AN OPTIMIZED BYTE PAIR ENCODING ALGORITHM FOR STRING MATCHING ON COMPRESSED TEXT

by

OLIVAR, RAVILO VEN

PASION, NICO ARCHELAUS

YAP, JESTONI MARK

ATENEO DE DAVAO UNIVERSITY

COMPUTER STUDIES DIVISION

DAVAO CITY

SEPTEMBER, 2010

AN OPTIMIZED BYTE PAIR ENCODING ALGORITHM FOR STRING MATCHING ON COMPRESSED TEXT

A Mini-Thesis

Presented to the

Undergraduate Faculty of the

Computer Studies Division

Ateneo de Davao University

In Partial Fulfillment

of the Requirements for the Degree

Bachelor of Science in Computer Science

by

OLIVAR, RAVILO VEN

PASION, NICO ARCHELAUS

YAP, JESTONI MARK

ATENEO DE DAVAO UNIVERSITY

COMPUTER STUDIES DIVISION

SEPTEMBER, 2010

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

CHAPTER 2

REVIEW OF RELATED LITERATURE AND WORKS

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

CHAPTER 4

THEORETICAL BACKGROUND

CHAPTER 5

RESULTS AND DISCUSSION

CHAPTER 6

CONCLUSION AND RECOMMENDATION

Chapter 1

INTRODUCTION

1.1 Background of the Study

The compressed pattern matching problem is a heavily tackled branch of string pattern matching - one of the operations of string processing. One major part involve in the compressed pattern matching problem is the compression algorithm. In this paper, we discuss the optimization of one compression algorithm – the Byte Pair Encoding Algorithm.

A study of Takeda et al. already acknowledges the Byte Pair Encoding as a text compression scheme that accelerates pattern matching. The Byte Pair Encoding has its perceived drawbacks, such as uncompetitive compression ratio and compression time in comparison to well-known compression algorithms.

The proponents of this paper are interested in the pursuit of this study to further improve not only the pattern matching speed but also to make the use of this compression algorithm practical – by improving its compression ratio and compression time.

1.2 Statement of the Problem

The study intends to improve the Byte Pair Encoding algorithm used for compressing files. To actually realize its potential improvements, the proponents through their study will test it by creating a pattern matching over compressed file tool implementing the Boyer-Moore pattern matching and Byte Pair Encoding compression algorithm.

The present study seeks to answer the following general problem: What optimization / enhancement could be done to improve the performance of the Byte Pair Encoding algorithm in relation to string matching on compressed text?

Specifically, it seeks to answer the following questions:

- What is the degree of improvement of this optimized / enhanced Byte Pair Encoding algorithm compared to the standard Byte Pair Encoding algorithm in terms of compression ratio and time?
- How can strings be found using Boyer-Moore pattern matching algorithm in files compressed using the optimized / enhanced Byte Pair Encoding algorithm?
- How does the string matching in the optimized / enhanced version of Byte Pair Encoding algorithm perform as compared to string matching in the standard version of Byte Pair Encoding algorithm?

1.3 Objectives of the Study

The main objective is to enhance or optimize the Byte Pair Encoding algorithm to achieve better pattern matching speed on compressed text. Furthermore, the proponents would also develop a tool to prove the potential benefit of the algorithm.

The specific objectives were to:

- Identify which part of the Byte Pair Encoding algorithm needs to be enhanced or optimized.
- Analyze and implement the Boyer-Moore pattern matching algorithm in both standard and enhanced Byte Pair Encoding algorithm compressed text through a pattern matching tool.
- Evaluate other compressed pattern matching tool and other existing approaches addressing the compressed pattern matching problem.
- To obtain and analyze the results and formulate a conclusion and draft recommendation for future studies.

1.4 Significance of the Study

The significance of the study lies in the fact that string pattern matching problem is still one of the most fundamental problems in the field of computer science. The study sheds light on how the compression algorithm can be optimized or enhanced to improve the performance of string pattern matching over compressed text. Moreover, the compression algorithm that is being

enhanced or optimized in this study – which is the Byte Pair Encoding algorithm – would be more practical to use in other compression related problems.

Furthermore, the study is also important for it will provide a platform or basis for future studies improving this certain type approach on pattern matching for compressed text and other related technologies.

1.5 Scope and Limitations of the Study

The scope of the study is to identify the weak points on certain parts of the standard Byte Pair Encoding algorithm and implement some possible solutions to come up with an enhanced / optimized Byte Pair Encoding algorithm. The study also includes the adaptation and implementation of the existing Boyer-Moore pattern matching algorithm on standard and enhanced / optimized Byte Pair Encoding algorithm compressed files. To test the results of the study, the proponents will design a pattern matching on a compressed text tool using the Boyer-Moore pattern matching algorithm on standard and optimized / enhanced Byte Pair Encoding compression algorithm.

1.6 Definition of Terms

LZGrep – Boyer-Moore String Matching Tool for Ziv-Lempel Compressed Text

BPE – Byte Pair Encoding

BM – Boyer Moore

LZ– Lempel-Ziv with LZW (Lempel-Ziv-Welch) as its most popular variant;