

เพิ่มมูลค่าข้อมูลทางธุรกิจด้วย Data mining



วิรัชคุณ์ เมาะพงษ์

คณาจารย์ภาควิชาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี



ต่อ จากฉบับที่แล้ว

ปัจจัยที่ทำให้ Data mining เป็นที่ได้รับความนิยม

จำนวนและขนาดข้อมูลขนาดใหญ่ถูกผลิตและขยายตัวอย่างรวดเร็ว การสืบค้นความรู้จะมีความหมายก็ต่อเมื่อ ฐานข้อมูลที่ใช้มีขนาดใหญ่มาก ปัจจุบันมีจำนวนและขนาดข้อมูลขนาดใหญ่ที่ขยายตัวอย่างรวดเร็ว โดยผ่านทาง Internet ดาวเทียมและแหล่งผลิตข้อมูลอื่นๆ เช่น เครื่องอ่าน Barcode, เครดิตการ์ด, อีคอมเมิร์ซ

● ข้อมูลถูกจัดเก็บเพื่อนำไปสร้างระบบการสนับสนุนการตัดสินใจ (Decision Support System) เพื่อเป็นการง่ายต่อการนำข้อมูล

มาใช้ในการวิเคราะห์เพื่อการตัดสินใจ ส่วนมากข้อมูลจะถูกจัดเก็บแยกมาจากระบบปฏิบัติการ (Operational System) โดยจัดอยู่ในรูปของคลัง หรือเหมืองข้อมูล (Data Warehouse) ซึ่งเป็นการง่ายต่อการนำเอาไปใช้ในการสืบค้นความรู้

● ระบบ Computer สมรรถนะสูง มีราคาต่ำลง เทคนิค Data mining ประกอบไปด้วย Algorithm ที่มีความซับซ้อนและความต้องการการคำนวณสูง จึงจำเป็นต้องใช้งานกับระบบ Computer สมรรถนะสูง ปัจจุบันระบบ Computer สมรรถนะสูงมีราคาต่ำลงพร้อมด้วยเริ่มมีเทคโนโลยีที่นำเครื่อง Microcomputer จำนวนมากมาเชื่อมต่อกันโดยเครือข่ายความเร็วสูง (PC Cluster) ทำให้ได้ระบบ Computer สมรรถนะสูงในราคาต่ำ

● การแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า เนื่องจากปัจจุบันมีการแข่งขันอย่างสูงในด้านอุตสาหกรรมและการค้า มีการผลิตข้อมูลไว้อย่างมากมายแต่ไม่ได้นำมาใช้ให้เกิดประโยชน์ จึงเป็นการจำเป็นอย่างยิ่งที่ต้องควบคุมและสืบค้นความรู้ที่ถูกซ่อนอยู่ในฐานข้อมูลความรู้ที่ได้รับ สามารถนำไปวิเคราะห์เพื่อการตัดสินใจในการจัดการในระบบต่างๆ ซึ่งจะเห็นได้ว่าความรู้เหล่านี้ถือว่าเป็นผลิตภัณฑ์ชิ้นหนึ่งเลยทีเดียว

ประเภทข้อมูลที่นำมาทำ Data mining

● Relational Database เป็นฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง โดยในแต่ละตารางจะประกอบไปด้วย แถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย entity-relationship (ER) model

● Data Warehouses เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ๆ เดียวกัน

● Transactional Database ประกอบด้วยข้อมูลในแต่ละ Transaction แทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงิน จะเก็บข้อมูลในรูปแบบ ชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น



● **Advanced Database** เป็นฐานข้อมูลที่จัดเก็บในรูปแบบอื่นๆ เช่น ข้อมูลแบบ object-oriented ข้อมูลที่เป็น text file ข้อมูลมัลติมีเดีย ข้อมูลในรูปแบบของ web

ลักษณะเฉพาะของข้อมูลที่นำมาทำ Data mining

● **ข้อมูลขนาดใหญ่** เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลได้ด้วยตาเปล่า หรือโดยการใช้ Database Management System (DBMS) ในการจัดการฐานข้อมูล

● **ข้อมูลที่มาจากหลายแหล่ง** โดยอาจรวบรวมมาจากหลายระบบปฏิบัติการ หรือหลาย DBMS เช่น Oracle, DB2, MS SQL, MS Access เป็นต้น

● **ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดเวลาที่ทำการ Mining** หากข้อมูลที่มีอยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหาขึ้นก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำ Mining แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จากการทำ Mining สมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้น เพื่อให้ได้ผลลัพธ์ที่มีความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำ Mining ใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

● **ข้อมูลที่มีโครงสร้างซับซ้อน** เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำ Mining ได้เช่นกันแต่ต้องใช้เทคนิคการทำ Data mining ขั้นสูง

เทคนิคต่างๆ ของ Data mining

1. **Association rule Discovery** เป็นเทคนิคหนึ่งของ Data mining ที่สำคัญและสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่างๆ หลักการทำงานของวิธีนี้ คือ การค้นหาความสัมพันธ์ของข้อมูล จากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์ หรือ

ทำนายปรากฏการณ์ต่างๆ หรือมาจากการวิเคราะห์การซื้อสินค้าของลูกค้าเรียกว่า “Market Basket Analysis” ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหาซึ่งการวิเคราะห์แบบนี้เป็นการใช้ “กฎความสัมพันธ์” (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติ ของ Amazon ข้อมูลการสั่งซื้อทั้งหมดของ Amazon ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือ ลูกค้าที่ซื้อหนังสือเล่มหนึ่งๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่า ควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน ตัวอย่างเช่น buys (x, database) -> buys (x, Data mining) [80%, 60%] หมายความว่า เมื่อซื้อหนังสือ database แล้วมีโอกาสที่จะซื้อหนังสือ Data mining ด้วย 60% และมีการซื้อทั้งหนังสือ database และหนังสือ Data mining พร้อมๆ กัน 80%

อีกตัวอย่าง คือ ในการซื้อสินค้าของลูกค้า 1 ครั้ง โดยไม่ต้องจำกัดว่าจะซื้อสินค้าในห้างร้าน หรือส่งผ่านทางไปรษณีย์ หรือการซื้อสินค้าจากร้านค้าเสมือนจริง (Virtual store) บน web โดยปกติเราจะต้องทราบราคาสินค้าโดยที่ลูกค้ามักซื้อด้วยกัน เพื่อนำไปพิจารณาปรับปรุงการจัดวางสินค้าในร้าน หรือใช้เพื่อหาวิธีวางรูปคู่กันโฆษณาสินค้า ก่อนอื่นขอกำหนดคำว่า กลุ่มรายการ (item set) หมายถึง กลุ่มสินค้าที่ปรากฏร่วมกัน เช่น {รองเท้า-ถุงเท้า}, {ปากกา-หมึก} หรือ {นม-น้ำผลไม้} โดยกลุ่มรายการดังกล่าวนี้ อาจจะจับคู่กลุ่มลูกค้ากับสินค้าก็ได้ เช่น วิเคราะห์หา “ลูกค้าที่ซื้อสินค้าบางชนิดซ้ำๆ กัน อย่างน้อย 5 ครั้งแล้ว” ในกรณีนี้ หากฐานข้อมูลเรามีการเก็บรายการซื้อขายเป็นจำนวนมากและคำถามข้างต้นนี้จำเป็นต้องค้นหาทุกๆ คู่ของลูกค้ากับสินค้า เช่น {คุณ ก, สินค้า A}, {คุณ ก, สินค้า B}, {คุณ ก, สินค้า C}, {คุณ ข, สินค้า B} เป็นต้น นับเป็นงานที่หนักพอสมควรสำหรับ DBMS

หลังจากที่ DBMS ประมวลผล เนื่องจากมีข้อมูลที่ต้องตรวจสอบมากมายหลายคู่และแต่ละคู่ต้องค้นหาจากฐานข้อมูล แต่ผลลัพธ์ของการค้นหาแบบนี้ มักจะมีจำนวนน้อยมาก เลยเรียกว่าเป็น “iceberg query” ซึ่งเปรียบกับสำนวนไทย คือ “งมเข็มในมหาสมุทร”

อ่าน ต่อฉบับหน้า

