มารู้จัก Data Mining กันเถอะ

ทิพาวรรณ ศิลวัฒนานุสรณ์ *

ปฏิเสธไม่ได้เลยว่า ณ ปัจจุบันนี้ ในองค์กรต่างๆ ไม่ว่าจะเป็นธุรกิจการเงิน การธนาคาร หน่วยงานราชการ หรือหน่วยงานอื่นๆ จะประสบกับปริมาณข้อมูลที่เพิ่มขึ้นอย่างมากมายมหาศาล จำนวนข้อมูลที่มีอยู่มากมายนั้น ผู้ใช้ได้ผ่านกระบวนการแปรรูปข้อมูลเหล่านั้นเพื่อให้ได้มาซึ่งสารสนเทศหรือยัง มั่นใจหรือไม่ว่าสารสนเทศที่ได้ตรงกับความต้องการและมีการนำมาใช้ให้เกิดประโยชน์อย่างแท้จริง ซึ่งในบทความนี้จะกล่าวพอสังเขปเพื่อให้ผู้อ่านได้ทราบถึงความหมายของดาต้าไมน์นิง เทคนิคต่างๆ ที่นำมาใช้ในดาต้าไมน์นิง รวมทั้งแนวโน้มของดาต้าไมน์นิง

Data Mining คืออะไร

หลายคนคงเคยได้ยินหรือคุ้นกับคำว่า Data Mining (ดาต้าไมน์นิง) อย่างไรก็ตามคงมีจำนวนอีกไม่น้อยที่ไม่เคยได้ยินหรือคุ้นกับคำๆ นี้ นิยามง่ายๆ สำหรับดาต้าไมน์นิงก็คือ ดาต้าไมน์นิงจะช่วยผู้ใช้ได้รับสารสนเทศที่มีประโยชน์จากฐานข้อมูลที่**มีขนาดใหญ่มากๆ** ประเด็นอยู่ตรงที่ขนาดของฐานข้อมูลที่มีปริมาณข้อมูลอยู่เป็นจำนวนมากๆ นั่นเองแล้วจะต้องมีปริมาณข้อมูลเท่าไร จึงจะกำหนดได้ว่าฐานข้อมูลนั้นมีขนาดใหญ่ ฐานข้อมูลที่มีปริมาณข้อมูลหรือจำนวนรายการเป็นล้านล้าน (เทอราไบต์) รายการ ถือได้ว่าเป็นฐานข้อมูลขนาดใหญ่ ถ้าในฐานข้อมูลมีจำนวนข้อมูลไม่มากพอเราก็ไม่จำเป็นที่ต้องพัฒนาเทคโนโลยีใหม่ๆ เช่น ดาต้าไมน์นิงเพื่อค้นหาสารสนเทศที่ตรงกับความต้องการและมีประโยชน์จากฐานข้อมูลนั้นๆ ตัวอย่างเช่น ย้อนอดีตไปเมื่อ 20 ปีที่แล้ว สมมติว่าคุณเป็นเจ้าของร้านขายสินค้าแห่งหนึ่งซึ่งดำเนินกิจการมาได้ระยะหนึ่ง คุณจะเก็บข้อมูลอะไรบ้างที่คุณสามารถนำมาใช้ประโยชน์ได้ และแน่นอนว่าข้อมูลสินค้า รวมทั้งข้อมูลลูกค้าจะเป็นสิ่งแรกที่คุณจะจัดเก็บ และไม่แน่ว่าคุณอาจจะจัดเก็บเพียงชื่อของลูกค้าเท่านั้น หรืออาจจะทราบว่าลูกค้าเป็นใครและสินค้าประเภทใดที่ลูกค้าซื้อเป็นประจำจากการวิเคราะห์ข้อมูลและคาดการณ์แนวโน้มการขาย จากการกระทำดังกล่าวจะเห็นได้ว่าคุณไม่จำเป็นต้องนำคอมพิวเตอร์เพื่อช่วยในการไมน์นิงข้อมูล เพราะการวิเคราะห์ข้อมูลและพยากรณ์แนวโน้มการขายสามารถทำได้โดยตัวคุณเอง อย่างไรก็ตาม ณ ปัจจ-

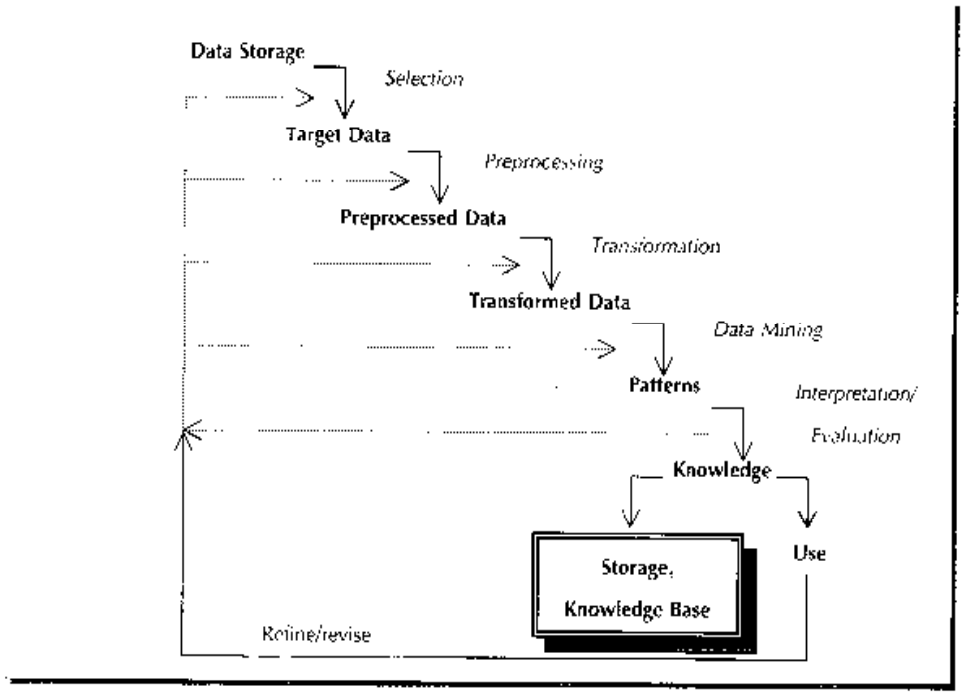
* อาจารย์ระดับ 6 ภาควิชาบรรณารักษศาสตร์และสารนิเทศศาสตร์
คณะมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

ปัจจุบัน ข้อมูลที่จัดเก็บอยู่ในฐานข้อมูลมีเป็นจำนวนล้าน ล้านรายการ และถูกจัดเก็บ ถูกรวบรวมไว้ในที่ซึ่งสามารถเข้าถึงได้อย่างรวดเร็ว โดยที่ข้อมูลอาจถูกจัดเก็บไว้ในโกดังข้อมูล (Data Warehouse) เพื่อที่จะได้ขุดหาสารสนเทศหรือความรู้ที่จำเป็นและมีประโยชน์ต่อการดำเนินการต่อไป ซึ่งกระบวนการขุดหาสารสนเทศหรือความรู้ดังกล่าวก็คือ Data Mining นั่นเอง

อีกนัยหนึ่งอาจกล่าวได้ว่า Data Mining เป็น "exploratory data analysis" คือกระบวนการของการสืบค้นและวิเคราะห์สกัดเอาสารสนเทศที่มีประโยชน์จากปริมาณข้อมูลจำนวนมากมายมหาศาลในฐานข้อมูลนั้น ๆ ในกระบวนการของการค้นพบความสัมพันธ์หรือสืบค้นความรู้แบบแผนและแนวโน้มที่มีประโยชน์จากข้อมูลที่มีปริมาณมหาศาล อาจใช้เทคโนโลยีในการรู้จำและวิเคราะห์ (Pattern Recognition Technologies) และเทคนิคทางด้านสถิติและคณิตศาสตร์ (Statistical and Mathematical Techniques) มาเกี่ยวข้องด้วย

จุดประสงค์หลักของ Data Mining มี 2 ประการใหญ่ ๆ คือ การทำนาย (Prediction) และการอธิบายความ (Description) โดยที่การทำนายนั้นจะเกี่ยวข้องกับการนำเอาตัวแปรบางตัวหรือเซตข้อมูลบางเซตข้อมูลในฐานข้อมูลนั้นมาทำนายค่าของตัวแปรตัวอื่น ๆ ที่น่าสนใจและทำนายแนวโน้มในอนาคตต่อตัวแปรอื่น ๆ ด้วยส่วนการอธิบายความจะเป็นในลักษณะของการหารูปแบบเพื่อที่จะอธิบายคุณลักษณะหรือคุณสมบัติโดยทั่วไปของข้อมูลต่าง ๆ ในฐานข้อมูล

นอกจากนี้ Data Mining ยังเป็นการค้นหารูปแบบที่น่าสนใจจากปริมาณข้อมูลจำนวนมหาศาล รูปแบบที่ว่านี้สามารถนำมาใช้เพื่อกำหนดกลยุทธ์ทางธุรกิจ หรือเพื่อระบุพฤติกรรมที่ไม่เป็นปกติได้ ตัวอย่างเช่น การจับจ่ายซื้อของผ่านบัตรเครดิตในลักษณะที่ยอดเงินมีการเพิ่มขึ้นอย่างรวดเร็วอาจจะคาดการณ์ได้ว่าผู้ที่ใช้บัตรเครดิตใบนั้นอาจจะไม่ใช่เจ้าของบัตรที่แท้จริงก็เป็นได้ บัตรเครดิตนั้นอาจถูกผู้ไม่หวังดีนำมาใช้ก็เป็นได้ นอกจากนี้คาดการณ์ไมน์นิงยังรวมถึงการทำนายแนวโน้มในปัจจุบันและในอนาคตบนพื้นฐานของประสบการณ์ที่ผ่าน ๆ มา เห็นได้ชัดเจนว่าปริมาณข้อมูลทั้งในอดีตและปัจจุบันจะถูกจัดเก็บจนกลายเป็นฐานข้อมูลที่ใหญ่ยิ่งไปกว่านั้นข้อมูลที่ถูกจัดเก็บดังกล่าวอาจจะมาจากหลาย ๆ แหล่งหรือหลาย ๆ โดเมน ทำให้ยากที่จะได้สารสนเทศที่ถูกต้องและเหมาะสมเพื่อนำมาใช้สนับสนุนในการวางแผนและการตัดสินใจ อย่างไรก็ตามคาดการณ์ไมน์นิงนี่เองที่สามารถสืบค้นความรู้ที่เป็นประโยชน์ ที่น่าสนใจ บนฐานข้อมูลขนาดใหญ่ ๆ ได้เป็นอย่างดี ทำให้ Data Mining เริ่มเป็นที่รู้จักและได้รับความสนใจและเป็นหัวข้อที่ได้รับความนิยมอย่างมากในปัจจุบัน



รูปที่ 1 แสดงการแปลงข้อมูลไปสู่ความรู้ (แหล่ง: ดัดแปลงจาก Fayyad et al., [1996] : 10)

จากรูปที่ 1 เป็นวิธีการค้นหาความรู้ (knowledge discovery process) ซึ่งประกอบไปด้วยการดำเนินการตามวัฏจักรขั้นตอน ดังนี้

Selection	การดึงหรือคัดเลือกข้อมูลที่มีความสัมพันธ์กันในฐานข้อมูลเพื่อเตรียมวิเคราะห์ต่อไป
Preprocessing	การจัดการกับข้อมูลที่ไม่สัมพันธ์กัน หรือข้อมูลที่มี noise หรือ missing
Transformation	การเปลี่ยนรูปข้อมูลให้อยู่ในรูปแบบที่เหมาะสมสำหรับการไมน์นิง
Data Mining	การประยุกต์ขั้นตอนหรือวิธีต่าง ๆ เพื่อสกัดให้ได้มาซึ่งแบบแผนข้อมูลหรือความรู้ที่ใช้ประโยชน์ได้
Interpretation/ Evaluation	การตีความแบบแผนที่ได้จากขั้นตอนของดาต้าไมน์นิงเพื่อให้ได้ความรู้ โดยนำความรู้ที่ได้ไปใช้หรืออาจจัดเก็บอยู่ในรูปฐานความรู้ต่อไป

จากลำดับขั้นตอนดังกล่าวจะพบว่า Data Mining เป็นขั้นตอนหนึ่งที่สำคัญของกระบวนการค้นหาความรู้ นอกจากนี้ยังมีคำนิยามอื่น ๆ ของคำว่า Data Mining ตัวอย่างเช่น คำว่า Knowledge Discovery, Knowledge Extraction, Data Dipping, Data Archeology, Data Exploration, Data Pattern Processing, Data Dredging, Pattern Discovery, Knowledge Mining และ Information Harvesting เป็นต้น ซึ่งอาจจะพบคำดังกล่าวในหนังสือหรือบทความอื่น ๆ ก็ให้เข้าใจว่าคำดังกล่าวกับดาต้าไมน์นิงไม่แตกต่างกันในความหมายมากนัก

แนวคิดและเทคนิคใน Data Mining

ในกระบวนการของดาต้าไมน์นิงจะพบว่ามีหลายขั้นตอนที่เกี่ยวข้องซึ่งรวมถึงการจัดการข้อมูล การเลือกเครื่องมือสำหรับการไมน์นิง การดำเนินการไมน์นิง การกรองและการสกัดสารสนเทศเพื่อให้ได้ความรู้ที่มีประโยชน์ และการประเมินค่าการกระทำเพื่อประโยชน์ในการตัดสินใจ ในส่วนนี้จะกล่าวถึงสารสนเทศอะไรบ้างที่ได้จากการไมน์นิงและเทคนิคของดาต้าไมน์นิง ดังรายละเอียดต่อไปนี้

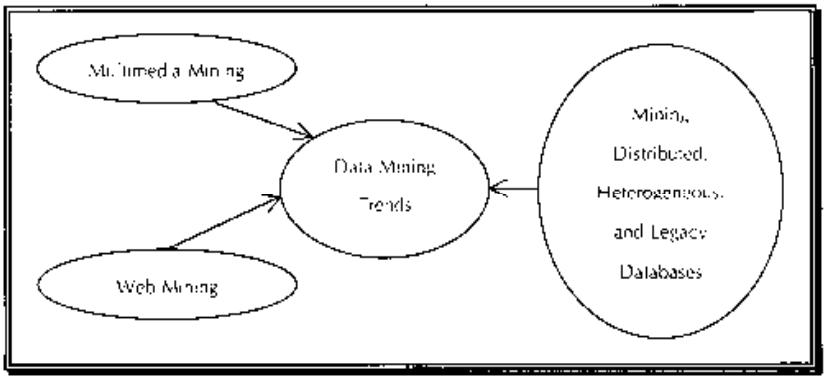
ตัวอย่างสารสนเทศที่ได้จาก Data Mining

Association	เป็นความสัมพันธ์กันระหว่างข้อมูล เช่น ถ้าลูกค้าซื้อรองเท้าแล้ว ถุงเท้าจะเป็นสิ่งที่ลูกค้าจะซื้อ
Classification	การแบ่งข้อมูลออกเป็นกลุ่มหรือหมวดหมู่ย่อยที่มีความหมาย ตัวอย่างเช่น บริษัทขายรถแห่งหนึ่งมีข้อมูลเกี่ยวกับลูกค้าซึ่งอาศัยในเมือง กขค ว่าจะเป็นเจ้าของรถซึ่งมีมูลค่ามากกว่า 200000 บาท สันนิษฐานได้ว่าใครก็ตามที่ไม่ได้มีรายชื่ออยู่ในกลุ่มนี้ แต่ได้อาศัยในเมือง กขค ก็จะเป็นเจ้าของรถที่มีมูลค่ามากกว่า 200000 บาทเช่นกัน ลักษณะเช่นนี้เราได้แบ่งกลุ่มข้อมูลตามลูกค้าที่อาศัยในเมือง กขค
Sequence Detection	ในช่วงระยะเวลาที่ต่อเนื่องกัน ลำดับของเหตุการณ์จะเกี่ยวข้องและสัมพันธ์กันในช่วงระยะเวลาดังกล่าว เช่น หลังจากที่สมชายไปธนาคาร สมชายมักจะไปร้านขายของชำ เป็นต้น
Data Dependency Analysis	การวิเคราะห์การขึ้นต่อกันของข้อมูล เพื่อดูแนวโน้มความสัมพันธ์หรือความเกี่ยวข้องกัน หรือดูการขึ้นต่อกันของข้อมูลที่ถูกสืบค้น เช่น ถ้าสมชาย สมศักดิ์ และสมศรี พบปะประชุมกันครั้งใดแล้วสมฤดีจะอยู่ในที่ประชุมนั้นเสมอ

นอกจากเทคนิคดังกล่าวที่กล่าวมาข้างต้น ยังมีเทคนิคอื่น ๆ ที่ใช้ในดาต้าไมน์นิง เช่น rough sets, fuzzy logic, inductive logic programming, neural networks และ statistical technique ซึ่งใน

ปัจจุบันจะพบงานวิจัยและการพัฒนาต้นแบบโดยใช้เทคนิคต่าง ๆ ของดาต้าไมน์นิงดังกล่าว

ทิศทางและแนวโน้มของ Data Mining



รูปที่ 2 แนวโน้ม Data Mining (แหล่ง: ดัดแปลงจาก Thuraisingham,Bhavani M. 1998 :7)

จากรูปที่ 2 จะพบว่าแนวโน้มของการพัฒนาเครื่องมือ Data Mining จะมีความเกี่ยวข้องกับฐานข้อมูลที่มีปริมาณข้อมูลมากขึ้นเรื่อย ๆ หรือข้อมูลที่มีรูปแบบซับซ้อนยิ่งขึ้น ดังรายละเอียดต่อไปนี้

- ฐานข้อมูลที่มีอยู่ในปัจจุบัน ส่วนใหญ่อยู่ในรูปของฐานข้อมูลแบบ Distributed หรือแบบ Heterogeneous ยิ่งไปกว่านั้นข้อมูลอาจจะพบได้ใน Legacy Databases ดังนั้นเทคนิคการไมน์นิงจึงเป็นสิ่งที่จำเป็นที่จะจัดการกับข้อมูลในฐานข้อมูลประเภทต่าง ๆ ดังกล่าว

- ปริมาณข้อมูลจำนวนมากยังคงเป็นแบบไม่มีโครงสร้าง เช่น ข้อมูลในฐานข้อมูลมัลติมีเดีย ส่วนใหญ่จะเป็นข้อมูลที่มีโครงสร้างและไม่มีโครงสร้าง ดังนั้นในอนาคตอันใกล้อาจมีการพัฒนาเครื่องมือ Data Mining สำหรับฐานข้อมูลมัลติมีเดียขึ้น

อย่างไรก็ตาม ข้อมูลและสารสนเทศที่อยู่บน World Wide Web ซึ่งมีปริมาณเป็นจำนวนมากและเพิ่มขึ้นอย่างรวดเร็ว ในขณะที่เครื่องมือที่จะดึงเอาข้อมูล/สารสนเทศ หรือเพื่อศึกษาพฤติกรรมการใช้ World Wide Web ยังมีประสิทธิภาพไม่เพียงพอ Web Mining จึงเป็นอีกหนึ่งแนวโน้มของเทคโนโลยี Data Mining ที่จะเกิดขึ้นในเร็ว ๆ นี้

นอกจากนี้ข้อมูลที่มีการแปรผัน ข้อมูลที่ไม่สมบูรณ์และข้อมูลที่มีไม่เพียงพอตลอดจนการกำหนด หรือเลือกชนิดของอัลกอริทึมใด ๆ ของ Data Mining และจะทำบนข้อมูลอะไร รวมทั้งการไมน์นิงในรูปแบบหลากหลายภาษาเป็นเรื่องที่ท้าทายและน่าสนใจที่จะศึกษาและพัฒนากันต่อไป เพื่อที่จะก่อให้เกิดสารสนเทศที่มีประโยชน์และนำมาใช้งานได้อย่างแท้จริง

เอกสารอ้างอิง

Berson, Alex and Smith Stephen J.. 1997. **Data Warehousing, Data Mining, and OLAP.** Boston : McGraw-Hill.

Chen, Lei-da, Sakaguchi, Toru, and Frolick, Mark N.. 2000. "Data Mining Methods, Applications, and Tools". **Information Systems Management,** 17(1), p. 65-70.

Cios, Krzysztof J.. 1998. **Data Mining Methods for Knowledge Discovery.** Boston : Kluwer Academic.

Date, C.J.. 2000. **An Introduction to Database Systems.** Reading : Addison-Wesley.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P.. 1996. **Advances in knowledge Discovery and Data Mining.** Menlo Park, Cal : AAA/MIT Press.

Thuraisingham, Bhavani. 1998. **Data Mining : Technologies, Techniques, Tools, and Trends.** Boca Raton : CRC Press.

Turban, Efraim. 1999. **Information Technology for Management : Making Connections for Strategic Advantage.** New York . John Wiley.