TEXT COMPRESSION BY SORTING TRANSFORMATION

SOMPHONG LERWONGRAT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

(COMPUTER SCIENCE)

IN

FACULTY OF GRADUATE STUDIES

MAHIDOL UNIVERSITY

1997

ชื่อวิทยานิพนธ์ การบีบอัดข้อมูลประเภท Text โดยการแปลงข้อมูลด้วยวิธีการจัดเรียงลำดับตัวอักษร

ผู้วิจัย สมพงษ์ เลอวงศ์รัตน์

ปริญญา วิทยาศาสตรมหาบัณฑิต (คอมพิวเตอร์)

คณะกรรมการควบคุมวิทยานิพนธ์

ดำรัส วงศ์สว่าง Ph.D.

ศุภชัย ตั้งวงศ์ศานต์ Ph.D.

วันที่สำเร็จการศึกษา 15 พฤษภาคม พ.ศ. 2540

บทคัดย่อ

ปัจจุบันรูปแบบที่มักจะพบในการบีบอัดข้อมูลที่เป็นตัวอักษรคือการ Modeling และการ Coding ซึ่งอาจจะถูกแทนที่ด้วยรูปแบบใหม่คือการ Transform, Modeling และ Codingแต่ในปัจจุบันได้มีการนำแนวความคิดดังกล่าวไปประยุกต์ใช้กับข้อมูลที่เป็นตัวอักษรเพื่อจะให้ได้ข้อมูลรูปแบบที่เหมาะสมกับการบีบอัดข้อมูลซึ่งเรามักจะพบในแนวความคิดนี้ในการบีบอัดข้อมูลที่เป็นรูปภาพเท่านั้นวิทยานิพนธ์ฉบับนี้จะกล่าวถึงวิธีการ Transfrom ข้อมูลที่เป็นตัวอักษรวิธีหนึ่งที่เรียกว่า Block-sorting Algorithm ซึ่งสามารถจะให้ Compression ratio ที่เทียบได้กับ Algorithm ที่ใช้อยู่ในปัจจุบันในกระบวนการ Transform นั้นจะทำโดยการจัดเรียงข้อมูลนำเข้าเสียใหม่โดยการทำ Permutation เช่น Rotation และการ Sorting ซึ่งผลลัพธ์ที่ได้นั้นจะทำให้ข้อมูลอยู่ในรูปที่สามารถบีบอัดได้ดีมากยิ่งขึ้นในแนวคิดเริ่มต้นได้มีการนำ Move-to-front Algorithm และ Huffman Algorithm มาทำการบีบอัดข้อมูลที่ได้จากการกระบวนการ Transform ซึ่งนอกจากจะทำให้ Compression ratio ดีขึ้นแล้วยังให้ความเร็วที่ดีอีกด้วย

วิทยานิพนธ์ฉบับนี้จะทำการศึกษาเกี่ยวกับแนวคิดของ Block-sorting Algorithm โดยละเอียดรวมทั้งผลกระทบของ Block size และการเปลี่ยนแปลงของ Entropy ใน Order ต่างๆในส่วนสุดท้ายจะทำการค้นหา Compressor ที่เหมาะสมกับข้อมูลที่ถูก Transform โดยแนวคิดนี้เพื่อให้การบีบอัดข้อมูลมีประสิทธิภาพดียิ่งขึ้นทั้งด้านอัตราการบีบอัดข้อมูลและความเร็วในการบีบอัดข้อมูล

Thesis Title Text Compression by Sorting Transformation

Name Somphong Lerwongrat

Degree Master of Science (Computer Science)

Thesis Supervisory Committee

Damras Wongsawang, Ph.D.

Supachai Tangwongsan, Ph.D.

Date of Graduation 15 May B.E. 2540 (1997)

ABSTRACT

The traditional compression paradigm; "Modeling and Coding" may be alternatively replaced by a new paradigm; "Transformation, Modeling and Coding". The recent works described the Block-sorting Transformation and lossless compression algorithms that give the compression ratio comparable to those algorithms currently in use. The transformation processes the raw input into a sequence of less disorder form to obtain more compressible form. It is already known that this is a context-based compressor of unbounded order. The originating of new paradigm starts to restructure contexts by sorting phase then processes the permuted text with the Move-to-front and finally the statistical compressor is applied. This technique not only gives good speed but also an excellent compression ratio.

This research studied this new technique in detail. The Block-sorting transformation was explored and studied in many aspects, the effects of those changed context and their entropy, several different orders context modeling. Finally various types of lossless compression were applied to improve the compression performance eventually yielding a compressor which was the best of this type. It was shown that the Block-sorting technique is comparable to other compressors in terms of compression ratio and speed.