การทำดาต้าไมนิ่งระบบขายแพคเกจทัวร์

Data Mining Application for a Travel Agency

โดย

นางสาวสุจิตรา มังคละไชยา

รหัส 44067419

อาจารย์ที่ปรึกษา

ดร. ภัทรชัย ลลิตโรจน์วงศ์

รายงานนี้เป็นส่วนหนึ่งของวิชาโครงการพัฒนาระบบงาน

หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ภาคเรียนที่ 2 ปีการศึกษา 2546

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อหัวข้อ	การทำคาต้าไมนิ่งระบบขายแพคเกจทัวร์
นักศึกษา	นางสาวสุจิตรา มังคละไชยา
อาจารย์ที่ปรึกษา	ดร.ภัทรชัย ลลิตโรจน์วงศ์
ระดับการศึกษา	วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
ปีการศึกษา	2546

บทคัดย่อ

การทำระบบขายทัวร์ จะต้องทราบถึงพฤติกรรมและความต้องการที่แท้จริงของลูกค้าในการซื้อทัวร์ เพื่อจะนำมาวิเคราะห์และวางแผนในการหาแนวโน้มความต้องการของลูกค้าในอนาคต จึงได้ทำการรวบรวมข้อมูลต่าง ๆ และนำเทคนิคการทำคาต้าไมนิ่ง มาใช้เพื่อวิเคราะห์คลังข้อมูลที่ได้ เพื่อหาแนวโน้มของความต้องการของลูกค้าในการซื้อทัวร์ และสถานที่ท่องเที่ยวที่ลูกค้าต้องการ เพื่อนำข้อมูลที่ได้ มาวางแผนและปรับปรุงระบบการจัดการกรุ๊ปทัวร์ และการโฆษณาให้ตรงกับความต้องการของลูกค้า และเพื่อนำไปวางแผนกลยุทธ์การขายต่อไป โดยในการพัฒนาระบบ ใช้การไมนิ่งแบบ Descriptive เพื่อหารูปแบบในการอธิบายข้อมูลโดยอาศัยหลักความสำคัญของ ข้อมูล โดยใช้แบบจำลองแบบ Database Segmentation และใช้วิธีการ Demographic Clustering เพื่อทำการแบ่งกลุ่มลูกค้าออกเป็นส่วน ๆ และนำผลที่ได้มาประกอบการตัดสินใจในการจัดแพคเกจทัวร์ตามกลุ่มลูกค้าได้

Title	Data Mining Application for a Travel Agency
Student	Ms.Sujittra Mungkalachaiya
Advisor	Dr.Pattarachai Lalitrojwong
Level of study	Master of Science in Information Technology
Major	Information Science
Academic Year	2003

Abstarct

In a traveling agency business, knowing customer behaviors and their requirements is one of the most important thing. For good analysis and strategic planning, they need to collect data and build a data warehouse. After that, they can employ data mining techniques to predict customer requirement trends, such as places to go. The knowledge from analysis can support for the marketing process.

กิตติกรรมประกาศ

โครงการพัฒนาโปรแกรมประยุกต์ สำหรับการทำคาต้าไมนิ่งระบบขายแพคเกจทัวร์ สำเร็จลงได้ด้วยคำแนะนำและความช่วยเหลือจาก ดร. ภัทรชัย ลลิตโรจน์วงศ์ ซึ่งเป็นอาจารย์ที่ปรึกษาของโครงการนี้ และขอขอบพระคุณ นาย กัน แจ ลี และเพื่อนๆ ทุกท่านที่ให้โอกาส และกรุณาสละเวลาในการให้คำปรึกษา และคำแนะนำที่มีประโยชน์ต่อการพัฒนาโครงการ

สุดท้ายนี้ขอขอบคุณกำลังใจจากครอบครัว ผู้จัดทำรู้สึกซาบซึ้งในความกรุณาของท่านเป็นอย่างยิ่ง และขอกราบขอบพระคุณไว้ ณ โอกาสนี้ด้วย

นางสาวสุจิตรา มังคละไชยา

27 มีนาคม 2547

สารบัญ

สารบัญตาราง

หน้า

ตารางที่

สารบัญรูป

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาของโครงการ

ในการประกอบธุรกิจนั้น ปัจจัยหนึ่งที่มีความสำคัญอย่างมากต่อความสำเร็จของธุรกิจก็คือความเข้าใจในตัวลูกค้า หรือกลุ่มลูกค้า ยิ่งได้รู้ข้อมูลของลูกค้ามาก ยิ่งเข้าใจความต้องการของกลุ่มลูกค้าที่สนใจได้มาก โอกาสที่จะทำธุรกิจให้ตรงกับความต้องการของตลาดจะมีมากขึ้น สามารถนำสร้างความได้เปรียบต่อคู่แข่งในการวางแผนกลยุทธ์ได้ เช่น ธุรกิจการท่องเที่ยว ซึ่งในปัจจุบัน มีอัตราการแข่งขันกันสูง เพราะฉะนั้น จึงมีความจำเป็นต้องเก็บข้อมูลเกี่ยวกับกลุ่มลูกค้าที่ต้องการและหาสารสนเทศมาสนับสนุนในวิเคราะห์และตัดสินใจในการวางแผนและจัดแพคเกจให้ตรงกับความต้องการของลูกค้าและการทำคาต้าไมนิ่งก็เป็นวิธีการหนึ่งที่สามารถนำมาช่วยในการวิเคราะห์ข้อมูลขนาดใหญ่ เพื่อทำนายแนวโน้ม และพฤติกรรมของข้อมูลในอนาคตเพื่อให้ทราบถึงความสัมพันธ์ในรูปแบบต่าง ๆ ได้ และเพื่อจำแนกกลุ่มลูกค้าตามความต้องการได้อย่างเหมาะสม เพื่อนำผลการวิเคราะห์มาสนับสนุนกลยุทธ์การขายโดยการทำคาต้าไมนิ่งสามารถตอบคำถามทางธุรกิจซึ่งเป็นคำถามที่มีการนำข้อมูลที่มีความสัมพันธ์กับเวลามาแก้ปัญหาซึ่งเป็นข้อจำกัดของระบบฐานข้อมูลธรรมดาบทความนี้ จะกล่าวถึงทฤษฎีที่เกี่ยวข้องได้แก่ การทำคาต้าไมนิ่ง การออกแบบคลังข้อมูล และการออกแบบคลังข้อมูลของระบบขายแพคเกจทัวร์

1.2 ปัญหาที่เกิดขึ้นในระบบงานเดิม

ในการประกอบธุรกิจด้านการท่องเที่ยวที่มีการแข่งขันสูง จำเป็นต้องมีการวิเคราะห์พฤติกรรมและความต้องการของลูกค้าเพื่อที่จะสามารถนำมาประกอบการกำหนดกลยุทธ์และการจัดการแพคเกจทัวร์ให้ตรงกับความต้องการของกลุ่มลูกค้าเป้าหมาย จากเดิมทำการคาดเดาและวิเคราะห์ข้อมูลโดยวิธีการทางสถิติซึ่งมีข้อจำกัดในการทำนายแนวโน้มและพฤติกรรมต่าง ๆ จึงได้นำเทคนิคการทำ คาต้าไมนิ่ง เข้ามาช่วยเพื่อทำนายแนวโน้ม และพฤติกรรมของข้อมูลในอนาคต เพื่อให้ทราบถึงความสัมพันธ์ในรูปแบบต่าง ๆ ได้

1.3 เป้าหมายในการดำเนินงาน

เพื่อวิเคราะห์พฤติกรรมและความต้องการในการซื้อแพคเกจทัวร์ของลูกค้าในกลุ่มที่สนใจจากคลังข้อมูลที่มีและนำมาใช้ในการพัฒนาระบบการตัดสินใจในการวางแผนกลยุทธ์ในการบริหารและจัดทำแพคเกจทัวร์ให้ตรงกับความต้องการของตลาด เพื่อเป็นการเพิ่มยอดขายผลกำไรทางการค้า

1.4 ขอบเขตของระบบงานใหม่

- รวบรวมมูลลูกค้าแล้วข้อมูลเกี่ยวกับความต้องการและพฤติกรรมการซื้อแพคเกจทัวร์
- จัดทำฐานข้อมูลที่เหมาะสมกับการวิเคราะห์ตามความต้องการ
- ทำการไมนิ่งข้อมูลลูกค้าเพื่อนำมาวิเคราะห์ผลลัพธ์
- ทำการวิเคราะห์และจัดกลุ่มลูกค้าประเภทต่าง ๆ และความต้องการที่สัมพันธ์กัน เพื่อนำข้อมูลมาสรุปและใช้ในการประกอบการตัดสินใจในการจัดรูปแบบแพคเกจทัวร์

1.5 ประโยชน์ของระบบงานใหม่

- ทำให้สามารถแยกกลุ่มลูกค้าได้อย่างชัดเจนเพื่อความสะดวกในการวิเคราะห์จัดหากลยุทธ์ในการจัดการธุรกิจท่องเที่ยว
- ทำให้ทราบถึงความสนใจและความต้องการของลูกค้าในแต่ละกลุ่มและสามารถนำมาช่วยในการตัดสินใจในการจัดการแพคเกจทัวร์ให้เหมาะสมกับกลุ่มลูกค้าที่เราต้องการได้

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1. ดาต้าไมนิ่ง

ดาต้าไมนิ่ง คือชุดซอฟแวร์วิเคราะห์ข้อมูลที่ถูกออกแบบมาเพื่อระบบสนับสนุนการตัดสินใจ ดาต้าไมนิ่งเป็นส่วนหนึ่งของกระบวนการค้นพบความรู้ในฐานข้อมูล (Knowledge Discovery in Database : KDD) การนำแนวโน้มของข้อมูลและสารสนเทศที่ซ่อนอยู่ ในฐานข้อมูลขนาดใหญ่เป็นสิ่งสำคัญ เพราะถ้าไม่รู้จักใช้ประโยชน์จากข้อมูลเหล่านั้น ก็จะเป็นการเก็บข้อมูลโดยสูญเปล่า ดาต้าไมนิ่งเป็นเครื่องมือที่สามารถค้นหาข้อมูล ในฐานข้อมูลขนาดใหญ่ หรือข้อมูลที่เป็นประโยชน์ที่อาจจะซ่อนอยู่ภายในฐานข้อมูล ซึ่งเป็นการเพิ่มคุณค่าให้กับฐานข้อมูลที่มีอยู่ โดยการทำดาต้าไมนิ่งมีลักษณะเหมือน กับการรวมนักวิเคราะห์กับผู้มีประสบการณ์ในด้านที่ ต้องการซึ่งสามารถหาปัจจัยที่มีอิทธิพล และแนวโน้มของข้อมูลได้ การเข้าถึงความรู้ จะเป็นการรวมรูปแบบทั้งหมดของข้อมูล โดยผู้ใช้สามารถที่จะเลือกรูปแบบที่ต้องการมาใช้งานได้ และสามารถเปลี่ยนรูปแบบในการ วิเคราะห์ได้ตามสถานการณ์ เช่นรูปแบบที่ใช้ในการวิเคราะห์ความต้องการของนักท่องเที่ยวสูงอายุ เป็นต้น

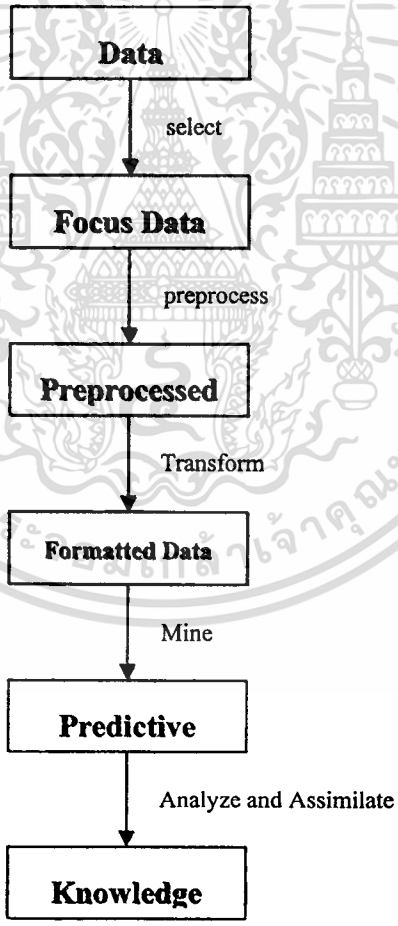
ดาต้าไมนิ่ง เป็นเทคนิคที่ใช้ในการค้นหาสารสนเทศเชิงพยากรณ์ที่มองไม่เห็น ออกมาจากฐานข้อมูลขนาดใหญ่ โดยเฉพาะคลังข้อมูล ซึ่งมีความจำเป็น ในการช่วยทำ ให้การวิเคราะห์ข้อมูลครอบคลุมถึงแนวโน้มใน ทุก ๆ กรณี ดาต้าไมนิ่ง ช่วยทำให้เกิดศักยภาพ ในการใช้ข้อมูลในฐานข้อมูล เพราะฉะนั้น ดาต้าไมนิ่งจึง เป็นทางเลือกอีกทางหนึ่ง ที่มีความสามารถในการวิเคราะห์ข้อมูลจำนวนมาก เพื่อสืบค้นหา แนวโน้มของข้อมูลที่มีอยู่แล้ว และนำไปสู่สารสนเทศในอนาคต ส่งผลให้สามารถนำไปใช้ ประโยชน์ได้ในการ ประยุกต์ใช้ ทางธุรกิจ

เทคนิคการทำดาต้าไมนิ่งใช้หลักการเข้าไปใกล้กับรายละเอียด คล้าย ๆ กับการสำรวจเข้าไปถึงความสัมพันธ์ในกลุ่มของคลังข้อมูล หรือฐานข้อมูลที่มีขนาดใหญ่ และ ดาต้าไมนิ่ง ยังมีการลำดับการแก้ปัญหาด้วยโปรแกรม หรือลำดับขั้นตอน การทำงานใกล้เคียง กับความเป็นจริง และมีการใช้แบบจำลองในการแก้ปัญหา โดยแบบจำลอง จะถูกออกแบบมา ให้อยู่บนรากฐานของความเป็นจริงของธุรกิจมากที่สุด

2.2. ขบวนการในการทำ ดาต้าไมนิ่ง

หากจะกล่าวถึงการทำดาต้าไมนิ่ง ส่วนใหญ่จะให้ความสำคัญกับการไมนิ่ง หรือการค้นหาลักษณะทิศทางของข้อมูล แต่แท้จริงแล้ว การไมนิ่งข้อมูล เป็นเพียงขั้นตอนหนึ่ง ในขบวนการทำดาต้าไมนิ่งเท่านั้น และการทำดาต้าไมนิ่ง เรียกได้อีกอย่างหนึ่งว่ากระบวนการ ค้นพบข้อมูลที่เป็นความรู้ในฐานข้อมูล (KDD)

แท้จริงแล้วดาต้าไมนิ่งเป็นการรวบรวมเทคนิคจากงานต่าง ๆ เช่น การจดจำรูปแบบการเรียนรู้ของเครื่องจักร หลักสถิติ และฐานข้อมูล เป็นต้น เพื่อนำมาค้นหารูปแบบความสัมพันธ์ของข้อมูล และวิเคราะห์หาข้อมูลที่เป็นประโยชน์ที่อาจจะซ่อนอยู่ภายใต้คลังข้อมูลขนาดใหญ่ เพื่อให้ได้ข้อมูลที่สามารถนำไปใช้งานได้จริงและเป็นประโยชน์ โดยสามารถแบ่งขั้นตอนการทำงานของดาต้าไมนิ่งได้ดังรูปที่ 2.1 (Han and Kamber.2001)



รูปที่ 2.1 กระบวนการทำดาต้าไมนิ่ง

การทำดาต้าไมนิ่งต้องอาศัยวัตถุประสงค์ทางธุรกิจเป็นพื้นฐานด้วยโดยเป็นส่วนที่บอกวัตถุประสงค์ที่ต้องการจากการทำดาต้าไมนิ่งรายละเอียดในการทำงานของดาต้าไมนิ่งประกอบด้วยหลายขั้นตอนที่มีการทำซ้ำ ๆ หรือต้องมีการวนกลับมาทำใหม่ โดยการทำดาต้าไมนิ่ง สามารถแบ่งการทำงานออกเป็น 5 ขั้นตอนดังนี้ (Berson and Smith.1997; Han and Kamber. 2001; Tutorial on High Performance Data Mining.1997)

2.2.1 กำหนดวัตถุประสงค์ (Bussiness Objectives Determination)

การกำหนดวัตถุประสงค์ทางธุรกิจ คือการกำหนดปัญหาและวัตถุประสงค์ทางธุรกิจ ขององค์กรให้ชัดเจน เพราะต้องเข้าใจถึงปัญหาและความต้องการทางธุรกิจประกอบด้วยการวิเคราะห์ข้อมูลเบื้องต้นว่ามีข้อมูลใดบ้าง

2.2.2 จัดเตรียมข้อมูล

การจัดเตรียมข้อมูล เป็นขั้นตอนสำคัญและใช้เวลานานที่สุดเนื่องจากต้องมีการคัดเลือกข้อมูลที่เหมาะสมและอยู่ในประเด็นที่ต้องการ ประกอบด้วยขั้นตอนย่อย 3 ขั้นตอน ดังนี้

2.2.2.1 การคัดเลือกข้อมูล เป็นการกำหนดรูปแบบข้อมูลที่ต้องการระบุลักษณะข้อมูลและเลือกข้อมูลที่ต้องการและนำข้อมูลที่ไม่ต้องการออกไป เป็นการเริ่มต้นของการเตรียมการไมนิ่ง การเลือกข้อมูลจะขึ้นอยู่กับวัตถุประสงค์ทางธุรกิจ การเลือกไม่ว่าจะเป็นตัวแปรความสัมพันธ์จำเป็นต้องเข้าใจความหมายประเภทข้อมูล และค่าที่สามารถเป็นไปได้ การเลือกข้อมูลต้องคำนึงถึงอายุข้อมูลด้วย

2.2.2.2 การประมวลผลข้อมูลเบื้องต้น เป็นขั้นตอนการประมวลผลข้อมูลเบื้องต้นจากเทคนิคที่เลือกไว้

2.2.2.3 การปรับเปลี่ยนรูปแบบของข้อมูล เป็นกระบวนการในการปรับข้อมูลให้อยู่ในรูปที่เหมาะสมต่อการนำไปวิเคราะห์

2.2.3 ดาต้าไมนิ่ง (Data Mining)

การทำดาต้าไมนิ่ง ทำโดยการเลือกวิธีการและอัลกอริทึมที่เหมาะสมเพื่อใช้กับข้อมูลที่เตรียมไว้ ขั้นตอนนี้มีความสัมพันธ์กับการวิเคราะห์ข้อมูลและขั้นตอนที่ผ่านมา โดยอาจจะย้อนกลับไปทำในขั้นตอนที่ 2 ใหม่

2.2.4 วิเคราะห์ผลลัพธ์ที่ได้

การวิเคราะห์ผลคือการแปลความหมายและประเมินผลที่ได้จากขั้นตอนที่ 3

2.2.5 นำไปเปรียบเป็นความรู้

เป็นการรวบรวมความเข้าใจเชิงธุรกิจที่ได้มาจากขั้นตอนการวิเคราะห์ผลลัพธ์เพื่อนำมาประยุกต์ใช้ให้ตรงกับความต้องการทางธุรกิจขององค์กร

2.3. เทคนิคและการจำลองแบบการทำดาต้าไมนิ่ง

การทำดาต้าไมนิ่ง สามารถแบ่งออกเป็น 2 ประเภท หลัก ๆ ดังนี้ (Berson and Smith. 1997; Han and Kamber. 2001)

1. Descriptive data mining

หารูปแบบในการอธิบายข้อมูลโดยอาศัยหลักความสัมพันธ์ของข้อมูล เช่นการหาความสัมพันธ์ของข้อมูลโดยการแบ่งกลุ่มข้อมูล

2. Predictive data mining

ใช้ตัวแปรเพื่อทำนายสิ่งที่ยังไม่รู้หรือเพื่อหาค่าในอนาคตจากตัวแปรการทำดาต้าไมนิ่งจะสำเร็จได้โดยกระบวนการจากแบบจำลองต่าง ๆ ซึ่งสามารถแบ่งเป็น 4 แบบได้แก่ (Han and Kamber.2001)

1. Clasification [Predictive]

เป็นกระบวนการสร้างแบบจำลองจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ เช่น แบ่งกลุ่มประเภทลูกค้าว่าเชื่อถือได้หรือไม่

2. Clustering [Descriptive]

เป็นเทคนิคลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน เช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าเป็นกลุ่มย่อยตามรายได้ของกลุ่มลูกค้า

3. Association Rule Discovery [Descriptive]

ใช้หลักการค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์หรือทำนายปรากฏการณ์ต่าง ๆ เช่น การวิเคราะห์การซื้อสินค้าของลูกค้า ซึ่งประเมินจากข้อมูลที่ได้รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นผลจากการ วิเคราะห์แบบกฎความสัมพันธ์ (Association Rule)

4. Sequential Pattern Discovery [Descriptive]

ให้กลุ่มของวัตถุที่มีความสัมพันธ์ระหว่างวัตถุนั้น ๆจะมี timeline of events เพื่อหากฎที่จะทำนายการขึ้นต่อกันระหว่างเหตุการณ์ที่ต่างกัน กฎถูกสร้างโดยรูปแบบแรกเหตุการณ์ที่เกิดขึ้นในรูปแบบต่าง ๆ ถูกดูแลโดยเกี่ยวข้องกับ เรื่องของเวลา

5. Regression [Prediction]

ทำนายค่าตัวแปรที่ต่อเนื่องบนพื้นฐานของตัวแปรตัวอื่น ๆ และนำมาตั้งสมมุติฐานการขึ้นต่อกัน

6. Deviation Detection [Predictive]

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐานหรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใดมักใช้วิธีการทางสถิติ วิธีนี้มักใช้ในการตรวจสอบ บัตรเครดิตหรือลายเซ็นปลอม

ทั้งนี้การจะนำข้อมูลมาผ่านกระบวนการดาต้าไมนิ่งได้นั้นต้องทำการรวบรวมข้อมูลเพื่อจะนำมาใช้ในการค้นหาสารสนเทศเชิงพยากรณ์ที่มองไม่เห็น โดยจัดเก็บในรูปแบบฐานข้อมูลที่ได้รวบรวมและจัดเก็บตามวัตถุประสงค์หลักในการวิเคราะห์ข้อมูล หรือคลังข้อมูล ซึ่งมีความจำเป็นในการช่วยทำให้การวิเคราะห์ข้อมูลครอบคลุมถึงแนวโน้มในทุก ๆ กรณี ซึ่งช่วยทำให้เกิดศักยภาพในการใช้ข้อมูลในฐานข้อมูลโดยการทำดาต้าไมนิ่งนั้นไม่มีเทคนิคหรือเครื่องมือเพียงชนิดเดียวของดาต้าไมนิ่งที่เหมาะสมกับงานทุกชนิดในแต่ละชนิด ก็จะมีเทคนิคของดาต้าไมนิ่งที่แตกต่างกันไปขึ้นอยู่กับชนิดของงาน

2.4 Database Segmentation (Han and Kamber. 2001)

Database Segmentation (Clustering) เป็นการแบ่งข้อมูลที่สนใจในฐานข้อมูลออกเป็นส่วนย่อย ๆ ตามกลุ่มของเรคคอร์ดที่คล้ายคลึงกันซึ่งเป็นเทคนิคลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วยกัน สามารถแบ่งวิธีการทำแบบจำลองแบบ Database Segmentation โดยแบ่งตามชนิดของข้อมูลที่เป็นอินพุท วิธีการคำนวณระยะระหว่างข้อมูลที่มากที่สุดและน้อยที่สุดที่เลือกให้อยู่ในกลุ่มเดียวกัน เช่นการแบ่งกลุ่ม คนอายุน้อยต้องทำการคำนวณว่าระยะระหว่างอายุในกลุ่มที่มากที่สุดและน้อยที่สุดเป็นเท่าไหร่ หรือข้อมูลที่มีค่าต่างกันเท่านี้ควรอยู่กลุ่มเดียวกันหรือคนละกลุ่ม และยังสามารถแยกได้ตามวิธีที่ใช้ ในการจัดการกับผลข้อมูลที่ได้แบ่งกลุ่มเพื่อการวิเคราะห์ โดยสามารถแบ่งแบบจำลอง Database Segmentation ได้ 2 วิธีดังนี้

1. Demographic Clustering ใช้หลักการทำงานบนเรคคอร์ดกับตัวแปรอินพุทแบบ categoric
2. Neural Clustering เป็นวิธีทางนิวรอลเน็ทเวอร์กซึ่งจะยอมรับเฉพาะตัวแปรอินพุทแบบตัวเลขแต่ก็จะสามารถใช้ตัวแปรแบบ categoric โดยจะต้องทำการแปลงข้อมูลให้อยู่ในลักษณะปริมาณก่อน

คลัสเตอร์ คือกลุ่มของข้อมูลที่มีลักษณะคล้ายกันหรือมีลักษณะร่วมกันการแบ่งคลัสเตอร์เป็นการจัดกลุ่มข้อมูลที่มีลักษณะร่วมกันไว้ด้วยกันและแยกข้อมูลที่มีลักษณะต่างไปออกไปไว้ยังคลัสเตอร์อื่น

การทำ clustering analysis ถูกนำมาใช้อย่างแพร่หลาย เช่นใช้ในการทำ pattern recognition, data analysis, image processing และ market research โดยการแบ่งคลัสเตอร์ยังสามารถบอกถึงความหนาแน่นหรือเบาบางของข้อมูลในแต่ละคลัสเตอร์ได้อีกด้วย

ในทางธุรกิจ สามารถนำเทคนิคการแบ่งคลัสเตอร์มาช่วยในการแบ่งกลุ่มลูกค้าตามรูปแบบการซื้อสินค้าได้ และยังนำมาใช้ในการแบ่งกลุ่มเอกสารเพื่อการค้นหาข้อมูลบนเว็บและในส่วนของการทำคาด้าไมนิ่งได้นำ clustering analysis มาใช้ในการแยกกลุ่มข้อมูลที่สนใจและพิจารณาการกระจายของข้อมูลให้ชัดเจนยิ่งขึ้น และสามารถค้นหาลักษณะเฉพาะของแต่ละคลัสเตอร์ที่สนใจเป็นพิเศษได้

2.5 ประเภทของ Clustering Method

ปัจจุบันมีอัลกอริทึมในการทำclusteringมากมายการเลือกใช้ก็แตกต่างกันไปตามวัตถุประสงค์และประเภทข้อมูลที่อัลกอริทึมนั้นรองรับ ซึ่งบางอัลกอริทึมต้องมีการนำหลาย ๆ อัลกอริทึมมาใช้ร่วมกัน

โดยทั่วไป สามารถแบ่งประเภทของการแบ่งคลัสเตอร์ได้เป็น 5 ประเภทได้แก่

1. **Partitioning methods**

การทำงานจะแบ่งจำนวนข้อมูลในฐานข้อมูล (จำนวนเรคคอร์ด) n ออกเป็นclusterต่าง ๆ จำนวน kคลัสเตอร์โดย k <= n และ

- แต่ละกลุ่มข้อมูลต้องมีอย่างน้อย 1 เรคอร์ด
- แต่ละเรคอร์ด ต้องถูกจัดอยู่ในกลุ่มใดกลุ่มหนึ่งอย่างน้อย 1 กลุ่ม

การแบ่งคลัสเตอร์วิธีนี้ ต้องกำหนดค่า k หรือจำนวนคลัสเตอร์ว่าต้องการจัดกลุ่มเป็นกี่กลุ่มจากนั้นอัลกอริทึมจะทำการสุ่มสร้างคลัสเตอร์เริ่มต้นขึ้นมาแล้วใช้เทคนิคที่เรียกว่า

"Iterative relocation"ในการวนลูปเพื่อย้ายเรคอร์ดที่มีคุณสมบัติใกล้เคียงกันไปยังกลุ่มต่าง ๆ ที่สร้างขึ้นมาจรดรบทั้งฐานข้อมูลกระบวนการนี้จะถูกทำซ้ำจนกระทั่งค่าที่ใช้วัดความเหมาะสมในการแบ่งคลัสเตอร์ที่เรียกว่าค่า "square-error" ที่เกิดขึ้นเป็นศูนย์ หรือเข้าใกล้ค่าที่สามารถยอมรับได้ที่ตั้งเอาไว้

อัลกอริทึมที่ได้รับความนิยมสำหรับเทคนิคนี้คือ "K-Means" ซึ่งใช้ mean เป็นค่าวัดความเหมือนหรือความต่างกันระหว่างข้อมูลแต่ละเรคอร์ดกับค่า mean ของ centroid ในแต่ละคลัสเตอร์

2. **Hierarchical methods**

เป็นการจัดกลุ่มข้อมูลโดยสร้างเป็นชั้น ๆ แบบลำดับขั้น ซึ่งมี 2 ประเภท ได้แก่

- Agglomerative หรือแบบ bottom-up คือ ทำการรวบรวมข้อมูลจากระดับล่างสุด (ทีละเรคอร์ด) ขึ้นมาเป็นลำดับที่สูงขึ้นจนสุดท้ายได้เป็นกลุ่มข้อมูลขนาดใหญ่
- Divisive หรือแบบ top-down ซึ่งจะมีการทำงานแบบตรงข้ามกับ Agglomerative คือทำจากระดับบนสุดและจะทำการแตกกลุ่มลงมาจนระดับล่างสุด

การวัดความเหมือนของแต่ละเรคอร์ด จะทำแบบทุกทิศทาง จากนั้น จึงทำการรวมกลุ่มเรคอร์ดที่มีค่าความเหมือนอยู่ในระดับเดียวกัน กล่าวคือ ถ้าฐานข้อมูลมี n เรคอร์ด แต่ละเรคอร์ดจะถูกเปรียบเทียบความเหมือน n-1 ครั้งจากนั้นเรคอร์ดที่มีค่าความเหมือนอยู่ในระดับเดียวกันจะถูกรวมอยู่ในกลุ่มเดียวกัน

3. **Density-based methods**

วิธีนี้จะจัดกลุ่มข้อมูลโดยการพิจารณาที่ความหนาแน่นของข้อมูลแทนการใช้ระยะห่างของแต่ละเรคอร์ดกับค่าตัวแทนของคลัสเตอร์โดยจะทำการขยายขนาดของคลัสเตอร์แต่ละคลัสเตอร์ไปเรื่อย ๆ ตราบใดที่คลัสเตอร์ใกล้ ๆ กัน (neighborhood) มีค่าเกินค่าหนึ่งที่ตั้งไว้ ตัวอย่างของอัลกอริทึมประเภทนี้ได้แก่ DBSCAN, OPTICS

4. **Grid-based method**

วิธีนี้จะทำการ Quantize ข้อมูลไปเรื่อย ๆ จนได้ cell จำนวนหนึ่งเพื่อใช้ในการสร้างโครงสร้าง grid ข้อดีของ metod นี้คือ ทำงานได้เร็ว ไม่ขึ้นกับจำนวนเรคอร์ดของฐานข้อมูล ตัวอย่างของอัลกอริทึมประเภทนี้ได้แก่ STING, CLIQUE, Wave cluster

5. **Model-based method**

วิธีนี้ จะทำการจัดกลุ่มข้อมูลโดยการเทียบเคียงกับ model ที่สร้างขึ้นมา ซึ่งจะมี 2 Approach หลัก ๆ คือ Statistical approach และ Neural network approach ตัวอย่างอัลกอริทึม ประเภทนี้ได้แก่ COBWEB, CLASSIT

2.6 อัลกอริทึมที่เลือกมาทำการ Clustering

อัลกอริทึมที่เลือกมาทำการแบ่งคลัสเตอร์ได้แก่ ROCK Algorithm(Robust Clustering using Links) ซึ่งเป็นอัลกอริทึมประเภท Hierachical

2.6.1.1 หลักการของ ROCK (Guha et.at.1999)

ในอัลกอริทึม ROCK จะแทนข้อมูลแต่ละเรคคอร์ดในฐานข้อมูลด้วยจุด เรียกว่าพอยท์โดยพอยท์ที่มีความเหมือนหรือมีความใกล้เคียงกัน จะจัดให้เป็น Neighbor ของกัน โดยความใกล้เคียงนี้ได้แก่ Similarity (Sim (p_i, p_j)) โดยค่า Similarity นี้ ใช้หาค่าความใกล้กันของพอยท์แต่ละคู่ ซึ่งแทนด้วยคู่ลำดับ (p_i, p_j) โดยจะมีค่าอยู่ระหว่าง 0 ถึง 1 หากค่า Similarity มีค่าเข้าใกล้ 1 มากเท่าใด แสดงว่าพอยท์คู่นั้นมีความเหมือนหรือสัมพันธ์กันมาก ส่วนพอยท์คู่ใดที่มีค่า Similarity เท่ากับ 0 หมายความว่าเป็นพอยท์ที่ไม่มีความสัมพันธ์กัน

โดยการพิจารณาค่า Similarity เพื่อหาว่า piont แต่ละคู่เป็น neighbor กันหรือไม่นั้น จะต้องมีการกำหนดค่า Threshold (θ)ที่สามารถยอมรับได้ขึ้นมาหนึ่งค่า หาก neighbor คู่ใด ๆ มีค่า Similarity มากกว่า θ สามารถเรียกเรคคอร์ดที่กำลังพิจารณานั้นว่าเป็น “common neighbor” กัน ดังนั้น สามารถพิจารณาค่า Similarity ได้ดังรูปที่ 2.2

$SIM(p_i, p_j) >= \theta$ (θ คือค่า Threshold ที่ยอมรับได้ที่ user กำหนดขึ้นมา)

รูปที่ 2.2 นิยามของ neighbor

ฟังก์ชันการหาค่า Similarity นั้น มีหลายฟังก์ชัน แต่โดยรวมทำงานเหมือนกัน ดังแสดงในรูปที่ 2.3

$$SIM(T1, T2) = \frac{|T1 \cap T2|}{|T1 \cup T2|}$$

รูปที่ 2.3 ฟังก์ชันในการหาค่า Similarity ระหว่างพอยท์

จากรูปที่ 2.3 T1 คือจำนวน item ในข้อมูลรายการ T1 ยิ่งจำนวน item ของรายการ T1 และ T2 สามารถเชื่อมโยงถึงกันได้มากเท่าไหร่หรือมีค่ามากเท่าไหร่ (T1 ∩ T2) แสดงว่า รายการ ทั้งสอง มีความเหมือนกันมากขึ้นเท่านั้น และมีโอกาสอยู่ในคลัสเตอร์เดียวกันมากขึ้น

การใช้จำนวนข้อมูลในเซ็ตของเรคคอร์ดที่กำลังพิจารณา (T1 ∪ T2) เป็นตัวหาร จำทำให้ค่า Sim มีค่าระหว่าง 0 ถึง 1 เท่านั้น ซึ่งสามารถนำไปเปรียบเทียบค่าความเหมือนกับรายการอื่น ๆ ต่อไปได้

2.6.2 Links

ในการพิจารณาว่ารายการต่างๆจัดอยู่ในคลัสเตอร์เดียวกันหรือไม่นั้นจะพิจารณาจากcommon neighbor ว่ามีลิงค์มากน้อยเพียงใดหากจำนวนลิงค์ของ common neighbor มีมาก โอกาสที่พอยท์คู่นั้น ๆ จะอยู่ในคลัสเตอร์เดียวกันก็จะมีมากขึ้น

จากนิยามข้างต้น แสดงให้เห็นว่า เราใช้จำนวน item ที่เหมือนกันระหว่างพอยท์ที่กำลังพิจารณา (p_i, p_j) ในการกำหนดความเหมือนกันของพอยท์เหล่านั้น โดยสามารถแทนด้วยสัญลักษณ์ Link (p_i, p_j) โดยยิ่งค่า Link (p_i, p_j) มีค่ามากขึ้นเท่าใด โอกาสที่ p_i และ p_jจะอยู่ในคลัสเตอร์เดียวกันมีมากขึ้น

2.6.3 Missing value

ข้อมูลประเภท Categorical ส่วนใหญ่จะเป็น Fixed dimension คือ มีจำนวนแอททริบิวท์ คงที่ เช่น ข้อมูลของนักท่องเที่ยวคนหนึ่งๆ จะประกอบด้วย เพศ อายุ อาชีพ รายได้ สถานที่ที่ไป แหล่งข้อมูลการท่องเที่ยว และกิจกรรมที่ทำ เป็นต้น โดยข้อมูลแบบ Fixed dimension จะแยก กลุ่มข้อมูลที่ขาดหายไป โดยไม่นำมาทำการคลัสเตอร์ เนื่องจาก missing value เหล่านี้ จะถูกจัดเป็นข้อมูลที่ไม่มีความเหมือนกัน ทำให้ไม่เกิดการจัดกลุ่มแบบผิด ๆ ทั้งนี้ขึ้นอยู่กับปริมาณของ missing value ด้วยว่ามีมากน้อยเพียงใด เพราะหากมีมากเกินไป กลุ่ม ข้อมูลที่ได้จากการคลัสเตอร์ก็อาจมีความสมเหตุสมผลน้อยลง

2.6.4 Criterion Function

กระบวนการทำคลัสเตอร์ เป็นกระบวนการทำงานแบบ iterative คือ วนการทำงานไปจนกระทั่งได้จำนวนคลัสเตอร์ที่เหมาะสม

ค่าที่ใช้ในการวัดความเหมาะสมในการทำคลัสเตอร์เรียกว่า "Criterion Function" ซึ่งวิธีการหานั้นจะแตกต่างกันไปตามอัลกอริทึมแต่ละประเภท โดยวิธีการของอัลกอริทึม ROCK นั้น จะพิจารณาที่ระดับความสัมพันธ์ Link(p_i, p_j) ของพอยท์ต่าง ๆ ในฐานข้อมูลว่ามีความเหมือน กันมากน้อยเพียงใด โดยในขั้นเริ่มต้น จะถือว่า แต่ละพอยท์เป็นแต่ละคลัสเตอร์ที่แยกจากกัน โดยสมการของ Criterion Function เป็นดังรูปที่ 2.4

$$Et = \sum_{i=1}^{k} ni \times \sum_{Pq,Pr \in Ci} \frac{link(Pq,Pr)}{ni^{1+2f(\theta)}}$$

โดย Ci คือ คลัสเตอร์ลำดับที่ i

Pi คือพอยท์ลำดับที่ i

n คือ จำนวนรายการภายในคลัสเตอร์ หรือภายในพอยท์นั้น ๆ

รูปที่ 2.4 Criterion Function

จากสมการข้างต้น พบว่าค่า $\sum_{i=1}^{k}\sum_{Pq,Pr \in Ci} link(Pq,Pr)$ แสดงถึงผลรวมของจำนวนลิงค์ระหว่างพอยท์ที่กำลังพิจารณาเพื่อใช้หาค่าความเหมือนกันของพอยท์ทั้งหมดในคลัสเตอร์ i การคำนวณดังกล่าว ยังไม่สามารถแยกพอยท์ที่มีจำนวนลิงค์น้อย ๆ ออกจากคลัสเตอร์ได้ อาจทำให้เกิดกรณีที่รายการต่าง ๆ ถูกรวมเป็นคลัสเตอร์เดียวกันทั้งหมด(ทั้งในกรณีที่มีลิงค์น้อยและมาก) ดังนั้นแต่ละคลัสเตอร์จะถูก normalize โดยทำการหารค่าข้างต้นด้วยจำนวนครอสลิงค์ทั้งหมดที่คาดว่าจะเกิดขึ้นของคลัสเตอร์ Ci จากนั้น ทำการถ่วงน้ำหนักด้วยจำนวนพอยท์(ni) ในคลัสเตอร์ Ci

จำนวนลิงค์ทั้งหมดที่คาดว่าจะเกิดขึ้น มีค่าเท่ากับ $n_i^{1+2f(\theta)}$ โดย $f(\theta)$ คือค่าเฉลี่ยความสมบูรณ์ของฐานข้อมูลที่แสดงถึงระดับความเหมือนกันของรายการทั้งหมดในฐานข้อมูล กล่าวคือ ถ้ามีความเหมือนกันมาก แสดงว่าเรคคอร์ดต่าง ๆ สามารถเชื่อมโยงถึงกันได้ จำนวนลิงค์ที่เป็นไปได้ก็จะลดลง ส่งผลให้ค่า Criterion สูงขึ้น (มีความเป็นไปได้ที่จะ อยู่คลัสเตอร์เดียวกันสูงขึ้น)

การคำนวนหาค่า f(θ) ที่เหมาะสมนั้นสามารถหาได้หลายวิธี แต่วิธีหนึ่งที่ง่าย และทำงานได้ดีคือการใช้ค่า f(θ) เท่ากับ (1-θ)/(1+θ) กล่าวคือ ถ้าค่าเป็น 1 หมายถึงข้อมูล ทั้งหมดในฐานข้อมูลมีความเหมือนกันทุกรายการ

ฟังก์ ชั่นที่ใช้วัดค่าความเหมือนกันของแต่ละคลัสเตอร์ นอกจากการใช้ Criterion Function แล้วยังมีอีกฟังก์ชั่นที่มีการทำงานในลักษณะเดียวกัน เรียกว่าค่า Goodness measure

ค่า Goodness measure นี้พัฒนาขึ้นมาโดยการตัดคู่ความสัมพันธ์ที่ไม่จำเป็นทิ้งไป โดยจำนวนครอสลิงค์ทั้งหมดที่เกิดขึ้น เท่ากับ $(n_i + n_j)^{1+2f(\theta)}$ โดย n_i และ n_j คือจำนวนเรคคอร์ดในคลัสเตอร์ i และ คลัสเตอร์ j ตามลำดับ จากนั้นทำการลบด้วยจำนวนครอสลิงค์ที่ลิงค์เข้าหาตัวเอง ซึ่งก็คือ $n_i^{1+2f(\theta)}$ และ $n_j^{1+2f(\theta)}$ การตัดคู่ความสัมพันธ์ที่ไม่จำเป็นนี้ เป็นการลดเวลาในการหาลิงค์ระหว่างพอยท์ทั้งหมดในฐานข้อมูล โดยสามารถแสดงสมการการคำนวณหาค่า Goodness measure ดังรูปที่ 2.5

$$G(Ci,Cj) = \frac{link[Ci,Cj]}{(n_i+n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}}$$

รูปที่ 2.5 สามการการคำนวณหาค่า Goodness measure

ค่า Link(Ci, Cj) คือจำนวนครอสลิงค์ระหว่าง คลัสเตอร์ Ci, Cj ซึ่งหา ได้จากผลรวมค่าลิงค์ของทุก ๆพอยท์ในคลัสเตอร์ i กับทุก ๆพอยท์ในคลัสเตอร์ j สามารถแสดงด้วยสมการดังรูปที่ 2.6

$$\sum_{Pq \in Ci, Pr \in Cj} link(Pq, Pr)$$

รูปที่ 2.6 การหาค่าลิงค์ระหว่างคลัสเตอร์ใดๆ

2.7.5 ขั้นตอนการทำงานของอัลกอริทึม ROCK

1. กำหนดค่า Threshold ของ Similarity(θ) เพื่อใช้วัดระดับความเหมือนของข้อมูลใน แต่ละคลัสเตอร์

2. คำนวณหาค่าลิงค์และค่า Similarity ของทุก ๆพอยท์ในฐานข้อมูล โดยใช้ Criterion Function
3. เก็บค่า Similarity ที่มากที่สุดของแต่ละพอยท์ไว้ในตารางแยกต่างหาก
4. พิจารณานำพอยท์คู่ที่มีค่าSimilarity มากที่สุดมาทำการรวมเพื่อสร้างเป็นคลัสเตอร์ใหม่
5. ปรับปรุงค่า Similarity ของคลัสเตอร์ใหม่นี้กับทุก ๆพอยท์ในฐานข้อมูล
6. ลบรายการของพอยท์ที่ถูกรวมแล้วออกจากฐานข้อมูล
7. วนลูปการทำงานจนกระทั่งไม่มีพอยท์คู่ใด ๆ มีค่า Similarity สูงกว่าค่า Threshold ที่กำหนดไว้

บทที่ 3
การวิเคราะห์และออกแบบระบบ

3.1. การวิเคราะห์ความต้องการของระบบ

ในกระบวนการวิเคราะห์และออกแบบการทำคาค้าไมนิ่งระบบขายแพคเกจทัวร์นั้น จำเป็นต้องมีการศึกษาความต้องการของลูกค้า ว่ามีความต้องการแบบใด ต้องการเที่ยวรูปแบบใด สถานที่ที่ชอบ โดยสามารถแยกตามประเภทของลูกค้าได้ เช่น แยกตามช่วงอายุ แยกตามรายได้ เป็นต้น ซึ่งผู้บริหาร อาจสามารถตั้งข้อสมมติฐานต่าง ๆ เพื่อนำข้อมูลที่ได้ มาเพื่อประกอบการตัดสินใจโดยสามารถยกตัวอย่างความต้องการสารสนเทศเพื่อการตัดสินใจของผู้บริหารได้แก่ ต้องการทราบปริมาณนักท่องเที่ยวที่มีช่วงอายุระหว่าง 15-24 ปี หรือ ความต้องการแพคเกจท่องเที่ยว ของลูกค้าที่มีรายได้ตั้งแต่ 1 หมื่นบาทต่อเดือน เป็นต้น ดังนั้น จึงต้องทำการรวบรวมข้อมูลจากกลุ่มลูกค้าประเภทต่าง ๆ และสร้างฐานข้อมูลมารองรับข้อมูล หลังจากนั้นข้อมูลที่นำมาทำการวิเคราะห์เพื่อแบ่งกลุ่มลูกค้าในการจัดหาแพคเกจทัวร์ ที่เหมาะสมนั้นได้แก่ พฤติกรรมในการซื้อทัวร์และความสนใจ รายได้ อายุ เพศ เป็นต้น

3.2 การออกแบบฐานข้อมูลในระบบขายแพคเกจทัวร์

ในการออกแบบฐานข้อมูลของระบบขายแพคเกจทัวร์นี้ได้ทำการออกแบบโดยคัดเลือกข้อมูลที่สนใจและเกี่ยวข้องกับการวิเคราะห์เพื่อหาความสัมพันธ์ โดยทำการเก็บข้อมูลต่าง ๆ ที่สนใจได้แก่ ข้อมูลของลูกค้า ข้อมูลของการท่องเที่ยวโดยข้อมูลเกี่ยวกับลูกค้าที่เราสนใจได้แก่ เพศ อายุ อาชีพ สถานะสมรส รายได้ กิจกรรมที่ทำ จุดประสงค์ในการเดินทางมา สถานที่ และแหล่งข้อมูลในการท่องเที่ยวครั้งนั้นๆ เพื่อนำไปจัดกลุ่มในการวิเคราะห์ ซึ่งในการออกแบบฐานข้อมูลในระบบขายแพคเกจทัวร์นั้น จะจัดเก็บข้อมูลไว้ในฐานข้อมูล ซึ่งได้ทำการจัดเก็บข้อมูลไว้ในตาราง โดยสนใจที่ข้อมูลของลูกค้าเป็นหลัก และมีรายละเอียดข้อมูลประเภทต่าง ๆ ที่ใช้ประกอบการวิเคราะห์ดังนี้

ตารางที่3.1 ข้อมูลลูกค้าที่นำมาทำการคลัสเตอร์

ฟิลด์	ประเภทข้อมูล	ความหมาย
ID	Numeric	ID ของลูกค้าแต่ละคน
Gender	Text	เพศ
Age	Numeric	อายุ
MarriageStatus	Text	สถานะสมรส
Occupation	Text	อาชีพ
Income	Numeric	รายได้
Activities	Text	กิจกรรมที่ทำ
SourceInfo	Text	แหล่งข้อมูลในการท่องเที่ยว
Location	Text	สถานที่ท่องเที่ยวที่ไป

และนอกจากนี้ ยังมีการออกแบบตารางที่จำเป็นในการเก็บข้อมูลของอัลกอริทึม ROCK โดยแปลงข้อมูลเดิมให้เป็นตัวเลขทางคณิตศาสตร์ เพื่อความสะดวกในการประมวลผลการหาค่าลิงค์ของแต่ละพอยท์ ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 ตาราง TransformItem การแปลงข้อมูลให้อยู่ในรูปที่เหมาะสม

ฟิลด์	ประเภทข้อมูล	ความหมาย
Item_Old	Text	ข้อมูลทั้งหมดของลูกค้า ได้แก่ เพศ อายุ สถานะสมรส อาชีพ รายได้ กิจกรรมที่ทำ แหล่งข้อมูลในการท่องเที่ยว และสถานที่ที่ไป
Item_new	Numeric	ข้อมูลทั้งหมดของลูกค้า จะถูกแปลงให้อยู่ในรูปตัวเลข เพื่อนำไปคำนวณตามอัลกอริทึมต่อไป

ตารางที่ 3.3 ตาราง TransformValue การแปลงค่าให้อยู่ในรูปที่เหมาะสม

ฟิลด์	ประเภทข้อมูล	ความหมาย
Value_Old	Text	ค่าที่เป็นไปได้ของข้อมูลทั้งหมดของลูกค้า ได้แก่ เพศ ได้แก่ ชาย หญิง เป็นต้น
Value_new	Numeric	ค่าใหม่ที่ถูกแปลงให้อยู่ในรูปตัวเลข เพื่อนำไปคำนวณตามอัลกอริทึมต่อไป

และหลังจากประมวลผล จะมีคลัสเตอร์ใหม่เกิดขึ้น ต้องมีการจัดเก็บค่าแอททริบิวท์ ของทุก ๆ พอยท์และจำถูกนำไปหาค่าลิงค์กับพอยท์อื่น ๆ และนำไปรวมกับคลัสเตอร์อื่นที่เหมาะสมต่อไป ซึ่งข้อมูลเหล่านี้จะถูกจัดเก็บในตาราง Local ดังแสดงในตารางที่ 3.4 ส่วนข้อมูลของคลัสเตอร์ใหม่ที่ถูกรวม (Merge) จะถูกเก็บไว้ในตาราง Temp เพื่อนำไปใช้ในการหาค่า Goodness measure ของทุก ๆ พอยท์ที่มีต่อคลัสเตอร์ใหม่นี้ต่อไป ส่วนผลลัพธ์จากการคำนวณค่า Goodness measure จะถูกจัดเก็บไว้ในตาราง Global ซึ่งจะกล่าวถึงต่อไป

ตารางที่ 3.4 ตาราง Local และตาราง Temp การแปลงค่าให้อยู่ในรูปที่เหมาะสม

ฟิลด์	ประเภทข้อมูล	ความหมาย
ID	Numeric	ID ของลูกค้าแต่ละคน
Item	Numeric	ชื่อ Item ต่าง ๆ ของลูกค้า เช่น เพศ อายุ อาชีพ รายได้ เป็นต้น
Value	Numeric	ค่าของแต่ละ Item เช่น เพศ มีค่าเป็น ชาย หรือ หญิง เป็นต้น

หลังจากได้คำนวณหาค่าลิงค์และค่า Similarity ของทุก ๆ พอยท์ในฐานข้อมูล โดยใช้ Criterion Function ซึ่งในที่นี้ใช้ Goodness measure จะจัดเก็บข้อมูลดังแสดงในตารางที่ 3.5

ตารางที่ 3.5 ตารางGlobal

ฟิลด์	ประเภทข้อมูล	ความหมาย
ID_Linking	Numeric	ID ที่ลิงค์กับพอยท์อื่น
ID_Linked	Numeric	ID ที่ถูกลิงค์จากลิงค์พอยท์ อื่น
SimValue	Numeric	ค่า Goodness measure ที่ได้จากการคำนวณระหว่างพอยท์ต่างๆ

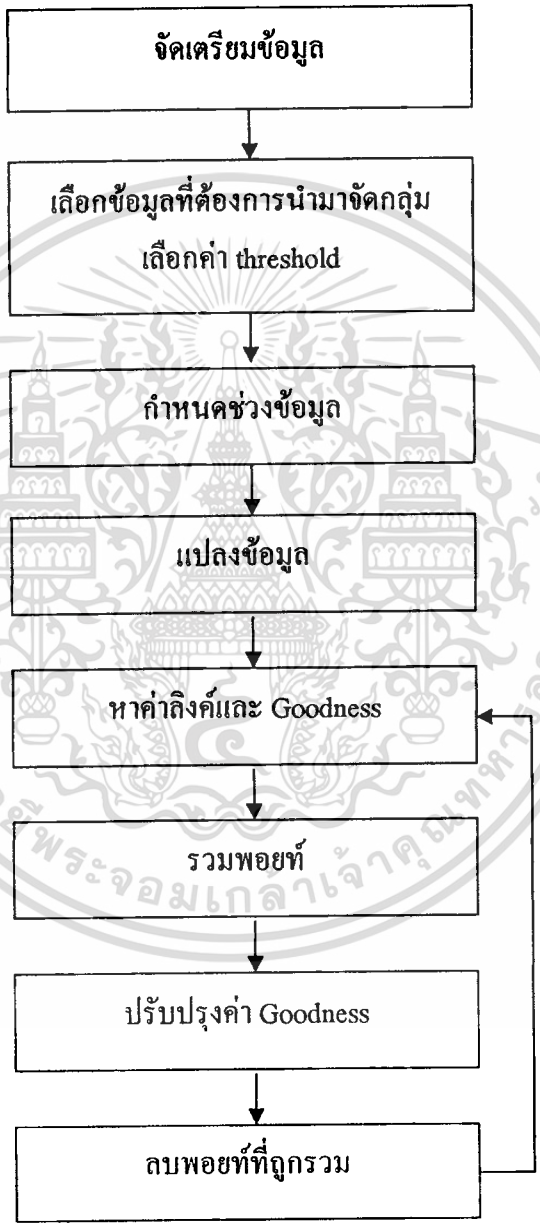
และนอกจากนั้น ยังมีการเก็บค่าจากการทำ Cartesian Product ของตาราง Local และ ตาราง Temp เพื่อแสดงจำนวนครอสลิงค์ระหว่างพอยท์ทั้งหมดในฐานข้อมูลกับคลัสเตอร์ใหม่ที่เกิดขึ้น ระหว่างการทำคลัสเตอร์ ดังแสดงในตารางที่ 3.6

ตารางที่ 3.6 ตาราง Link ที่เกิดจากการทำ Cartesian Product

ฟิลด์	ประเภทข้อมูล	ความหมาย
ID_Linking	Numeric	ID ที่ลิงค์กับพอยท์อื่น
ID_Linked	Numeric	ID ที่ถูกลิงค์จากลิงค์พอยท์อื่น
Item_Linking	Numeric	Item ของพอยท์ที่ลิงค์ไปยังพอยท์อื่น
Item_Linked	Numeric	Item ของพอยท์ที่ถูกลิงค์
Value_Linking	Numeric	Value ของ Item ที่ ลิงค์ไปยังพอยท์อื่น
Value_Linking	Numeric	Value ของ Item ที่ถูกลิงค์
Diff_ID	Numeric	ความแตกต่างของค่า ID ของคู่ พอยท์ ที่ลิงค์ กัน
Diff_Item	Numeric	ความแตกต่างของค่า Item ของคู่ พอยท์ที่ลิงค์กัน
Diff_Value	Numeric	ความแตกต่างของค่า Value ของคู่ พอยท์ที่ลิงค์กัน

3.3 รายละเอียดการทำงานของโปรแกรม

ระบบงานนี้แบ่งการทำงานออกเป็น 3 ส่วนหลัก คือ ส่วนที่ผู้ใช้ ทำการเลือกข้อมูลที่ต้องการนำไปทำการคลัสเตอร์ ส่วนการดำเนินการตามอัลกอริทึม ROCK และส่วนการจัดทำรายงาน โดยแต่ละส่วน มีขั้นตอนดังแสดงในรูปที่ 3.1



รูปที่ 3.1 ขั้นตอนการทำงานของโปรแกรม

จากรูปที่ 3.1 สามารถแสดงรายละเอียดการทำงานของโปรแกรมได้ดังนี้

3.3.1 การเลือกข้อมูล

- ผู้ใช้ทำการเลือกข้อมูลที่ต้องการนำไปทำคลัสเตอร์
- การ Transform ข้อมูลให้อยู่ในรูปที่เหมาะสมในการทำคลัสเตอร์

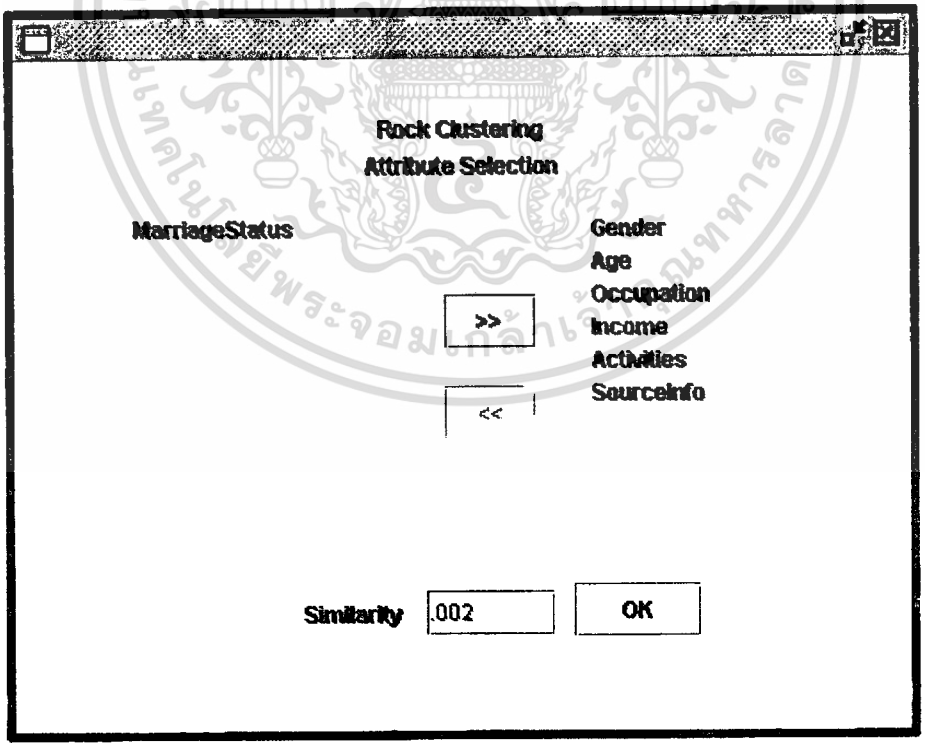
3.3.2 การดำเนินการตามอัลกอริทึม ROCK

- สร้างและปรับปรุงลิงค์ของแต่ละพอยท์
- คำนวณหาค่า Goodness measure ระหว่างพอยท์และคลัสเตอร์
- รวมพอยท์เพื่อสร้างคลัสเตอร์ใหม่ (Merging)

3.3.3 จัดทำรายงานและสรุปผลจากการคลัสเตอร์

3.3.1 การเลือกข้อมูล

เป็นขั้นตอนที่ผู้ใช้ ต้องทำการคัดเลือกแอททริบิวท์ที่ต้องการนำมาทำการคลัสเตอร์ จากแอททริบิวท์ทั้งหมดที่มีอยู่ และกำหนดค่า Threshold เพื่อใช้วัดระดับความเหมือนของข้อมูลในแต่ละคลัสเตอร์ ซึ่งสามารถแสดงตัวอย่างการเลือกข้อมูลได้ดังรูปที่ 3.2



รูปที่ 3.2 การเลือกข้อมูลและกำหนดค่า Threshold

และหลังจากนั้น ทำลดกลุ่มแอททริบิวท์ก่อนนำไปดำเนินการตามอัลกอริทึม เช่น จัดการแบ่งกลุ่มอายุของลูกค้า เพื่อเป็นแปลงข้อมูลตัวเลขที่มีความต่อเนื่องให้เป็นข้อมูลแบบ Categorical และแบ่งเป็นช่วง ๆ เช่น อายุ รายได้ เป็นต้น ขั้นตอนนี้ เป็นขั้นตอนที่จำเป็นอีกขั้นตอนหนึ่ง เรียกว่า การทำ Data Discretization เนื่องจาก อัลกอริทึม ROCK เหมาะสมสำหรับข้อมูลประเภท Categorical โดยพิจารณาค่าความเหมือนกันของแต่ละพอยท์จากค่าลิงค์ของแต่ละพอยท์ดังนั้น ข้อมูลประเภทมีค่าความต่อเนื่องมาก จะมีโอกาสน้อยที่จะลิงค์กันได้ ซึ่งตัวอย่างการทำ Discretization สามารถแสดงได้ตามตารางที่ 3.7

ตารางที่ 3.7 การทำ Data Discretization ข้อมูลอายุลูกค้า

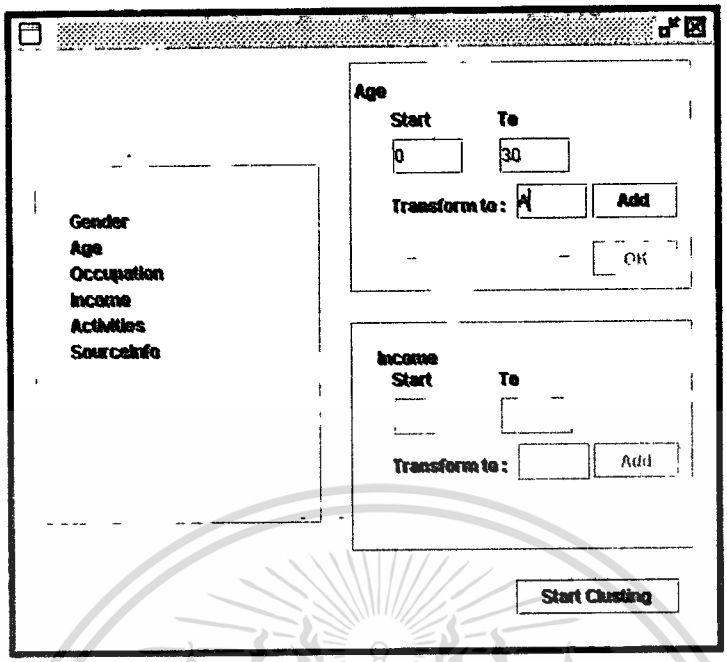
ข้อมูลเดิม	ข้อมูลใหม่
13	A_<15
25	B_15-30
30	C_30-45
70	E_60+

ซึ่งการทำ Data Discretization นี้ ผู้ใช้สามารถกำหนดช่วงของข้อมูลของแต่ละรายการได้เองดังแสดงในรูปที่ 3.3

หลังจากที่ได้ทำการคัดเลือกและ Transform ข้อมูลแล้วจะนำข้อมูลมาเก็บไว้ในตาราง Local และ ตาราง Temp ซึ่งเป็นตารางหลักในการทำงานตามอัลกอริทึม ROCK โดยขั้นตอนนี้เป็นขั้นตอนการทำงานในโปรแกรม

1. การสร้างตาราง Local

ตารางนี้ เกิดจากการหมุนตารางข้อมูลลูกค้า (Tourist) โดยจะแปลง Item ต่าง ๆ ซึ่งเป็นชื่อฟิลด์ของตารางข้อมูลหลัก ให้กลายเป็น Value ของ Item ใหม่ ที่มีชื่อว่า "Item" ส่วนค่าของแต่ละฟิลด์เดิมนั้น จะถูกหมุนให้กลายเป็น Value ของฟิลด์ใหม่ชื่อว่า "Value" เช่น ลูกค้าคนหนึ่ง เพศชาย อายุ 19 อาชีพนักศึกษา รายได้ 8000 สามารถแปลงได้ดังรูปที่ 3.4



รูปที่ 3.3 การทำ Data Discretization ข้อมูล

ID	Age	Gender	Occupation	Income	Activities	SourceInfo
1	19	M	Student	8000	Adventure	Website
2	21	M	Student	10000	Adventure	Website
3	24	F	Individual	10000	Sightseeing	Magazine

ID	Item	Value
1	Age	19
1	Gender	M
1	Occupation	Student
1	Income	8000
1	Activities	Adventure
1	SourceInfo	Website
2	Age	21
2	Gender	M
2	Occupation	Student
2	Income	108000
2	Activities	Adventure
2	SourceInfo	Website

รูปที่ 3.4 การหมุนตารางต้นทางเป็นตาราง Local

การสร้างคลัสเตอร์ใหม่ตามอัลกอริทึม ROCK นั้น จะเกิดจากการทำ Join Operation ของ Item ต่าง ๆ ของแต่ละพอยท์เกิดเป็นเซ็ตข้อมูลใหม่ การใช้ตารางในลักษณะนี้จะช่วยให้การเพิ่มเซ็ตข้อมูลของคลัสเตอร์ใหม่ทำได้ง่าย รวมถึงสามารถสรุปเซ็ตข้อมูลของแต่ละคลัสเตอร์ได้ง่ายขึ้นด้วย เช่น นำพอยท์ 1 รวมกับพอยท์ 2 เพื่อสร้างคลัสเตอร์ใหม่ สุดท้าย จะได้ตารางใหม่ ดังรูปที่ 3.5

ID	Item	Value
1	Age	19
1	Gender	M
1	Occupation	Student
1	Income	8000
1	Activities	Adventure
1	SourceInfo	Website
2	Age	21
2	Gender	M
2	Occupation	Student
2	Income	10000
2	Activities	Adventure
2	SourceInfo	Website
3	Age	24
3	Gender	F
3	Occupation	Individual
3	Income	10000
3	Activities	Sightseeing
3	SourceInfo	Magazine

ก่อนรวม

ID	Item	Value
1	Age	19
12	Gender	M
12	Occupation	Student
1	Income	8000
12	Activities	Adventure
12	SourceInfo	Website
2	Age	21
23	Income	10000
3	Age	24
3	Gender	F
3	Occupation	Individual
3	Activities	Sightseeing
3	SourceInfo	Magazine

หลัง รวม

รูปที่ 3.5 การสร้างคลัสเตอร์ใหม่

หลังจากนั้น จะทำการแทนข้อมูลในตาราง Local ให้เป็นตัวเลขทั้งหมด เพื่อนำไปคำนวณหาค่าลิงค์ต่อไป

ส่วนตาราง Temp จะเป็นตารางชั่วคราวในการเก็บพอยท์ต่าง ๆ ของคลัสเตอร์ใหม่เพื่อนำไปสร้างความสัมพันธ์ของแต่ละ พอยท์ ในคลัสเตอร์นี้ กับ พอยท์ ต่าง ๆ ในตาราง Local ซึ่งความสัมพันธ์นี้ ก็คือลิงค์ระหว่างพอยท์ ต่าง ๆ นั่นเอง

เมื่อเริ่มการคลัสเตอร์นั้น จะถือว่า แต่ละพอยท์ก็คือแต่ละคลัสเตอร์ ดังนั้นก่อนการดำเนินการตามขั้นตอนของอัลกอริทึม ROCK ต้องทำการคัดลอกข้อมูลจากตาราง Local มาเก็บไว้

ในตาราง Temp เพื่อสร้าง ลิงค์ ของทุก ๆ พอยท์ ในฐานข้อมูล และใช้คำนวณหาค่า Goodness ที่จะใช้ในการสร้างคลัสเตอร์ต่อไป และในรอบการวนลูปถัดไป ตาราง Temp นี้ จะเก็บเฉพาะ Item และ Value ของคลัสเตอร์ใหม่ที่เกิดขึ้น เพื่อทำการหาค่า Goodness ที่สัมพันธ์กับคลัสเตอร์ใหม่เท่านั้น

3.3.2 การดำเนินการตามอัลกอริทึม ROCK

การทำงานของอัลการิทึม ROCK ในการคลัสเตอร์ ข้อมูลเป็นการทำงานแบบ iterative คือ วนลูปการทางานไปเรื่อย ๆ ตราบใดที่ค่า Goodness ของคลัสเตอร์ยังมีค่าอยู่ในเกณฑ์ที่กำหนดไว้ ดังหน้าที่หลักในส่วนนี้ สามารถแบ่งย่อยได้ดังนี้

- สร้างและปรับปรุงข้อมูลลิงค์ของแต่ละพอยท์
- คำนวณหาค่า Goodness ระหว่างพอยท์และคลัสเตอร์
- รวมพอยท์เพื่อสร้างคลัสเตอร์ใหม่ (Merging)

3.3.2.1 การสร้างและปรับปรุงข้อมูลลิงค์ของแต่ละพอยท์

การสร้างลิงค์ในระบบงานนี้ ใช้วิธีการทำ Cartesian product operation ระหว่างพอยท์ต่าง ๆ ในตาราง Local และตาราง Temp และเมื่อทำ Cartesian product แล้ว จะทำให้ทุก ๆ Value เกิดการลิงค์กันในทุกทิศทาง ผลลัพธ์จากการทำ Cartesian product จะเก็บไว้ในตาราง Link ซึ่งต้องมีขั้นตอนการกำจัดลิงค์ที่ไม่ถูกต้อง อันเกิดจากการลิงค์ของ Item คนละประเภทกันออกไป เช่น ลิงค์ของ Item "Gender" กับ Item "Occupation" เป็นต้น ซึ่งลิงค์ที่ไม่ถูกต้องเหล่านี้ หาได้จากการนำ Value ของ Item ของแต่ละพอยท์มาลบกัน ถ้าได้ค่าเป็น 0 แสดงว่าเป็นการลิงค์ของ Item ประเภทเดียวกัน หากมีค่าไม่เท่ากับ 0 แสดงว่าเป็นการลิงค์ของ Item คนละประเภทกัน ถือได้ว่าเป็นลิงค์ที่ไม่ถูกต้อง

3.3.2.2 การคำนวณหาค่า Goodness ระหว่างพอยท์และคลัสเตอร์

หลังจากคำนวณลิงค์ระหว่างพอยท์ต่าง ๆ แล้วต้องมาทำการคำนวณหาค่า Goodness ระหว่างพอยท์ต่าง ๆ จากข้อมูลในตารางลิงค์ ผลลัพธ์ที่ได้ จะเก็บไว้ในตาราง Global ซึ่งจะนำไปใช้ในการพิจารณาหาคลัสเตอร์ใหม่ที่เหมาะสมในการรวมต่อไป

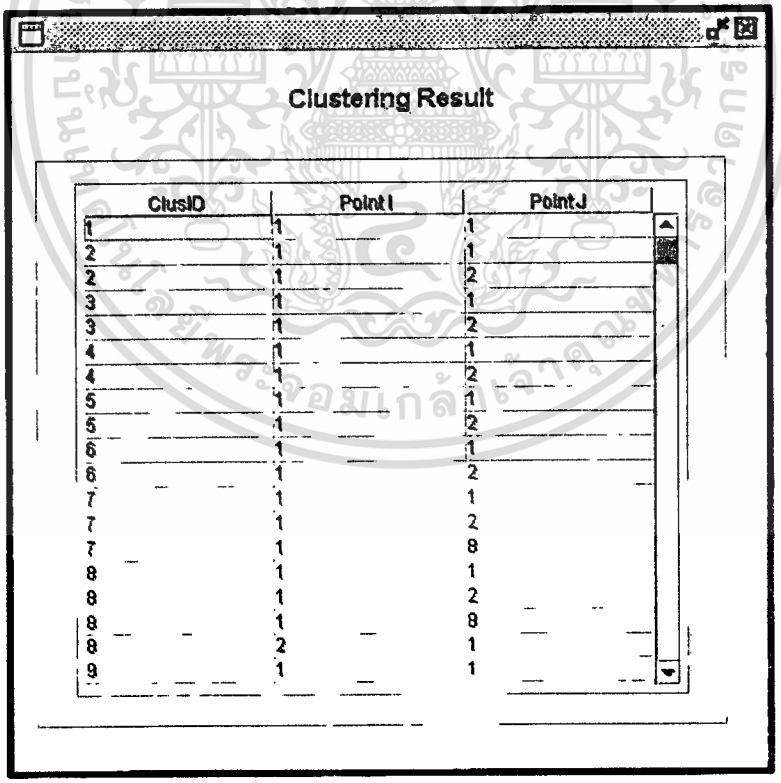
3.3.2.3 การรวมพอยท์เพื่อสร้างคลัสเตอร์ใหม่ (Merging)

1. พิจารณาค่า Goodness ของคู่พอยท์จากตาราง Global ที่มีค่ามากที่สุดที่มากกว่าค่า Threshold ที่กำหนดไว้ จากนั้นทำการรวมพอยท์ของคลัสเตอร์คู่นั้น
2. รวมพอยท์คู่นั้นแล้วสร้างคลัสเตอร์ใหม่ขึ้นมา

3. บันทึกข้อมูลคลัสเตอร์ใหม่นี้ในตาราง Local พร้อมทั้งลบเรคคอร์ดที่ถูกรวมแล้วออกจากตาราง
4. ทำการปรับปรุงตาราง Global ใหม่โดยลบทุกเรคคอร์ดที่ลิงค์ไปยังพอยท์ที่ถูกรวม
5. บันทึกข้อมูลของคลัสเตอร์ใหม่นี้ลงในตาราง Temp
6. วนลูประหว่างการหาค่า Goodness ระหว่างพอยท์และคลัสเตอร์เพื่อคำนวณค่า Goodness ของทุกพอยท์กับคลัสเตอร์ใหม่นี้ แล้วปรับปรุงค่า Goodness ในตาราง Global
7. การทำงานนี้ จะสิ้นสุดเมื่อไม่มีค่า Goodness ของพอยท์คู่ใดที่มากกว่าค่า Threshold ที่ตั้งไว้ ผลลัพธ์ที่ได้จากการทำงาน จะเก็บไว้ในตาราง Local

3.3.3 การจัดทำรายงานและสรุปผล

เมื่อทำการคลัสเตอร์ตามอัลกอริทึมจนเสร็จสมบูรณ์แล้ว จำนวนเรคคอร์ที่เหลือทั้งหมดในตาราง Local ก็คือจำนวนคลัสเตอร์ที่ได้จากการดำเนินการ ดังแสดงในรูปที่ 3.6



รูปที่ 3.6 ผลจากการแบ่งกลุ่ม

หลังจากนั้นทำการแปลงค่าต่าง ๆ ให้กลับไปเป็นข้อมูลเดิม และจัดเป็นรูปแบบรายงานเพื่อให้สะดวกในการวิเคราะห์และพิจารณาต่อไป ดังตารางที่ 3.8

ตารางที่ 3.8 ผลการวิเคราะห์ผลลัพธ์

ClusterNo	Gender	Age	Occupation	Income	Activities	SourceInfo
1	M	A(0-30)	Student, individual	A(0-10000), B(10001-20000)	Adventure, Camping	Website
2	M,F	A(0-30), B(31-60)	Student, individual, Government	A(0-10000), B(10001-20000), C(20001-30000)	Camping, Sightseeing	Website, Magazine, Tour Agency

จากตารางที่ 3.8 สามารถอธิบายโดยใช้ตารางประกอบการอธิบายได้ดังนี้คือ สามารถจัดกลุ่มลูกค้าจากการเลือกข้อมูลในการจัดกลุ่มได้แก่ เพศ อายุ อาชีพ รายได้ กิจกรรมที่จำ และแหล่งข้อมูลการท่องเที่ยว ได้ 2 กลุ่ม คือ

กลุ่มที่ 1 เพศชาย อายุไม่เกิน 30 ปี อาชีพ นักศึกษาและ ทำงานเอกชน และมีรายได้ไม่เกิน 20000 บาท กิจกรรมที่ทำได้แก่ แนวผจญภัยและตั้งแคมป์ โดยมักหาข้อมูลจากเว็บไซท์ เพราะฉะนั้น แพคเกจทัวร์ที่ควรจัดให้ลูกค้ากลุ่มนี้ ควรมุ่งเป้าหมายไปที่ลูกค้าเพศชาย อายุไม่เกิน 30 ปี และควรจัดให้มีกิจกรรมแนวผจญภัยและตั้งแคมป์ไว้ในแพคเกจ

กลุ่มที่ 2 เพศชายและหญิง อายุไม่เกิน 60 ปี อาชีพ นักศึกษา ทำงานเอกชน และรับราชการ และมีรายได้ไม่เกิน 30000 บาท กิจกรรมที่ทำได้แก่ ตั้งแคมป์และชมวิว โดยมักหาข้อมูลจากเว็บไซท์ หนังสือนำเที่ยวและบริษัททัวร์ เพราะฉะนั้นสามารถวิเคราะห์ได้ว่าควรมีการจัดแพคเกจสำหรับลูกค้าทั่วไปไว้รองรับลูกค้ากลุ่มนี้ โดยจัดกิจกรรมการตั้งแคมป์และชมวิวไว้ในแพคเกจ

บทที่ 4
บทสรุป

4.1 สรุปโครงการ

การนำ Data Mining มาใช้จัดกลุ่มข้อมูลหรือการทำคลัสเตอร์นั้น จะทำให้ผู้ประกอบกิจการทราบถึงพฤติกรรมโดยรวมของกลุ่มลูกค้า และทราบว่า ควรจะจัดลูกค้าประเภทใดไว้ด้วยกันบ้าง เพื่อที่จะจัดการแพคเกจการท่องเที่ยวให้เหมาะสมกับความต้องการของลูกค้าในแต่ละกลุ่ม พร้อมทั้งยังสามารถนำมาปรับปรุงกลยุทธ์ทางการตลาดให้สามารถเข้าถึงกลุ่มลูกค้าได้ตรงตามความต้องการมากขึ้นทั้งในด้าน ราคา สถานที่และกิจกรรมที่จัดไว้ภายในแพคเกจทัวร์ เป็นต้น

4.2 ข้อควรพิจารณาเพิ่มเติม

การทำงานตามอัลกอริทึมนี้ จำเป็นต้องใช้พื้นที่ในหน่วยความจำค่อนข้างมาก ในการเก็บค่า link ทั้งหมดที่เป็นไปได้ของทุก ๆ เรคคอร์ดในฐานข้อมูล หากข้อมูลที่ต้องการพิจารณามีจำนวน Item มาก ก็จะเกิดจำนวน link ที่เป็นไปได้มากยิ่งขึ้น และใช้พื้นที่หน่วยความจำมากขึ้นตามไปด้วย

บรรณานุกรม

กิตติ ภักดีวัฒนะกุล. 2546a. **คัมภีร์ JAVA.** เล่ม 1. กรุงเทพฯ:เคทีพี คอมแอนด์คอนซัลท์.

กิตติ ภักดีวัฒนะกุล. 2546b. **คัมภีร์ JAVA.** เล่ม 2. กรุงเทพฯ:เคทีพี คอมแอนด์คอนซัลท์.

Berson, Alex and Smith, Stepen J. 1997. **Data Warehousing, Data Mining & OLTP**. San Francisco. McGraw-Hill.

Guha, S., Rastogi, R. and Shim, K. 1999. "ROCK: A Robust Clustering Algorithm for Categorical Attributes". **Information Systems**. 1(25):345-366.

Han ,J. and Kamber, M. 2001. **Data Mining : Concepts and Techniques**. San Francisco: Simon Fraser University.

Silberschatz,A, Korth, H. and Sudarshan, S. 2002. **Database System Concepts**. 4th Edition. New York. Mc Graw Hill.

Tuterial on High Performance Data Mining. 2003 [Online] Available: http://www.users.cs.umn.edu/~mjoshi/hpdmtut/

ประวัติผู้เขียน

ชื่อผู้เขียน	นางสาวสุจิตรา มังคละไชยา
วันเดือนปีเกิด	8 กันยายน 2522
สถานที่เกิด	อำเภอเมือง จังหวัดเลย
ที่อยู่ปัจจุบัน	31/1 เพชรเกษม 36 บางหว้า ภาษีเจริญ กรุงเทพฯ 10160
วุฒิการศึกษาปริญญาตรี	วท.บ.(วิทยาการคอมพิวเตอร์ประยุกต์)
สถานที่สำเร็จการศึกษาระดับปริญญาตรี	คณะวิทยาศาสตร์ประยุกต์ สถาบันเทคโนโลยี พระจอมเกล้าพระนครเหนือ
ปีที่สำเร็จการศึกษา	ปีการศึกษา 2543