TEXT COMPRESSION
BY
CHARACTER SEQUENCE ANALYSIS METHOD

SONGKRIT KRITSADEERATTANAMANEE

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2001

ISBN 974-665-372-5

3936987 SCCS/M : MAJOR : COMPUTER SCIENCE ;
M.Sc. (COMPUTER SCIENCE)
KEY WORD : TEXT COMPRESSION

SONGKRIT KRITSADEERATTANAMANEE : TEXT COMPRESSION BY CHARACTER SEQUENCE ANALYSIS METHOD. THESIS ADVISOR : DAMRAS WONGSAWANG, Ph.D., THANWADEE SUNETNANTA, Ph.D. 124 P.
ISBN 974-665-372-5

Text compression is one of the interesting topics among groups of researchers due to its wide variety of applications. However, searching directly in the compressed text is still problematic. Many compression schemes that provide search capability have been proposed and implemented by many researchers. This study also aims at obtaining effective search capability.

The present work is based on the research entitled "Text Compression Scheme that allows Fast Searching Directly in The Compressed File" by Udi Manber of the University of Arizona. This study tries to develop and improve the above scheme further by having the target of more than 30% compression rate with the capability of direct searching in compressed files. The basic idea of the newly developed scheme, called CSAM (Character Sequence Analysis Method), has been proposed by the researcher of this study. The CSAM scheme is similar to the Udi Manber's scheme in that it still applies the pattern substitution method. However, CSAM looks into texts in more detail and carefully analyzes character sequence appearing in the actual text to find the best substitution. CSAM can achieve more than 30% for the compression saving, in average, while the ability of pattern searching without decompression is still provided. The speed of compression may not be attractive for general applications at this stage of the development. However, this scheme may be suitable for applications which need to read quite often but seldom write.

In this Thesis, the CSAM scheme was described, the prototype developed, tested, and implemented. Moreover, the performance and the experimental results have been presented and discussed. Finally, further improvements of the scheme have also been suggested.

3936987 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม.(วิทยาการคอมพิวเตอร์)
ทรงกฤต กฤษฎีรัตนมณี : การบีบอัดข้อมูลประเภทตัวอักษร โดยวิธีการวิเคราะห์อนุกรมตัวอักษร (TEXT COMPRESSION BY CHARACTER SEQUENCE ANALYSIS METHOD) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D., ธันวดี สุเนตนันท์,Ph.D. 124 หน้า.
ISBN 974-665-372-5

การบีบอัดข้อมูลประเภทตัวอักษร เป็นหนึ่งในหัวข้อที่ได้รับการสนใจในหมู่นักวิจัยทั้งหลาย เนื่องจากความหลากหลายของการประยุกต์ใช้งาน แต่อย่างไรก็ตาม การค้นหาข้อมูลโดยตรงในข้อมูลที่ถูกบีบอัดอยู่ยังเป็นปัญหาที่สำคัญ และยังไม่สามารถแก้ไขให้เป็นที่หน้าพอใจได้ และนี่ก็คือเหตุผลที่ทำให้การบีบอัดข้อมูล แบบที่ยังสามารถให้มีการค้นหาข้อมูลโดยตรงในชุดข้อมูลที่ถูกบีบอัดอยู่ กลายเป็นหัวข้อที่ถูกนำเสนอ และพัฒนาเพื่อใช้งานจริงโดยเหล่านักค้นคว้าวิจัยทั้งหลาย

จากแนวความคิดของ Udi Manber เกี่ยวกับงานวิจัยที่ชื่อว่า Text Compression Scheme that allows Fast Searching Directly in The Compressed File ทำให้เกิดความสนใจในการพัฒนา และค้นคว้าหาวิธีการใหม่ๆ ที่จะทำให้การบีบอัดข้อมูลเกิดการประหยัดเนื้อที่ได้มากกว่า30% และในขณะเดียวกันก็ยังสามารถทำการค้นหาข้อมูลในชุดข้อมูลที่ถูกบีบอัดอยู่ได้โดยตรง ซึ่งวิธีการที่คิดค้นพัฒนาขึ้นมาใหม่นี้ชื่อว่า การบีบอัดข้อมูลประเภทตัวอักษร โดยวิธีการวิเคราะห์อนุกรมตัวอักษร (CHARACTER SEQUENCE ANALYSIS METHOD – CSAM) วิธีการบีบอัดข้อมูลแบบCSAM ได้ประยุกต์ใช้หลัก Pattern Substitution ซึ่งเป็นหลักการเดียวกันกับที่ใช้ในงานของ Udi แต่วิธีการแบบCSAMนี้ จะทำการวิเคราะห์อนุกรมของตัวอักษรอย่างละเอียดเป็นสำคัญ CSAMสามารถบีบอัดข้อมูลได้มากกว่า30%โดยเฉลี่ย และในขณะเดียวกัน ก็ยังคำรงไว้ซึ่งความสามารถในการค้นหาข้อมูลในชุดข้อมูลที่ถูกบีบอัดอยู่ได้ ความเร็วในการบีบอัดข้อมูลอาจจะยังไม่เหมาะสมที่จะนำไปประยุกต์ใช้งานทั่วๆไปในตอนนี้ แต่อย่างไรก็ตามCSAMสามารถใช้ได้ดีกับข้อมูลที่ไม่มีการเปลี่ยนแปลงบ่อยครั้งนัก แต่ต้องการการบีบอัดเพื่อการประหยัดเนื้อที่ และมีการเรียกใช้บ่อยๆ

ในวิทยานิพนธ์ฉบับนี้ ได้นำเสนอวิธีการบีบอัดข้อมูลแบบCSAM โปรแกรมต้นแบบได้ถูกพัฒนาขึ้น และทำการทดสอบ ประสิทธิภาพและผลการทดลองถูกวิเคราะห์และอภิปรายผล นอกจากนี้ยังได้ให้ข้อเสนอแนะในสิ่งที่ควรจะพัฒนาต่อเพื่อเป็นการเพิ่มประสิทธิภาพของโปรแกรม