RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

CHAIVATNA SUMETPHONG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)
FACULTY OF GRADUATE STUDIES
MAHIDOL UNIVERSITY
2011

RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

CHAIVATNA SUMETPHONG 4838800 ITCS/D

Ph.D. (COMPUTER SCIENCE)

THESIS ADVISORY COMMITTEE: SUPACHAI TANGWONGSAN, Ph.D., DAMRAS WONGSAWANG, Ph.D., SUKANYA PHONGSUPHAP, Ph.D.

ABSTRACT

In this research, we focus on solving the problem of recognizing broken characters that are found abundantly in historical documents. Most OCR systems are designed to correctly recognize finely printed documents in both Thai and English scripts. When tested with degraded documents, the accuracy of these systems drops drastically. One of the most important problems that decreases the accuracy of OCR systems is the broken characters found in the document image. Although humans can read broken characters without difficulty, it is far more complicated for computers to recognize them.

We propose a strategy that aims to find the optimal set-partition of the pieces of the broken characters based on a probability functions model, thus yielding a sequence of characters. To counter the inevitability of recognition errors and the need to group the characters as words, we employ an N-grams Graph generated from a dictionary. To demonstrate the generality of the method, we performed experiments on a Thai historical document, an English historical document and a Thai fax document. The results obtained are very promising and this research could be extremely useful for researchers engaged in recognizing historical documents degraded due to the presence of broken characters.

To find optimal solutions for set-partition problems, the obvious method would be to adopt the brute-force enumerative approach. However, this is drastically slow when the number of elements in the set increases. We propose a Partition-Growing Algorithm that is capable of generating optimal set-partitions of the broken characters in a smaller timeframe by using heuristics based on characteristics of probability functions. The algorithm is further extended towards solving other problem domains that require finding an optimal set-partition. We demonstrate the applicability of the algorithm to other set partitioning problems (SPP) that might be linear or non-linear in nature by modeling these problems based on probability functions. The experiments performed show that the algorithm has a potential to be adapted to general problems that require partitioning a set optimally.

KEY WORDS : HISTORICAL DOCUMENTS / BROKEN CHARACTERS / OPTIMAL SET-PARTITIONING / N-GRAMS GRAPH / PARTITION-GROWING ALGORITHM

83 pages

การรู้จำตัวอักษรขาดในเอกสารทางประวัติศาสตร์และการแก้ปัญหาการแบ่งเซททางคณิตศาสตร์

RECOGNIZING BROKEN CHARACTERS IN HISTORICAL DOCUMENTS AND SOLVING OTHER SET-PARTITIONING PROBLEMS

ชัยวัฒน์ สุเมธพงศ์ 4838800 ITCS/D

ปร.ด. (วิทยาศาสตร์คอมพิวเตอร์)

คณะกรรมการที่ปรึกษาวิทยานิพนธ์: สุภชัย ตั้งวงศ์ศานต์, Ph.D.; ดำรัส วงศ์สว่าง, Ph.D.; สุกัญญา พงษ์สุภาพ, Ph.D.

บทคัดย่อ

ในงานวิจัยนี้ เราเน้นการแก้ปัญหาของการรู้จำตัวอักษรขาด (Broken Characters Recognition) ที่ปรากฏเป็นจำนวนมากในเอกสารทางประวัติศาสตร์ ปัจจุบันระบบการรู้จำตัวอักษร (OCR) ส่วนใหญ่จะถูกออกแบบให้จัดการกับเอกสารสิ่งพิมพ์ที่มีตัวอักษรทั้งไทยและอังกฤษได้ผลเป็นอย่างดี อย่างไรก็ตามเมื่อนำมาทดสอบกับเอกสารที่มีปัญหา เช่น เก่า เหลือง ตัวอักษรขาด ประสิทธ์ผลในความแม่นยำของระบบจะลดลงอย่างมากโดย เฉพาะกรณีของตัวอักษรขาดและไม่สมบูรณ์ แม้มนุษย์เราจะสามารถอ่านตัวหนังสือที่มีตัวอักษรขาดได้ไม่ยากนัก แต่สำหรับคอมพิวเตอร์แล้วก็เป็นเรื่องที่ค่อนข้างจะยุ่งยากด้วยความไม่สมบูรณ์ของตัวอักษร

ในงานวิจัยนี้ เราจะได้นำเสนอวิธีการแบ่งเซทเหมาะที่สุด (Optimal Set-Partition) ในการจัดการกับตัวอักษรขาดที่เกิดขึ้นในเอกสารโดยทั่วไป เพื่อให้ได้ผลลัพธ์ที่ดีด้วยแบบของฟังก์ชันความน่าจะเป็น (Probability Functions) นอกจากนี้เพื่อให้ได้ผลลัพธ์ที่ดีขึ้นในเรื่องความถูกต้องแม่นยำ เราจะตรวจคำศัพท์ของผลลัพธ์กับคำที่มีในพจนานุกรมด้วยการสร้างเอ็นแกรมกราฟ (N-grams Graph) เมื่อได้ทำการทดสอบประสิทธ์ผลของระบบงานด้วยเอกสารทางประวัติศาสตร์ฉบับภาษาไทย ฉบับภาษาอังกฤษ และฉบับโทรสารภาษาไทย พบว่าความแม่นยำอยู่ในเกณฑ์ที่น่าพอใจ จึ่งมีความเชื่อว่าผลงานนี้จะเป็นประโยชน์ต่องานการเก็บรักษาเอกสารทางประวัติศาสตร์ ของไทยและของต่างประเทศที่มีตัวอักษรขาดปรากฏโดยทั่วไป

การศึกษาหาผลลัพธ์ที่เหมาะที่สุดสำหรับปัญหาของการแบ่งเซทนั้น วิธีการที่ง่ายและตรงที่สุดเห็นจะเป็นการแยกประเมินทีละส่วนจากหัวจดท้ายจนหมดรายการ แต่วิธีนี้ต้องใช้เวลามากและไม่เหมาะสมในกรณีที่ขนาดเซทเพิ่มมากขึ้นซึ่งจะทำให้ได้คำตอบที่ช้ามากจนไม่สามารถนำไปใช้ได้ในทางปฏิบัติ ในงานวิจัยนี้เราจึงได้นำเสนออัลกอริทัมใหม่ด้วยการขยายการแบ่งส่วน (Partition-Growing) ที่สามารถแบ่งเซทได้อย่างดีที่สุดของตัวอักษรขาดภายในช่วงเวลาอันสั้น โดยอาศัยคุณลักษณะเฉพาะของฟังก์ชันความน่าจะเป็นเป็นตัวหลักในการแก้ปัญหา นอกจากนี้เรายังนำอัลกอริทัมนี้ไปแก้ปัญหาอื่นๆที่เกี่ยวกับการแบ่งเซท (SPP) ทั้งที่เป็นลักษณะเชิงเส้นและไม่เชิงเส้น ด้วยแบบจำลองตัวระบบของฟังก์ชันความน่าจะเป็น ผลการทดลองพบว่าอัลกอริทัมนี้ทำงานอย่างได้ผลสำหรับปัญหาของการแบ่งเซทที่ดีที่สุดจริง

83 หน้า