รายงานการวิจัย

SUT Miner: ระบบเหมืองข้อมูลที่มีประสิทธิภาพ

(SUT Miner: An efficient data mining system)

ผู้วิจัย

ผู้อำนวยการชุดโครงการวิจัย

รองศาสตราจารย์ ดร.กิตติศักดิ์ เกิดประสพ

สาขาวิชาวิศวกรรมคอมพิวเตอร์

สำนักวิชาวิศวกรรมศาสตร์

หัวหน้าโครงการวิจัยย่อย

รองศาสตราจารย์ ดร. กิตติศักดิ์ เกิดประสพ สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์

รองศาสตราจารย์ ดร.นิตยา เกิดประสพ สาขาวิชาวิศวกรรมคอมพิวเตอร์ สำนักวิชาวิศวกรรมศาสตร์

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ พ.ศ. 2548 และ 2549

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

มิถุนายน 2553

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณมหาวิทยาลัยเทคโนโลยีสุรนารีและสำนักงานคณะกรรมการวิจัยแห่งชาติ ที่ได้จัดสรรงบประมาณในการทำวิจัยให้ในปีงบประมาณ 2548-2549 โครงการนี้ยังได้รับงบประมาณบางส่วน รวมถึงความร่วมมือในการดำเนินงานจากหน่วยปฏิบัติการวิจัยด้านวิศวกรรมข้อมูลและการค้นหาความรู้ (Data Engineering and Knowledge Discovery -- DEKD -- Research Unit) ขอขอบคุณผู้ทรงคุณวุฒิที่ได้เสียสละเวลาทำหน้าที่ตรวจข้อเสนอโครงการ ตรวจร่างรายงานการวิจัยฉบับสมบูรณ์ และให้ข้อเสนอแนะที่เป็นประโยชน์อย่างมากต่อการปรับปรุงรายงานการวิจัยฉบับบสมบูรณ์นี้ ให้มีความถูกต้องครบถ้วนมากยิ่งขึ้น

บทคัดย่อภาษาไทย

การทำเหมืองข้อมูลเป็นเทคโนโลยีใหม่ของการวิเคราะห์ข้อมูลอัตโนมัติ เพื่อค้นหาความรู้ที่จะเป็นประโยชน์แก่การวางแผนดำเนินการของหน่วยงานเจ้าของข้อมูล ความรู้ที่ค้นพบนี้เป็นได้หลายรูปแบบ เช่น แพทเทิร์นที่เกิดขึ้นภายในกลุ่มข้อมูลที่สามารถใช้ทำนายลักษณะที่จะเกิดขึ้นในอนาคตหรือทำนายแนวโน้มการเปลี่ยนแปลงของข้อมูล ลักษณะที่สัมพันธ์เชื่อมโยงกันของข้อมูล ลักษณะที่เบี่ยงเบนไปจากข้อมูลกลุ่มใหญ่ และรูปแบบประเภทอื่นๆ อีกหลากหลาย ปัจจัยสำคัญของการทำเหมืองข้อมูลให้ได้ความรู้ที่มีประโยชน์ คือ อัลกอริทึมสังเคราะห์ความรู้ และ ตัวข้อมูล ถ้าอัลกอริทึมไม่มีประสิทธิภาพก็จะไม่สามารถค้นหาความรู้ที่แฝงอยู่ในข้อมูลได้ หรือถ้าข้อมูลไม่มีคุณภาพเพียงพอก็จะไม่ช่วยให้สังเคราะห์ความรู้ใดออกมาได้ งานวิจัยนี้จึงเสนอขึ้นเพื่อพัฒนาซอฟต์แวร์ชื่อเอสยูทีไมเนอร์ที่มีส่วนประกอบทั้งส่วนฟิลเตอร์ ส่วนค้นหาแพทเทิร์น และส่วนประเมินแพทเทิร์น เอสยูทีไมเนอร์คือระบบเหมืองข้อมูลที่พัฒนาขึ้น เพื่อใช้เป็นเครื่องมือวิเคราะห์ข้อมูลอัจฉริยะสำหรับค้นหารูปแบบหรือสารสนเทศที่เป็นประโยชน์ จากข้อมูลที่เก็บอยู่ในฐานข้อมูล ผู้วิจัยได้นำเสนอวิธีการออกแบบและการพัฒนาเอสยูทีไมเนอร์ ที่จัดเป็นระบบเหมืองข้อมูลสมบูรณ์แบบ เนื่องจากได้รวมส่วนการประมวลผลก่อนและหลังการทำเหมืองข้อมูลเข้าไว้ในระบบเดียวกัน ในรายงานนี้ผู้วิจัยได้นำเสนอกรอบแนวคิดของระบบ และวิธีการพัฒนาระบบในส่วนของโปรแกรมการทำเหมือง โดยใช้วิธีการโปรแกรมเชิงตรรกะด้วยภาษาโปรล็อก วิธีการโปรแกรมเชิงประกาศของโปรล็อกช่วยให้การเขียนคำสั่งทำได้สั้นและชัดเจน นอกจากนี้การสนับสนุนการทำแพทเทิร์นแมทชิงของภาษาโปรล็อกยังเป็นข้อได้เปรียบอย่างมากสำหรับงานการค้นหารูปแบบ ผลสำเร็จของการพัฒนาซอฟต์แวร์นี้จะเป็นจุดเริ่มต้นของการสร้างระบบทำเหมืองข้อมูลสมรรถนะสูงที่ทำงานกับข้อมูลขนาดใหญ่มากและมีลักษณะการเกิดขึ้นอย่างต่อเนื่องได้ต่อไปในอนาคต

บทคัดย่อภาษาอังกฤษ

Data mining is a new technology in automatic data analysis. It is the search for valuable information, or knowledge, in large volumes of data. The discovered knowledge can greatly benefit the organization that owns the data in many ways, for example, to aid decision-making, to reduce the risk of business planning or to project the revenue in future investment. The common types of knowledge discovery are prediction, deviation detection, database segmentation, clustering, association rules and link analysis. Despite the many forms and functions of discovered knowledge, the most important factors in mining valuable knowledge are the learning algorithm and data. The practical data mining system needs an efficient algorithm and high quality data. This research project presents the design and implementation of a complete data mining system, called SUT Miner. The SUT Miner is a data mining system developed as an intelligent data analysis tool to discover patterns and extract useful information from facts stored in databases. In addition to the mining engine, our system incorporates the pre-mining and post-mining parts. In this report, we describe a framework of SUT Miner and present the implementation scheme on the mining engine part using a logic programming paradigm, a Prolog language in particular. A high-level declarative style of Prolog facilitates a clear and concise coding. The language also supports pattern matching, which is a big advantage for a task of pattern discovery. The success of this project will be the preliminary step toward the future development of a high-performance data mining system that can handle very large volume of continuously generated data.

สารบัญ

หน้า

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

หน้า

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาที่ทำการวิจัย

การขุดค้นและวิเคราะห์ข้อมูลในฐานข้อมูลเพื่อค้นหาความรู้ เรียกว่า *การทำเหมืองข้อมูล* (data mining) เป็นเทคโนโลยีใหม่ของการประยุกต์ใช้ข้อมูลที่เก็บอยู่ในฐานข้อมูลให้เกิดประโยชน์สูงสุดแก่หน่วยงานที่เป็นเจ้าของข้อมูล การประยุกต์ใช้ข้อมูลที่กล่าวถึงนี้มีได้หลายแนวทาง แต่โดยทั่วไปมักจะเป็นการสรุปภาพรวมของข้อมูลในฐานข้อมูล, การจัดกลุ่มข้อมูลอัตโนมัติ, การวิเคราะห์แนวโน้มการเปลี่ยนแปลงของข้อมูล, หรือ การค้นหาความสัมพันธ์ที่ซ่อนอยู่ภายในกลุ่มของข้อมูล เช่น จากข้อมูลการซื้อสินค้าของลูกค้าซุปเปอร์มาเก็ตในรอบหนึ่งเดือนที่ผ่านมา เมื่อวิเคราะห์หาความสัมพันธ์แล้วพบว่า *"ถ้าลูกค้าซื้อนมและขนมปังแล้ว ลูกค้าจะซื้อบะหมี่สำเร็จรูปด้วย"* ผลลัพธ์ที่ได้นี้คือ ความรู้เกี่ยวกับพฤติกรรมของผู้บริโภค ที่จะช่วยให้ผู้จัดการซุปเปอร์มาเก็ตสามารถวางแผนการจัดวางสินค้าและการจัดโปรแกรมกระตุ้นยอดขายสินค้าบางรายการได้อย่างมีประสิทธิภาพ

การวิเคราะห์ข้อมูลด้วยโปรแกรมช่วยงาน เช่น SPSS, SAS นักวิเคราะห์ข้อมูลจะต้องเป็นผู้กำหนดว่าจะศึกษาลักษณะใดจากข้อมูลและจะใช้ข้อมูลส่วนใดบ้าง แต่โปรแกรมทำเหมืองข้อมูลจะกระทำขั้นตอนต่างๆ เหล่านี้ให้โดยอัตโนมัติ โปรแกรมทำเหมืองข้อมูลมีความสามารถที่จะค้นหาแนวโน้ม รูปแบบร่วม หรือลักษณะอื่นๆที่น่าสนใจ โดยไม่ต้องพึ่งพาการสั่งงานทุกขั้นตอนจากนักวิเคราะห์ข้อมูล นอกจากนี้ยังอาจจะสามารถค้นพบลักษณะที่น่าสนใจจากข้อมูลซึ่งนักวิเคราะห์ข้อมูลไม่ได้คาดหมายมาก่อน

ระบบเหมืองข้อมูล (data mining system) มีความแตกต่างจากระบบผู้เชี่ยวชาญ (expert system) ตรงที่ฐานความรู้ของระบบเหมืองข้อมูลได้จากการสังเคราะห์จากข้อมูลโดยตรง สามารถปรับปรุงฐานความรู้ของตัวเองได้อัตโนมัติตามข้อมูลใหม่ที่ได้รับเพิ่มขึ้น ซึ่งจะแตกต่างจากระบบผู้เชี่ยวชาญที่ฐานความรู้ถูกป้อนเข้ามาในระบบโดยวิศวกรความรู้ (knowledge engineer) และจะคงตัวอยู่เช่นนั้นตลอดการใช้งานจนกว่าวิศวกรความรู้จะปรับปรุงฐานความรู้

การประยุกต์ใช้เทคโนโลยีการทำเหมืองข้อมูลมีได้หลากหลายลักษณะ สามารถจัดกลุ่มกว้างๆได้เป็นสองกลุ่ม คือ กลุ่มที่ใช้การทำเหมืองข้อมูลเพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย

การทำการทำเหมืองข้อมูลเพื่อการทำนาย เป็นการนำความรู้ที่เรียนรู้มาจากข้อมูลที่มีอยู่เพื่อประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต เช่น จากข้อมูลลูกค้าของแผนกสินเชื่อ

ของธนาคารที่ได้มีการจัดกลุ่มของลูกค้าไว้แล้วว่าใครเป็นลูกค้าชั้นดี ใครเป็นลูกค้าในระดับปานกลาง และใครเป็นลูกค้าที่มักจะผิดนัดชำระหนี้ โปรแกรมการทำเหมืองข้อมูลสามารถเรียนรู้จากข้อมูลเหล่านี้และค้นหาโมเดลที่สามารถใช้อธิบายลักษณะของลูกค้าชั้นดี ลูกค้าระดับปานกลาง และลูกค้าที่ไม่เป็นที่ต้องการ จากโมเดลที่ได้นี้สามารถนำไป ใช้ทำนายลูกค้าใหม่ที่มาขอสินเชื่อได้ว่าเขาน่าจะเป็นลูกค้าประเภทใด

การทำเหมืองข้อมูลเพื่อการอธิบาย เป็นการค้นหารูปแบบหรือแพทเทิร์นบางอย่างที่น่าสนใจจากกลุ่มข้อมูล รูปแบบนี้มักจะเป็นความสัมพันธ์ หรือลักษณะที่เชื่อมโยงกันของข้อมูล การทำเหมืองข้อมูลแบบนี้ต่างจากแบบแรกตรงที่ ผู้ใช้ไม่ได้กำหนดล่วงหน้าว่าจะให้โปรแกรมทำเหมืองข้อมูลค้นหารูปแบบหรือโมเดลของอะไร แต่ให้ค้นหาทุกรูปแบบที่น่าสนใจจากข้อมูล

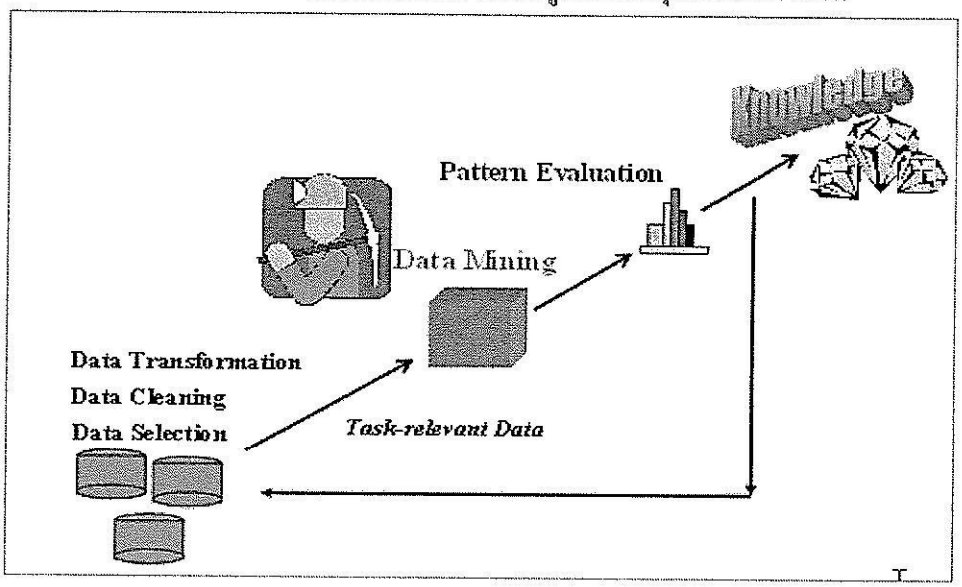
การทำเหมืองข้อมูลทั้งแบบเพื่อการทำนายและเพื่อการอธิบาย จัดเป็นการเรียนรู้ของเครื่องแบบที่ต้องรับการชี้แนะ (supervised learning) การเรียนรู้ของเครื่องในแบบที่ยากขึ้น คือ แบบที่ไม่ต้องชี้แนะ (unsupervised learning) ได้แก่ การจัดกลุ่มข้อมูลอัตโนมัติ (clustering)

งานทำเหมืองข้อมูลที่มีได้หลากหลายรูปแบบนี้ เกิดขึ้นเนื่องจากความรู้ที่ต้องการได้จากข้อมูลมีได้หลายลักษณะ ตัวอย่างต่อไปนี้แสดงผลสำเร็จส่วนหนึ่งของการนำเทคโนโลยีการทำเหมืองข้อมูลไปประยุกต์ใช้

- ด้านการแพทย์: ใช้โปรแกรมทำเหมืองข้อมูลค้นหาผลข้างเคียงที่ไม่ได้คาดหมายมาก่อนของการใช้ยา โดยอาศัยข้อมูลจากแฟ้มประวัติผู้ป่วย, ใช้ในการวิเคราะห์หาความสัมพันธ์ของสารพันธุกรรม
- ด้านการเงิน: ใช้โปรแกรมทำเหมืองข้อมูลช่วยในการตัดสินใจว่าควรจะอนุมัติเครดิตให้ลูกค้ารายใดบ้าง, ใช้ในการคาดหมายความน่าจะเป็นว่าธุรกิจนั้นๆมีโอกาสที่จะล้มละลายหรือไม่, ใช้คาดหมายการขึ้น/ลงของหุ้นในตลาดหลักทรัพย์
- ด้านการเกษตร: ใช้โปรแกรมทำเหมืองข้อมูลจำแนกประเภทของโรคพืชที่เกิดกับถั่วเหลืองและมะเขือเทศ
- ด้านวิศวกรรม: ใช้โปรแกรมทำเหมืองข้อมูลวิเคราะห์และวินิจฉัยสาเหตุการทำงานผิดพลาดของเครื่องจักรกล
- ด้านอาชญวิทยา: ใช้โปรแกรมทำเหมืองข้อมูลวิเคราะห์หาเจ้าของลายนิ้วมือ
- ด้านอวกาศ: ใช้โปรแกรมทำเหมืองข้อมูลวิเคราะห์ข้อมูลที่ส่งมาจากดาวเทียมขององค์การนาซา

กระบวนการทำเหมืองข้อมูล เปรียบได้กับการทำเหมืองแร่ที่เราใช้เครื่องจักรคัดแยกแร่ที่มีค่าออกจากกองหิน กรวด ดินที่ปะปนมากับสายแร่ เพียงแต่ในกระบวนการทำเหมืองข้อมูลสิ่งที่เราได้จากกองข้อมูลมหาศาล คือ ความรู้ (knowledge) ที่ซ่อนอยู่ในกองข้อมูล กระบวนการนี้แสดงเป็นแผนภาพได้ดังรูปที่ 1.1

ความรู้ที่ได้นี้จะช่วยให้เราเข้าใจลักษณะของข้อมูล และเข้าใจปัจจัยที่ทำให้เกิดลักษณะบางอย่างขึ้นในข้อมูลบางกลุ่ม ซึ่งจะช่วยให้เราสามารถทำนายแนวโน้มของข้อมูลใหม่ที่จะเกิดขึ้นในอนาคตได้ รวมถึงเข้าใจความสัมพันธ์ที่เชื่อมโยงข้อมูลแต่ละกลุ่มย่อยเข้าด้วยกัน



รูปที่ 1.1 กระบวนการทำเหมืองข้อมูลเพื่อคัดแยกความรู้ออกจากข้อมูลดิบ

โดยทั่วไปกระบวนการทำเหมืองข้อมูล จะประกอบด้วย 4 ขั้นตอนใหญ่ๆคือ

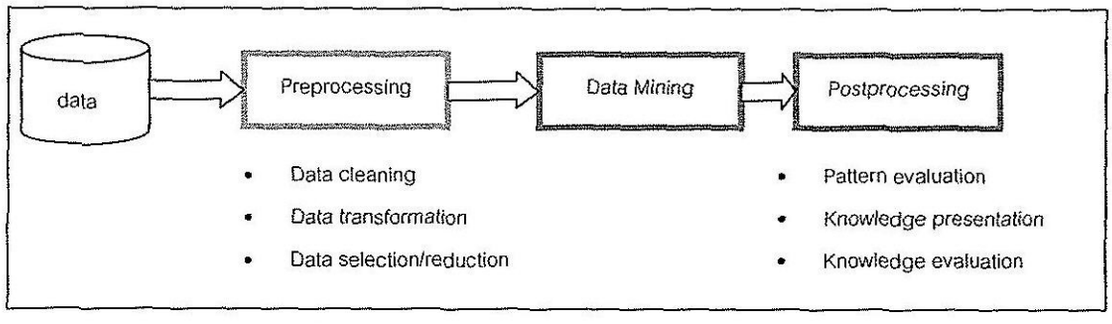
1) **เตรียมข้อมูล (data preparation)** : ประกอบด้วยขั้นตอนย่อย คือ การคัดเลือกข้อมูล (data selection) เพื่อให้ได้เฉพาะข้อมูลที่จะเป็นประโยชน์ต่อการทำเหมืองข้อมูล การคัดเลือกข้อมูลอย่างเหมาะสม จะช่วยให้ค้นหาความรู้ที่เป็นประโยชน์ได้อย่างรวดเร็ว นอกจากนี้ยังต้องมีการปรับปรุงคุณภาพข้อมูล (data cleaning) ซึ่งจำเป็นในกรณีที่ข้อมูลดิบมีส่วนที่ไม่สมบูรณ์ หรือมีข้อมูลบางส่วนบกพร่อง มีข้อมูลรบกวน (noise)ปะปนอยู่ และถ้าข้อมูลไม่อยู่ในรูปแบบที่ถูกต้องหรือเหมาะสม จะต้องมีการปรับเปลี่ยนรูปแบบข้อมูล (data transformation) ให้อยู่ในรูปแบบที่โปรแกรมทำเหมืองข้อมูลจะเรียกใช้งานได้
2) **ค้นหาโมเดลจากข้อมูล (mining)** : กระบวนการค้นหาโมเดลหรือความสัมพันธ์ จะเริ่มจากข้อมูลเริ่มต้นจำนวนไม่มากนัก จากนั้นนำผลที่ได้จากอัลกอริทึมค้นหาโมเดล (mining algorithm) ไปยืนยันกับข้อมูลทดสอบ ถ้าผลที่ได้ยังไม่น่าพอใจอาจจะต้อง

ปรับค่าพารามิเตอร์บางตัวของ mining algorithm และเริ่มกระบวนการค้นหาใหม่กับข้อมูลจำนวนมากขึ้น จนกว่าผลที่ได้มีความถูกต้องอยู่ในระดับที่ยอมรับได้ จึงจะจบกระบวนการค้นหา

3) **ตรวจสอบและวิเคราะห์ผล (pattern evaluation)** : โมเดลหรือความสัมพันธ์ที่หามาได้ในขั้นตอนที่ 2 จะต้องถูกนำมาทดสอบอัตราความผิดพลาดและวิเคราะห์ความซับซ้อนของรูปแบบโมเดล ถ้าอัตราความผิดพลาดยังสูงเกินไป อาจจะต้องย้อนกลับไปที่ขั้นตอนที่ 2 อีกครั้ง เพื่อปรับปรุงโมเดลให้ถูกต้องยิ่งขึ้น ในทำนองเดียวกัน ถ้าโมเดลที่หามาได้มีรูปแบบที่ซับซ้อนเกินไปจนยากต่อการทำความเข้าใจ อาจจะต้องย้อนกระบวนการกลับไป เพื่อให้หาโมเดลใหม่ที่มีความถูกต้องเท่าเดิมแต่มีรูปแบบที่ซับซ้อนน้อยลง

4) **คัดเลือกโมเดลและกำหนดวิธีแสดงผล (knowledge processing and representation)** : โมเดลหรือความรู้ที่ได้จาก mining algorithm มักจะมีปริมาณมาก และรูปแบบการแสดงความรู้ที่ยากต่อการทำความเข้าใจ จึงต้องมีกระบวนการภายหลังการทำเหมืองข้อมูล (postprocessing) เพื่อประเมินและคัดเลือกความรู้นำเสนอต่อผู้ใช้

จากทั้งสี่ขั้นตอนในกระบวนการทำเหมืองข้อมูล เริ่มต้นตั้งแต่การนำเข้าข้อมูลจนกระทั่งได้ความรู้ที่นำไปใช้ประโยชน์ได้ สรุปภาพรวมของกระบวนการได้ดังรูปที่ 1.2



รูปที่ 1.2 กระบวนการทำเหมืองข้อมูลโดยสรุป

จากภาพรวมของกระบวนการทำเหมืองข้อมูล จะเห็นได้ว่าการทำเหมืองข้อมูลเป็นศาสตร์ที่ต้องรวมความก้าวหน้าจากหลายสาขา ได้แก่ database technology, artificial intelligence, machine learning, statistics, neural networks, pattern recognition, knowledge-based systems, knowledge acquisition, information retrieval, high-performance computing และ data visualization การทำเหมืองข้อมูลเป็นศาสตร์แขนงใหม่ที่เริ่มเป็นที่รู้จักในช่วงปลายทศวรรษที่ 1980 Piatetsky-Shapiro และ Frawley (1991) ได้รวบรวมงานวิจัยในยุคแรกของศาสตร์นี้ไว้ในหนังสือชื่อ

Knowledge Discovery in Databases และได้เริ่มมีการรวมตัวของนักวิจัยในสาขานี้จัดการประชุมทางวิชาการเพื่อเสนอความก้าวหน้าของงานวิจัยภายใต้ชื่อการประชุม International Conference on Knowledge Discovery and Data Mining (KDD) ตั้งแต่ปี 1995 ปัจจุบันศาสตร์ทางด้านการทำเหมืองข้อมูลได้รับความสนใจจากนักวิจัยจำนวนมากทั่วโลกในสาขาต่างๆ ที่เกี่ยวข้อง ค้นคว้าเพื่อหาแนวทางที่จะทำให้กระบวนการทำเหมืองข้อมูลในทุกขั้นตอนมีประสิทธิภาพสูงขึ้น และทำงานกับข้อมูลขนาดใหญ่มากได้

การทำเหมืองข้อมูลให้มีประสิทธิภาพ จะต้องเริ่มตั้งแต่การเตรียมข้อมูลให้ได้ข้อมูลที่มีคุณภาพ นักวิจัยจำนวนมาก เช่น Redman (1992), Wang et al. (1995), Wand และ Wang (1996), Kennedy et al. (1998), Weiss และ Indurkhya (1998), Pyle (1999), Ballou และ Tayi (1999) ได้อธิบายลักษณะของข้อมูลที่มีคุณภาพและเสนอแนะเทคนิคที่สามารถนำมาใช้ช่วยเพิ่มคุณภาพข้อมูล ลักษณะที่มักจะพบบ่อยในกลุ่มของข้อมูลที่ไม่มีคุณภาพ ได้แก่ ข้อมูลบางส่วนขาดหายไป (missing values) มีผู้เสนอแนวทางจัดการกับกรณีเช่นนี้ไว้หลายแนวทางดังปรากฏในงานวิจัยของ Friedman (1977), Breiman et al. (1984) และ Quinlan (1989)

ประโยชน์ของการทำเหมืองข้อมูลจะเห็นได้ชัดเจนเมื่อข้อมูลมีปริมาณมาก แต่ปัญหาที่เกิดตามมาคือระบบเหมืองข้อมูลไม่สามารถรองรับข้อมูลขนาดมหาศาลนั้นได้ แนวทางแก้ปัญหานี้คือใช้วิธีลดขนาดข้อมูล เทคนิคการลดขนาดข้อมูลมักจะกระทำในสองแนวทาง คือ ลดลักษณะ (attributes or features) ที่อธิบายข้อมูลแต่ละรายการให้เหลือเฉพาะลักษณะหลักที่สำคัญ งานวิจัยในแนวทางนี้จะอยู่ในกลุ่มที่เรียกว่า feature selection รายละเอียดปรากฏในหนังสือและงานวิจัยจำนวนมาก (Neter et al., 1996; Dash and Liu, 1997; Kohavi and John, 1997; Dash et al.,1997; Liu and Motoda, 1998)

แนวทางที่สองในการลดขนาดของข้อมูล ใช้วิธีการลดจำนวนรายการข้อมูลให้เหลือเฉพาะข้อมูลที่สามารถเป็นตัวแทนของข้อมูลกลุ่มใหญ่ได้ เทคนิคที่นิยมใช้ลดจำนวนข้อมูลคือ การสุ่ม (sampling) การจัดกลุ่ม (clustering) และการใช้ฮิสโตแกรม (histogram) John และ Langley (1996) ได้ศึกษา static และ dynamic sampling ในขณะที่ Josien และคณะ (2001) ได้ทดลองศึกษาพฤติกรรมของโปรแกรมทำเหมืองข้อมูลเมื่อข้อมูลถูกลดขนาดลง Kivinen และ Mannila (1993) ได้วิเคราะห์ในเชิงทฤษฎีถึงจำนวนข้อมูลที่สามารถเป็นตัวแทนของข้อมูลกลุ่มใหญ่ได้ Barbara และคณะ (1997) รวมถึง Devore และ Peck (1997) ได้นำเทคนิคฮิสโตแกรมมาใช้วิเคราะห์การกระจายของข้อมูลเพื่อสุ่มข้อมูลจากแต่ละกลุ่ม

งานวิจัยและพัฒนาทั้งหลายที่เกี่ยวข้องกับการเตรียมข้อมูลมักจะทำเพียงเฉพาะบางส่วนของการเตรียมข้อมูล เช่น แก้ไขในกรณีข้อมูลบางส่วนสูญหาย (missing values) หรือเสนอเทคนิคเฉพาะบางเทคนิคเพื่อให้ข้อมูลมีขนาดที่เหมาะสมกับโปรแกรมทำเหมืองข้อมูล แต่โครงการวิจัยนี้

ได้พัฒนาทั้งสามขั้นตอนของการเตรียมข้อมูลคือ data transformation, data cleaning และ data reduction ข้อมูลที่ผ่านการเตรียมแล้วจะสามารถถูกส่งต่อให้กับระบบเหมืองข้อมูลขั้นต่อไปได้

การพิจารณาความน่าสนใจ (interestingness) ของความรู้ที่ค้นพบโดยการทำเหมืองข้อมูลเป็นอีกประเด็นปัญหาสำคัญ ที่ได้รับความสนใจจากนักวิจัยจำนวนมากตั้งแต่ยุคเริ่มต้นของงานวิจัยด้านการทำเหมืองข้อมูล (Piatetsky-Shapiro and Matheus, 1994; Piatetsky-Shapiro et al., 1994) เนื่องจากได้เห็นถึงความสำคัญว่าความรู้ที่จะนำไปใช้ประโยชน์ได้ จะต้องเป็นความรู้ที่เป็นการค้นพบใหม่และน่าสนใจ โดย Piatetsky-Shapiro และ Matheus (1994) ได้พัฒนาระบบชื่อ KEFIR ขึ้นมาเพื่อใช้กับงานด้านการประกันสุขภาพ ระบบจะพิจารณาข้อมูลที่เบี่ยงเบนไปจากข้อมูลปกติ และตัดสินว่าข้อมูลที่เบี่ยงเบนนี้เป็นข้อมูลที่ควรค่าแก่การให้ความสนใจ การพิจารณาเกณฑ์ความน่าสนใจโดยระบบนี้จึงเป็นการพิจารณาในขอบเขตที่จำกัดมาก ในช่วงปีต่อมา Silberschatz และ Tuzhilin (1995; 1996) ได้ร่วมมือกันเสนอวิธีการพิจารณาความน่าสนใจของความรู้ที่ค้นพบโดยใช้ probabilistic belief เป็นแนวทางในการพิจารณาความน่าจะเป็นว่าความรู้ใดที่ผู้ใช้จะเห็นว่าน่าสนใจ

ปัจจัยที่นักวิจัยใช้พิจารณาประเด็นความน่าสนใจมีได้หลายปัจจัย ได้แก่ จำนวนข้อมูลที่ความรู้นั้นสามารถอธิบายได้ หรือที่เรียกว่า ความครอบคลุม (coverage) ความถูกต้องของความรู้ (confidence) ความไม่แปรผันของความรู้ (strength) ความมีนัยสำคัญของความรู้ (significance) ความเข้าใจง่ายของความรู้ (simplicity) ความใหม่ของการค้นพบ (unexpectedness) และการเป็นความรู้ที่นำไปสู่การปฏิบัติได้ (actionability) ปัจจัยเหล่านี้ถูกหยิบยกขึ้นมาพิจารณาโดยนักวิจัยหลายคณะ (Major and Mangano 1993; Piatetsky-Shapiro and Matheus 1994; Quinlan 1992)

ปัจจัยในด้าน coverage, confidence, strength, significance, simplicity เป็นปัจจัยที่สามารถสร้างเทคนิคขึ้นมาจัดการได้โดยไม่ต้องใช้ข้อมูลหรือคำแนะนำประกอบจากผู้ใช้ หรือจากความรู้อื่นที่เกี่ยวเนื่องกัน (domain knowledge) แต่ปัจจัยด้าน unexpectedness และ actionability เป็นปัจจัยที่พิจารณาตัดสินได้ยาก จึงจำเป็นต้องใช้ฮิวริสติก (heuristics) ช่วยในการพิจารณา

ในส่วนของการจัดการกับปริมาณความรู้ ที่เป็นผลผลิตของกระบวนการทำเหมืองข้อมูลและมักจะได้ความรู้ที่ไม่เป็นประโยชน์ปะปนออกมาเป็นจำนวนมาก เป็นปัญหาที่มีผลกระทบอย่างมากกับงานทำเหมืองข้อมูลโดยเฉพาะงานประเภท association (Bayardo et al., 1999; Brin et al., 1991; Frawley et al., 1991; Klemettinen et al., 1994; Suzuki 1997) นักวิจัยส่วนใหญ่จะใช้วิธีจัดการระหว่างกระบวนการทำเหมืองข้อมูล เช่นใช้วิธีตัดทิ้ง (pruning) เส้นทางค้นหาที่คาดว่าจะนำไปสู่ความรู้ที่ไม่เกิดประโยชน์ นักวิจัยในกลุ่มนี้ได้แก่ Quinlan (1992), Breiman และคณะ (1984), Clark และ Matwin (1993) นักวิจัยในอีกกลุ่มจะใช้วิธีนำความรู้อื่น (domain knowledge) มาใช้ช่วยในระหว่างการค้นหาความรู้ โดยใช้สมมุติฐานว่าความรู้อื่นๆที่เกี่ยวข้องจะช่วยในการ

ตัดสินใจขจัดทิ้งผลลัพธ์ของการทำเหมืองข้อมูลที่จะไม่ก่อประโยชน์กับงาน นักวิจัยในกลุ่มนี้ได้แก่ Ortega และ Fisher (1995), Clark และ Matwin (1993), Pazzani และ Kibler (1992)

ถึงแม้แนวทางของเทคนิค pruning และเทคนิคการใช้ domain knowledge จะช่วยลดปริมาณความรู้ที่ไม่ก่อประโยชน์ แต่วัตถุประสงค์ของนักวิจัยในทั้งสองกลุ่มนี้มุ่งเน้นไปที่การลดปริมาณความรู้เพื่อเพิ่มประสิทธิภาพของกระบวนการค้นหาความรู้ (mining process) มากกว่าจะเพื่อเอื้อประโยชน์ให้กับผู้ใช้ ให้สามารถใช้ประโยชน์จากผลลัพธ์ของการทำเหมืองข้อมูลได้อย่างเต็มประสิทธิภาพและเกิดประโยชน์ในเวลาอันรวดเร็ว

ในระยะหลังของการวิจัยในสาขาการทำเหมืองข้อมูล นักวิจัยเริ่มตระหนักมากขึ้นว่ากระบวนการจัดการกับผลลัพธ์ (หรือ ความรู้) ที่ได้หลังจากการทำเหมืองข้อมูลเป็นสิ่งจำเป็นเพื่อให้การใช้ประโยชน์จากการทำเหมืองข้อมูลเป็นอัตโนมัติและทันต่อความต้องการมากขึ้น ดังจะเห็นได้จากงานวิจัยของ Wang และคณะ (1998), Adomavicius และ Tuzhilin (2001)

ซอฟต์แวร์ระบบเหมืองข้อมูลที่มีอยู่ในปัจจุบัน ส่วนใหญ่จะเน้นการพัฒนาเฉพาะส่วนค้นหาความรู้ (mining algorithms) โดยให้ขั้นตอนการเตรียมข้อมูลและการคัดเลือกข้อมูล (preprocessing) เป็นความรับผิดชอบของผู้ใช้ ซึ่งผู้ใช้ส่วนใหญ่จะต้องใช้เวลามากกว่า 80% ของการทำเหมืองข้อมูลในการเตรียมและคัดเลือกข้อมูล นอกจากนี้ซอฟต์แวร์ระบบเหมืองข้อมูลเกือบทั้งหมดจะแสดงผลลัพธ์ในปริมาณที่มากเกินไป ทำให้ผู้ใช้ต้องเป็นผู้เชี่ยวชาญด้านการวิเคราะห์ข้อมูลจึงจะสามารถแปลผล และคัดเลือกผลลัพธ์เพื่อนำไปใช้ประโยชน์ในงานบริหารจัดการหรือการตัดสินใจใดๆต่อไปได้

คณะผู้วิจัยจึงมีแนวคิดที่จะริเริ่มโครงการขนาดใหญ่นี้ขึ้น ด้วยวัตถุประสงค์เพื่อนำไปสู่การพัฒนาระบบเหมืองข้อมูลที่มีขีดความสามารถสูงในชื่อของระบบเอสยูทีไมเนอร์ (SUT Miner) ที่มีความสมบูรณ์ครบถ้วน ในแง่ของการเตรียมฟังก์ชันที่จะอำนวยความสะดวกให้กับผู้ใช้ในการนำเข้าข้อมูล การเตรียมและคัดเลือกข้อมูลอย่างมีประสิทธิภาพ (ส่วน Preprocessing ในรูปที่ 1.2) มีฟังก์ชันการทำเหมืองข้อมูลแบบจัดกลุ่ม (clustering) และมีฟังก์ชันการทำเหมืองข้อมูลแบบจำแนก (classification) ที่สามารถนำไปใช้ประโยชน์ได้ในงานเพื่อการทำนาย (prediction) นอกจากนี้ยังได้จัดเตรียมฟังก์ชันกลั่นกรองความรู้ที่ได้จาก mining algorithms ให้ได้ผลลัพธ์ที่ตรงกับความต้องการของผู้ใช้มากที่สุดและง่ายต่อการแปลผลเพื่อนำไปสู่การใช้ประโยชน์จริง (ส่วน Postprocessing ในรูปที่ 1.2)

ระบบเหมืองข้อมูล SUT Miner ได้รับการพัฒนาขึ้นในลักษณะของโอเพนซอร์ส หรือเป็นซอฟต์แวร์ที่เปิดเผยซอร์สโค้ดและเป็นฟรีแวร์ เพื่อให้ผู้ใช้โดยทั่วไปสามารถนำไปใช้งาน หรือปรับปรุงซอร์สโค้ดให้ตรงกับความต้องการได้โดยไม่มีค่าใช้จ่าย ภาษาที่ใช้ในการพัฒนาโปรแกรมคือภาษาโปรล็อก ในรูปแบบของ SWI Prolog ที่เป็นฟรีแวร์เช่นเดียวกัน (ดาวน์โหลดได้จาก

www.swi-prolog.org) ผลสำเร็จของชุดโครงการนี้จึงคาดว่าจะช่วยให้ประเทศชาติประหยัดค่าใช้จ่ายในด้านการจัดซื้อซอฟต์แวร์ระบบเหมืองข้อมูลได้อย่างมาก และยังเป็นการพัฒนาเทคโนโลยีการทำเหมืองข้อมูลให้เอื้อประโยชน์ต่อระบบสนับสนุนการตัดสินใจ (decision support system) ได้มากขึ้น เนื่องจากผู้บริหารโดยทั่วไปที่ไม่ใช่นักวิเคราะห์ข้อมูลและนักคอมพิวเตอร์ จะสามารถใช้งานระบบเหมืองข้อมูลนี้ได้ง่ายขึ้นกว่าซอฟต์แวร์ที่มีอยู่ในปัจจุบัน

ผลสำเร็จของการพัฒนาส่วนกลั่นกรองความรู้ (postprocessing) จะเป็นนวัตกรรมใหม่ของเทคโนโลยีการทำเหมืองข้อมูล ให้ก้าวไปสู่การสร้างระบบผู้เชี่ยวชาญ (expert system) อย่างเป็นกระบวนการอัตโนมัติมากขึ้น เนื่องจากความรู้ที่เป็นผลลัพธ์จากการทำเหมืองข้อมูลจะได้รับการประเมินและกลั่นกรองความรู้ที่เป็นประโยชน์และแม่นตรง ส่งไปเก็บไว้ในฐานความรู้ (knowledge base) ของระบบผู้เชี่ยวชาญ ซึ่งจะช่วยให้งานของวิศวกรความรู้ (knowledge engineer) ง่ายขึ้นและรวดเร็วมากขึ้น

การดำเนินงานในชุดโครงการวิจัยนี้ นอกจากจะเป็นการสร้างนวัตกรรมด้านระบบเหมืองข้อมูล และช่วยพัฒนาความสามารถและทักษะของนักวิจัยแล้ว ยังเป็นการช่วยพัฒนาผู้ช่วยวิจัยให้เป็นบุคลากรที่มีความรู้ ความเชี่ยวชาญทางด้านเทคโนโลยีการทำเหมืองข้อมูลและการวิเคราะห์ข้อมูลอัตโนมัติ ซึ่งจะเป็นกำลังสำคัญในการพัฒนาประเทศชาติได้อย่างยั่งยืนต่อไป

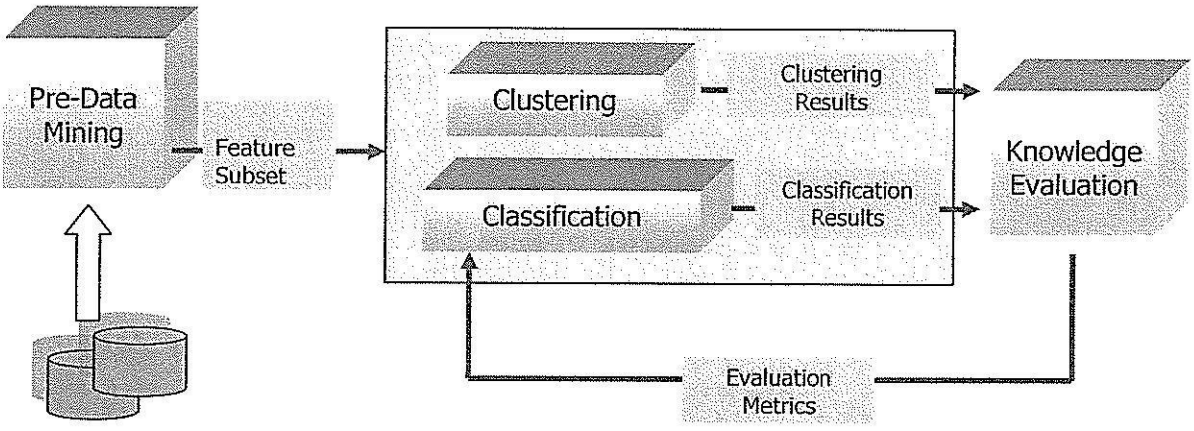
1.2 วัตถุประสงค์หลักของแผนงานวิจัย

- เพื่อพัฒนาระบบเหมืองข้อมูลที่มีความสมบูรณ์ครบถ้วนในลักษณะ integrated system ที่มีทั้งส่วน preprocessing ส่วน mining และส่วน postprocessing
- เพื่อออกแบบและพัฒนากระบวนการเตรียมข้อมูล และคัดเลือกข้อมูลให้เป็นวิธีการอัตโนมัติมากขึ้น
- เพื่อพัฒนา mining algorithms ประเภท clustering และ classification
- เพื่อพัฒนาแนวทางใหม่ในการคัดเลือกและกลั่นกรองความรู้ที่เป็นผลลัพธ์เบื้องต้นจากการทำเหมืองข้อมูล
- เพื่อสร้างเทคโนโลยีขึ้นใช้เองโดยไม่ต้องซื้อจากต่างประเทศ และเพิ่มความสามารถและทักษะให้กับนักวิจัยไทยให้สามารถแข่งขันได้ในระดับโลก

1.3 กรอบของแผนงานวิจัย

แผนงานวิจัยที่เสนอขึ้นนี้จะเน้นการพัฒนาระบบเหมืองข้อมูลให้เป็นลักษณะระบบรวม (integrated system) กล่าวคือจะเป็นระบบที่ครอบคลุมส่วนประกอบหลักครบทั้งสามส่วน คือ ส่วน

การเตรียมข้อมูล ซึ่งเป็นขั้นตอนก่อนการทำเหมืองข้อมูล(pre-data mining) ส่วนการทำเหมืองข้อมูล(mining) ที่ทำหน้าที่ค้นหาความรู้ที่ซ่อนอยู่ในข้อมูลโดยจะมีวิธีการค้นหาความรู้ในสองรูปแบบ คือ แบบจัดกลุ่มข้อมูล (clustering) และแบบจำแนกประเภทข้อมูล (classification) ส่วนกลั่นกรองความรู้ (knowledge evaluation) จะเป็นขั้นตอนหลังการทำเหมืองข้อมูล (post-data mining) โดยแสดงส่วนประกอบทั้งหมดนี้เป็นแผนภาพได้ดังรูปที่ 1.3



รูปที่ 1.3 แผนภาพรวมของระบบเหมืองข้อมูล SUT Miner

ในการพัฒนาระบบเหมืองข้อมูลให้มีส่วนประกอบครบถ้วนสมบูรณ์ดังแผนภาพ จำเป็นต้องประกอบด้วยการพัฒนาในส่วนย่อย คือ

ส่วนที่ 1 Pre-data mining ทำหน้าที่คัดเลือกข้อมูลและปรับปรุงคุณภาพข้อมูล (โครงการวิจัยที่ 1)

ส่วนที่ 2 Clustering data mining ทำหน้าที่จัดกลุ่มข้อมูลโดยอัตโนมัติ (โครงการวิจัยที่ 2)

ส่วนที่ 3 Classification data mining ทำหน้าที่ค้นหารูปแบบที่สามารถใช้จำแนกข้อมูล (โครงการวิจัยที่ 3)

ส่วนที่ 4 Post-data mining ทำหน้าที่ประเมินผลลัพธ์ที่ได้และคัดเลือกแสดงเฉพาะผลลัพธ์ที่คาดว่าจะเป็นประโยชน์และน่าสนใจสำหรับผู้ใช้ (โครงการวิจัยที่ 4)

ซอฟต์แวร์ระบบเหมืองข้อมูลที่ดี ที่มีใช้อยู่ในปัจจุบันจะมีส่วนที่ 1, 2, และ 3 การพัฒนาเพิ่มเติมส่วนที่ 4 ในงานวิจัยนี้จะเป็นแนวทางใหม่ของการสร้างระบบเหมืองข้อมูล ที่คาดหมายว่าจะเป็นประโยชน์สำหรับผู้ใช้มากขึ้น