ENTROPY AND COMPRESSION OF THAI TEXT

PENSRI WANGCHAROEN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE (COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
1998

ISBN 974-589-368-4

3837391 SCCS/M : MAJOR : Computer Science : M.Sc. (Computer Science)
KEY WORD : TEXT COMPRESSION/ INFORMATION THEORY
PENSRI WANGCHAROEN : ENTROPY AND COMPRESSION OF THAI TEXT. THESIS ADVISOR : DAMRAS WONGSAWANG Ph.D., SUPACHAI TANGWONGSAN Ph.D. 64 P. ISBN 974-589-368-4

Text data compression is the process to produce a smaller representation from which the original or its approximation can be recovered at a later time. There are many text data compression algorithms currently in use. The most well-known are Huffman Coding, Arithmetic Coding and Lempel-Ziv methods. All of them can be used to compress any text file in any language. The compression ratio obtained from these algorithms is approximately 50 % depending on text types. However the compression ratio can be improved as long as it does not go beyond its entropy, the compression boundary. An approach for text data compression performance improvement is to take advantage of the language characteristics.

This research will emphasize the study and improvement of Thai text data compression. Some characteristics of text and the most frequent Thai words will be looked for from groups of Thai text samples. Then they will be coded to form a Predefined-Thai Dictionary to be used in a new LZW data compression called PTD-LZW. For better performance, the Flush technique will also be brought to use with PTD-LZW. These ideas can be applied not only to Thai text but also to texts of any language.

The study showed that the entropy of the sample Thai texts is approximately 3 bits per character, and the compression ratio of original LZW is 50% on average, but PTD-LZW gives approximately 1-3% compression rate improvement. Finally if we use PTD-LZW with flush technique, approximately 3-5% improvement is obtained.

3837391 SCCS/M : สาขาวิชา : วิทยาการคอมพิวเตอร์ ; วท.ม. (วิทยาการคอมพิวเตอร์)
ศัพท์สำคัญ : การบีบอัดข้อมูลประเภทข้อความ

เพ็ญศรี วังเจริญ : เอนโทรปี และการบีบอัดข้อมูลประเภท Text ภาษาไทย (Entropy and Compression of Thai Text) คณะกรรมการควบคุมวิทยานิพนธ์ : ดำรัส วงศ์สว่าง, Ph.D. , ศุภชัย ตั้งวงศ์ศานต์, Ph.D., 64 หน้า. ISBN 974-589-368-4

การบีบอัดข้อมูลประเภทตัวอักษร จะเป็นการบีบอัดข้อมูลที่ไม่ยอมให้เกิดข้อผิดพลาด เมื่อมีการขยายข้อมูลกลับคืนมา มีวิธีการต่างๆ หลายวิธีที่ใช้อยู่ในปัจจุบัน อาทิเช่น Huffman Coding, Arithmetic Coding, LZ ตระกูลต่างๆ วิธีการดังกล่าวเป็นวิธีการบีบอัดข้อมูลที่สามารถใช้ได้กับทุกๆ ภาษา ไม่ว่าจะเป็นภาษาอังกฤษ, ภาษาไทย หรือภาษาอื่นๆ ซึ่งจะให้อัตราการบีบอัดข้อมูลโดยเฉลี่ยประมาณ 50 % แล้วแต่ชนิดของข้อมูล อย่างไรก็ตามการบีบอัดข้อมูลก็อาจจะสามารถทำได้มากขึ้น ตราบใดก็ตามที่ยังไม่เกิน Entropy อันเป็นขอบเขตของการบีบอัดข้อมูล แนวทางหนึ่งในความพยายามที่จะปรับปรุงอัตราการบีบอัดข้อมูลให้มีค่าสูงขึ้น สามารถทำได้โดยการพยายามนำข้อได้เปรียบของลักษณะเฉพาะของภาษานั้นๆ มาใช้ โดยการนำเอาลักษณะเฉพาะนั้นเข้าไปมีส่วนช่วยในการบีบอัดข้อมูลจะทำให้อัตราการบีบอัดข้อมูลดีขึ้น

งานวิจัยฉบับนี้ จะศึกษาถึงการปรับปรุงวิธีการบีบอัดข้อมูลภาษาไทย โดยดึงเอาลักษณะเฉพาะของภาษาไทยออกมา ซึ่งลักษณะเฉพาะของภาษาไทยที่นำมาใช้ในที่นี้คือ การหาคำต่างๆ ในภาษาไทยที่มีความถี่ในการใช้งานค่อนข้างสูง จากข้อความประเภทต่างๆ และนำคำต่างๆ เหล่านั้นมาสร้างเป็น Pre-defined Dictionary เพื่อนำไปใช้กับวิธีการบีบอัดข้อมูลแบบ LZW ซึ่งจะเรียกว่า PTD-LZW ผสมกับเทคนิคในการล้างข้อมูลใน Dictionary เมื่อ Dictionary ที่ใช้อยู่เต็ม หรือที่เรียกว่า Flush Technique เพื่อให้ได้มาซึ่งวิธีการบีบอัดข้อมูลภาษาไทยที่มีอัตราการบีบอัดข้อมูลที่ดีขึ้น ซึ่งแนวคิดต่างๆ ข้างต้นนี้สามารถนำไปประยุกต์ใช้กับข้อมูลตัวอักษรภาษาอื่นๆ ได้เช่นกัน

จากการวิจัยได้ข้อสรุปดังนี้คือ ผลการวัด Entropy ของกลุ่มตัวอย่างข้อความภาษาไทยจะได้ประมาณ 3 bits ต่อตัวอักษร ผลการบีบอัดข้อมูลแบบ LZW โดยทั่วๆ ไปจะได้อัตราการบีบอัดประมาณ 50 % ในขณะที่เมื่อใช้ PTD-LZW จะสามารถบีบอัดข้อมูลเพิ่มขึ้นได้อีก 1-3% สำหรับแฟ้มข้อมูลขนาดเล็ก และเมื่อมีการนำเอาเทคนิคในการล้าง Dictionary มาใช้ร่วมกับ PTD-LZW ในการบีบอัดแฟ้มข้อมูลขนาดใหญ่ จะสามารถเพิ่มการบีบอัดข้อมูลได้ประมาณ 3-5%