
# IMPROVEMENT ON PATTERN MATCHING ALGORITHMS USING CHARACTERS DISTRIBUTION PROPERTY OF A LANGUAGE

PORNTEP SUKSRIVILAIKUL



A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2002

ISBN 974-04-1869-4

4037530 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc. (COMPUTER SCIENCE)
KEY WORD : PATTERN MATCHING
PORNTEP SUKSRIVILAIKUL : IMPROVEMENT ON PATTERN MATCHING ALGORITHMS USING CHARACTERS DISTRIBUTION PROPERTY OF A LANGUAGE. THESIS ADVISORS : DAMRUS WONGSAWANG, Ph.D., SUKANYA PHONGSUPHAP, Ph.D. 159 p. ISBN 974-04-1869-4.

This research, we study and analyze an improvement on pattern matching algorithms, focusing on reducing the number of characters comparisons. An interesting concept is using characters distribution property of a language. This method, called CFDC (Characters Frequency Distribution Based Comparison), is used for improvement on the original algorithms, by combining with the original jumping mechanism. These improved algorithms are called Pattern Matching Algorithms with CFDC.

The exact pattern matching, or as some researchers call it exact string searching, is a significant part of many applications, including text editing, data retrieval, and symbol manipulation. In spite of the use of indices for searching large amounts of text, string matching may help in an information retrieval (IR) system which nowadays plays as a key role of an organization.

We found a problem occurring in comparisons of matched characters before a mismatch found in each matching attempt, that is the waste times for such useless comparisons. From the review we noticed that generally, the improvements for a better performance in pattern matching found in the present research works can be focused on two aspects: the fewer number of characters comparisons in the checking step and the farther distance in the jumping step. Most research work focus on jumping improvement. There is an alignment of characters in pattern for jumping distance computation, such as Boyer Moore's Jumping Mechanism. The characters comparison in the conventional algorithms is ordered sequentially. By testing on general text files, articles, HTMLs, technical manuals, news, research papers as well as short stories in our corpus, with general word patterns, the experimental results show the number of characters comparisons, performed in the improved algorithms, are less than those performed in the original algorithms for 0.1 to 24.9 percents on average.

We can conclude in the final of this research work that our proposed model, Pattern Matching Algorithms with CFDC, is significant improvement, in terms of number of characters comparisons, compared with the original algorithms.

4037530 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)

พรเทพ สุขศรีวิไลกุล : การปรับปรุงวิธีการค้นหาสายอักขระโดยใช้คุณสมบัติการกระจายของตัวอักษรในภาษา (IMPROVEMENT ON PATTERN MATCHING ALGORITHMS USING CHARACTERS DISTRIBUTION PROPERTY OF A LANGUAGE) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D., สุกัญญา พงษ์สุภาพ, Ph.D. 159 หน้า.
ISBN 974-04-1869-4.

การค้นหาสายอักขระ (Pattern) บนข้อความ (Text) มีบทบาทสำคัญต่อระบบการสืบค้นข้อมูลสารสนเทศ ตัวอย่างเช่นเป็นตัวช่วยค้นหาคำสำคัญอย่างละเอียด ในข้อมูลหรือเอกสารที่สืบค้นมาได้ เป็นการกลั่นกรองข้อมูลที่เราต้องการอีกชั้นหนึ่ง ซึ่งในปัจจุบันมีข้อมูลสารสนเทศอยู่ในระบบเป็นจำนวนมาก การมีวิธีการค้นหาสายอักขระที่มีประสิทธิภาพจึงเป็นสิ่งสำคัญ

ในการปรับปรุงวิธีการค้นหาสายอักขระที่เราพบในปัจจุบันจะกระทำกันอยู่สองลักษณะคือ หาวิธีลดจำนวนการเปรียบเทียบตัวอักษร (Number of Characters Comparisons) และหาวิธีการเคลื่อนย้ายตัวแบบสายอักขระไปที่ตำแหน่งของการเปรียบเทียบถัดไปให้ได้ระยะไกลที่สุด หรือเรียกอีกอย่างหนึ่งว่าการกระโดด (Jumping) อย่างไรก็ตามงานวิจัยส่วนใหญ่ที่พบ ต่างก็มุ่งที่จะปรับปรุงในเรื่องของวิธีการเคลื่อนย้ายตัวแบบสายอักขระดังกล่าว มีการใช้การจัดวางตัวอักษรบนสายอักขระ มาช่วยในการคำนวณระยะกระโดด เช่นที่เรารู้จักกันดีในชื่อของ Boyer Moore's Jumping Mechanism ในขณะที่วิธีการจัดลำดับในการเปรียบเทียบตัวอักษรที่ใช้อยู่ในปัจจุบันซึ่งเป็นลักษณะเรียงไปตามลำดับตำแหน่งของตัวอักษร (Sequential Ordering) มีปัญหาเกิดขึ้นคือ มีการเปรียบเทียบตัวอักษรที่เข้าคู่กัน (Match) ก่อน ทั้งๆที่มีตัวอักษรที่ไม่เข้าคู่กัน (Mismatch) ปรากฏอยู่ในสายอักขระที่เรากำลังทดสอบอยู่ ทำให้สูญเสียเวลาในการเปรียบเทียบตัวอักษรเหล่านั้นโดยไม่จำเป็น

ในงานวิจัยฉบับนี้จึงทำการศึกษาการปรับปรุงวิธีการค้นหาสายอักขระดังกล่าว โดยมุ่งเน้นที่การลดจำนวนการเปรียบเทียบตัวอักษรบนสายอักขระกับข้อความให้น้อยลง จากแนวคิดที่เราสนใจคือการนำเอาคุณสมบัติการกระจายของตัวอักษรในภาษามาใช้ เราเรียกวิธีนี้ว่า การเปรียบเทียบโดยอาศัยการกระจายความถี่ของตัวอักษร (Characters Frequency Distribution Based Comparison - CFDC) ซึ่งจะใช้ปรับปรุงวิธีการค้นหาสายอักขระที่มีอยู่เดิม โดยการนำเอาวิธีนี้มาใช้ร่วมกับกลไกการกระโดด (Jumping Mechanism) ของวิธีการค้นหาเดิม ซึ่งเรียกวิธีที่ถูกปรับปรุงแล้วว่าวิธีการค้นหาสายอักขระด้วยการจัดลำดับการเปรียบเทียบแบบอาศัยการกระจายความถี่ของตัวอักษร (Pattern Matching Algorithms with CFDC) จากการทดลองพบว่า จำนวนการเปรียบเทียบตัวอักษรในวิธีที่ถูกปรับปรุงลดน้อยลงโดยเฉลี่ย ร้อยละ 0.1 ถึง 24.9 เมื่อเทียบกับวิธีการค้นหาสายอักขระแบบเดิม