PARTIAL STRING MATCHING USING BIT-SLICED SIGNATURE FILES

ATHIWAT ARPAPONGSAK G4037547

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1999

ISBN 974-662-307-9

4037547SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc. (COMPUTER SCIENCE)
KEY WORDS : SIGNATURE FILE / WILDCARD / TEXT PATTERN SEARCH
ATHIWAT ARPAPONGSAK : PARTIAL STRING MATCHING USING BIT-SLICED SIGNATURE FILES. THESIS ADVISOR : DAMRAS WONGSAWANG Ph.D.
82 p. ISBN 974-662-307-9

The partial string searching using text pattern matching in unformatted data normally requires much processing time because it compares an indicated query with all data which are usually in large size. A signature file represents an actual file in searching. Since signature file size is much smaller than an actual file's size, the processing time is faster. However, signature file algorithm cannot be directly used in partial searching. This thesis proposes a new approach to the use of signature file algorithm in partial string searching. We develop a searching algorithm called Wildcard Searching with Signature File (WSSF). WSSF creates two signature files that will be used in partial string searching. From theoretical analysis and experimentation of WSSF with actual data, we found that WSSF is more efficient than any existing text pattern matching algorithms when applied to partial string searching.

This thesis presents how WSSF works and its procedures in detail. The researches on WSSF, its experimentation with real data, and its results are discussed. Moreover, suggestions and comments for improving WSSF in using this algorithm in the real world are also presented.

4037547SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์; วท.ม (วิทยาการคอมพิวเตอร์)

อธิวัฒน์ อาภาพงศ์ศักดิ์ : การค้นหาคำแบบบางส่วนโดยการใช้แฟ้มข้อมูลลายเซ็นแบบลำดับ (PARTIAL STRING MATCHING USING BIT-SLICED SIGNATURE FILES) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง Ph.D. , ธันวดี ธนิตสุขการ Ph.D., ธีรเกียรติ์ เกิดเจริญ Ph.D. 82 หน้า. ISBN 974-662-307-9

ปกติการค้นหาบางส่วนของคำ (Partial String Searching) ในข้อมูลที่ไม่มีรูปแบบต้องใช้เวลาในการประมวลผลมาก เมื่อใช้เทคนิคของการค้นหาแบบเปรียบเทียบกับข้อมูลโดยตรง (Text Pattern Matching) เพราะจะต้องค้นหาเปรียบเทียบกับข้อมูลทั้งหมดซึ่งมักจะมีขนาดใหญ่ แฟ้มข้อมูลลายเซ็นจะทำหน้าที่เป็นตัวแทนของข้อมูลจริง ซึ่งสามารถจะใช้ในการค้นหาได้เช่นเดียวกับข้อมูลจริง แต่จะมีขนาดเล็กกว่าข้อมูลจริงมาก จึงทำให้การค้นหาทำได้เร็วกว่า อย่างไรก็ตาม แฟ้มข้อมูลลายเซ็นไม่สามารถจะใช้กับการค้นหาแบบบางส่วนได้โดยตรง งานวิทยานิพนธ์นี้จะได้นำเสนอวิธีการที่จะใช้แฟ้มข้อมูลลายเซ็นในการค้นหาแบบบางส่วน วิธีการค้นหาที่เรียกว่า WSSF (Wildcard Searching with Signature Files) ได้ถูกพัฒนาขึ้น WSSF จะสร้างแฟ้มข้อมูลลายเซ็น 2 แฟ้ม ซึ่งจะสนับสนุนการค้นหาแบบบางส่วนได้ จากการวิเคราะห์และทดลองกับข้อมูลจริงพบว่า WSSF ให้ประสิทธิภาพในการค้นหาแบบบางส่วนได้ดีกว่าวิธีการปกติ

วิทยานิพนธ์ฉบับนี้จะได้นำเสนอวิธีการของ WSSF โดยละเอียด พร้อมกับการวิเคราะห์ การนำไปทดลองกับข้อมูลจริง พร้อมทั้งผลการทดลองที่ได้ นอกจากนี้ยังจะได้นำเสนอข้อแนะนำต่าง ๆ ในการที่จะปรับปรุง WSSF ให้ดีขึ้นสำหรับการใช้งานจริง