field, such as computer engineering, computer science, bio-science database query and so on, text matching processing is essential and therefore applied frequently. To check whether this pattern is a substring of the text, the algorithm will compares a short string called pattern with a long string called text.

1.1 PROBLEM BACKGROUND

The most important problem in text searching is the exact text matching [1]. This can be describes in finding an exact related text to search an exact pattern input by user which expect the relevant output from their search in the shortest time. For example, there are given a text, T of length m and pattern, P of length n in matching text problem. The output from matching algorithm is either an indication that the pattern P does not exist in T or the starting index in T is a substring matching P. Hence, want to find whether P is a substring of T starting from some index that matches pattern P.

1.2 PROBLEM STATEMENT

Conventional method of search algorithm has raising some problems such as inefficiency of pattern search techniques. The problem consists of:

1.2.1 The amount of memory space needed by the computations.

1.2.2 The matching accuracy between pattern input and text in database.

1.3 OBJECTIVES

Based on the problems statement, the objectives of this study are:

i. To develop a searching function using Boyer-Moore Horspool algorithm based on keyword in a database.
ii. To provide an automatic search routine in any searching application.
iii. To evaluate performance of the proposed searching algorithm through character match percentage.

1.4 SCOPE

The scopes for this project research are:

i. Search algorithm for pattern matching in a text

Study and explore details for search algorithm to find advantages and disadvantages of selected technique which is Boyer Moore Horspool algorithm. The keyword search can be input maximum only one word and search only for symptoms field.

ii. Applying search algorithm techniques in medical database.

This selected technique will be applied to medical database and do some analysis regarding to character match between the input and output display.

iii. Medical database collection

For this study, we collected 1000 data for medical database. Based on this data we divided it into 5 dataset which each dataset hold for 200 data.

1.5 THESIS ORGANIZATION

Chapter 1 begins by giving an introduction part that explains the project overview of the research, the main objectives of the project where those objectives are approaching from the problem statements. As well, it clearly defined the scopes of the project to be focus on this system.

Chapter 2 offers a literature review. It is focusing on the studies and discussion of the existing system. It also will extract idea from the existing systems that will be used in this system. The comparison is declared in this chapter. Besides, software and hardware use also will discuss in this chapter.

Chapter 3 outlines the method about the methodology terms to be discussed on in developing the project. Through this chapter, it wills reviews all the technologies and methods that are used to develop and implement the project.

Next, in Chapter 4 which is implementation describes the implementation for the project development where focusing on construction phase.

Chapter 5 offers a testing and result discussion for the project. This chapter to show the result experiment of selected technique based on its accuracy.

Chapter 6 is a discussion on conclusion of the project research in searching algorithm and it future work.

CHAPTER 2

LITERATURE REVIEW

2.0 INTRODUCTION

Search algorithm is a list of data documents that match the predefined matching criteria which is the outcome of search forms. It also related to the act of seeking for more valuable information from available dataset [2]. Based on [3], without searching algorithm, we would not be possible to find different phrase based upon a certain pattern.

This chapter presents a description of a few standard algorithms used in text searching. It is to find within a text T and match for pattern P, where the text, pattern and match can be interpreted in different ways. This research focuses on when the text is fixed and not known in progress. Text searching is commonly to retrieve many documents that are relevant to the intended search question by the user. Based on [1], this matching algorithm scans the text with the aid of the sliding window mechanism.

In this research, a study on different text search techniques is explored and defined. The focus is on finding the first, or all the occurrences of a keyword in a text-string searching, arises in many computing applications that can be applying in medical database. This search technique is so fundamental that most large computer programs use it in one form or another. It is used to find the matches between the text and specified pattern that enter by the user to search selected terms related to the medical database. Hence, there is some comparison to find the best search algorithm between selected techniques in this research so that can be implemented in medical database. There are several basic techniques for search algorithm based on text searching [2]:

A. Full text search

This type of search automatically indexes all words in each document which is contrast with keyword search. Commonly, it is widely used for internet information retrieval in automated search engines. Due to its completed document indexing, this search presents a clear advantage that relevant document are rarely overlooked.

B. Keyword search

Keyword search allows user to find phrase and words located everywhere in the record since it is widely used in library management systems.

2.1 SEARCH ALGORITHM TECHNIQUES

This technique of text matching algorithms is based on character comparisons between the character in text and character in the pattern. There are five types of techniques that for comparison to this study. The following are some examples of search algorithm.

2.1.1 Boyer Moore Algorithm

Boyer Moore (BM) algorithm was discovered in 1997 is considered as the most efficient string matching algorithm in unusual application. In order to calculate the new comparing position, Boyer Moore algorithm uses both good-suffix function and bad-char function, [6]. The good-suffix function looks up string leader character from right to the left. In addition, the bad-character shift consists in aligning the text character T with rightmost occurrence in P. Normally, this algorithm work fast in the larger alphabet case.

The Boyer Moore algorithm requires O(nm) time complexity for the searching phase. However, for the heuristics' pre-processing phases it requires $O(m+\sigma)$ time and space complexity [5]. Furthermore, times complexity is O(n/m) under best performance, however, under the worst performance, times complexity is O(mn).

2.1.2 Boyer Moore Horspool Algorithm

Boyer Moore Horspool (BMH) algorithm is a simplification of the Boyer Moore algorithm. This algorithm used to solves the problem of string matching where to find an occurrence of a pattern (a string) in a text (another string), or to decide that none exists. The Boyer Moore Horspool algorithm is efficient due to standard benchmark for practical string search literature. Besides, the scanning pattern from right to left method will use when trying to match the pattern against the text [4].

This algorithm proposed to only use the bad-char function of the rightmost character of the window to compute the shift in the BM algorithm [3]. Furthermore, for mismatch case, the shift value is determined by the bad char value of last character instead of character that caused mismatch. So that more jump is archived using bad char. The concept of good-suffix from original Boyer Moore algorithm is removed so that easy to implement. This choice will lead to a better performance in searching the text.

Furthermore, practical applications show that this algorithm is much more efficient compared to Boyer Moore Algorithm .On average, and in worst cases, BMH algorithm is faster than BM algorithm. BMH algorithm require O(m+ σ)time complexity in preprocessing phase and O(mn) time complexity in searching phase since it is a heuristic. In the best performance, it time complexity is O (n/m).

2.1.3 Brute-Force Algorithm

The Brute Force algorithm is the most basic method of approaching the problem of pattern matching. It is also known as Naive String Matching algorithm. This algorithm compares the pattern to the text which is one character at the time until mismatching characters are found. It is very easy and simple to follow.

This algorithm uses a comparing technique which is scanning left to right, one character at a time and requires no preprocessing phase or extra space [1]. This algorithm can be a best choice when need to finding a small pattern in a small text [3]. However, it has no preprocessing phase and did not required extra space [4].

For every possible substring in the text, the Brute Force algorithm takes O(m) time to find whether the pattern appears in the text. Hence, the time complexity for this algorithm is O(nm). Once a mismatch occurs, the sliding window will shifts one position to the left and next restarts to match from the first position of the pattern makes this algorithm not efficiency [1].

2.1.4 Knuth-Morris-Pratt Algorithm

Knuth-Morris-Pratt algorithm (Knuth *et al*, 1977) was introduced by Don Knuth, Jim Morris, and Vaughan Pratt. This algorithm is an improvement of the Brute Force (BF) algorithm [1]. This algorithm was developing in order to speed up the procedure of exact pattern matching by improving the length of the shift [4]. It is consider the first linear string matching algorithm and use a shift function based on the notion of the prefixes of the pattern.

It is the classical algorithm that implements efficiently the left-to-right scan strategy which is quite similar to the Brute Force approach [6]. This algorithm states that the most we can shift the pattern in order to avoid redundant comparisons is namely the length of the next function. In addition, to achieve this task, Knuth-Morris-Pratt preprocessor the pattern to find matches of prefixes of the pattern with the pattern itself.

The Knuth-Morris-Pratt algorithm require O(m) time and space complexity for the preprocessing phase, while it requires O(n+m) time complexity for the searching phase [1]. Disable to work well when the size alphabet increases is the strongest weakness of this algorithm [3]. It will lead to higher chance of mismatch.