

การเก็บและการค้นคืนสารสนเทศโดยใช้แนวความคิดของแฟ้มข้อมูลผกผัน

นางสาว พรทิพย์ บัวสาม



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2535


ISBN 974-581-197-1

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

018733

๒๕๓๕

INFORMATION STORAGE AND RETRIEVAL USING INVERTED FILE



MISS PORNTIP BUASAM

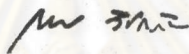
A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science
Department of Computer Engineering
Graduate School
Chulalongkorn University

1992

ISBN 974-581-197-1

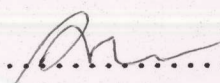
หัวข้อวิทยานิพนธ์ การเก็บและการคั่นคืนสารสนเทศโดยใช้แนวความคิดของแฟ้มข้อมูลผกผัน
โดย นางสาว พรทิพย์ บัวสาม
ภาควิชา วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษา อาจารย์ จารุมาตร ปิ่นทอง

บัณฑิตวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้หัวข้อวิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาโทบัณฑิต




..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร.ถาวร วิษราภิษ)

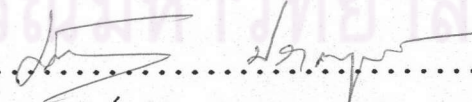
คณะกรรมการสอบวิทยานิพนธ์



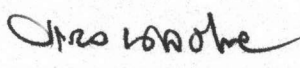
..... ประธานกรรมการ
(รองศาสตราจารย์ เตือน สันตพันธ์ประทุม)



..... อาจารย์ที่ปรึกษา
(อาจารย์ จารุมาตร ปิ่นทอง)



..... กรรมการ
(รองศาสตราจารย์ มัทนา ปราการสมุทร)



..... กรรมการ
(อาจารย์ ดร.สรพงษ์ เต็งอำนวยการ)



พรทิพย์ บัวสาม : การเก็บและการค้นคืนสารสนเทศโดยใช้แนวความคิดของแฟ้มข้อมูลผกผัน
(INFORMATION STORAGE AND RETRIEVAL USING INVERTED FILE)
อาจารย์ที่ปรึกษา : อาจารย์ จารุมাত্র ปิ่นทอง, 77 หน้า ISBN 974-581-197-1

วัตถุประสงค์ของวิทยานิพนธ์นี้ เพื่อศึกษาวิธีการเก็บ และการค้นคืนสารสนเทศที่เป็นข้อความ โดยอาศัยแนวความคิดของแฟ้มข้อมูลผกผันเป็นพื้นฐาน และพัฒนาโปรแกรมคอมพิวเตอร์โดยใช้ภาษาซี การออกแบบการทำงานของโปรแกรม แบ่งออกเป็น 2 ส่วน คือ ส่วนของการสร้างฐานข้อมูล และส่วนของการค้นคืนฐานข้อมูล ส่วนของการสร้างฐานข้อมูล จะทำการอ่านข้อความในแฟ้มข้อมูลข้อความที่ผ่านการแบ่งคำแล้ว โดยที่แฟ้มข้อมูลข้อความสามารถสร้างโดยโปรแกรมเวิร์ดโปรเซสเซอร์ทั่วไป และทำการสร้างแฟ้มข้อมูลดัชนีนารี เพื่อเก็บค่าที่เป็นดัชนี และสร้างแฟ้มข้อมูลผกผัน เพื่อเก็บรายการหมายเลขอ้างอิงตำแหน่งของเอกสารที่สัมพันธ์กับดัชนี และส่วนของการค้นคืนฐานข้อมูล จะเป็นการค้นคืนเอกสารที่ต้องการ โดยในการค้นคืน เอกสารที่ต้องการจะถูกระบุโดยดัชนี ดัชนีจะถูกค้นหาในแฟ้มข้อมูลดัชนีนารี และเอกสารในแฟ้มข้อมูลข้อความ ที่ถูกระบุโดยหมายเลขอ้างอิงที่สัมพันธ์กับดัชนี ก็จะถูกอ่านออกมาจากแฟ้มข้อมูลข้อความ ในการออกแบบแฟ้มข้อมูล จะใช้โครงสร้างข้อมูลแบบบิตรี แทนแฟ้มข้อมูลดัชนีนารี และใช้โครงสร้างข้อมูลแบบลิงคัลิส แทนแฟ้มข้อมูลผกผัน ผลที่ได้จากการทำงานของโปรแกรมนี้ ผู้ใช้สามารถค้นคืนเอกสารที่ต้องการได้สะดวก รวดเร็ว และถูกต้อง

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2534

ลายมือชื่อนิสิต
ลายมือชื่ออาจารย์ที่ปรึกษา
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

C116893 : MAJOR COMPUTER SCIENCE

KEY WORD : INFORMATION RETRIEVAL/INVERTED FILE

PORNTIP BUASAM : INFORMATION STORAGE AND RETRIEVAL USING INVERTED FILE.

THESIS ADVISOR : CHARUMATR PINTHONG, 77 pp. ISBN 974-581-197-1

The purpose of this thesis is to study text information storage and retrieval based on inverted file concept, and to develop computer program using C language compiler. The program consists of two subsystem : the database creation subsystem, and the database retrieval subsystem. Database creation subsystem scans a word separated text file, created by using any word processor, and then creates a dictionary file for storing the allowable indexing terms, and creates an inverted file for storing an associated list of document reference numbers of each index. Database retrieval subsystem retrieves documents from the stored text file. Required documents are identified by an arbitrary index, index requires a dictionary search to find the associated document reference numbers, and the identified documents are selected from the text file. The dictionary file is implemented by using B-tree and the inverted file using linked list. This program can retrieve stored documents conveniently, rapidly, and accurately.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ภาควิชา วิศวกรรมคอมพิวเตอร์

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์

ปีการศึกษา 2534

ลายมือชื่อนิติ *Porntip Buasam*

ลายมือชื่ออาจารย์ที่ปรึกษา *[Signature]*

ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ ได้สำเร็จลุล่วงไปได้ด้วยความช่วยเหลืออย่างยิ่ง ของอาจารย์
จารย์มาตร ปิ่นทอง อาจารย์ที่ปรึกษาวิทยานิพนธ์ ซึ่งท่านได้ให้คำแนะนำและข้อคิดเห็นต่างๆ
ของการวิจัยมาด้วยดีตลอด

ขอขอบคุณเจ้าหน้าที่ บริษัทซัมมิตคอมพิวเตอร์ ที่ได้เอื้อเพื่อให้ใช้โปรแกรมสำหรับ
ตัดคำเอกสารภาษาไทย และคำแนะนำที่เป็นประโยชน์ต่อการวิจัยในครั้งนี้

เนื่องจากทุนการวิจัยครั้งนี้บางส่วนได้รับมาจาก ทุนอุดหนุนการวิจัยของบัณฑิตวิทยาลัย
จึงขอขอบพระคุณบัณฑิตวิทยาลัยมา ณ ที่นี้ด้วย

ขอขอบคุณ พี่ๆ น้องๆ และเพื่อนๆ ที่ช่วยเหลือ และให้กำลังใจแก่ผู้วิจัยตลอดมา
ท้ายนี้ ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งสนับสนุนและให้กำลังใจแก่ผู้วิจัย
เสมอมา จนสำเร็จการศึกษา



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญภาพ	ญ

บทที่

1. บทนำ	1
ความเป็นมาของปัญหา	1
วัตถุประสงค์ของการวิจัย	2
ขอบเขตของการวิจัย	2
ขั้นตอนของการวิจัย	2
ประโยชน์ที่คาดว่าจะได้รับ	2
2. แนวความคิดและทฤษฎี	3
การค้นหาข้อมูล	3
การค้นหาแบบภายใน	4
การค้นหาแบบภายนอก	4
การค้นหาแบบลำดับ	4
การค้นหาเชิงดัชนี	6
การค้นหาแบบทวิภาค	7
การประมวลผลเพิ่มข้อมูลข้อความ	8
โครงสร้างข้อมูล	13
โครงสร้างข้อมูลแบบลิงค์ลิส	13
โครงสร้างข้อมูลแบบทรี	13
โครงสร้างข้อมูลแบบไบนารีทรี	14
โครงสร้างข้อมูลแบบ AVL ทรี	17
โครงสร้างข้อมูลแบบเพจไบนารีทรี	19
ปัญหาในการสร้างเพจทรีจากบนลงล่าง	21
โครงสร้างข้อมูลแบบบีทรี	22
การเพิ่มข้อมูลเข้าไปในบีทรี	24

3.	การออกแบบโครงสร้างของแฟ้มข้อมูลสำหรับระบบ	27
	แฟ้มข้อมูลนำเข้า	27
	แฟ้มข้อมูลที่สร้างขึ้นจากระบบ	29
	แฟ้มข้อมูลรวม	29
	แฟ้มข้อมูลเฉพาะ	29
	แฟ้มข้อมูล filename.DIC	30
	แฟ้มข้อมูล filename.INX	31
4.	การพัฒนาโปรแกรมสำหรับระบบ	34
	ลักษณะของภาษาซี	34
	การกำหนดตัวแปร	36
	ขั้นตอนการทำงานของโปรแกรม	38
	การสร้างฐานข้อมูล	38
	ฟังก์ชัน create	38
	ฟังก์ชัน insert และ ฟังก์ชัน ins	40
	ฟังก์ชัน binsearch	44
	การค้นหาข้อมูล	47
	ฟังก์ชัน search	47
	การเพิ่มข้อมูล	48
	ฟังก์ชัน appendtext	48
	การปรับปรุงแฟ้มข้อมูลสารบัญ	49
5.	ผลการทดสอบโปรแกรม	50
6.	สรุปผลการวิจัยและข้อเสนอแนะ	55
	สรุปผลการวิจัย	55
	ปัญหาและข้อเสนอแนะ	56
	เอกสารอ้างอิง	57
	ภาคผนวก	59
	ประวัติผู้เขียน	77

สารบัญภาพ

หน้า

รูปที่ 2.1	เพิ่มข้อมูลดัชนี แบบสปาร์สอินเด็กซ์ และฟูลอินเด็กซ์	7
รูปที่ 2.2	รายการของค่า ดัชนีข้อความ และเพิ่มข้อมูลข้อความ	10
รูปที่ 2.3	รายการของค่าที่เรียงลำดับแล้ว	11
รูปที่ 2.4	เพิ่มข้อมูลดัชนีนารี และเพิ่มข้อมูลผกผัน	12
รูปที่ 2.5	แสดงโครงสร้างข้อมูลแบบลิงค์ลิสต์	13
รูปที่ 2.6	แสดงโครงสร้างข้อมูลแบบทรี	13
รูปที่ 2.7	แสดงโครงสร้างข้อมูลแบบไบนารีเสิร์ชทรี	14
รูปที่ 2.8	โครงสร้างข้อมูลแบบไบนารีทรีที่มีการเก็บในหน่วยความจำ	15
รูปที่ 2.9	ไบนารีเสิร์ชทรีหลังจากเพิ่มคีย์ LV	15
รูปที่ 2.10	ไบนารีเสิร์ชทรีหลังจากเพิ่มคีย์ใหม่ 8 คีย์	16
รูปที่ 2.11	แสดง AVL Tree ที่มีความสูงสมดุล 1	17
รูปที่ 2.12	แสดงทรีที่ไม่เป็น AVL	17
รูปที่ 2.13	แสดงทรีที่มีความสูงสมดุลแบบสมบูรณ์	18
รูปที่ 2.14	แสดง AVL Tree	18
รูปที่ 2.15	แสดงโครงสร้างข้อมูลแบบเพจไบนารีทรี	20
รูปที่ 2.16	เพจไบนารีทรีที่สร้างจากคีย์ที่รับมาในลักษณะสุ่ม	21
รูปที่ 2.17	แสดงโครงสร้างข้อมูลแบบบีทรีลำดับ 5	23
รูปที่ 2.18	แสดงทรีที่มีการค้นหาข้อมูลได้ 5 ทาง	24
รูปที่ 2.19	แสดงการเจริญเติบโตของบีทรี	26
รูปที่ 3.1	แผนภาพแสดงขั้นตอนการทำงานของระบบ	33

จุฬาลงกรณ์มหาวิทยาลัย