Comparison of Boyer-Moore and Knuth-Morris-Pratt Algorithms

By

CHOWDHURY MD TASNIM

Project Paper Submitted in Partial Fulfillment as the Requirement for the Master in Information Technology in the Faculty of Creative Media & Innovative Technology

IUKL

2018

Abstract of project paper presented to the Senate of Infrastructure University Kuala Lumpur in partial fulfillment of the requirement for the degree of Master in Information Technology.

Comparison of Boyer-Moore and Knuth-Morris-Pratt Algorithms

By

Chowdhury Md Tasnim

March/2018

Chair: Dr. Abudhahir Buhari

Faculty: Faculty of Creative Media and Innovative Technology

Sequence parallelization systems are an essential class of sequence parallelization that attempt to discover a location where one or a few sequences are found (likewise called contents) are found inside a bigger string or in sequences . The essential string parallelization issue is characterized as takes after given twice of sequences in content and the content to be found, decide if the content shows up in the sequence. Sequence parallelization algorithms are connected in numerous applications of computer and the relevant gadgets. For example, in the field of information preparation, images and voice acknowledgment, data recovery, computational science content to be matched for the formation of sequence . Besides, sequence parallelization systems have turned into a critical segment of uses which are utilized to look nucleotide or amino corrosive succession designs in natural grouping databases as of late. For instance, when proteomics information is utilized for genome explanation in a process called proteogenomic mapping where an arrangement of peptide recognizable pieces of proof got utilizing mass spectrometry is coordinated against an objective genome deciphered in each of the six perusing outlines in the multiple content parallelization systems. Among twice of them popularly utilized Boyer-Moore and Knuth-Morris-Pratt (KMP) algorithms. Here discussed about the reinforced algorithm which would act promptly for the support of those project. Also, in this project there taken twice number of cases of finding the efficiency of algorithm. They are accuracy and execution time .For the experiment of

twice number of conditions are tested in also in twice number of paragraphs which are large in size. After that the result section shows that, Boyer-Moore algorithm revealed out as faster system in terms of efficiency.

ACKNOWLEDGEMENT

First, my thanks and respect to almighty for blessing me with the capability to do this work. I am especially grateful to my supervisor Dr. Abudhahir Buhari for helping me greatly to extend my educational experience. Without his guidance, it was impossible for me to complete this work. I am also greatly indebted to my Examiner Dr.Elisha Tadiwa Nyamasvisva to grant me this opportunity. I am also grateful to my parents for their support, prayer and inspiration to extend my study.

APPROVAL

This Project paper was submitted to the Senate of Infrastructure University Kuala Lumpur (IUKL) and has been accepted as partial fulfillment of the requirement for the degree of Masters in Information Technology in the Faculty of Creative Media & Innovative Technology. The members of the project paper Examination Committee were as follows:

Dr. Abudhahir Buhari
Faculty: Faculty of Creative Media and Innovation Technology
University: Infrastructure University Kuala Lumpur (IUKL)
(Supervisor)

Dr. Elisha Tadiwa Nyamasvisva
Faculty: Faculty of Creative Media and Innovation Technology
University: Infrastructure University Kuala Lumpur (IUKL)
(Internal Examiner)

Assoc. Prof. Dr. Manal Mohsen Abood
Director
Centre for Postgraduate Studies
Infrastructure University Kuala Lumpur (IUKL)
Date:

DECLARATION

I declare that the thesis is my original work based on some concept of others except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Infrastructure University Kuala Lumpur.

Signature …….....
CHOWDHURY MD TASNIM
June 06, 2018

TABLE OF CONTENT

LIST OF FIGURES

Page

LIST OF TABLES

LIST OF ABBREVIATIONS

BM	Boyer-Moore
KMP	Knuth-Morris-Pratt
AC	Aho Corasick
WM	Wu Manber
SWBM	Set Wise Boyer-Moore
BOM	Backward Oracle Matching
NW	Needleman Wunsch
SW	Smith Waterman
WM	Wu Manber
UNIX	Uniplexed Information and Computing System
BNDM	Backward Non Deterministic Acylic finite state automation Matching
BDM	Backward Deterministic Matching
SBOM	Set Backward Oracle Matching
IDE	Integrated Development Environment
IBM	International Business Machines
TRF	Turbo Reverse Factor
RF	Reverse Factor
RK	Robin Karp

CHAPTER 1

INTRODUCTION

1.1 Introduction

As a critical area in the field of science in computer, scientists around the world improve numerous sequence parallelization algorithms. Each having their own forte as far as effectiveness, dependability, execution and so forth. A tremendous and imperative region it is, string parallelization system are classified in numerous areas. At present, there are rundown of string parallelization algorithms. Each has possessed conduct with respect to capacity, execution, preparing time, calculation time intricacy and most pessimistic scenario situations. As the disclosure procedure of new natural succession increments with the innovative head ways keep on progressing, interest for the examination of arrangement of string parallelization getting more utilized(Custom et al., 2016).Sequence parallelization algorithms have twice of modus. They are- exact parallelization and approximate parallelization. In exact string parallelization, the content is fully paralleled with the distinct sequence window of input sequence and it exhibits from the beginning or introductory position of record or index. The algorithms which belong to this type are Knuth-Morris-Pratt (KMP), Needleman Wunsch (NW), Dynamic Programming, Boyer Moore and Smith Waterman (SW). In approximate sequence parallelization, if certain section of the content paralleled with the selective sequence window then at once it exhibits the output. Examples of these classes are Brute Force, Fuzzy sequence searching and Rabin Karp (Janani & Vijayarani, 2016).Among twice of them popularly utilized Boyer-Moore and Knuth-Morris-Pratt (KMP) systems. Here discussed about the reinforced system which would act promptly for the support of thosealgorithm.

REFERENCES

Al-mamory, S. O., Hamid, A., Abdul-razak, A., & Falah, Z. (2010). String Matching Enhancement for Snort IDS (pp. 1020–1023). https://doi.org/10.1109/ICCIT.2010.5711211

Bhardwaj, V. (2015). A Comparative Study of Wu Manber String Matching Algorithm and its Variations, *132*(17), 34–38. Retrieved from https://pdfs.semanticscholar.org/4e22/0ec3678890e7e94e9277d85dc7086fe46f08.pdf

Custom, C., Service, W., Issues, S., Topics, O., Profiles, C., Writing, C., & Now, O. (2016). A study on string matching algorithm politics. Retrieved from http://essaymonster.net/politics/83620-a-study-on-string-matching-algorithm-politics.html

Janani, R., & Vijayarani, S. (2016). An Efficient Text Pattern Matching Algorithm for Retrieving Information from Desktop. *Indian Journal of Science and Technology*, *9*(43), 1. https://doi.org/10.17485/ijst/2016/v9i43/95454

Juilee. (2014). Join the Ques10 Community. Retrieved from http://www.ques10.com/p/9332/to-implement-the-knuth-morris-pratt-string-matchin/

Keim, G. (1997). Boyer-Moore Algorithm. Retrieved from https://www2.cs.duke.edu/courses/cps130/fall97/lectures/lect14/node15.html

Kelly, J. (2006). *An Examination of Pattern Matching Algorithms for Intrusion Detection Systems*. Carleton University. Retrieved from https://pdfs.semanticscholar.org/4f2f/8e96de6774813501ae94021f9c36d475eddb.pdf

Lecroq, C. C.-T. (1997). Boyer-Moore algorithm. Retrieved from http://www-igm.univ-mlv.fr/~lecroq/string/index.html

Lecroq, C. C.-T. (1997). Karp-Rabin algorithm. Retrieved from http://www-

igm.univ-mlv.fr/~lecroq/string/index.html

M. Crochemore, A. Czumaj, L. Gasieniec, s S. Jarominek, T. Lecroq, W. Plandowski, W. R. (2010). A Composite Boyer-Moore Algorithm for the String Matching Problem, 492–496. https://doi.org/10.1109/PDCAT.2010.58

Pandiselvam, P. (2014). A comparative study on string matching algorithms of biological sequences, 1–5. Retrieved from https://arxiv.org/abs/1401.7416

Plandowski. (1994). Speeding Up Two String-Matching Algorithms, *12*(4–5), 247–267.

Sandeep Jain, G. G. G. (2006). *RESEARCH ARTICLES TVSBS : A fast exact pattern matching algorithm for biological sequences*. Retrieved from http://www.jstor.org/stable/24094174

Sandeep Jain, G. G. G. (2009)Pattern Searching :Set 7 (Boyer Moore Algorithm – Bad Character Heuristic). Retrieved from https://www.geeksforgeeks.org/pattern-searching-set-7-boyer-moore-algorithm-bad-character-heuristic/

Sandeep Jain, G. G. G. (2009). Searching for Patterns : Set 2 (KMP Algorithm). Retrieved from https://www.geeksforgeeks.org/searching-for-patterns-set-2-kmp-algorithm/

Ui, F. (2009). Pattern Matching. Retrieved from https://www.techopedia.com/definition/8801/pattern-matching