การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลในระบบจัดการฐานข้อมูล

เกียรติศักดิ์ จันทร์หอม

วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย
มหาวิทยาลัยเชียงใหม่
พฤษภาคม 2554

การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล

เกียรติศักดิ์ จันทร์หอม

วิทยานิพนธ์นี้เสนอต่อบัณฑิตวิทยาลัยเพื่อเป็นส่วนหนึ่ง
ของการศึกษาตามหลักสูตรปริญญา
วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์

บัณฑิตวิทยาลัย
มหาวิทยาลัยเชียงใหม่
พฤษภาคม 2554

การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล

เกียรติศักดิ์ จันทร์หอม

วิทยานิพนธ์นี้ได้รับการพิจารณาอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษา
ตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
อาจารย์ ดร.ลัชนา ระมิงค์วงศ์

..... กรรมการ
ผู้ช่วยศาสตราจารย์ ดร.จักรพงศ์ นาทวิชัย

..... กรรมการ
อาจารย์ ดร.พฤษภ์ บุญมา

..... กรรมการ
รองศาสตราจารย์ขจรศักดิ์ คันธพนิต

อาจารย์ที่ปรึกษาวิทยานิพนธ์

.....
ผู้ช่วยศาสตราจารย์ ดร.จักรพงศ์ นาทวิชัย

27 พฤษภาคม 2554

กิตติกรรมประกาศ

วิทยานิพนธ์นี้สำเร็จลงได้ด้วยความกรุณาจาก ผู้ช่วยศาสตราจารย์ ดร. จักรพงศ์ นาทวิชัย อาจารย์ที่ปรึกษาวิทยานิพนธ์ผู้ซึ่งกรุณาให้ความรู้ คำแนะนำ คำปรึกษา และตรวจแก้ไขจนวิทยานิพนธ์เสร็จสมบูรณ์ ผู้เขียนขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอกราบขอบพระคุณคณะกรรมการสอบทุกท่าน ที่รับเป็นกรรมการสอบวิทยานิพนธ์ และให้คำปรึกษาเป็นอย่างดี

ขอกราบขอบพระคุณอาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์ที่กรุณาให้คำแนะนำ และขอกราบขอบพระคุณเจ้าหน้าที่ห้องธุรการประจำภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่านที่ให้ความช่วยเหลือคอยช่วยเหลือในด้านงานเอกสารจนวิทยานิพนธ์นี้เสร็จสมบูรณ์

ท้ายสุดนี้ หากมีสิ่งใดขาดตกบกพร่อง หรือมีข้อผิดพลาดประการใด ผู้เขียนขออภัยเป็นอย่างสูงในความผิดพลาดและข้อบกพร่องนั้น และผู้เขียนหวังว่าวิทยานิพนธ์นี้จะมีประโยชน์บ้างไม่มากก็น้อย สำหรับหน่วยงานที่เกี่ยวข้อง ตลอดจนผู้สนใจที่จะศึกษารายละเอียดเกี่ยวกับวิทยานิพนธ์นี้ต่อไป

เกียรติศักดิ์ จันทร์หอม

ชื่อเรื่องวิทยานิพนธ์ การพัฒนาแนวทางการค้นหาความซ้ำซ้อนของข้อมูลบนระบบจัดการฐานข้อมูล

ผู้เขียน นายเกียรติศักดิ์ จันทร์หอม

ปริญญา วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.จักรพงศ์ นาทวิชัย

บทคัดย่อ

ในปัจจุบันนี้ข้อมูลต่างๆมีอยู่เป็นจำนวนมากมาย ไม่ว่าจะเป็นข้อมูลที่เป็นข่าวสาร ข้อมูลที่เกี่ยวกับการสมัครสมาชิกต่างๆ เพื่อที่จะใช้ระบบสารสนเทศต่างๆ ข้อมูลที่เป็นเอกสาร ข้อมูลที่เป็นภาพหรือว่าข้อมูลที่เป็นเสียง ข้อมูลต่างๆเหล่านี้จะถูกจัดเก็บอยู่ในรูปฐานข้อมูล เพื่อทำให้เกิดความสะดวกในการเข้าถึงข้อมูล และสะดวกในการจัดเก็บข้อมูล ข้อมูลต่างๆที่ได้อธิบายไปเหล่านี้มีอยู่เป็นจำนวนมหาศาลในโลกปัจจุบันซึ่งถูกเรียกว่าเป็นยุคของข้อมูลข่าวสารต่างๆ จากข้อมูลที่มีอยู่เป็นจำนวนมากเหล่านี้อาจทำให้เกิดปัญหาความซ้ำซ้อนของข้อมูลได้

จากการศึกษาแนวทางในการแก้ไขปัญหาความซ้ำซ้อนของข้อมูล จึงทำให้ทราบถึงปัญหาในขั้นตอนการค้นหาความซ้ำซ้อนของข้อมูล คือ การค้นหาความซ้ำซ้อนของข้อมูลในวิธีการเก่าๆ จำเป็นต้องมีการนำข้อมูลออกจากฐานข้อมูล เพื่อนำไปทำการจำแนกความซ้ำซ้อนของข้อมูล เมื่อทำการค้นหาความซ้ำซ้อนของข้อมูลเสร็จจึงนำข้อมูลที่ผ่านการจำแนกความซ้ำซ้อนของข้อมูลกลับเข้ามาใส่ในฐานข้อมูล จากวิธีการดังกล่าวแสดงให้เห็นว่าเป็นขั้นตอนที่มีความยุ่งยาก และทำให้เสียเวลาในการนำข้อมูลออก และนำข้อมูลกลับเข้ามาในฐานข้อมูล

ดังนั้นวิทยานิพนธ์ฉบับนี้จึงได้เสนอกระบวนการในการค้นหาความซ้ำซ้อนของข้อมูล เพื่อแก้ไขปัญหาดังกล่าว อีกทั้งทำการเพิ่มประสิทธิภาพและประสิทธิผลในขั้นตอนการค้นหาความซ้ำซ้อน โดยเสนอให้ใช้ภาษาเอสคิวแอลในการอิมพลีเมนต์ และเสนอให้ใช้วิธีการเขียนฟังก์ชัน

ขึ้นมาใช้เอง (UDF) เพื่อใช้ในการค้นหาความซ้ำซ้อนเพื่อทำการแก้ไขปัญหาในการนำข้อมูลเข้าและส่งข้อมูลออก อีกทั้งยังมีการทำดัชนีข้อมูลเพื่อเพิ่มประสิทธิผลในการค้นหาความซ้ำซ้อนให้มีความรวดเร็วมากขึ้น

Thesis Title	Development of Data-Duplication Detection Approach on Database Management System
Author	Mr. Kiettisak Chanhom
Degree	Master of Engineering (Computer Engineering)
Thesis Advisor	Asst. Prof.Dr. Juggapong Natwichai

ABSTRACT

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.

สารบาญ

สารบาญภาพ