

ระบบคั่นคีนสารสนเทศแบบจัดลำดับและแบบคั่นคีนย้อนกลับบนโครงสร้างแถวลำดับแพ้ด

นายมานพ จงเจริญใจ



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2541

ISBN 974-331-780-5

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

AN INFORMATION RETRIEVAL SYSTEM USING RANKING
AND RELEVANCE FEEDBACK ON THE PAT ARRAY

Mr. Manop Jongcharoenjai

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science

Department of Computer Engineering

Graduate School

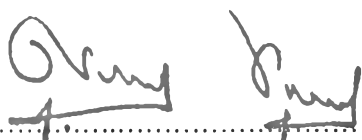
Chulalongkorn University

Academic Year 1998

ISBN 974-331-780-5

หัวข้อวิทยานิพนธ์ ระบบคั่นคืนสารสนเทศแบบจัดลำดับและแบบคั่นคืนย้อนกลับบน
โครงสร้างแถวลำดับแพ็ค
โดย นายมานพ จงเจริญใจ
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา อาจารย์ จารุมาตร ปิ่นทอง
อาจารย์ที่ปรึกษาร่วม อาจารย์ ดร. ธาราทิพย์ สุวรรณศาสตร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้บัณฑิตวิทยาลัยฉบับนี้เป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต



..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ นายแพทย์ สุภวัฒน์ ชุตินวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. สมชาย ประสิทธิ์จตุระกุล)

..... อาจารย์ที่ปรึกษา
(อาจารย์ จารุมาตร ปิ่นทอง)

..... อาจารย์ที่ปรึกษาร่วม
(อาจารย์ ดร. ธาราทิพย์ สุวรรณศาสตร์)

..... กรรมการ
(รองศาสตราจารย์ ดร. วันชัย ธีรไพบูลย์)

..... กรรมการ
(อาจารย์ ดร. บุญเสริม กิจศิริกุล)

มานพ จงเจริญใจ : ระบบค้นคืนสารสนเทศแบบจัดลำดับและแบบค้นคืนย้อนกลับบน โครงสร้าง
แถวลำดับแพ้ (AN INFORMATION RETRIEVAL SYSTEM USING RANKING AND
RELEVANCE FEEDBACK ON THE PAI ARRAY) อ.ที่ปรึกษา : อ.จารุมาศ ปิ่นทอง,
อ.ที่ปรึกษาร่วม : อ. ดร. ธราทิพย์ สุวรรณศาสตร์ ; 54 หน้า. ISBN 974-331-780-5.

วิทยานิพนธ์นี้เสนอการพัฒนาระบบค้นคืนสารสนเทศในรูปแบบเอกสารที่เป็นข้อความด้วยวิธีค้น
คืนแบบจัดลำดับและแบบค้นคืนย้อนกลับโดยใช้แถวลำดับแพ้เป็นดัชนีเพื่อใช้ในการค้นคืน แถวลำดับแพ้
เป็นโครงสร้างที่เหมาะสมกับข้อความภาษาไทยที่การแบ่งคำยังไม่ถูกต้องสมบูรณ์ แถวลำดับแพ้จัดเก็บดัชนีใน
รูปของสายอักขระแบบกึ่งอนันต์ ที่เรียกว่าซิสตริง

การพัฒนาโปรแกรมค้นคืนแบ่งออกเป็น 3 ส่วนคือ ส่วนของการสร้างดัชนีของนำหน้าคำ ส่วน
ของการจัดลำดับผลการค้นคืน และส่วนของการค้นคืนย้อนกลับ สำหรับส่วนของการสร้างดัชนีของนำหน้า
คำ จะเก็บค่าตัวชี้ตำแหน่งซิสตริงที่ไม่ซ้ำกันและค่าความถี่ของแต่ละซิสตริงในเอกสารทั้งหมดไว้ในแถวลำดับ
แพ้เพื่อลดขั้นตอนการประมวลผลในช่วงค้นคืน การค้นคืนจะเปรียบเทียบคิวิที่ผู้ใช้ป้อน กับค่าที่ได้จาก
ซิสตริงซึ่งเป็นค่าที่ถูกต้องตามหลักภาษาศาสตร์สำหรับส่วนของการจัดลำดับผลการค้นคืนนั้น เมื่อได้ผลลัพธ์
การค้นคืน จะนำผลลัพธ์นั้นมาคำนวณหาค่าตามสูตรคำนวณนำหน้าคำ เพื่อให้ได้ค่านำหน้าคำรวมของแต่ละ
เอกสาร แล้วนำผลนำหน้าคำที่ได้มาทำการจัดลำดับตามค่านำหน้าคำ และส่วนของการค้นคืนย้อนกลับจะนำ
เอกสารที่ผู้ใช้แสดงว่าเอกสารนั้นตรงตามต้องการมาใช้สร้างคำใหม่ เพื่อให้ผู้ใช้นำคำใหม่นี้ไปใช้ค้นคืนซ้ำอีก
ครั้ง เพื่อให้ผลการค้นคืนใหม่มีค่าความถูกต้องสูงขึ้นกว่าเดิม

ในการวิจัยนี้ได้เลือกสูตรคำนวณนำหน้าคำมาทั้งหมด 5 สูตร และจากผลการทดลองการค้นคืน
แบบจัดลำดับโดยใช้สูตรคำนวณนำหน้าคำ 5 สูตร พบว่ามี 2 สูตรที่ให้ผลเฉลี่ยค่าความถูกต้องสูงสุดคือ สูตร
คำนวณนำหน้าคำที่ประกอบไปด้วยค่าความถี่ของคำที่ปรากฏในเอกสาร และสูตรคำนวณนำหน้าคำที่
ประกอบไปด้วยค่าความถี่ของคำที่ปรากฏในเอกสารคูณกับค่าความถี่เอกสารแบบผกผัน ส่วนผลการทดลอง
การค้นคืนแบบค้นคืนย้อนกลับ พบว่าการเลือกใช้คำที่มีค่าความถี่อยู่ในช่วงขีดจำกัดที่เหมาะสม ช่วยให้
ระบบเสนอคิวิใหม่ที่ช่วยให้ผลการค้นคืนมีผลเฉลี่ยค่าความถูกต้องสูงขึ้นกว่าเดิมได้

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2541

ลายมือชื่อนิสิต มานพ จงเจริญใจ
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม ธารทิพย์ สุวรรณศาสตร์

C818343 MAJOR COMPUTER SCIENCE
KEY WORD: RANKING / RELEVANCE FEEDBACK / PAT ARRAY / INFORMATION RETRIEVAL

MANOP JONGCHAROENJAI : AN INFORMATION RETRIEVAL SYSTEM USING
RANKING AND RELEVANCE FEEDBACK ON THE PAT ARRAY. THESIS ADVISOR :
CHARUMATR PINTHONG. THESIS COADVISOR: TARATIP SUWANNASART, Ph.D.
54 pp. ISBN 974-331-780-5.

This thesis presents a development of information retrieval system using ranking and relevance feedback on PAT arrays which are used as index for retrieval. A PAT array is a structure that fits for Thai text which Thai text which is not completely segmented. PAT arrays store index in semi-infinite strings (sistrings).

The development consists of 3 subsystems : the index term weights creation subsystem, the ranking subsystem, and the relevance feedback subsystem. The index term weights creation subsystem stores unique sistrings and frequency of each sistring from all documents in a PAT array in order to decrease retrieval time. The retrieval compares between user's query and words from sistrings which follow the linguistics rules. After getting the retrieval results, the ranking subsystem calculates term weights for each document, and sort the documents in descending order using the term weights as a key. The relevance feedback subsystem allows the user to select relevant documents, and enter new query in order to improve the results.

This research experienced five term weight formulas. The experiments showed that there are two formulas that give the best results. One of the two formulas consists of term frequency in its formula and the other consists of term frequency multiplied by inverse document frequency. For the relevance feedback, the experiments showed that choosing the appropriate threshold help the system promote new queries that help improve better results.

ภาควิชา..... วิศวกรรมคอมพิวเตอร์.....
สาขาวิชา..... วิทยาศาสตร์คอมพิวเตอร์.....
ปีการศึกษา..... 2541.....

ลายมือชื่อนิสิต..... มานพ จงจรจ.....
ลายมือชื่ออาจารย์ที่ปรึกษา.....
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม.....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างดียิ่งของอาจารย์จารย์มาตร ปิ่นทอง อาจารย์ที่ปรึกษาวิทยานิพนธ์และอาจารย์ ดร.ธราทิพย์ สุวรรณศาสตร์ อาจารย์ที่ปรึกษา วิทยานิพนธ์ร่วม ซึ่งท่านได้ให้คำแนะนำ ข้อคิดเห็นต่างๆ ในการวิจัยด้วยดีตลอดมา และขอ ขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สมชาย ประสิทธิ์จตุระกุล ที่ได้ให้คำแนะนำที่เป็นประโยชน์ใน เรื่องขั้นตอนและวิธีที่ใช้ในค้นคว้าสารสนเทศที่นำมาใช้ในงานวิจัย ขอขอบพระคุณห้องปฏิบัติการ วิศวกรรมระบบสารสนเทศ (Information Systems Engineering Laboratory) ที่เอื้อเฟื้ออุปกรณ์ใน การทำงานวิจัย รวมทั้งพี่ ๆ และเพื่อน ๆ ที่ได้ให้คำปรึกษาและความช่วยเหลือในด้านต่าง ๆ ซึ่งทำ ให้การทำงานวิจัยเป็นไปอย่างราบรื่น

ทำยนี้ ผู้วิจัยใคร่ขอกราบขอบพระคุณบิดา-มารดา ซึ่งสนับสนุนในด้านการเงินและให้กำลังใจ แก่ผู้วิจัยเสมอมาจนสำเร็จการศึกษา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญ	ช
สารบัญตาราง	ฅ
สารบัญภาพ	ญ
บทที่	
1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 การวิจัยที่เกี่ยวข้อง	2
1.3 วัตถุประสงค์ของการวิจัย	3
1.4 ขอบเขตของการวิจัย	3
1.5 ขั้นตอนและวิธีดำเนินการวิจัย	3
1.6 ประโยชน์ที่คาดว่าจะได้รับ	4
2 แนวคิดและทฤษฎี	5
2.1 แนวคิดและทฤษฎี	5
2.2 โครงสร้างข้อมูลที่ใช้ในระบบค้นคืนสารสนเทศ	6
2.3 โครงสร้างต้นไม้แฟ้ม	7
2.4 สายอักขระแบบเซมิอินไฟไนต์หรือซิสตริง	10
2.5 โครงสร้างแถวลำดับแฟ้ม	10
2.6 การวัดประสิทธิภาพของระบบค้นคืนสารสนเทศ	11
2.7 ความสำคัญของคำ	12
2.8 การให้น้ำหนักคำ	13
2.9 ขั้นตอนการหาควิรีที่เหมาะสมที่สุด	18

2.10 การคืนคืนย้อนกลับ	19
3 การวิเคราะห์และการออกแบบ	23
3.1 การพัฒนาส่วนจัดเก็บดัชนีเพื่อเก็บค่าน้ำหนักคำ	23
3.2 การคืนคืนและจัดเก็บผลการคืนคืน	26
3.3 ขั้นตอนการคืนคืนย้อนกลับ	30
4 การพัฒนาโปรแกรม	32
4.1 โปรแกรมจัดเก็บดัชนีเพื่อเก็บค่าน้ำหนักคำ	32
4.2 โปรแกรมคืนคืนเอกสาร	33
4.3 การเปรียบเทียบกับระบบเดิม	40
5 การทดสอบโปรแกรม	41
5.1 รูปแบบเอกสารที่ใช้	41
5.2 การวัดประสิทธิภาพระบบ	42
5.3 ผลการทดสอบการจัดลำดับ	44
5.4 ผลการทดสอบการคืนคืนย้อนกลับ	46
6 บทสรุปและข้อเสนอแนะ	51
6.1 บทสรุป	51
6.2 ข้อเสนอแนะ	51
รายการอ้างอิง	52
ประวัติผู้เขียน	54

สารบัญตาราง

	หน้า
ตารางที่ 2.1 แสดงผลการคั่นคืน	12
ตารางที่ 4.1 แสดงตัวอย่างผลการคั่นคืนสำหรับสูตรนำหน้าค้ำแบบที่ 1	36
ตารางที่ 4.2 แสดงตัวอย่างผลการคั่นคืนสำหรับสูตรนำหน้าค้ำแบบที่ 2	37
ตารางที่ 4.3 แสดงตัวอย่างผลการคั่นคืนสำหรับสูตรนำหน้าค้ำแบบที่ 3	37
ตารางที่ 4.4 แสดงตัวอย่างผลการคั่นคืนสำหรับสูตรนำหน้าค้ำแบบที่ 4	37
ตารางที่ 4.5 แสดงตัวอย่างผลการคั่นคืนสำหรับสูตรนำหน้าค้ำแบบที่ 5	38
ตารางที่ 5.1 แสดงผลการคั่นคืน	42-43
ตารางที่ 5.2 แสดงค่าความถูกต้องและค่าเรียกคืน	44
ตารางที่ 5.3 แสดงผลค่าเฉลี่ยความถูกต้องของฐานข้อมูลทศด้อย	46
ตารางที่ 5.4 แสดงผลค่าเฉลี่ยความถูกต้องของฐานข้อมูลข่าว	46
ตารางที่ 5.5 แสดงค่าเฉลี่ยความถูกต้องของการคั่นคืนย้อนกลับของ Q1	47
ตารางที่ 5.6 แสดงค่าเฉลี่ยความถูกต้องของการคั่นคืนย้อนกลับของ Q2	48
ตารางที่ 5.7 แสดงค่าเฉลี่ยความถูกต้องของการคั่นคืนย้อนกลับของ Q3	48
ตารางที่ 5.8 แสดงค่าเฉลี่ยความถูกต้องของการคั่นคืนย้อนกลับของ Q4	49
ตารางที่ 5.9 แสดงค่าเฉลี่ยความถูกต้องของการคั่นคืนย้อนกลับจากทุกคิวรี	49

สารบัญภาพ

	หน้า
รูปที่ 2.1 แสดงองค์ประกอบของระบบคั่นคืนข้อมูล	6
รูปที่ 2.2 แสดงต้นไม้ตัดสินใจ	8
รูปที่ 2.3 แสดงทรี	8
รูปที่ 2.4 แสดงต้นไม้แพ้	9
รูปที่ 2.5 กราฟแสดงความสัมพันธ์ระหว่างความสำคัญและความถี่ของคำ	22
รูปที่ 2.6 แสดงผลการปรับค่าขีดจำกัด	22
รูปที่ 3.1 แสดงโครงสร้างแถวลำดับแพ้	24
รูปที่ 3.2 แสดงโครงสร้างแถวลำดับแพ้ใหม่	25
รูปที่ 3.3 แสดงการคั่นคืนบนแถวลำดับแพ้	26
รูปที่ 3.4 แสดงการแฮชเพื่อหาตำแหน่งเก็บข้อมูล	27
รูปที่ 3.5 แสดงขั้นตอนการคั่นคืนแบบจัดลำดับ	27
รูปที่ 3.6 แสดงขั้นตอนการเรียงลำดับข้อมูลด้วยฮีฟ	28
รูปที่ 3.7 แสดงโครงสร้างฮีฟ	29
รูปที่ 3.8 แสดงขั้นตอนการทำงานของระบบ	31
รูปที่ 4.1 แสดงโปรแกรมสร้างดัชนีเพื่อเก็บค่าน้ำหนักคำ	32
รูปที่ 4.2 แสดงโปรแกรมส่วนรับคิวรี	33
รูปที่ 4.3 แสดงโปรแกรมส่วนแสดงผลการคั่นคืนแบบจัดลำดับ	34
รูปที่ 4.4 แสดงโปรแกรมส่วนเลือกเอกสารที่ตรงตามต้องการ	38
รูปที่ 4.5 แสดงโปรแกรมส่วนแสดงคำใหม่ที่ได้จากการคั่นคืนย้อนกลับ	39
รูปที่ 4.6 แสดงโปรแกรมส่วนแสดงผลการคั่นคืนใหม่ที่ได้จากการคั่นคืนย้อนกลับ..	40