STRING MATCHING ALGORITHMS: TO SEARCH AND COMPARE LONGEST COMMON SUBSEQUENCE STRINGS RETRIEVED FROM WAGNER & FISCHER AND HIRSCHBERG ALGORITHMS ON MALAY DICTIONARY WORDS

MOHD RAZIF BIN MOHD GHAZI

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY
AND QUANTITATIVE SCIENCES
UNIVERSITI TEKNOLOGI MARA

SHAH ALAM

1999

ACKNOWLEDGEMENTS

اَلسَّلَامُ عَلَيْكُمْ وَرَحْمَةُ اللهِ

First and foremost I would like to thank Almighty God, Allah S.W.T for His Mercy and Compassion that enabled me to complete this thesis, Alhamdulillah.

Next, I would like to thank to my lecturer cum supervisor, Assoc. Prof. Dr. Zainab Abu Bakar, for giving me valuable guidance without which this thesis could not have been completed successfully and within the time frame allocated. She also gave valuable insight and comments that helped me enormously in the preparation of this thesis. Special thanks is also due to Dean of Information Technology and Quantitative Science, MARA Institute of Technology, Assoc. Prof. Azizi Ngah Tasir, who permitted me the usage of the computer facilities in the faculty.

Credit is also due to the Course tutor, Tuan Haji Mohd Zaki Ghazalee, for his valuable help to me in the preparation of this thesis.

I would like also to thank my family and friends for their support to me throughout my studies in ITM. Without their support I believe that none of my dreams or any of my work could have been realized. Thank you.

ABSTRACT

Information Technology has enabled information that can be in many forms such as text, image or sound, to be accessed widely using search terms via a computer. Due to this type of popularity and advancement in technology, there is an increased interest in searching Malay text to enable scholars and researchers to access the data on-line. This thesis studies the method of a string searching algorithm. There are two methods being evaluated in this research. These are the Wagner & Fischer and Hirschberg algorithms. These methods were chosen because they are based on dynamic programming. Dynamic programming is used to solve both the String Distance Problem and Longest Common Subsequence (LCS) problem. The results obtained shows searching and comparison of LCS retrieved by both algorithms. Comparing of words is carried out based on Dice coefficient.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

String search algorithms are used to find a string of characters from an existing string. It has important applications to data processing, information retrieval, text editing and word processing, linguistic analysis, and also areas of molecular biology such as genetics sequence analysis. Examples of string searching algorithms are Brute Force, Boyer Moore, Knuth Morris Pratt, and many others. This project is to experiment and study two string searching algorithms. They are the Hirschberg and Wagner & Fischer algorithms.

1.2 PROBLEM DESCRIPTION

This study was conducted to overcome the problems that occur on string searching algorithms. The problems are:

- Many researchers used English words for their string searching experiments. String searching is also needed for Malay words. So this project was formulated to help solve the problem of the string searching process on Malay dictionary words. It is also expected to improve performance of the Malay word retrieval system.