

การค้นคืนสารสนเทศโดยใช้แนวคิดแบบจำลองปริภูมิเวกเตอร์

นาย สำโรช เมาลานนท์



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษิตตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

พ.ศ. 2535

ISBN 974-579-986-6

ลิขสิทธิ์ของบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย

018732

117149563

INFORMATION RETRIEVAL USING VECTOR SPACE MODEL CONCEPT

MR. SAROCH MOULANONT

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 974-579-986-6

หัวข้อวิทยานิพนธ์

การค้นคว้าสารสนเทศโดยใช้แนวคิดแบบจำลองปริภูมิเวกเตอร์

โดย

นาย สำโรช เมลาเนา

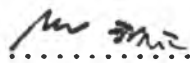
ภาควิชา

วิศวกรรมคอมพิวเตอร์


อาจารย์ที่ปรึกษา


อาจารย์ จารุมาตร ปันทอง

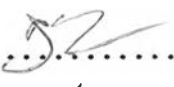
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้วิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาโทบัณฑิต



..... คณบดีบัณฑิตวิทยาลัย
(ศาสตราจารย์ ดร. กวาร์ วัชรภักย์)

คณะกรรมการสอบวิทยานิพนธ์


..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ วิชาญ เลิศวิภาตระกุล)


..... อาจารย์ที่ปรึกษา
(อาจารย์ จารุมาตร ปันทอง)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ สุเมธ วัชรระชัยสุรพล)


..... กรรมการ
(ผู้ช่วยศาสตราจารย์ วันพร ปันแก้ว)

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว



สารบัญช เมลลนนท : การคณคณสารสนเทศโดยใชแนวกคคแบบจลองปรกวมเวกเตอร์
(INFORMATION RETRIEVAL USING VECTOR SPACE MODEL CONCEPT)
อ.พรกษา อจจรยจจรมาตร ปนทอง. 89 ทน. ISBN 974-579-986-6

จคประสงคของวิทยนพนธฉบับน เพื่อศกษากลไกทใชในการคณคณสารสนเทศ โดยอศยแนวกคคแบบจลองปรกวมเวกเตอร์เป็นพณฐาน พรอมคสรองโปรแกรมเพื่อใชในการทอสบประเมณผล ในวิทยนพนธฉบับนค้กลวถงคเป็นมของกรวจย ทฤษฎทเกยวของกบกรวจย ค้แก คเป็นมของระบบกรจคกรขอความ ระบบฐานขอมูลขอความ กรประยคททางกรจคกรขอความ กรบวณกรคณคณสารสนเทศทงแบบสถณยม และระบบกรคณคณชนสูง กรพฒนโปรแกรมโดยใชภาษาซีเป็นเครองมือ กรออกแบบโปรแกรมค้คณถงสวณประกอบพณฐานของระบบฐานขอมูล ค้แก แบบจลองขอมูล บทนยมขอมูล และกลไกกรคณคณสารสนเทศ ผลกรทอสบโปรแกรมปรกฐว กรคณวณคาควมไกลเคยงระหวงเวกเตอร์ของเอกสรกบประกษอสบถมถูคองตรงตามผลทคณวณค้คด้วยมือ และสมารถคณคณเอกสรออกมค้คตามลค้คคาควมไกลเคยงจกมกไปทนอย สวณผลกรทอสบระบบกรคณคณสารสนเทศโดยรวมพว กรทคธรชนเพื่อเป็นทวแทนของเอกสร มค้คควมสค้คต่อประลทธภพของระบบกรคณคณสารสนเทศอยงมก

ภควช วิศวกรรมคณพวเตอร์
สขววช วิทยาศาสตร์คณพวเตอร์
ปกรศกษ 2534

ลายมือชอนลต
ลายมือชออจจรยทปรกษ
ลายมือชออจจรยทปรกษรวม

พิมพ์ต้นฉบับบทคัดย่อวิทยานิพนธ์ภายในกรอบสี่เหลี่ยมนี้เพียงแผ่นเดียว

C017067 : MAJOR COMPUTER SCIENCE

KEY WORD : INFORMATION RETRIEVAL/VECTOR SPACE MODEL

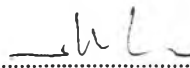

SAROCH MOULANONT : INFORMATION RETRIEVAL USING VECTOR SPACE MODEL

CONCEPT. THESIS ADVISOR : CHARUMATR PINTHONG. 89 pp.

ISBN 974-579-986-6

The objective of this thesis is to study the mechanism of information retrieval based on vector space model concept and develop computer program, bibliographic database, for experimental evaluation. This thesis composes of the background research, theory related to the researches such as background of text management, text database, application of text management, conventional and advance information retrieval. The computer programs are written in C language. The design considers the basic components of database system that are data model, data definition and retrieval mechanism. The result of program testing can be summarized as follows :- the computational similarity between the vector of documents and query compare with manual calculation are correct. The documents can be arranged by descending order of corresponding similarity with the query. The total results of the information retrieval system show that the indexing of documents representation is very significant for the efficiency of retrieval system.

ภาควิชา วิศวกรรมคอมพิวเตอร์
สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์
ปีการศึกษา 2534

ลายมือชื่อนิสิต 
ลายมือชื่ออาจารย์ที่ปรึกษา 
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณท่านอาจารย์ที่ปรึกษาวิทยานิพนธ์ อาจารย์จารุมาศ ปิ่นทอง ที่ได้ให้คำปรึกษาแนะนำแนวทางที่เป็นประโยชน์ต่อการวิจัย พร้อมทั้งคอยผลักดันให้วิทยานิพนธ์ฉบับนี้ได้สำเร็จลุล่วงด้วยดีมาตลอด ขอขอบพระคุณท่านอาจารย์ที่ปรึกษานิสิต รศ. สมชาย ทานอง ที่ให้ความกรุณา ติดตามสอบถามความคืบหน้าของวิทยานิพนธ์อยู่เสมอมา และขอขอบพระคุณท่านอาจารย์ทุกท่านในภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ได้ประสิทธิ์ประสาทวิชาความรู้ให้

นอกจากนี้ขอขอบคุณ คุณศุภกร โสภณวสุ ที่ได้ช่วยเหลือจัดเตรียมแผ่นใส และให้กำลังใจในระหว่างการทำวิทยานิพนธ์ ขอขอบคุณบรรดาเพื่อน ๆ พี่ ๆ น้อง ๆ ที่คอยให้ความช่วยเหลือ และกำลังใจต่อผู้วิจัยเสมอมา

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณ บิดา มารดา ซึ่งเป็นผู้มีพระคุณสูงสุดอันหาที่เปรียบมิได้ที่ได้ให้การอุปการะและกำลังใจต่อผู้วิจัยตลอดมา

สารบัญ

	หน้า
บทคัดย่อภาษาไทย.....	ง
บทคัดย่อภาษาอังกฤษ.....	จ
กิตติกรรมประกาศ.....	ฉ
สารบัญประกอบ.....	ญ

บทที่

1. บทนำ.....	1
1.1 ความเป็นมา.....	1
1.2 วัตถุประสงค์.....	2
1.3 ขอบเขตของการวิจัย.....	3
1.4 ขั้นตอนของการวิจัย.....	3
1.5 ประโยชน์ที่คาดว่าจะได้รับ.....	4
2. แนวคิดและทฤษฎีที่เกี่ยวข้อง.....	5
2.1 การจัดการข้อความ และฐานข้อมูลข้อความ.....	5
2.2 ระบบการค้นคืนสารสนเทศ.....	11
2.3 คุณลักษณะของฐานข้อมูลข้อความ.....	12
2.4 กระบวนการกระจายค่า.....	16
2.5 กระบวนการค้นคืน.....	17
2.5.1 ระบบการค้นคืนแบบสัญญาณ.....	18
2.5.2 ระบบการค้นคืนขั้นสูง.....	23

สารบัญ (ต่อ)

3.	การออกแบบส่วนประกอบของโปรแกรม.....	28
3.1	การออกแบบฐานข้อมูล.....	28
3.1.1	แบบจำลองข้อมูล.....	28
3.1.2	บทนิยามข้อมูล.....	29
3.1.3	กลไกการค้นคืนสารสนเทศ.....	32
3.2	การออกแบบโครงสร้างข้อมูล.....	32
3.2.1	แฟ้มข้อมูลเก็บเอกสาร.....	32
3.2.2	แฟ้มข้อมูลเวกเตอร์.....	32
3.2.3	แฟ้มข้อมูลค่าสำคัญ.....	33
3.3	การออกแบบรายละเอียดโปรแกรม.....	34
3.3.1	ฟังก์ชันพื้นฐานของวินโดว์.....	35
3.3.2	ฟังก์ชันพื้นฐานของป๊อปอัพ.....	36
3.3.3	โครงสร้างของโปรแกรม.....	36
4.	ผลการพัฒนาและการทดสอบโปรแกรม.....	53
4.1	ลักษณะระบบโปรแกรมและวิธีการใช้.....	53
4.2	แฟ้มข้อมูลที่เกิดขึ้นในระบบ.....	61
4.3	ระบบคอมพิวเตอร์ที่นำมาใช้ในการพัฒนาระบบ.....	62
4.4	การทดสอบโปรแกรม.....	62
5.	สรุปผลการวิจัยและข้อเสนอแนะ.....	65
5.1	สรุปผลการวิจัย.....	65
5.2	ข้อเสนอแนะ.....	66

สารบัญ (ต่อ)

เอกสารอ้างอิง.....	68
--------------------	----

ภาคผนวก

ภาคผนวก ก.	71
ภาคผนวก ข.	76
ภาคผนวก ค.	80
ภาคผนวก ง.	81

ประวัติหน้าวิจัย.....	89
-----------------------	----

สารบัญรูปประกอบ

รูปที่		หน้า
2.1	กราฟแสดงความสัมพันธ์ระหว่างค่า RECALL และ PRECISION..	8
2.2	องค์ประกอบระบบการจัดการฐานข้อความ.....	9
2.3	การประยุกต์ทางการจัดการข้อความ.....	10
2.4	ตัวอย่างระเบียบฐานข้อมูลบรรณานุกรม.....	12
2.5	ตัวอย่างระเบียบฐานข้อมูลข้อความเต็ม.....	14
2.6	ตัวอย่างระเบียบฐานข้อมูลตัวอย่างอิง.....	15
2.7	ระเบียบในแฟ้มข้อมูลลำดับ.....	19
2.8	แฟ้มตรรกษณ์.....	20
2.9	แผนภาพเวนนแสดงตัวดำเนินการบูล.....	22
3.1	ตัวอย่างแฟ้มเอกสารของฐานข้อมูลบรรณานุกรม.....	29
3.2	การคอมไพล์โปรแกรมกับคำสั่งเบสิกมีมา.....	30
3.3	DATA BASE SCHEMA C SOURCE CODE.....	31
3.4	แฟ้มข้อมูลเวกเตอร์.....	33
3.5	ตัวอย่างแผนภาพโครงสร้างกระบวนการ.....	36
3.6	แผนภาพโครงสร้างกระบวนการแบบลำดับ.....	36
3.7	แผนภาพโครงสร้างกระบวนการแบบทางเลือก.....	37
3.8	แผนภาพโครงสร้างกระบวนการแบบวนซ้ำ.....	37
3.9	แผนภาพโครงสร้างโปรแกรมหลักเชิงลำดับชั้น.....	38

สารบาณรูปประกอบ (ต่อ)

3.10	แผนภาพโครงสร้างมอดูลกระบวนการนำเอกสารเข้าระบบและ สร้างเวคเตอร์สำหรับเอกสาร.....	41
3.10.1	กระบวนการสร้างรายการโยงคำตรรกะของเอกสาร.....	42
3.10.2	กระบวนการแสดงวินโดว์คำตรรกะ.....	43
3.10.3	กระบวนการสร้างเวคเตอร์ของเอกสาร.....	44
3.10.4	กระบวนการนำรายการโยงคำตรรกะนี้ใส่ลงในบิตรี.....	45
3.10.5	กระบวนการค้นหาเนื้อที่รายการโยงคำตรรกะ.....	46
3.11	แผนภาพโครงสร้างมอดูลกระบวนการข้อความเพื่อ ค้นคืนสารสนเทศ.....	48
3.11.1	กระบวนการรับข้อความในกรอบวินโดว์.....	49
3.11.2	กระบวนการแบ่งแยกคำ.....	50
3.11.3	กระบวนการสอบถามสารสนเทศ.....	51
4.1	ลักษณะจอภาพรายการเลือกหลักและกรอบแสดงโลโก้ของระบบ...	55
4.2	ลักษณะจอภาพการนำเอกสารเข้าระบบ.....	56
4.3	ลักษณะจอภาพแสดงการวิเคราะห์คำของแฟ้มเอกสาร.....	57
4.4	ลักษณะจอภาพการสอบถามสารสนเทศ.....	58
4.5	ลักษณะจอภาพแสดงการวิเคราะห์คำของประโยคสอบถาม.....	59
4.6	ลักษณะจอภาพแสดงผลค้นคืนจากการสอบถามสารสนเทศ.....	60