

248419

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการการวิจัยแห่งชาติ



248419

การพัฒนาแนวทางการค้นหาความถี่ซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล

เกียรติศักดิ์ จันทร์หอม

วิศวกรรมศาสตรมหาบัณฑิต  
สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย  
มหาวิทยาลัยเชียงใหม่  
พฤษภาคม 2554

600253334

ห้องสมุดงานวิจัย สำนักงานคณะกรรมการวิจัยแห่งชาติ



248419

การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล



เกียรติศักดิ์ จันทร์หอม

วิทยานิพนธ์นี้เสนอต่อบัณฑิตวิทยาลัยเพื่อเป็นส่วนหนึ่ง  
ของการศึกษาตามหลักสูตรปริญญา  
วิศวกรรมศาสตรมหาบัณฑิต  
สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย  
มหาวิทยาลัยเชียงใหม่  
พฤษภาคม 2554

# การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล

เกียรติศักดิ์ จันทร์หอม

วิทยานิพนธ์นี้ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษา  
ตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต  
สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะกรรมการสอบวิทยานิพนธ์

อาจารย์ที่ปรึกษาวิทยานิพนธ์

.....ประธานกรรมการ

.....

อาจารย์ ดร.ลิชณา ระมิงค์วงศ์

ผู้ช่วยศาสตราจารย์ ดร.จักรพงษ์ นาทวิชัย

..... กรรมการ

ผู้ช่วยศาสตราจารย์ ดร.จักรพงษ์ นาทวิชัย

..... กรรมการ

อาจารย์ ดร.พงษ์ บุญมา

..... กรรมการ

รองศาสตราจารย์เกียรติศักดิ์ คันธพนิต

27 พฤษภาคม 2554

© ลิขสิทธิ์ของมหาวิทยาลัยเชียงใหม่

### กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร. จักรพงษ์ นาทวีชัย อาจารย์ที่ปรึกษาวิทยานิพนธ์ผู้ซึ่งกรุณาให้ความรู้ คำแนะนำ คำปรึกษา และตรวจแก้ไขจนวิทยานิพนธ์เสร็จสมบูรณ์ ผู้เขียนขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณะกรรมการสอบทุกท่าน ที่รับเป็นกรรมการสอบวิทยานิพนธ์ และให้คำปรึกษาเป็นอย่างดี

ขอกราบขอบพระคุณอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ที่กรุณาให้คำแนะนำ และขอกราบขอบพระคุณเจ้าหน้าที่ห้องธุรการประจำภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่านที่ให้ความช่วยเหลือคอยช่วยเหลือในดำเนินงานเอกสารจนวิทยานิพนธ์นี้เสร็จสมบูรณ์

ท้ายสุดนี้ หากมีสิ่งใดขาดตกบกพร่อง หรือมีข้อผิดพลาดประการใด ผู้เขียนขออภัยเป็นอย่างสูงในความผิดพลาดและข้อบกพร่องนั้น และผู้เขียนหวังว่าวิทยานิพนธ์นี้จะมีประโยชน์บ้างไม่มากก็น้อย สำหรับหน่วยงานที่เกี่ยวข้อง ตลอดจนผู้สนใจที่จะศึกษารายละเอียดเกี่ยวกับวิทยานิพนธ์นี้ต่อไป

เกียรติศักดิ์ จันทร์หอม

ชื่อเรื่องวิทยานิพนธ์	การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล
ผู้เขียน	นายเกียรติศักดิ์ จันทร์หอม
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผศ.ดร.จักรพงษ์ นาทวีชัย

บทคัดย่อ

248419

ในปัจจุบันนี้ข้อมูลต่างๆมีอยู่เป็นจำนวนมากมาย ไม่ว่าจะเป็นข้อมูลที่เป็นข่าวสาร ข้อมูลที่เกี่ยวข้องกับการสมัครสมาชิกต่างๆ เพื่อที่จะใช้ระบบสารสนเทศต่างๆ ข้อมูลที่เป็นเอกสาร ข้อมูลที่เป็นภาพหรือว่าข้อมูลที่เป็นเสียง ข้อมูลต่างๆเหล่านี้จะถูกจัดเก็บอยู่ในฐานข้อมูล เพื่อทำให้เกิดความสะดวกในการเข้าถึงข้อมูล และสะดวกในการจัดเก็บข้อมูล ข้อมูลต่างๆที่ได้อธิบายไปเหล่านี้มีอยู่เป็นจำนวนมากศาลในโลกปัจจุบันซึ่งถูกเรียกว่าเป็นยุคของข้อมูลข่าวสารต่างๆ จากข้อมูลที่มีอยู่เป็นจำนวนมากเหล่านี้อาจทำให้เกิดปัญหาความซ้ำซ้อนของข้อมูลได้

จากการศึกษาแนวทางในการแก้ไขปัญหาความซ้ำซ้อนของข้อมูล จึงทำให้ทราบถึงปัญหาในขั้นตอนการค้นหาความซ้ำซ้อนของข้อมูล คือ การค้นหาความซ้ำซ้อนของข้อมูลในวิธีการเก่าๆ จำเป็นต้องมีการนำข้อมูลออกจากฐานข้อมูล เพื่อนำไปทำการจำแนกความซ้ำซ้อนของข้อมูล เมื่อทำการค้นหาความซ้ำซ้อนของข้อมูลเสร็จจึงนำข้อมูลที่ผ่านมาจากการจำแนกความซ้ำซ้อนของข้อมูลกลับเข้ามาใส่ในฐานข้อมูล จากวิธีการดังกล่าวแสดงให้เห็นว่าเป็นขั้นตอนที่มีความยุ่งยาก และทำให้เสียเวลาในการนำข้อมูลออก และนำข้อมูลกลับเข้ามาในฐานข้อมูล

ดังนั้นวิทยานิพนธ์ฉบับนี้จึงได้เสนอกระบวนการในการค้นหาความซ้ำซ้อนของข้อมูล เพื่อแก้ไขปัญหาดังกล่าว อีกทั้งทำการเพิ่มประสิทธิภาพแลประสิทธิภาพในขั้นตอนการค้นหาความซ้ำซ้อน โดยเสนอให้ใช้ภาษาเอสคิวแอลในการอิมพลีเมนต์ และเสนอให้ใช้วิธีการเขียนฟังก์ชัน

จีนมาใช้เอง (UDF) เพื่อใช้ในการค้นหาความซ้ำซ้อนเพื่อทำการแก้ไขปัญหาในการนำข้อมูลเข้า  
และส่งข้อมูลออก อีกทั้งยังมีการทำดัชนีข้อมูลเพื่อเพิ่มประสิทธิภาพในการค้นหาความซ้ำซ้อนให้มี  
ความรวดเร็วมากขึ้น

**Thesis Title**                      Development of Data-Duplication Detection Approach on Database  
Management System

**Author**                                Mr. Kiattisak Chanhom

**Degree**                                Master of Engineering (Computer Engineering)

**Thesis Advisor**                    Asst. Prof.Dr. Juggapong Natwichai

**ABSTRACT**

248419

Data-duplication is one of the important issues in the context of information system management. Instead of storing a single real-world object as an entity in an information system. The duplication, storing more than one entities representing a single object, can be occurred. This problem can decrease the quality of service of information systems. In this paper, we propose an efficient approach to detect the duplication based on the RDBMS foundation. Our approach is based on the assumption that the data to be processed have been stored in the RDBMS at the first place. Thus, the proposed approach does not require the data to be imported/exported from the storage. Also, such approach will benefit from the query optimizer of the RDBMS. The experiment results on the real-life datasets have been presented to validate such proposed work.

สารบัญ

	หน้า
กิตติกรรมประกาศ	ก
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	ฉ
สารบัญภาพ	ณ
บทที่ 1 บทนำ	1
1.1 ที่มาและปัญหาของการศึกษา	1
1.2 แนวทางการแก้ปัญหา	3
1.3 วัตถุประสงค์ของการศึกษา	4
1.4 ประโยชน์ที่ได้รับจากการศึกษาเชิงทฤษฎีและเชิงประยุกต์	4
1.5 ขอบเขตการทำวิจัย	4
1.6 วิธีการทำวิจัย	4
1.7 เครื่องมือที่ใช้ในการพัฒนา	5
บทที่ 2 ทฤษฎีที่เกี่ยวข้อง	6
2.1 ทฤษฎีในการเลือกเดสคริปชัน	6
2.2 ทฤษฎีในการจำแนกข้อมูล	9
2.3 การปรับปรุงประสิทธิภาพการสอบถาม	15
บทที่ 3 ปัญหาและวิธีการแก้ปัญหา	19
3.1 การเลือกเดสคริปชัน	21
3.1.1 การเลือกตามแนวคิดเฮอริสติกในระดับสกีมา (Schema-base Heuristics) และการกำหนดเงื่อนไข (Condition)	21
3.1.2 การประยุกต์ใช้ตามแนวคิดอินสแตนเบสเฮอริสติก (Instance Based Heuristics) และการเลือกโดยใช้โดเมนของความรู้ (Domain-knowledge)	24
3.2 การจำแนกข้อมูลภายในฐานข้อมูล	25
3.2.1 การจำแนกข้อมูลที่เป็นลบ	25
3.2.2 การจำแนกข้อมูลที่เป็นบวก	26

3.2.3	การจำแนกข้อมูลที่มีความคล้ายคลึงกัน	27
3.3	การทำดัชนีข้อมูล	28
3.3.1	การทำดัชนีข้อมูลเพื่อใช้สำหรับวิธีการที่ให้ผู้สร้างฟังก์ชัน การทำงานขึ้นมาเอง	28
3.3.2	การทำดัชนีข้อมูลเพื่อใช้สำหรับวิธีการอิมพลีเมนต์บน ระบบจัดการฐานข้อมูลด้วยภาษาแอลเอสคิวแอลทั้งหมด	29
บทที่ 4	การทดลองและผลการทดลอง	32
4.1	ข้อมูลที่ใช้ในการทดลอง	33
4.2	ผลการทดลอง	35
4.2.1	แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาในทุกขั้นตอน	35
4.2.2	แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาเฉพาะขั้นตอนการค้นหาความซ้ำซ้อนเท่านั้น	37
4.2.3	แนวคิดในการค้นหาความซ้ำซ้อนของข้อมูล โดยจะทำการจับเวลาเมื่อฐานข้อมูลมีการเปลี่ยนแปลง	38
4.3	สรุปผลการทดลอง	39
บทที่ 5	สรุปผลการวิจัยและข้อเสนอแนะ	40
5.1	สรุปผลการวิจัย	40
5.2	ข้อเสนอแนะและงานวิจัยในอนาคต	41
	เอกสารอ้างอิง	42
	ประวัติผู้เขียน	44

## สารบัญภาพ

รูป	หน้า	
2.1	การเก็บข้อมูลของบุคคลทั้งในแบบสก็มา และตารางการเก็บข้อมูล	7
2.2	การจำแนกข้อมูลด้วยการจำแนกในแบบต่างๆ	10
3.1	สก็มาของฐานข้อมูลที่ซีพีเอส	20
3.2	สก็มาของฐานข้อมูลที่ซีพีเอสที่ผ่านกระบวนการตามแนวคิด เฮร์ริสติกในระดับสก็มา	22
3.3	สก็มาที่ได้ผ่านกระบวนการในการเลือกเดสคริปชันแล้ว	24
3.4	แสดงลำดับของเงื่อนไขที่ใช้ในการจำแนกความซ้ำซ้อนของข้อมูล	27
3.5	แสดงสก็มาที่ได้ผ่านการจอยกันแล้ว	28
3.6	สก็มาที่จะใช้การเก็บค่ารูปแบบที่มีความซ้ำซ้อนกันของข้อมูล	29
3.7	แสดงสก็มาของแอตทริบิวต์ที่ได้ผ่านการตัดคำแล้ว	30
4.1	กราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูล กับขนาดที่แตกต่างกัน โดยจะทำการจับเวลาในทุกๆขั้นตอน	35
4.2	กราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูล กับขนาดที่แตกต่างกัน โดยจะทำการจับเวลาในทุกๆขั้นตอน	37
4.3	กราฟเปรียบเทียบเวลาที่ใช้ในการค้นหาความซ้ำซ้อนของข้อมูล กับขนาดที่แตกต่างกัน โดยจะทำการจับเวลาเมื่อมีการเปลี่ยนแปลง ของข้อมูลภายในฐานข้อมูล	38