วิทยานิพนธ์

ระบบวิเคราะห์คำถาม และตอบปัญหาการใช้งานอินเตอร์เน็ต ด้วยเทคนิคการทำเหมืองข้อมูล

QUESTION ANALYSIS AND ANSWERING SYSTEM FOR THE INTERNET USAGE BY USING DATA MINING TECHNIQUES

นายนฤป รักงาม

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2550

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)

ปริญญา

วิทยาการคอมพิวเตอร์	วิทยาการคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง ระบบวิเคราะห์คำถาม และ ตอบปัญหาการใช้งานอินเตอร์เน็ตด้วยเทคนิคการทำเหมืองข้อมูล

Question Analysis and Answering System for the Internet Usage by Using Data Mining Techniques

นามผู้วิจัย นายนฤป รักงาม

ได้พิจารณาเห็นชอบโดย

ประธานกรรมการ .....

(รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D.)

กรรมการ .....

(อาจารย์ชาคริต วัชโรภาส, Ph.D.)

กรรมการ .....

(ผู้ช่วยศาสตราจารย์สมชาย นำประเสริฐชัย , Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์อุมาพร ศิรธรานนท์, M.S.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์วินัย อาจคงหาญ, M.A.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 5 **เดือน** มิถุนายน **พ.ศ.** 2550

วิทยานิพนธ์

เรื่อง

ระบบวิเคราะห์คำถาม และตอบปัญหาการใช้งานอินเตอร์เน็ตด้วยเทคนิคการทำเหมืองข้อมูล

Question Analysis and Answering System for the Internet Usage by Using Data Mining Techniques

โดย

นายนฤป รักงาม

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
พ.ศ. 2550

นฤป รักงาม 2550: ระบบวิเคราะห์คำถาม และตอบปัญหาการใช้งานอินเตอร์เน็ต ด้วยเทคนิคการทำเหมืองข้อมูล ปริญญาวิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์) สาขาวิทยาการคอมพิวเตอร์ ภาควิชาวิทยาการคอมพิวเตอร์ ประธานกรรมการที่ปรึกษา: รองศาสตราจารย์อนงค์นาฏ ศรีวิหค, Ph.D. 63 หน้า

งานวิจัยนี้นำเสนอระบบการวิเคราะห์คำถามอัตโนมัติ เพื่อสร้างระบบที่สามารถทำนายสาเหตุของปัญหาการใช้งานอินเตอร์เน็ตของลูกค้า และนำเสนอแนวทางแก้ไข โดยประยุกต์ใช้เทคนิคการสกัดคำถาม (Data Extraction) ใช้รวมกับเทคนิคทางการทำเหมืองข้อมูล (Data mining) เทคนิคการสกัดคำถามนั้นใช้สำหรับสกัดข้อมูลที่สัมพันธ์กันจากคำถามที่เกี่ยวข้องกับการใช้งานอินเตอร์เน็ตของลูกค้า เพื่อใช้เป็นข้อมูลนำเข้าในขั้นตอนการทำเหมืองข้อมูลต่อไป ขั้นตอนของการทำเหมืองข้อมูลของงานวิจัยนี้แบ่งเป็น 2 กรณี คือ (1) กรณีที่สามารถเชื่อมต่ออินเตอร์เน็ตได้ (online) จะใช้การจัดกลุ่มข้อมูลแบบ 2 ขั้นตอน (SOM และ K-Means อัลกอริทึม) ในการจัดกลุ่มของลูกค้าและใช้ OLAP Cube ในการวิเคราะห์เพื่อหาความน่าจะเป็นของสาเหตุที่ทำให้เกิดปัญหา (2) กรณีที่ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้ (offline) จะใช้โปรแกรมตัวแทน (Software Agent) ในการตรวจสอบปัญหาและใช้ต้นไม้ตัดสินใจ (Decision Tree) เพื่อหาคำตอบของปัญหาที่เกิดขึ้น ในงานวิจัยนี้ใช้ค่า Precision ในการวัดความแม่นยำ โดยผลที่ได้จากการใช้ OLAP Cube จะให้ค่าความแม่นยำร้อยละ 70.46 ส่วนการใช้ Decision tree ได้ให้ค่าความแม่นยำร้อยละ 87.56 ผลลัพธ์ที่ได้จากการใช้ระบบนี้นอกจากจะช่วยลดปริมาณการโทรสอบถามมายังพนักงาน Call center แล้ว ยังเอื้อประโยชน์แก่ลูกค้าทั่วไปที่ไม่มีความรู้ในเรื่องอินเตอร์เน็ตให้สามารถตรวจสอบแก้ไขปัญหาด้วยตนเองได้

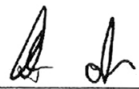
นฤป รักงาม
ลายมือชื่อนิสิต

ลายมือชื่อประธานกรรมการ

29 / 05 / 2550

Narroup Ruk-ngam 2007: Question Analysis and Answering System for the Internet Usage by Using Data Mining Techniques. Master of Science (Computer Science), Major Field: Computer Science, Department of Computer Science. Thesis Advisor: Associate Professor Anongnart Srivihok, Ph.D. 63 pages.

This study proposes a system for an automatic question analysis and answering which applies information extraction techniques for extracting the relevant information from questions of the Internet connections. The system uses data extractions and data mining techniques. User profiles and their questions on the Internet usage are used as input for the system. The system processes are divided in two categories due to the type of internet connection cases. First, online case: two steps clustering by SOM and K-Means algorithms are used to segment customer data. Then outputs from clustering and question extraction are used for OLAP to analyze for cause of the Internet problems. Second, offline case: software agent is applied; this case a decision tree is built which each node contains cause of problems. The system will give the answers for solving the Internet connections in both online and offline cases. The results from this study reveal that it is possible to develop an automatic question answering system. The proposed system had been tested with sample questions from the Internet service provider: True corporation Call Center. Precision of the system is about 70.46% for online and 87.56% for offline which is good. This study offers useful information regarding the areas of data mining for Call Center or Web Based Support Systems.

Narroup Rak-ngam  29 / 05 / 2007

Student's signature Thesis Advisor's signature

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร. อนงค์นาฏ ศรีวิหค ประธานกรรมการที่ปรึกษา สำหรับความช่วยเหลือ ความห่วงใย และคำแนะนำที่มีให้ตลอดมา ขอกราบขอบพระคุณอาจารย์ชาคริต วัชโรภาส กรรมการวิชาเอก และผู้ช่วยศาสตราจารย์สมชาย นำประเสริฐชัย กรรมการวิชารอง ที่กรุณาให้ความช่วยเหลือแต่งตั้งคณะกรรมการวิทยานิพนธ์ให้สำเร็จลุล่วงไปด้วยดี และขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์อุศนา ตัณฑุลเวศม์ ผู้แทนบัณฑิตวิทยาลัยที่กรุณาให้คำแนะนำต่อข้าพเจ้า

ขอกราบขอบพระคุณคณาจารย์ทุกท่าน ที่แนะนำ สั่งสอน และให้ความรู้แก่ข้าพเจ้าตลอดระยะเวลาการศึกษา อันเป็นประโยชน์ต่อการนำมาใช้ในวิทยานิพนธ์

ขอกราบขอบพระคุณ คุณพ่อ คุณแม่ และคนในครอบครัวของข้าพเจ้า ที่คอยให้กำลังใจพร้อมทั้งให้ความช่วยเหลือเสมอมา

ขอบคุณเพื่อน ๆ ทุกคน พี่โอ๋ พี่ตุ้ม ป๊อบ นัด และเพื่อนในบริษัททั้งหลายที่ให้กำลังใจและสนับสนุนด้านการเรียนมาตลอด เพื่อนๆในภาควิชาวิทยาการคอมพิวเตอร์ พี่พัช กุ้ง และเอื้อย ที่คอยให้กำลังใจและช่วยเหลือให้คำแนะนำในการทำวิทยานิพนธ์ตลอดมา

สุดท้ายขอขอบคุณภาควิชาวิทยาการคอมพิวเตอร์ที่ให้โอกาสข้าพเจ้าได้เรียนรู้จนสำเร็จการศึกษา

งานวิทยานิพนธ์นี้ ข้าพเจ้าหวังเป็นอย่างยิ่งว่าจะเป็นประโยชน์ต่อผู้ศึกษา ค้นคว้าและสนใจ หากมีข้อผิดพลาดประการใด ข้าพเจ้าขอน้อมรับไว้เพื่อนำไปใช้ในการปรับปรุงให้วิทยานิพนธ์นี้มีความสมบูรณ์ยิ่งขึ้น สำหรับความดีที่ได้รับจากวิทยานิพนธ์นี้ข้าพเจ้าขอมอบให้แก่ผู้มีพระคุณทุกท่าน

นฤป รักงาม

พฤษภาคม 2550

สารบัญ

หน้า

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

ระบบวิเคราะห์คำถาม และตอบปัญหาการใช้งานอินเตอร์เน็ตด้วยเทคนิคการทำเหมืองข้อมูล

Question Analysis and Answering System for the Internet Usage by Using Data Mining Techniques

คำนำ

ปัจจุบันปริมาณการใช้งานอินเตอร์เน็ตมีจำนวนเพิ่มมากขึ้นอย่างรวดเร็ว ปัญหาการใช้งานอินเตอร์เน็ตก็มีเพิ่มขึ้นด้วย ทำให้บริษัทผู้ให้บริการต้องเสียค่าใช้จ่ายในการติดตามและแก้ไขปัญหาตามมากขึ้น จากสถิติที่ได้รวบรวมเกี่ยวกับปัญหาการใช้งานอินเตอร์เน็ต พบว่าปัญหาที่เกี่ยวข้องกับการใช้งานอินเตอร์เน็ต 80 เปอร์เซ็นต์ของปัญหาเป็นปัญหาที่ง่ายต่อการแก้ไข และเป็นปัญหาที่เกิดจากด้านของตัวผู้ใช้บริการเอง ตัวอย่างเช่น การใส่ Username/Password ผิด การเสียบสายโทรศัพท์กับ Modem ไม่แน่น หรือการที่ลูกค้า Disable LAN ทำให้ไม่สามารถใช้งานอินเตอร์เน็ตได้ แต่ในปัจจุบันพบว่ากรณีที่เกิดปัญหาการใช้งานขึ้นกับลูกค้าที่ไม่มีความรู้ด้านการใช้อินเตอร์เน็ต จะโทรเข้ามาแจ้งและสอบถามวิธีการแก้ไขปัญหาที่เกิดขึ้นจากเจ้าหน้าที่ Call center ซึ่งเจ้าหน้าที่ศูนย์บริการของบริษัทก็จะต้องสอบถามถึงอาการที่พบ ตรวจสอบ และเสนอแนะวิธีการแก้ไขปัญหาทีละขั้นตอนทำให้ใช้เวลาในการแก้ไขปัญหาของลูกค้า

ในงานวิจัยนี้มุ่งเน้นเพื่อศึกษา และค้นคิดวิธีการสร้างระบบที่สามารถทำนายได้ว่าปัญหาการใช้งานอินเตอร์เน็ตที่เกิดขึ้นกับลูกค้ามีสาเหตุเกิดจากอะไร และมีวิธีการแก้ไขปัญหาการใช้งานได้อย่างไร ระบบนี้นอกจากจะช่วยลดปริมาณการโทรสอบถามมาที่พนักงาน Call center แล้วยังสามารถเอื้อประโยชน์แก่ลูกค้าทั่วไปที่ไม่มีความรู้ในเรื่องอินเตอร์เน็ตให้สามารถตรวจสอบแก้ไขปัญหาได้เอง ซึ่งในการสร้างระบบดังกล่าวจะแบ่งวิธีการแก้ไขปัญหาออกเป็น 2 ส่วนหลัก ๆ ตามปัญหาที่พบคือ

1) ปัญหาการใช้งานที่เกิดขึ้นทำให้ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้ (Offline) โดยจะสร้าง Software agent ตรวจสอบอุปกรณ์ต่าง ๆ ที่มีผลต่อการใช้งานอินเตอร์เน็ตเพื่อวิเคราะห์ปัญหาและหาสาเหตุในการแก้ไข

2) ปัญหาการใช้งานที่เกิดในกรณีการเชื่อมต่ออินเตอร์เน็ตได้ (Online) จะประยุกต์ระบบอัตโนมัติ คือทำการจับกลุ่มประเภทลูกค้าที่ใช้งานว่าในแต่ละกลุ่มกรณีเกิดปัญหาการใช้งานในแต่ละอาการมีวิธีการแก้ไขปัญหาอย่างไร เพื่อจะได้นำเสนอวิธีการแก้ไขได้ตรงกลุ่ม

วัตถุประสงค์

1. ศึกษาเทคนิคและวิธีการที่ใช้ในการทำนายสาเหตุของการเกิดปัญหาการใช้งานอินเตอร์เน็ต

2. เพื่อเสนอเทคนิคในการสร้างระบบอัตโนมัติที่นำเสนอวิธีการแก้ไขปัญหาการใช้งานอินเตอร์เน็ต

การตรวจเอกสาร

ความรู้เบื้องต้นเกี่ยวกับงานวิจัย

1. เทคนิคออนโทโลจี (Ontology)

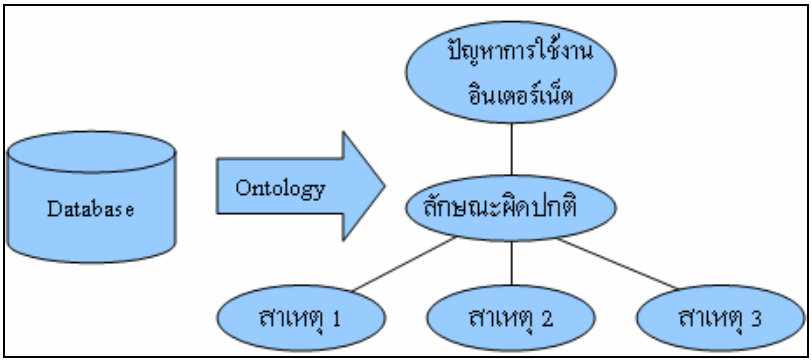
เทคนิคออนโทโลจี (Ontology) เป็นเทคนิคที่ใช้โครงสร้างของต้นไม้หรือกราฟเป็นรูปแบบ ของการจำแนกแบบโครงสร้างแบบลำดับชั้น นิยามของออนโทโลจี คือ รูปแบบที่แสดงถึงคุณสมบัติเฉพาะที่มีแนวคิดส่วนร่วมเดียวกัน (Gruber, 93) ดังนั้นออนโทโลจีจึงมีความสามารถที่จะแยกแยะความต้องการที่แตกต่างกันของผู้ใช้งาน โดยลักษณะของออนโทโลจีนี้สามารถแบ่งออกได้เป็น 2 ประเภทที่สำคัญ คือ

- Real-world Ontology

ใช้สำหรับจำแนกความหมายของคำในสภาพจริงปัจจุบัน เช่น WordNet เป็นเทคนิคที่ใช้ในการสกัดความหมายของคำ (Imsombut and Kawtrakul 2005)

- Task-Oriented Ontology

ใช้สำหรับการติดตามรอยความรู้ (Knowledge Tracking) เช่น ดูความเชื่อมโยงของข้อมูลที่ซ่อนอยู่ ภายในโดยใช้ออนโทโลจีมาช่วยในการสกัดหาสาเหตุของปัญหาในการทำงาน เช่น ปัญหาการใช้ระบบอินเทอร์เน็ตของผู้ใช้



ภาพที่ 1 ตัวอย่างรูปแบบของออนโทโลจีแบบ Task-Oriented Ontology

ภาพที่ 1 แสดงตัวอย่างการแบ่งปัญหาการใช้งานอินเตอร์เน็ตโดยใช้เทคนิคออนโทโลจี จากรูปจะแสดงถึงการแบ่งลักษณะผิดปกติที่ลูกค้ามีโอกาสพบและสาเหตุของการเกิดลักษณะผิดปกติตามอาการที่พบเป็นต้น

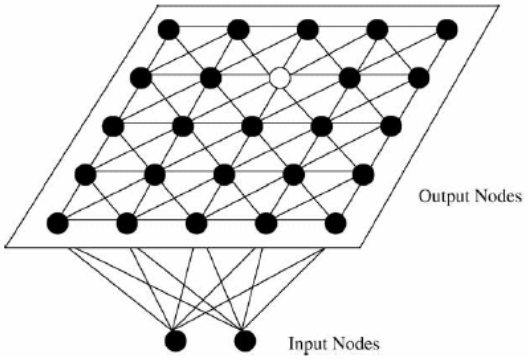
2. Latent Semantic Analysis (LSA)

Latent Semantic Analysis เป็นเทคนิคที่ใช้สำหรับสกัดข้อความในเอกสาร โดยใช้หลักการหาความหมายของคำและ การวัดความเหมือนกันของความหมายของคำ โดยสามารถแบ่งแยกความหมายของคำได้ดังนี้

1. Polynymous หมายถึง คำเดียวอาจมีความหมายได้หลายอย่าง เช่น chip ที่หมายถึงอุปกรณ์คอมพิวเตอร์ และ chip ที่หมายถึงมันฝรั่งทอด
2. Synonymous หมายถึง คำหลายคำมีความหมายเดียวกัน เช่น motorcycle หรือคำว่า bike ต่างก็หมายถึงรถจักรยานยนต์เหมือนกัน

3. Kohonen's Self-Organizing Maps (SOM)

Kohonen's Self - Organizing Maps เป็นอัลกอริทึมของโครงข่ายประสาทเทียมที่นิยมใช้ในการจัดกลุ่ม (Clustering) ซึ่งเป็นโครงข่าย ที่ไม่มีชั้นฮิดเดน (Hidden Layer) ฉะนั้นโครงข่ายชนิดนี้จึงมีเพียงแค่ 2 ชั้นคือชั้นอินพุต (Input Layer) และชั้นเอ๊าท์พุต (Output Layer) ดังภาพต่อไปนี้



ภาพที่ 2 ลักษณะของ Self Organizing Maps (SOM)
ที่มา : Jain and Dubes (1988)

แนวคิดของ Kohonen's Self-Organizing Maps Neural Network คือ การทำซ้ำของข้อมูลในแต่ละข้อมูล (Jain and Dubes, 1998) เพื่อที่จะได้หาค่าของน้ำหนักของข้อมูลที่มีอยู่ทั้งหมดตามจำนวนกลุ่มที่ต้องการแบ่ง มีวิธีการดังนี้

1. กำหนดน้ำหนัก (w) และระยะทาง (*D*) ให้กับข้อมูล โดยให้ค่า α มีค่าตั้งแต่ 0 ถึง 1
2. นำค่าของข้อมูลมาคำนวณด้วยสูตร Distance คือ

$$D(j) = \sum_{i=1}^{n} \left(w_{i,j} - x_i \right)^2 \qquad (1)$$

โดยที่ *i* คือ ชิ้นข้อมูล (Data Item) ที่ใช้ในการจัดกลุ่มข้อมูล (Input)
j คือ จำนวนกลุ่มข้อมูลที่ต้องการค้นหา (Output)
D คือ ระยะทางระหว่างจุดข้อมูลกับจุดกึ่งกลางของกลุ่ม
x คือ ค่าของแอททริบิวต์ (Attribute)

3. เปรียบเทียบกับ *D(j)* ของแต่ละกลุ่มข้อมูลว่าค่าของกลุ่มไหนมีค่าน้อยที่สุด
4. นำข้อมูลของกลุ่มที่มีค่าน้อยที่สุดมาคำนวณหาน้ำหนักใหม่ โดยใช้สูตรดังนี้

$$w_{i,j}(new) = w_{i,j}(old) + \alpha\left(x_i - w_{i,j}(old)\right) \qquad (2)$$

โดยที่ *i* คือ ชิ้นข้อมูล (Data Item) ที่ใช้ในการจัดกลุ่มข้อมูล (Input)
j คือ จำนวนกลุ่มข้อมูลที่ต้องการค้นหา (Output)
w(new) คือ ค่าน้ำหนักใหม่
w(old) คือ ค่าน้ำหนักเก่า
α คือ ค่าความผิดพลาดที่ยอมรับได้มีค่าตั้งแต่ 0 ถึง 1

5. ทำการลดค่า α ลง
6. ทำซ้ำตั้งแต่ข้อ 2 ใหม่จนกว่าจะไม่มีการเปลี่ยนแปลงค่า *D(j)* ในแต่ละกลุ่ม

ข้อเด่นของเทคนิคการจัดกลุ่มแบบ SOM

1. สามารถจัดการกับข้อมูลจำนวนมากได้

ข้อด้อยของเทคนิคการจัดกลุ่มแบบ SOM

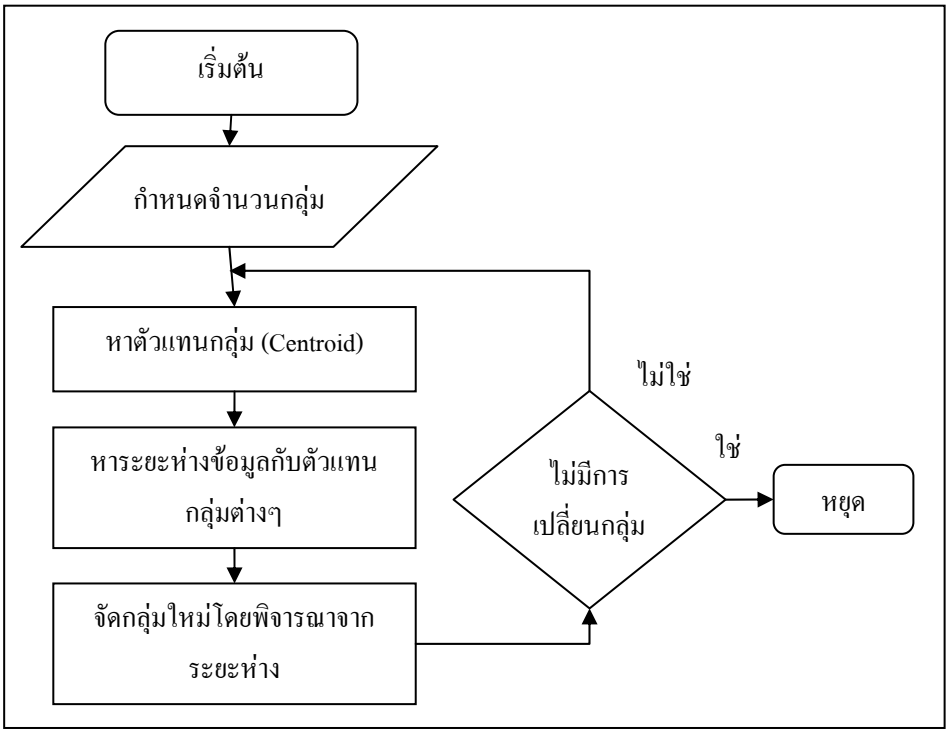
1. ใช้เวลาในการคำนวณสูง
2. การใช้ตัวแปรต่างกันก็จะให้ผลแตกต่างกัน

4. K-Means

K-Means เป็นเทคนิคการจัดกลุ่มข้อมูลโดยอัลกอริทึมเค-มีน (K-Means Algorithm) ซึ่งมีการพัฒนาและนำเสนอโดย Mac Queen ในปี 1967 แสดงให้เห็นถึงอัลกอริทึมในการจัดกลุ่มของสมาชิกภายในแต่ละกลุ่มนั้นว่าจะมีระยะใกล้จุดศูนย์กลาง หรือตัวแทนของกลุ่ม (Mean) มากน้อยเพียงใด โดยขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-Means อัลกอริทึมนั้นจะประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อมูลแต่ละตัวเข้ากลุ่ม และสุดท้ายการปรับปรุงตัวแทนกลุ่มในแต่ละกลุ่ม ปัญหาที่พบของอัลกอริทึม K-Means คือ ปัญหาการกำหนดจำนวนกลุ่มเริ่มต้น ซึ่งส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มนั้นมีจำนวนสมาชิกน้อยเกินไป หรืออาจจะไม่มีสมาชิกในกลุ่มเลย จึงได้มีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดไปได้นั้น จะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา (Bradley, 1998)

ขั้นตอนการจัดกลุ่มโดยเทคนิค K-Means อัลกอริทึม มีดังนี้

1. กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับ จำนวนกลุ่ม โดยที่ในทุกๆ กลุ่มนั้นจะต้องมีสมาชิกภายในกลุ่มเสมอ
2. คำนวณหาตัวแทนกลุ่ม หรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม
3. จัดกลุ่มข้อมูลใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อมูลในกลุ่มกับตัวแทนของกลุ่มต่างๆ ว่าข้อมูลนั้นสมควรที่จะอยู่กลุ่มใด
4. ทำการปรับปรุงการจัดกลุ่มโดยย้อนกลับไปทำข้อ 2 และจะหยุดเมื่อข้อมูลสมาชิกในกลุ่มแต่ละกลุ่มนั้นไม่มีการเปลี่ยนแปลงแล้ว



ภาพที่ 3 ขั้นตอนการจัดกลุ่มข้อมูลโดยใช้ K-means อัลกอริทึม

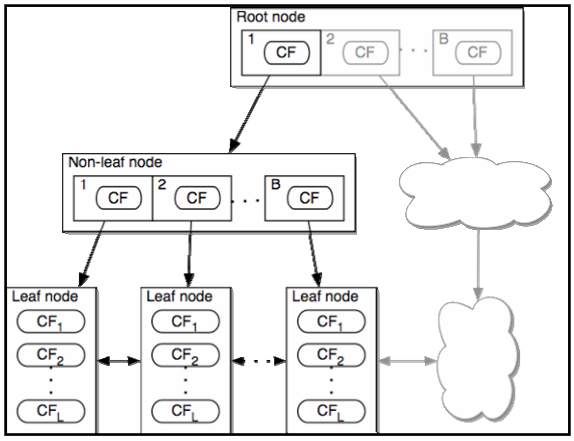
5. การจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical

เทคนิคการจัดกลุ่มข้อมูล (Clustering) ในการทดลองเลือกใช้อัลกอริทึมของการจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical เป็นอัลกอริทึมในการจัดกลุ่มข้อมูล ซึ่งสามารถรองรับข้อมูลจำนวนมากได้อย่างมีประสิทธิภาพและรวดเร็ว ซึ่งการทำงานในขั้นต้นนั้นแบ่งออกเป็น 2 ขั้นตอน คือ

1. Pre-cluster คือ การจัดกลุ่มข้อมูลแบ่งออกเป็นกลุ่มย่อยจำนวนหลายๆกลุ่ม ในขั้นตอนนี้ใช้เทคนิค “Cluster feature (CF) tree” ซึ่งเป็นเทคนิคหนึ่งของ Sequential Clustering โดยใช้หลักการพิจารณาจากแต่ละระเบียนของข้อมูล (record) ว่าควรที่จะรวมกับข้อมูลกลุ่มแรก หรือ เริ่มจัดแบ่งเป็นกลุ่มข้อมูลใหม่โดยพิจารณาจากระยะห่างระหว่างข้อมูล
2. Cluster คือ เมื่อได้ขั้นตอนที่ 1 แล้วนำผลที่ได้มาจัดแบ่งกลุ่มใหม่อีกครั้งตามจำนวนกลุ่มที่ต้องการ โดยนำข้อมูลจากผลการแบ่งกลุ่มย่อยจากข้อ 1 มาใช้ ซึ่งเทคนิคที่ใช้ในขั้นตอนนี้คือ “Agglomerative Hierarchical Clustering”

5.1 Clustering Feature Tree (CF-Tree)

CF-Tree คือ ต้นไม้ใช้ในการจัดเก็บข้อมูลเป็นกลุ่มย่อย (Sub-cluster) เพื่อใช้ในการจัดกลุ่มข้อมูลโดยจะใช้หลักการพิจารณาแต่ละระเบียนข้อมูล (Record) ว่าควรที่จะอยู่ในกลุ่มย่อยใด หรือจะรวมกับกลุ่มย่อยเดิม หรือเริ่มจัดเป็นกลุ่มย่อยใหม่โดยพิจารณาจากระยะห่างระหว่างข้อมูล โดยใช้ BIRCH Algorithm (Zhang *et al*., 1996)



ภาพที่ 4 ลักษณะของ Clustering Feature Tree (CF-Tree)

สูตรที่ใช้ในการคำนวณ

$$CF = (N, \vec{LS}, SS) \tag{3}$$

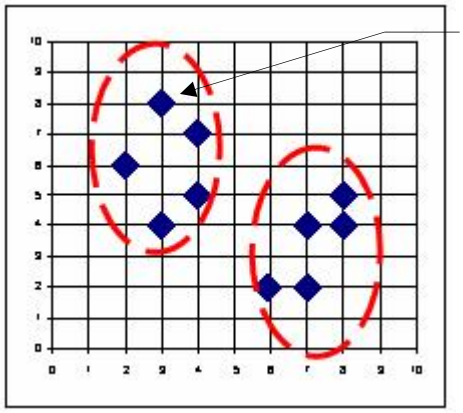
N คือ จำนวนข้อมูลทั้งหมดในกลุ่มข้อมูล

$\vec{LS} = \sum_{i=1}^{N} \vec{X}_i$ คือ ผลรวมของข้อมูลในแต่ละมิติ

$SS = \sum_{i=1}^{N} \vec{X}_i^{\,2}$ คือ ผลรวมของกำลังสองของข้อมูลในแต่ละมิติ

$$CF_1 + CF_2 = (N_1 + N_2, \vec{LS}_1 + \vec{LS}_2, SS_1 + SS_2) \tag{4}$$

ข้อมูล 1 – (3, 4)
ข้อมูล 2 – (2, 6)
ข้อมูล 3 – (4, 5)
ข้อมูล 4 – (4, 7)
ข้อมูล 5 – (3, 8)



CF = (5, (16, 30), (54,190))

ภาพที่ 5 ลักษณะของการจัดกลุ่มโดย Cluster Feature

5.2 Hierarchical Algorithm

Hierarchical Algorithm เป็นทฤษฎีการจัดกลุ่มแบบลำดับขั้น (Hierarchical Clustering) ซึ่งเป็นทฤษฎีที่นิยมใช้กันมากที่สุดสำหรับนำไปประยุกต์ใช้ในการจัดกลุ่มข้อมูลที่มีคุณลักษณะซ้อน (Nested) เช่น ข้อมูลประเภท Micro Array และ Gene Expression (Eisen *et al.*, 1998) และข้อมูลที่มีลักษณะค่อนข้างซับซ้อน

ตัวอย่างการจัดกลุ่มแบบลำดับขั้น Hierarchical Clustering สามารถแบ่งออกเป็น 2 ประเภทคือ

1. Agglomerative (Bottom-Up) มีหลักการในการทำงาน คือ เริ่มจากการจัดกลุ่มของข้อมูลออกเป็นจำนวน n กลุ่มจากข้อมูลทั้งหมด n ตัว ให้เป็นบัพภายนอก (External node) แล้วทำการรวมข้อมูลที่มีค่าความเหมือนใกล้เคียงกันมากที่สุดครั้งละ 2 ตัว ทำซ้ำจนกระทั่งได้ข้อมูลสุดท้าย 1 กลุ่มซึ่งเป็นบัพราก (Root node) (Enrique, 2001)
2. Divisive (Top-Down) มีหลักการในการทำงาน เหมือนวิธีของ Agglomerative แต่ต่างกันตรงที่วิธี Divisive นี้จะทำงานจาก Root node ไปสู่ External node หรือจะทำงานจากบนลงล่างนั่นเอง แต่วิธีนี้ไม่เป็นที่นิยมเนื่องจากใช้เวลาในการคำนวณนานกว่าแบบ Agglomerative

วิธีการคำนวณของ Agglomerative Hierarchical clustering มีวิธีการดังนี้ (Ward, 1963)

1. แบ่งข้อมูลออกเป็น n กลุ่ม (cluster) โดยในแต่ละ กลุ่ม มีสมาชิก 1 ตัว

$$L = S_1, S_2, S_3, S_4, ..., S_{n-1}, S_n$$

โดยที่ S = ข้อมูลที่ใช้ในการจัดกลุ่มข้อมูล (Input)

n = จำนวนข้อมูลทั้งหมด

L = จำนวนกลุ่มของข้อมูล

2. คำนวณหาค่าความเหมือนของข้อมูลทีละคู่ (S_i,S_j) ทุกๆ ข้อมูลใน L Cluster โดยการคำนวณหาค่าความเหมือนสามารถคำนวณจากสูตรต่อไปนี้

2.1. Single Linkage คือ คำนวณความเหมือนโดยการหาระยะทางระหว่าง 2 คลัสเตอร์ที่ใกล้กันที่สุดซึ่งคำนวณได้จากสูตร

$$d(r,s) = \min(dist(x_{ri}, x_{si})) \quad (5)$$

โดยที่

d คือ ระยะทางคำนวณจากวิธี Euclidian distance

r คือ กลุ่มข้อมูลของ Cluster R

s คือ กลุ่มข้อมูลของ Cluster S

dist คือ ค่าของระยะทางระหว่าง 2 จุดที่หา

x_{ri} คือ ค่าของข้อมูล Cluster R ตำแหน่งที่ i

x_{sj} คือ ค่าของข้อมูล Cluster S ตำแหน่งที่ j

2.2. Complete Linkage วิธีนี้จะตรงข้ามกับวิธี Single Linkage นั่นคือ การหาระยะทางระหว่าง 2 cluster ที่ห่างกันที่สุดแต่วิธีนี้ไม่เป็นที่นิยมเนื่องจากมี Noise มากทำให้การคำนวณใช้เวลานาน ซึ่งคำนวณได้จากสูตร

$$d(r,s) = \max(dist(x_{ri}, x_{si})) \quad (6)$$

2.3. Average Linkage เป็นวิธีการที่ผสมระหว่าง 2 วิธีแรก คือ Single Linkage กับ Complete Linkage แล้วเลือกค่าเฉลี่ยของระยะห่างระหว่าง 2 คลัสเตอร์ ซึ่งคำนวณได้จากสูตร

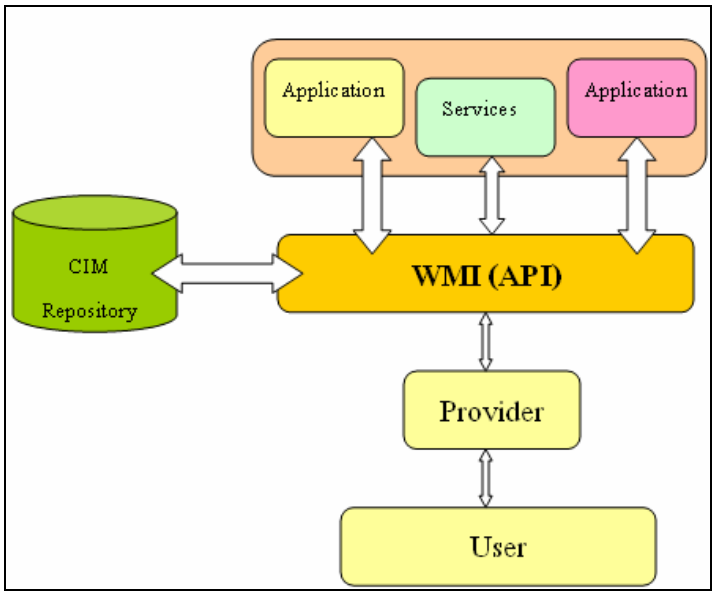
$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri}, x_{sj}) \quad (7)$$

3. ทำการรวม (Merge) คู่ของข้อมูลที่คำนวณได้จากขั้นตอนที่ 2 (S_i,S_j) และทำการสร้าง Tree ต้นไม้ โดยที่ node S_{ij} จะเป็น Parent node ให้กับ node S_i และ S_j โดยใช้วิธีการ Merge จากสมการใดสมการหนึ่งจากสมการในข้อ 2

4. กลับไปทำซ้ำขั้นตอนที่ 2 จนกว่าจะเหลือเพียง 1 Cluster นั่นคือ Root node

6. Window Management Instrument (WMI)

WMI เป็นเทคนิคที่พัฒนาขึ้นสำหรับการจัดการระบบการทำงานที่ซับซ้อน ซึ่งเทคนิค WMI จัดการกับข้อมูลปริมาณมากมายของระบบ ที่เก็บไว้ในรูปแบบฐานข้อมูลเชิงสัมพันธ์ การค้นคืนจะใช้ชุดคำสั่ง SQL เพื่อเรียกชุดข้อมูลที่สนใจมาแสดงผล



ภาพที่ 6 ลักษณะการทำงานของ WMI architecture

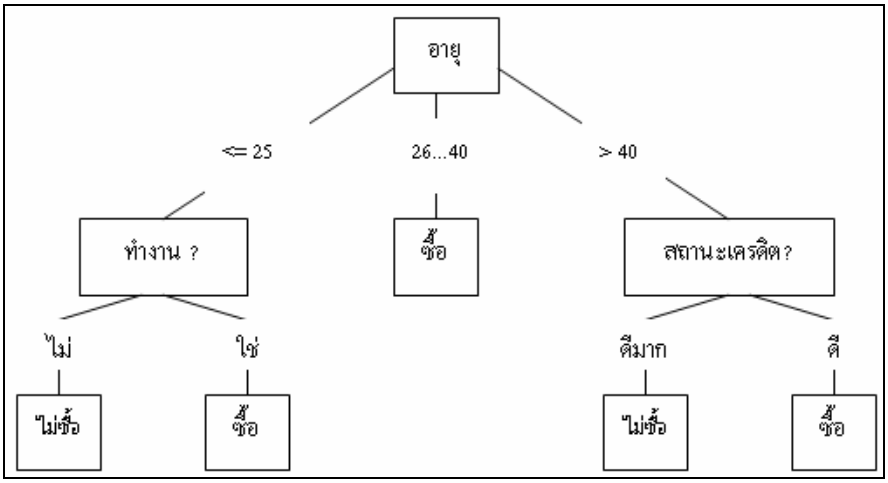
ตัวอย่าง Algorithm ของ WMI กรณีตรวจสอบ Network Adapter

```
Set objWMIService = GetObject("winmgmts:\\" & strComputer & "\root\CIMV2")
Set colItems = objWMIService.ExecQuery("SELECT * FROM Win32_NetworkAdapter", _
"WQL", wbemFlagReturnImmediately + wbemFlagForwardOnly)
```

จากภาพที่ 6 แสดงลักษณะการทำงาน โดยจากชั้นล่างสุดเป็นชั้นของ provider ซึ่งถือว่าเป็นตัวแทนที่อยู่ระหว่างระบบปฏิบัติงาน (เช่น ระบบปฏิบัติการ, โปรแกรมประยุกต์ หรือ ค่าของอุปกรณ์ต่าง ๆ) กับผู้ใช้งาน โดยลักษณะคำสั่งของ WMI จะเป็นประเภทของชุดคำสั่ง API (Application Program Interface) คือ วิธีการเฉพาะสำหรับเรียกใช้แอพพลิเคชั่นอื่น ๆ หรือชุดรหัสคอมพิวเตอร์ซึ่งทำหน้าที่เชื่อมต่อการทำงานระหว่างแอพพลิเคชั่นกับระบบปฏิบัติการ โดยคำสั่งทั้งหมดจะอยู่ภายในที่เก็บข้อมูลของ CIM (Common Information Model) ที่จะบรรจุข้อมูลที่จำเป็นของการค้นคืนข้อมูล เช่น เส้นทางการตรวจสอบข้อมูล เป็นต้น

7. Decision Tree

Decision tree หรือ ต้นไม้การตัดสินใจ เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลสำหรับใช้ในการจำแนกประเภทของข้อมูล โดยต้นไม้การตัดสินใจจะมีลักษณะคล้ายโครงสร้างต้นไม้ ที่ในแต่ละโหนดจะแสดงคุณลักษณะ (attribute) แต่ละกิ่งจะแสดงผลในการทดสอบ และลีฟโหนด (leaf node) แสดงกลุ่มที่กำหนดไว้ ซึ่งต้นไม้การตัดสินใจนี้จะง่ายต่อการปรับเปลี่ยนเป็นกฎการจำแนกประเภทของข้อมูล (classification rule)



ภาพที่ 7 ตัวอย่างการทำงานของ ต้นไม้ตัดสินใจของลูกค้าในการซื้อสินค้า

จากภาพที่ 7 เป็นรูปแบบ Model ทำนายการซื้อสินค้าโดยดูจากอายุของลูกค้าว่าอยู่ในกลุ่มใดการจัดกลุ่มตามอายุมี 3 กลุ่ม คือ (1) อายุน้อยกว่า 25 ปี (2) อายุระหว่าง 26 ถึง 40 ปี และ (3) อายุมากกว่า 40 ปี จากนั้นไปที่

- กลุ่มอายุ $\leq$ 25 ปี ทดสอบว่าลูกค้าทำงานหรือไม่
  - ถ้า ไม่ทำงาน ผล คือ จะไม่ซื้อ
  - ถ้า ทำงาน ผล คือ จะซื้อ
- กลุ่ม อายุ 26 – 40 ปี ลูกค้าทุกคน จะซื้อสินค้า
- กลุ่มอายุ >40 ต้องดูว่าเครดิตของลูกค้าเป็นอย่างไร
  - ถ้า เครดิตของลูกค้าอยู่ในขั้น ดีมาก ลูกค้าจะไม่ซื้อ
  - ถ้า เครดิต ของลูกค้าอยู่ในขั้น ดี ลูกค้าจะซื้อ

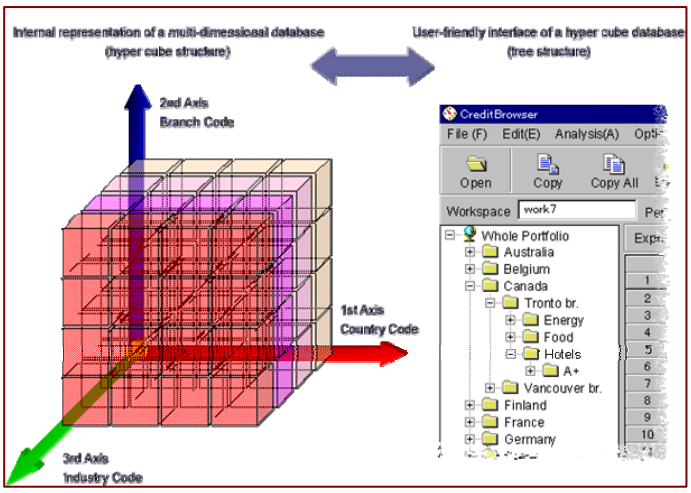
จากตัวอย่างข้างต้นจะเห็นได้ว่าความสามารถของ Decision Tree สามารถนำมาใช้ในการทำนายถึงเหตุการณ์ต่าง ๆ โดยดูจากข้อมูลที่เก็บไว้ในต้นไม้

8. On-Line Analytical Processing (OLAP)

On-Line Analytical Processing (OLAP) เป็นการประมวลผลเชิงวิเคราะห์ออนไลน์ ของข้อมูลที่มีการเก็บแบบหลายมิติ (multidimensional) โดยข้อมูลเหล่านี้ได้จากรายการ Transaction ที่เก็บในฐานข้อมูลหลายฐาน หรือ คลังข้อมูล ในช่วงระยะเวลาหนึ่ง โดย OLAP สามารถค้นคืนข้อมูล และ นำมาวิเคราะห์แนวโน้มในอนาคต จากเงื่อนไขต่างๆของข้อมูลนำเข้าที่กำหนด วิธีนี้เป็นการสนับสนุนกระบวนการเข้าถึงข้อมูลและการประมวลผลข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลหลายมิติ

ข้อดีของ OLAP คือ

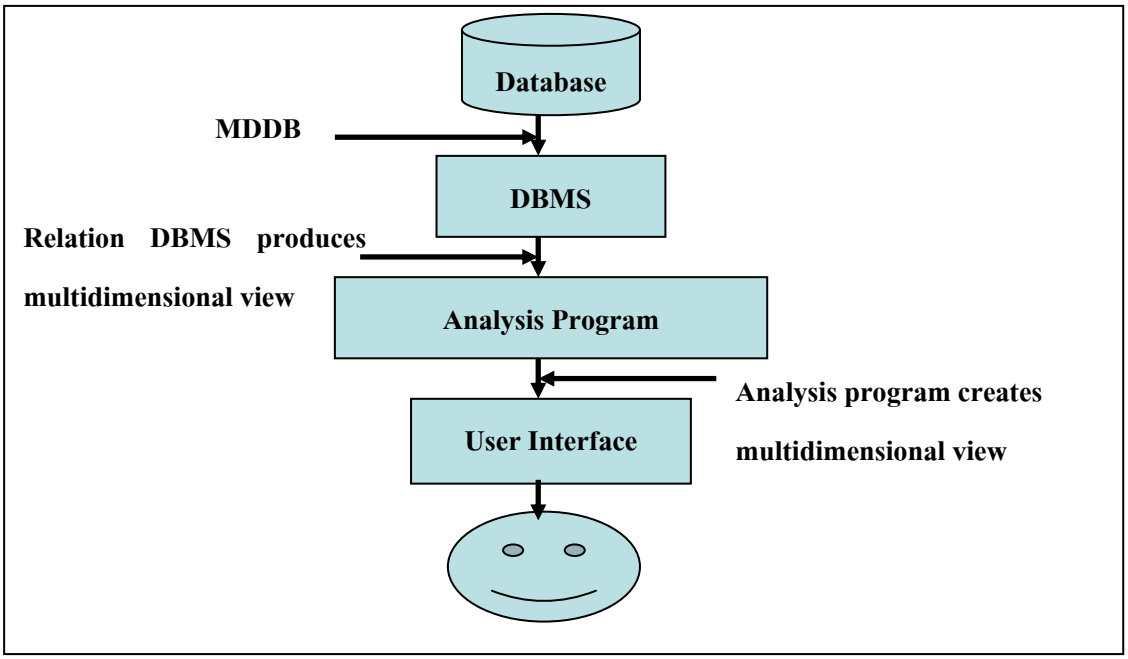
1. ค้นหาข้อมูลได้รวดเร็ว
2. หาผลรวมได้ง่าย และมีประสิทธิภาพ
3. เรียก ดูข้อมูลได้อย่างรวดเร็ว



ภาพที่ 8 ตัวอย่างรูปแบบของ OLAP

ที่มา: Numerical Technologies Incorporated (2007)

การออกแบบโครงสร้างของโปรแกรม OLAP นั้นในขั้นตอนแรกจะมีการออกแบบฐานข้อมูลแบบหลายมิติ (Multidimensional database (MDDB)) และการเชื่อมโยงความสัมพันธ์ระหว่าง Entity ตามที่ได้ออกแบบไว้ ขั้นตอนถัดไปก็เน้นการพัฒนาโปรแกรมตามที่ได้ออกแบบไว้ (Mallach, 2000)



ภาพที่ 9 โครงสร้างการออกแบบโปรแกรม OLAP

9. การวัดประสิทธิภาพการแบ่งกลุ่มข้อมูล

การวัดประสิทธิภาพของการแบ่งกลุ่มข้อมูล ความสามารถที่จะวัดประสิทธิภาพของการแบ่งกลุ่มข้อมูลได้ดีหรือไม่นั้น สามารถทำได้ 2 วิธีการดังต่อไปนี้

1. การวัดค่าความแตกต่างของข้อมูลภายในกลุ่ม (Root Mean Square Standard Deviation : RMSSTD)

ค่าความแตกต่างของข้อมูลภายในกลุ่มนั้นแสดงให้เห็นถึงประสิทธิภาพของการจัดกลุ่มว่ามีมากน้อยเพียงใด ซึ่งหากค่าความแตกต่างของสมาชิกภายในกลุ่มนั้นมีน้อยนั้น ย่อมหมายถึงการแบ่งกลุ่มที่ดี (Jain *et al.,* 1999)

สูตรการคำนวณค่าความแตกต่างภายในกลุ่ม

$$RMSSTD = \sqrt{\frac{\sum_{i=1..c}\sum_{j=1..d}\sum_{k=1}^{n_{ij}}(x_k - \bar{x}_j)^2}{\sum_{i=1..c}\sum_{j=1..d}(n_{ij} - 1)}} \qquad (8)$$

โดยที่

c คือ จำนวนกลุ่มที่แบ่งได้ทั้งหมด

d คือ จำนวนคอลัมน์ทั้งหมดภายในชุดข้อมูล

$\bar{x}_j$ คือ ค่าเฉลี่ยของข้อมูลคอลัมน์ที่ j

n_{ij} คือ จำนวนข้อมูลในกลุ่มที่ i คอลัมน์ที่ j

2. การวัดค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R Squared : RS)

ค่าความแตกต่างของข้อมูลระหว่างกลุ่มนั้น เป็นค่าหนึ่งซึ่งแสดงให้เห็นถึงประสิทธิภาพของการจัดกลุ่มว่ามีมากน้อยเพียงใด ซึ่งหากค่าความแตกต่างระหว่างกลุ่มมีมากนั้นย่อมหมายถึงการแบ่งกลุ่มที่ดี (Jain *et al*., 1999) แต่ละกลุ่มจะมีลักษณะที่แตกต่างกัน โดยที่ค่า RS นี้จะอยู่ระหว่าง 0 ถึง 1

$$RS = \frac{SS_t - SS_w}{SS_t} \tag{9}$$

$$SS_t = \sum_{j=1}^{d} \sum_{k=1}^{n_j} (x_k - \bar{x}_j)^2 \text{ และ } SS_w = \sum_{\substack{i=1..c \\ j=1..d}} \sum_{k=1}^{n_{ij}} (x_k - \bar{x}_j)^2$$

โดยที่

SS_t คือ ผลรวมของผลต่างกำลังสองทุกข้อมูลภายในกลุ่ม

SS_w คือ ผลรวมของผลต่างกำลังสองทุกข้อมูลภายในกลุ่ม

c คือ จำนวนกลุ่มที่แบ่งได้ทั้งหมด

d คือ จำนวนตัวแปรทั้งหมด

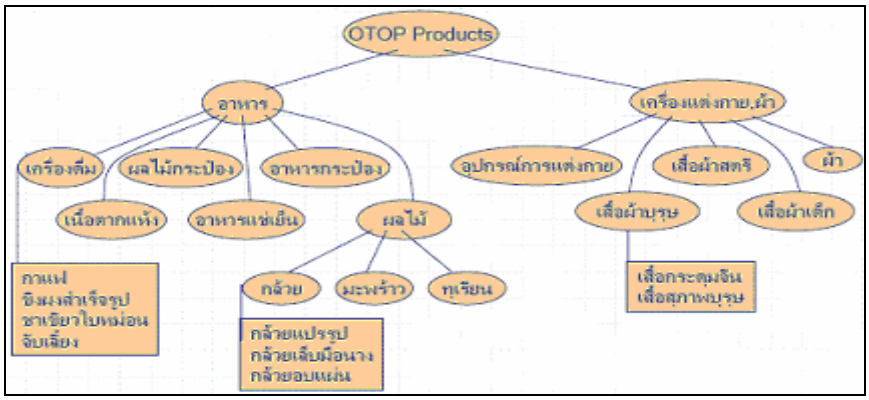
$\bar{x}_j$ คือ ค่าเฉลี่ยของข้อมูลคอลัมน์ที่ j

n_{ij} คือ จำนวนข้อมูลในกลุ่มที่ i คอลัมน์ที่ j

งานวิจัยที่เกี่ยวข้อง

1. งานวิจัยที่เกี่ยวข้องกับออนโทโลจี

จากการศึกษางานวิจัยเพิ่มเติมของนภดล หมื่นไพ และ อัศนีย์ ก่อตระกูล (2003) ที่ได้ใช้เทคนิคออนโทโลจีช่วยในการจัดประเภทของสินค้าOTOP ซึ่งออนโทโลจีเป็นรูปหนึ่งของต้นไม้ ที่มีโครงสร้างแบบลำดับชั้นที่ได้ศึกษามาแล้วว่ามีความสามารถในการจำแนกเอกสารได้ดีโดยข้อดีของออนโทโลจีเป็นการจำแนกคุณสมบัติเฉพาะออกมาได้โดยแน่นอน ซึ่งสามารถนำออนโทโลจีมาเป็นโครงสร้างสำหรับฐานความรู้ได้ (Knowledge base)



ภาพที่ 10 ตัวอย่างการแบ่งประเภทสินค้า OTOP ของนภดล หมื่นไพ และ อัศนีย์ ก่อตระกูล (2003)

2. งานวิจัยที่เกี่ยวข้องกับ Latent Semantic Analysis (LSA)

จากการศึกษางานวิจัยของ Jui-Feng Yeh และคณะ (2003) ที่ได้ทำการวิเคราะห์อาการป่วยโดยการวิเคราะห์คำถามก่อนที่จะทำการค้นคืนข้อมูลที่เป็นคำตอบ Yeh จัดกลุ่มประเภทของคำถามที่เข้ามาโดยใช้ทฤษฎี Latent Semantic Analysis (LSA) ทฤษฎีนี้ช่วยในการแยกประเภทของคำถามที่เข้ามาว่าเป็นคำถามชนิดใด และการตอบคำถามจะตอบในลักษณะใด วิธีนี้ทำให้สามารถแก้ปัญหาที่เกิดจากการใช้คำถามที่มีความหมายเดียวกันแต่ใช้คำศัพท์แตกต่างกันโดยทำการจัดกลุ่มคำถามตามความหมายของคำ

Type	Function words	Intension
1	哪(what for singular) 哪些(what for plural) 那些(what) 哪樣(which one) 甚麼(what) 什麼(what) 何許(what) 何種(what kind of) 哪種(which kind of) 那種(what kind of) 何謂(what means) 什麼是(what means) 定義(definition)	What 類型
2	多久(How long) 何時(what time) 幾時(what time)	When 時間
3	哪兒(where) 哪裡(what place) 何在(where) 那裡(where) 何處(where)	Where 地點
4	爲何(for what) 爲啥(for what) 爲什麼(why) 爲甚麼(for what) 緣何(what is the reason) 幹嗎(for what) 何以(the reason) 何故(what is the reason) 原因(reason)	Why 道理
5	怎(How) 怎麼(how) 怎樣(how) 如何(how) 怎的(how) 方法(the method is) 怎麼辦(what way to) 怎麼樣(how)	How 方式, 方法
6	多(how) 多麼(how) 何等(what degree is) 何其(how)	Degree 程度
7	若干(how about) 多少(how many/much) 幾(what is the number) 幾多(how many/much) 幾何(how many) 幾許(how many/much)	Quantity 數量值
8	可否(may) 是否(is … or not) 有無(is …..or not)	Whether 正反疑問
9	有關(relational to) 關係(have the relation with) 關聯(relational to) 關連(have the relation woth)	Relation 關係
10	能否(have the capacity to) 能夠(be able to) 可能(can) 能不能(can ……or not)	Capability 能力

ภาพที่ 11 ตัวอย่างการแบ่งประเภทคำถาม ของ Jui-Feng Yeh และคณะ (2003)

3. งานวิจัยที่เกี่ยวข้องกับ Decision Tree

จากการศึกษางานวิจัยของกฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์ (2001) ที่ได้ใช้หลักการของการทำเหมืองข้อมูล โดยมีการประยุกต์ใช้เทคนิค Decision Tree ในระบบพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์ ข้อมูลที่ใช้ทดลองเป็นข้อมูลการลงทะเบียนเรียนของนิสิตคณะวิศวกรรมศาสตร์มหาวิทยาลัยเกษตรศาสตร์ ผลการทดลองทำให้ทราบว่า Model decision tree มีผลลัพธ์เป็นที่น่าพึงพอใจโดยมีความแม่นยำถึงร้อยละ 84.58

4. งานวิจัยที่เกี่ยวข้องกับ Collaborative Filtering

จากการศึกษางานวิจัยของ Weng และ Liu (2004) ที่ประยุกต์ใช้หลักการของ Collaborative Filtering เป็นเทคนิคที่ใช้ในการเปรียบเทียบลักษณะของผู้ใช้ปัจจุบัน (Active user) กับกลุ่มผู้ใช้งานในอดีต ซึ่งมีลักษณะคล้ายคลึงกัน ในงานวิจัยนี้ได้ใช้เทคนิคการจัดกลุ่มข้อมูลลูกค้าที่ซื้อสินค้าใน

ร้านค้าแห่งหนึ่ง เพื่อใช้ในการนำเสนอสินค้าให้ตรงกับกลุ่มลูกค้า ผลการทดลองพบว่าระบบสามารถนำเสนอสินค้าได้ตรงกับความต้องการแม่นยำร้อยละ 61.7 แต่เมื่อระบบมีการเรียนรู้ผู้ใช้งาน ผลลัพธ์จะแม่นยำเพิ่มขึ้น และได้วัดความแม่นยำอีกครั้งในการเข้าใช้งานในครั้งที่ 3 ผลลัพธ์มีความแม่นยำถึงร้อยละ 78.57

จากงานวิจัยของ Waminee Niyagas และคณะ (2006) ได้ใช้เทคนิคการจัดกลุ่มข้อมูล โดยเปรียบเทียบระหว่างเทคนิคการแบ่งกลุ่มแบบ 2 ขั้นตอน 2 เทคนิค คือ เทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ SOM และ K-means อัลกอริทึม และเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical โดยข้อมูลที่ใช้ทดลองเป็นข้อมูลลูกค้าที่ใช้ e-banking ซึ่งรวมค่า RFM วิเคราะห์พฤติกรรมการใช้บริการของลูกค้าอินเทอร์เน็ตแบงก์กิ้ง ซึ่งผลลัพธ์ของการทดลองพบว่าเทคนิคการจัดกลุ่ม 2 ขั้นตอนแบบ SOM และ K-Means มีการจัดกลุ่มได้ดีกว่าการใช้ Clustering Feature Tree และ Agglomerative Hierarchical

5. งานวิจัยที่เกี่ยวข้องกับ Windows Media Instruction (WMI)

จากงานวิจัยของ Jayaputera และคณะ (2004) ได้ประยุกต์ใช้เทคนิค .Net-base system (บนพื้นฐาน WMI) ในการควบคุมดูแลระบบการจองโรงแรม โดยในงานวิจัยได้ทำการทดสอบประสิทธิภาพของการทำงานของโปรแกรม ความน่าเชื่อถือของโปรแกรม และประโยชน์ที่ได้รับจากโปรแกรม ซึ่งจากงานวิจัยนี้ทำให้ทราบได้ว่าเทคนิค WMI มีความสามารถในการตรวจสอบค่าต่าง ๆ และการทำงานของระบบคอมพิวเตอร์และโปรแกรมประยุกต์ต่าง ๆ ได้

จากงานวิจัยของ Poernomo และคณะ (2005) ใช้เทคนิคของ WMI ตรวจสอบเวลาการตอบสนองของโปรแกรม ผลลัพธ์ของงานวิจัยนี้เพื่อทดสอบความน่าเชื่อถือของโปรแกรม และประสิทธิภาพของโปรแกรม โดยการทดลองจะเริ่มวัดประสิทธิภาพโปรแกรมตอนเริ่มทำงานและสิ้นสุดเมื่อโปรแกรมทำงานเสร็จสิ้น ผลจากการศึกษางานวิจัยนี้สามารถทำให้ทราบว่ามีการใช้เทคนิค WMI ในการตรวจสอบลักษณะการทำงานของระบบและโปรแกรมประยุกต์ได้

6. งานวิจัยที่เกี่ยวข้องกับ Online analysis processing (OLAP)

Markl และคณะ (1996) นำเสนอวิธีการเพิ่มประสิทธิภาพการทำงานของ เทคนิค OLAP Cube ให้ใช้เวลาสั้นลง โดยทดลองนำข้อมูลของสินค้าประเภทเครื่องดื่มจัดกลุ่มข้อมูลด้วย อัลกอริทึม Hierarchical Clustering ก่อนนำข้อมูลวิเคราะห์โดย OLAP งานวิจัยนี้พบว่าการจัดกลุ่มข้อมูลสามารถเพิ่มประสิทธิภาพการวิเคราะห์ข้อมูล โดย สามารถลดเวลาในการทำงาน เพราะมีการแบ่งส่วนข้อมูล เนื่องจาก OLAP ใช้เวลาวิเคราะห์ข้อมูลขนาดใหญ่นาน กว่าข้อมูลขนาดเล็ก

Li และคณะ (2007) นำเสนอเทคนิคการนำเสนอข้อมูล M-Commerce ให้ตรงกับผู้ใช้งาน โดยพิจารณาข้อมูล 3 ส่วนคือ User Information, Content Information และ Context Information มาพิจารณา โดยผลลัพธ์ของการทดลองในงานวิจัยนี้ได้ค่า precision อยู่ที่ประมาณร้อยละ 69

งานวิทยานิพนธ์นี้ ผู้วิจัยได้นำเสนอวิธีสร้างระบบการวิเคราะห์คำถาม และตอบปัญหาการใช้งานอินเตอร์เน็ตโดยการประยุกต์ใช้เทคนิคการทำเหมืองข้อมูล การทำงานของระบบนี้จะแบ่งตามปัญหาการใช้งานอินเตอร์เน็ตออกเป็น 2 ส่วน คือ

ส่วนที่หนึ่ง การตอบปัญหากรณีที่ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้ ประยุกต์ใช้เทคนิคการสร้าง Software agent เพื่อใช้ในการวิเคราะห์คำถามและตรวจสอบค่าและอุปกรณ์ โดยการสร้าง Software agent นั้นใช้หลักการของเทคนิค WMI เพราะจากงานวิจัยที่ศึกษา (Jayaputera *et al.*, 2004) พบว่าเทคนิค WMI สามารถนำมาช่วยในการตรวจสอบระบบการทำงานของโปรแกรมต่าง ๆ ได้ ขั้นตอนการตรวจสอบสาเหตุของปัญหาจาก Software agent นั้น จะได้มาจากการสร้างกฎเกณฑ์จากเทคนิค Decision tree เพราะข้อดีของ Decision tree จะง่ายต่อการสร้างและปรับกฎเกณฑ์การจำแนกได้

ส่วนที่สอง การตอบปัญหากรณีใช้งานอินเตอร์เน็ตได้ปกติ สำหรับในงานวิจัยนี้ได้มีการวิเคราะห์สาเหตุของปัญหา พบว่าอุปกรณ์และค่าที่จำเป็นต่อการเชื่อมต่ออินเตอร์เน็ตจะมีค่าปกติทำให้สามารถเชื่อมต่ออินเตอร์เน็ตได้ ดังนั้นจึงประยุกต์เทคนิค Collaborative Filtering มาช่วยในการวิเคราะห์ปัญหา เพราะจากงานวิจัยที่ศึกษา (Weng and Liu,2004) พบว่าเทคนิค Collaborative สามารถนำมาใช้ในการแนะนำโดยการวิเคราะห์ข้อมูลในอดีตของกลุ่มของผู้ใช้งาน และสาเหตุการเกิดปัญหา พบว่าลักษณะของปัญหาที่พบและวิธีแก้ไขปัญหาอาจเกิดขึ้นกับกลุ่มที่มีลักษณะการใช้

งานเหมือนกัน การประยุกต์เทคนิคการจัดกลุ่มนั้นได้นำเสนอการจัดกลุ่มแบบ2 ขั้นตอนโดยใช้เทคนิค SOM และ K-means เพื่อเพิ่มประสิทธิภาพในการจัดกลุ่ม ซึ่งมีการเปรียบเทียบกับเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอน โดยใช้ Clustering Feature (CF) Tree และ Agglomerative Hierarchical ซึ่งเมื่อทำการจัดกลุ่มเรียบร้อยแล้ว จากนั้นจะนำข้อมูลของกลุ่มมาวิเคราะห์ถึงความน่าจะเป็นของสาเหตุที่เกิดปัญหาของลูกค้าแต่ละกลุ่ม ด้วยเทคนิค OLAP เพราะจากงานวิจัยที่ศึกษา (Markl *et al.*, 1996) พบว่าเทคนิค OLAP สามารถใช้ดูข้อมูลในลักษณะเป็นหลายมิติได้ และผลจากการวิเคราะห์ OLAP สามารถใช้ในการตัดสินใจดูแนวโน้มของเนื้อหาต่าง ๆ ได้

อุปกรณ์และวิธีการ

อุปกรณ์

1. Hardware

    1.1 คอมพิวเตอร์แลปท๊อป: Pentium [R] M 1.40 GHz, 256 MB of RAM, 40 Gigabytes of Hard disk

2. Software

    2.1 Java Script

    2.2 Visual Basic

    2.3 HTML

    2.4 PHP Tool: Dream weaver โปรแกรมทางสถิติ

    2.5 MATLAB

3. Database

    3.1 MySQL version 4.0.21

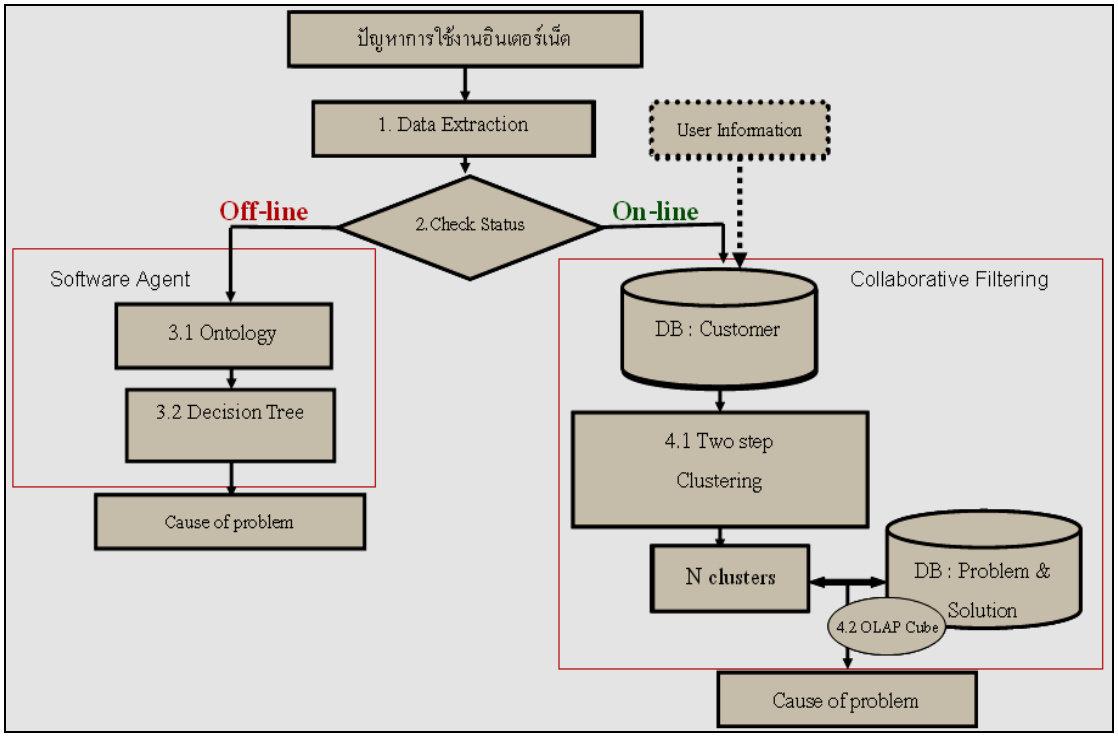
4. Operating System

    4.1 Window XP

วิธีการ

โครงสร้างการทำงานของระบบวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ต

ในการทดลองครั้งนี้ มีโครงสร้างการทำงานของระบบการวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ต ดังภาพที่ 12



ภาพที่ 12 โครงสร้างการทำงานทั้งหมดของงานวิจัยการแก้ปัญหาการใช้งานอินเตอร์เน็ต

ข้อมูลสำหรับทำการทดลอง (Data set)

ข้อมูลที่ใช้ในการทดลองแบ่งออกเป็น 2 ชุดข้อมูล คือข้อมูลที่ใช้ทดลองกับการแก้ปัญหาจากกรณีที่ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้ และข้อมูลที่ใช้ทดลองกับการแก้ปัญหาจากกรณีที่สามารถเชื่อมต่ออินเตอร์เน็ตได้

ชุดที่ 1 ข้อมูลการสอบถามปัญหากรณีเชื่อมต่ออินเตอร์เน็ตไม่ได้ ข้อมูลการทำรายการตั้งแต่เดือนพฤศจิกายน – ธันวาคม ในปี 2549 จำนวน 10,000 รายการ

ชุดที่ 2 ข้อมูลการสอบถามปัญหากรณีใช้อินเตอร์เน็ตได้ ข้อมูลการทำรายการตั้งแต่เดือนพฤศจิกายน – ธันวาคม ในปี 2549 จำนวน 7,110 รายการ

ขั้นตอนการทดลองกลุ่มปัญหาการเชื่อมต่ออินเตอร์เน็ตไม่ได้

1. Data Cleaning และ Extraction ในขั้นตอนนี้จะนำคำถามของลูกค้าเรื่องปัญหาการใช้งานอินเทอร์เน็ต มาตัดคำโดยใช้โปรแกรม Swath (ไพศาล,1997) จากนั้นเริ่มการจัดกลุ่มคำตามอาการที่ผิดปกติ นำคำที่มีความหมายตรงกัน จัดไว้ให้อยู่ในกลุ่มเดียวกัน เช่น มีคำว่า ติด…ดับ ในประโยคก็อาจสกัดได้ว่า ปัญหาที่พบเป็นอาการเกิดจากการเชื่อมต่ออินเตอร์เน็ตหลุดบ่อย หรือคำถามมีคำว่า Error อาจจะเกิดได้จากมี Message Error ขึ้นแสดงบนหน้าจอ

Word synonymy	Intension
Connect, เชื่อมต่อ , ต่อเน็ต , Connected	การเชื่อมต่อ / Connect
ติด....ดับ , หลุด , เน็ตตัด ,disconnected	หลุดบ่อย / Frequently disconnect
เออเรอ, Error, Errors , ข้อความ, message	Message Error
.....ช้า, เร็ว…k, speed,….MB, slow,อืด, สปีด ,low speed	ความเร็ว / Low Speed
…นาที ,Minute, min, time, hour, ตลอด,บ่อย	ช่วงเวลา / Time
web…ไม่ได้, The page …	เปิด Web ไม่ได้ / Can't open web
Mail…ไม่ได้, ส่ง…mail, send/receive…mail	รับ/ส่ง mail ไม่ได้

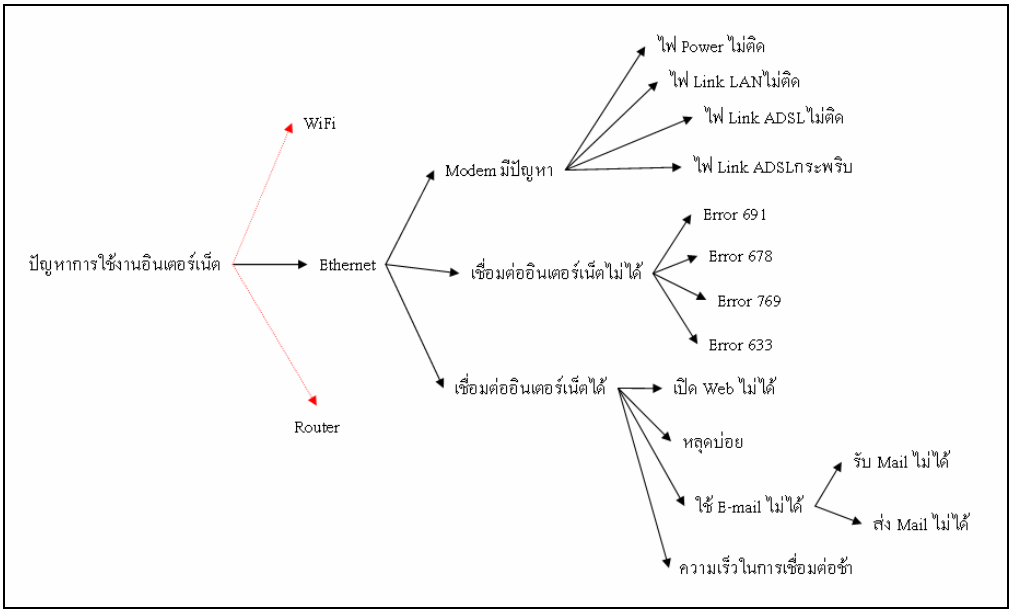
ภาพที่ 13 การจำแนกคำที่มีความหมายเหมือนกัน

2. Check Status ในขั้นตอนนี้ได้สร้าง Software agent เพื่อตรวจสอบสถานะของเครือข่ายอินเตอร์เน็ตในปัจจุบันว่าสามารถเชื่อมต่อได้หรือไม่ เนื่องจากสาเหตุของการเกิดปัญหาการใช้งานอินเตอร์เน็ตจะใช้เทคนิคการวิเคราะห์ปัญหาที่แตกต่างกัน เช่น กรณีข้อความ Error หลังการเชื่อมต่ออินเตอร์เน็ต กรณีที่ไม่สามารถเชื่อมต่อได้ ณ เวลาที่ค้นหาในระบบนั้น สาเหตุอาจเกิดจากอุปกรณ์หรือค่าคอมพิวเตอร์ในปัจจุบันมีอาการผิดปกติ แต่กรณีสอบถามอาการ Error จากกรณีเชื่อมต่อได้นั้น ผู้ใช้งานอาจประสบปัญหาในอดีตและได้ค้นหาดังนั้นจึงมีการแบ่งวิธีการวิเคราะห์ปัญหา

3. การวิเคราะห์ปัญหาการใช้งานกรณีเชื่อมต่ออินเตอร์เน็ตไม่ได้

3.1. การรวบรวมและแก้ปัญหาโดยใช้หลัก Ontology การทดลองจะเริ่มจากการจัดเก็บข้อมูลเรื่องอาการผิดปกติที่เกิดจากการใช้งานอินเตอร์เน็ตไว้ในฐานข้อมูล ซึ่งได้ถูกจำแนกตามลักษณะของอาการที่ถูกสอบถามทาง Call center แล้วจัดเก็บไว้ในรูปแบบของออนโทโลจี โดยในงาน วิจัยนี้ได้แยกประเภทตามลักษณะอาการผิดปกติที่เหมือนกันมาไว้ในกลุ่มเดียวกัน

ภาพที่ 14 แสดงปัญหาจากการใช้งานอินเตอร์เน็ตประเภท Ethernet Modem สามารถแบ่งออกได้เป็น 3 ประเภท คือ (1) กรณีไม่สามารถเชื่อมต่อได้, (2) กรณีเชื่อมต่อเครือข่ายได้แต่พบปัญหาการใช้อินเทอร์เน็ต และ (3) กรณีตรวจสอบพบว่า Modem มีปัญหาการใช้งาน ซึ่งอาการที่พบในแต่ละกรณีสามารถแบ่งแยกตามคำถามของลูกค้าออกมาได้อีก เช่น ปัญหาการเชื่อมต่อไม่ได้ อาจสังเกตได้จากมีข้อความแจ้งเตือนว่าเกิด Error 678 หรือ Error 691 เป็นต้น



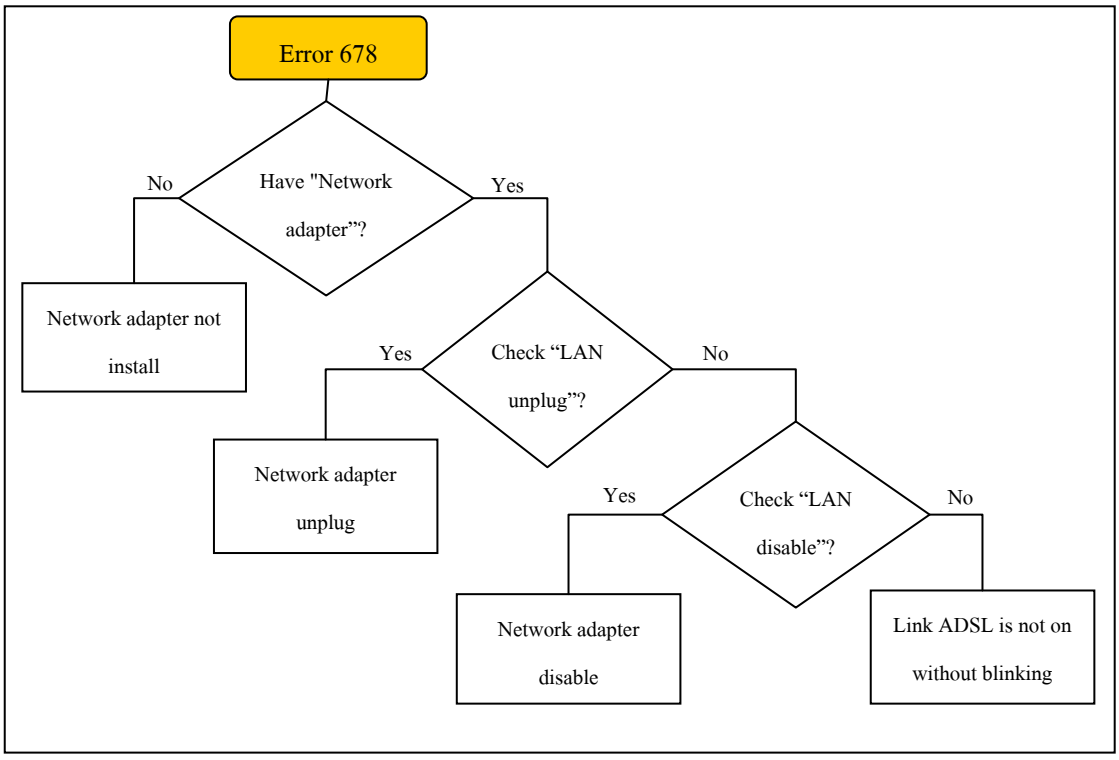
ภาพที่ 14 การแบ่งปัญหาการใช้งานอินเตอร์เน็ตโดยใช้หลักการของออนโทโลยี

ในขั้นตอนของออนโทโลจีได้พัฒนา Software agent ตรวจสอบคำสำคัญที่สกัดได้จากคำถามที่สอบถามเข้ามา โดยตรวจสอบว่าคำถามที่สกัดได้เป็น Leaf node ของ tree หรือไม่ กรณีที่เป็น Leaf node จะทำการวิเคราะห์อาการผิดปกติด้วยเทคนิค Decision tree กรณีที่ไม่เป็น Leaf node นั้น จะวิเคราะห์ Child node จาก Node ที่สกัดได้

ตัวอย่าง การสกัดอาการผิดปกติโดยเทคนิคออนโทโลจี

อาการที่สกัดได้จากคำถาม คือ เชื่อมต่ออินเตอร์เน็ตไม่ได้ จากภาพที่ 14 พบว่าอาการเชื่อมต่ออินเตอร์เน็ตไม่ได้นั้นไม่ได้เป็น Leaf node ของต้นไม้ ดังนั้นในทำงานจะตรวจสอบทุก Child node ที่อยู่ภายใต้ Node การเชื่อมต่ออินเตอร์เน็ตไม่ได้ ต่อไปจนถึง Leaf node และนำ node ที่ได้วิเคราะห์ด้วย Decision tree ต่อไป เป็นต้น

3.2. การวิเคราะห์สาเหตุของปัญหาโดยใช้หลัก Decision Tree โดยวิธีการตรวจสอบอุปกรณ์และค่าต่าง ๆ นั้น ประยุกต์ใช้เทคนิค WMI มาช่วยในการตรวจสอบค่า โดยขั้นตอนและกฎเกณฑ์ในการตรวจสอบนั้นจะใช้หลักการของ Decision tree ที่ได้มาจากผู้เชี่ยวชาญทำการวิเคราะห์และจัดรูปแบบของต้นไม้

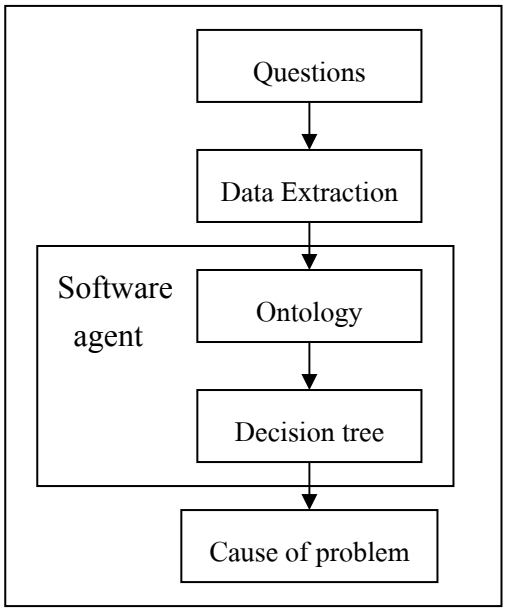


ภาพที่ 15 ตัวอย่างโครงสร้างการวิเคราะห์คำถามกรณีเกิดปัญหา (กรณี Error 678)

จากภาพที่ 15 แสดงตัวอย่างขั้นตอนการตรวจสอบโดย Decision tree โดยดูจากสาเหตุการเกิดปัญหา Error 678 ในขั้นตอนที่ 1 ตรวจสอบ Network adapter ว่าที่เครื่องคอมพิวเตอร์ได้มีการติดตั้งแล้วหรือไม่ กรณีที่ตรวจสอบไม่พบ จะแสดงสาเหตุของปัญหาและวิธีการติดตั้ง

Network adapter ก่อนใช้งาน กรณีที่ตรวจสอบพบอุปกรณ์ Network adapter จะไปตรวจสอบในขั้นตอนที่ 2 ตรวจสอบสถานะ LAN ว่าได้เสียบใช้งานกับคอมพิวเตอร์หรือไม่ กรณีตรวจสอบแล้วพบว่าไม่ได้เชื่อมต่อสาย LAN กับคอมพิวเตอร์จะแสดงสาเหตุของปัญหาและคำแนะนำให้เสียบสาย LAN กับคอมพิวเตอร์ กรณีตรวจสอบพบว่าสายLAN เชื่อมต่อปกติ จะไปตรวจสอบในขั้นตอนที่ 3 ตรวจ สอบสถานะ LAN ว่าถูก Disable หรือไม่ กรณีตรวจสอบพบว่า LAN ถูก Disable จะแสดงสาเหตุของปัญหาและวิธีการ EnableLAN ให้ปกติ กรณีที่ตรวจสอบสถานะ LAN ปกติจะสรุปสาเหตุของ Error 678 ว่าเกิดจาก Modem ทำงานผิดปกติ ไฟ ADSL ไม่ทำงานจากนั้นนำเสนอวิธีการแก้ปัญหาให้ลูกค้าตรวจสอบเป็นต้น

โครงสร้างการทำงานของระบบวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ตกรณีเชื่อมต่ออินเตอร์เน็ตไม่ได้ (กรณี Offline)



ภาพที่ 16 โครงสร้างของงานวิจัยสำหรับวิเคราะห์ปัญหากรณีที่เชื่อมต่ออินเตอร์เน็ตไม่ได้ (กรณี Offline)

ขั้นตอนการทดลองกลุ่มปัญหาการเชื่อมต่ออินเตอร์เน็ตได้

1. การขุดค้นข้อมูล (Data Mining) การขุดค้นข้อมูลได้เลือกใช้เทคนิคของการจัดกลุ่มข้อมูล (Clustering) เข้ามาใช้ในการทดลองโดยทำการเปรียบเทียบระหว่างอัลกอริทึมการจัดกลุ่มแบบ 2 ขั้นตอน (Two Step Clustering) 2 เทคนิคด้วยกัน คือ

1.1 การจัดกลุ่มแบบ 2 ขั้นตอนโดย Clustering Feature Tree และ Hierarchical Algorithm ในโปรแกรมทางสถิติ ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

1.2 การจัดกลุ่มแบบ 2 ขั้นตอนโดย SOM กับ K-Means อัลกอริทึม

ขั้นตอนที่ 1

ขั้นตอนนี้จะทำการจัดกลุ่มข้อมูลโดยใช้ SOM อัลกอริทึมจัดกลุ่มข้อมูลตั้งแต่ 2 -10 กลุ่ม จากนั้นทำการวัดคุณภาพของการจัดกลุ่มโดยการหาค่าความแตกต่างของข้อมูลภายในกลุ่ม (RMSSTD) และค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (RS) จากจำนวนกลุ่มที่แบ่งทั้งสิ้น 2-10 กลุ่ม เพื่อเลือกจำนวนกลุ่มที่ให้คุณภาพการจัดกลุ่มที่ดีที่สุด คือ ค่าความแตกต่างของข้อมูลภายในกลุ่มที่มีค่าน้อย และค่าความแตกต่างของข้อมูลระหว่างกลุ่มที่มีค่ามาก ซึ่งกลุ่มที่ได้นั้น จะถูกใช้เป็นข้อมูลสำหรับการจัดกลุ่มในขั้นตอนที่ 2 ต่อไป

ขั้นตอนที่ 2

ในขั้นตอนนี้จะใช้ เค-มีน อัลกอริทึม (K-Means Algorithm) ทำการจัดกลุ่มโดยจำนวนกลุ่มหรือค่า K นั้น จะเป็นค่าที่ได้มาจากข้อมูลในขั้นตอนที่ 1 ซึ่งหลังจากผ่านขั้นตอนที่ 2 แล้วผลที่ได้ก็คือ ข้อมูลลักษณะลูกค้าตามตัวแปรที่พิจารณา โดยข้อมูลของแต่ละกลุ่มจะถูกแบ่งตามลักษณะการใช้บริการและปัจจัยอื่น ๆ

2. การวิเคราะห์ข้อมูลโดยใช้หลัก OLAP Cube การวิเคราะห์ข้อมูลในงานวิจัยนี้ได้ใช้เทคนิค OLAP Cube มาช่วยโดยมี Attribute ที่ใช้ในการวิเคราะห์คือ กลุ่มประเภทลูกค้าที่เป็นผลลัพธ์จากการจัดกลุ่มปัญหาการใช้งานและวิธีการแก้ไข ปัญหา เพื่อหาความสัมพันธ์ของข้อมูลต่อไป

ตารางที่ 1 ตัวอย่างข้อมูลลูกค้าที่ใช้จัดกลุ่ม

Login	First use	Speed	OS	Modem	Grade	Last use
labtest1	02/05/06 – 16:00:00	1 Mb	Windows XP	Zyxel	B	10/04/07 – 18:00:00
labtest2	13/02/07 – 8:10:00	256 Kb	Windows XP	Huawei	A	18/04/07 – 16:07:00
labtest3	10/03/07 – 23:08:00	512 Kb	Windows XP	Huawei	A	18/04/07 – 21:30:00

ตารางที่ 2 ตัวอย่างข้อมูลอาการผิดปกติและสาเหตุของปัญหา

Login	ปัญหา 1		ปัญหา 2	
	รหัสปัญหา	รหัสสาเหตุของปัญหา	รหัสปัญหา	รหัสสาเหตุ
labtest1	1001	202	1004	201
labtest2	1001	202	-	-
labtest3	1004	201	-	-

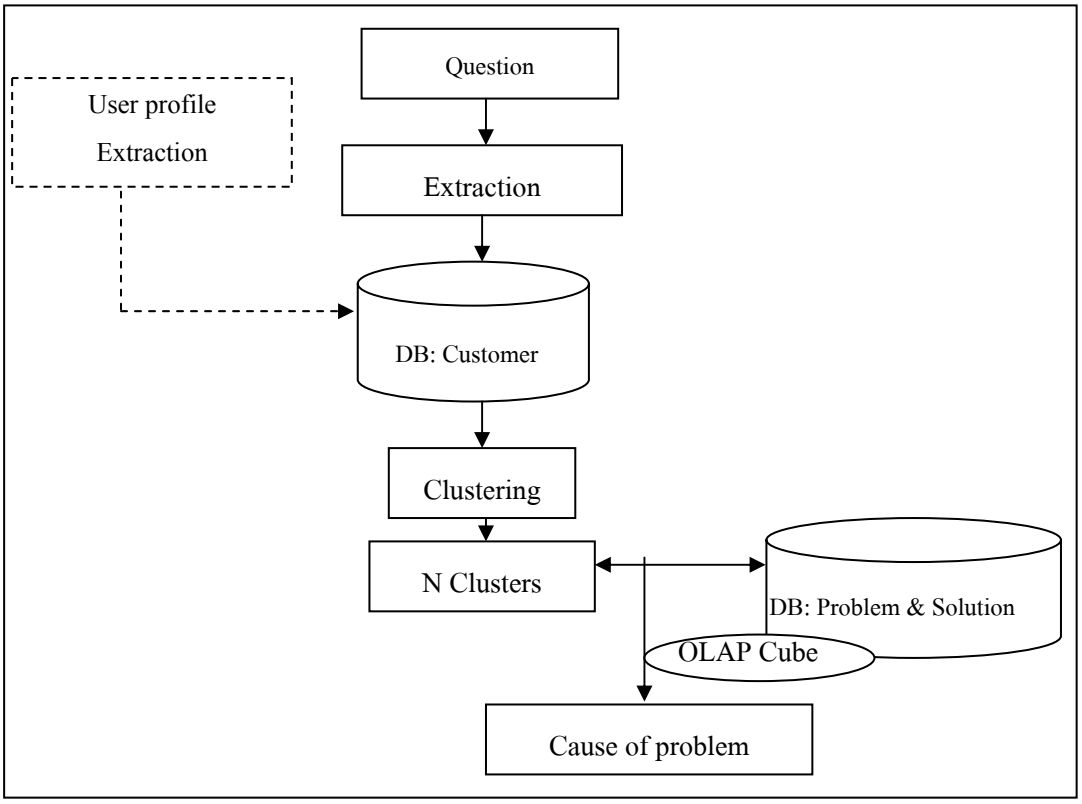
ตารางที่ 3 ตัวอย่างข้อมูลอาการผิดปกติของปัญหา

รหัสปัญหา	คำอธิบาย
1001	หลุดบ่อย
1002	เปิด web ไม่ได้
1003	ปัญหาการรับ/ส่ง Mail

ตารางที่ 4 ตัวอย่างตารางแสดงสาเหตุอาการผิดปกติ

รหัสสาเหตุ	คำอธิบาย
201	ไวรัส
202	สายชำรุด

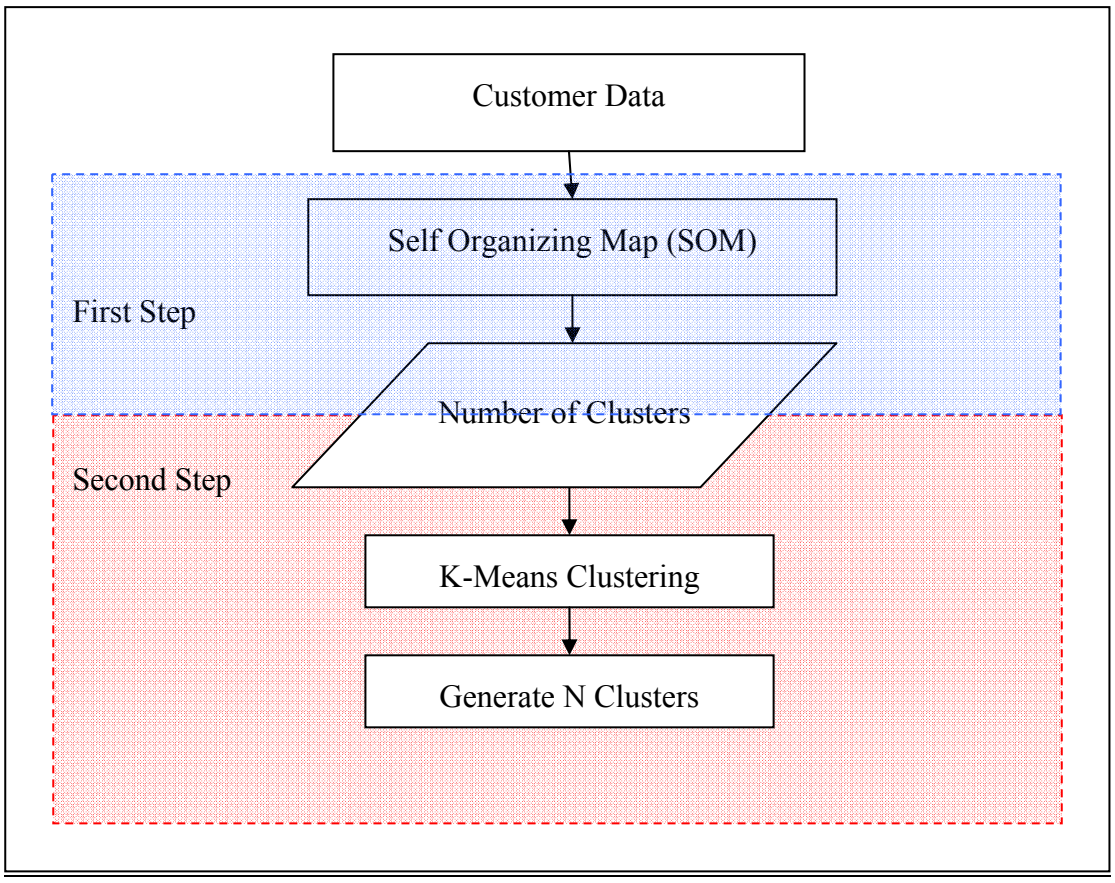
โครงสร้างการทำงานของระบบวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ตกรณีเชื่อมต่ออินเตอร์เน็ตได้ (กรณี Online)



ภาพที่ 17 โครงสร้างของงานวิจัยสำหรับวิเคราะห์ปัญหากรณีเกิดขึ้นภายหลังจากการเชื่อมต่ออินเตอร์เน็ตได้(กรณี Online)

เทคนิคการแบ่งกลุ่มแบบ 2 ขั้นตอนโดยใช้ SOM และ K-means อัลกอริทึม

การจัดแบ่งกลุ่มข้อมูลแบบ 2 ขั้นตอนโดยใช้ SOM และ K-means นั้น ในขั้นตอนแรกจะนำข้อมูลของลูกค้ามาทำการจัดกลุ่มโดยใช้ SOM อัลกอริทึม เพื่อหาจำนวนกลุ่มที่ดีที่สุด จากนั้นขั้นตอนที่สอง จะทำการจัดกลุ่มใหม่โดยใช้ K-means อัลกอริทึมทำการจัดกลุ่มลูกค้าตามข้อมูลจำนวนกลุ่มที่ได้รับจากขั้นตอนแรก (ดังภาพที่ 18)



ภาพที่ 18 ภาพแสดงเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดยอัลกอริทึม SOM และ K-means

ในการสร้างตัวแบบสำหรับการแบ่งกลุ่มข้อมูลลูกค้าอินเตอร์เน็ตนั้น ใช้ตัวแปร 6 ตัวดังต่อไปนี้

ตารางที่ 5 ตัวแปรทั้งหมดที่จะใช้ในการสร้างตัวแบบการจัดกลุ่มลูกค้าที่ใช้งานอินเตอร์เน็ต

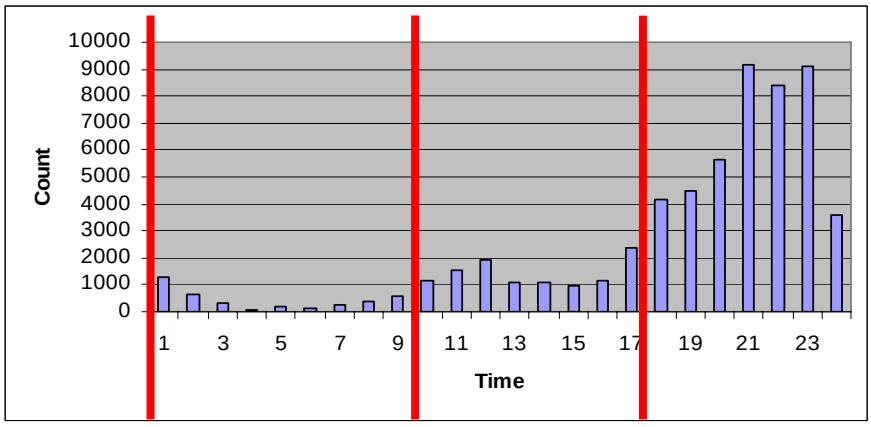
ข้อมูล	ชื่อตัวแปร
ช่วงเวลาที่ใช้งาน	Time
ระบบปฏิบัติการที่ใช้	Operating System
ประเภทของลูกค้า	Type Customer
ความสามารถด้าน IT	IT _literacy
Modem ที่ใช้งาน	Modem
ความเร็วอินเตอร์เน็ตที่ขอใช้งาน	Speed

1) ช่วงเวลาที่ใช้งาน (Time) โดยจำนวนกลุ่มจะได้จากสถิติการใช้งานจริง จากภาพที่ 19 ช่วงเวลาการใช้งานอินเตอร์เน็ตสามารถแบ่งเป็น 3 กลุ่ม

กลุ่มที่1 0:00 am. - 9:59 am.

กลุ่มที่2 10:00 am. – 17.59 pm.

กลุ่มที่3 18:00 pm. - 23:59 pm.



ภาพที่ 19 แสดงปริมาณการใช้งานอินเตอร์เน็ตช่วงวันที่ 15-31 มกราคม 2550

2) ระบบปฏิบัติการ (Operating System) จะแบ่งประเภทตามที่มีผู้ใช้งานจริงโดยแบ่งเป็น 5 กลุ่ม

กลุ่มที่1 Windows XP

กลุ่มที่2 Windows 2000

กลุ่มที่3 Windows ME

กลุ่มที่4 Windows 98

กลุ่มที่ 5 Other(OS Macintosh and Windows Vista)

3) ประเภทของลูกค้า (Type Customer) ลักษณะการแบ่งจะดูระยะเวลาการเริ่มใช้งานอินเตอร์เน็ตของลูกค้าโดยแบ่งเป็น 2 กลุ่ม

กลุ่มที่1 ลูกค้าใหม่ (น้อยกว่า 3 เดือน)

กลุ่มที่2 ลูกค้าเก่า

4) ความรู้ความสามารถเรื่องอินเตอร์เน็ต (IT_literacy) โดยข้อมูลได้มาจาก Call Center โดยดูตามลักษณะของผู้ใช้งานโดยแบ่งเป็น 3 กลุ่ม

กลุ่มที่1 ความรู้เกรด A (ลูกค้าที่มีความรู้สามารถตรวจสอบค่าต่าง ๆ ได้เองและใช้ศัพท์เทคนิคทาง ITได้)

กลุ่มที่2 ความรู้เกรด B (สามารถแก้ปัญหาตามที่ Call Center แนะนำและทราบข้อมูลเบื้องต้น)

กลุ่มที่3 ความรู้เกรด C (ลูกค้าไม่สามารถปฏิบัติตามคำแนะนำได้)

5) ลักษณะของ Modem ที่ใช้งาน (Modem) โดยแบ่งตามรุ่น Modem ที่เป็นที่นิยมในตลาดโดยแบ่งออกเป็น 4 กลุ่ม

กลุ่มที่1 Zyxel

กลุ่มที่2 Billion

กลุ่มที่3 Huawei

กลุ่มที่4 Other

6) ความเร็วอินเตอร์เน็ตที่ขอใช้งาน(Speed) แบ่ง ออกเป็น 4 กลุ่ม

กลุ่มที่1 128 Kb และ 256 Kb

กลุ่มที่2 512 Kb

กลุ่มที่3 1 Mb

กลุ่มที่4 มากกว่า 1 Mb

ผลและวิจารณ์

ผล

1. ผลการทดลองกรณี Offline

การวัดประสิทธิภาพในส่วนนี้จะใช้เทคนิค Precision จากสูตร

$$\text{Precision} = \frac{\text{จำนวนคำตอบที่ถูกค้นคืนถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมดที่ถูกค้นคืน}}$$

จากข้อมูลที่ใช้ในการทดลองจำนวน 10,000 รายการ พบว่ามีความถูกต้องในการค้นคืนถึง 8,756 รายการ ดังนั้นค่า Precision ที่วัดได้จากการทดลองนี้เท่ากับ 87.56 % ซึ่งได้ผลลัพธ์ออกมาอยู่ในระดับดี

2. ผลการทดลองกรณี Online

2.1 ขั้นตอนแรก ทำการจัดกลุ่มข้อมูลสำหรับ Training Data ซึ่งมีจำนวนทั้งสิ้น 6,399 จากข้อมูลทั้งหมด 7,110 รายการ คิดเป็น 90 % โดยใช้ SOM อัลกอริทึมตั้งแต่ 2 – 10 กลุ่ม ได้ผลการทดลองดังตารางที่ 6

ตารางที่ 6 ผลการวัดประสิทธิภาพการจัดกลุ่มลูกค้าโดย SOM อัลกอริทึมตั้งแต่ 2-10 กลุ่ม

จำนวนกลุ่ม (Cluster)	ค่าความแตกต่างภายในกลุ่ม	ค่าความแตกต่างระหว่างกลุ่ม
2	0.961	0.262
3	0.936	0.300
4	0.825	0.456
5	0.745	0.556
6	0.730	0.574
7	0.721	0.584
8	0.634	0.679
9	0.705	0.603
10	0.743	0.559

จากตารางที่ 6 แสดงให้เห็นถึงค่าความแตกต่างของข้อมูลภายในกลุ่ม และค่าความแตกต่างของข้อมูลระหว่างกลุ่ม เมื่อทำการจัดกลุ่มข้อมูลโดยใช้ SOM อัลกอริทึม ระบุจำนวนกลุ่มในการจัดกลุ่มตั้งแต่ 2 - 10 กลุ่ม หากแต่เมื่อพิจารณาจากค่าความแตกต่างของข้อมูลภายในกลุ่ม และค่าความแตกต่างของข้อมูลระหว่างกลุ่มแล้วนั้น พบว่าเมื่อทำการระบุจำนวนกลุ่มในการจัดกลุ่มเท่ากับ 8 กลุ่ม จะทำให้ค่าความแตกต่างของข้อมูลภายในกลุ่ม และค่าความแตกต่างของข้อมูลระหว่างกลุ่มดีที่สุด คือ จะให้ค่าความแตกต่างของข้อมูลภายในกลุ่มน้อยที่สุดในการจัดกลุ่มตั้งแต่ 2 – 10 กลุ่ม โดยมีค่าเพียง 0.634 ในขณะที่ค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อระบุจำนวนกลุ่มเท่ากับ 8 กลุ่มนั้น มีค่ามากที่สุด คือ มีค่าเท่ากับ 0.679

จากผลลัพธ์ดังกล่าวนี้ ทำให้สามารถสรุปได้ว่าจำนวนกลุ่มเท่ากับ 8 กลุ่มนั้น เป็นผลลัพธ์ที่ดีที่สุดที่ได้จากขั้นตอนที่ 1 ซึ่งจะนำไปใช้ในขั้นตอนที่ 2 คือการจัดกลุ่มโดย K-Means อัลกอริทึมต่อไป

2.2 ขั้นตอนที่ 2 นี้จะทำการจัดกลุ่มโดย K-Means อัลกอริทึม โดยกำหนดจำนวนกลุ่มให้เท่ากับผลลัพธ์ที่ได้จากการจัดกลุ่มโดยใช้ SOM อัลกอริทึม ในขั้นตอนที่ 1

ตารางที่ 7 ผลการวัดประสิทธิภาพการจัดกลุ่มลูกค้าโดย K-Means อัลกอริทึมเมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม

วิธีการแบ่งกลุ่ม	จำนวนกลุ่ม (Cluster)	ความแตกต่างภายในกลุ่ม	ความแตกต่างระหว่างกลุ่ม
SOM อัลกอริทึม	8	0.634	0.679
K-means อัลกอริทึม	8	0.616	0.697

จากตารางที่ 7 แสดงให้เห็นถึงผลการจัดกลุ่มโดย K-Means อัลกอริทึมเมื่อกำหนดจำนวนกลุ่มเท่ากับ 8 กลุ่ม ซึ่งจะเห็นว่าค่าความแตกต่างของข้อมูลภายในกลุ่ม และค่าความแตกต่างของข้อมูลระหว่างกลุ่มเมื่อเปรียบเทียบกับการจัดกลุ่มโดย SOM อัลกอริทึมนั้นมีค่าดีขึ้น คือ ค่าความแตกต่างของข้อมูลภายในกลุ่มลดลงจาก 0.634 เหลือเพียง 0.616 ในขณะที่ค่าความแตกต่างของข้อมูลระหว่างกลุ่มดีขึ้น คือ เพิ่มจาก 0.679 เป็น 0.697

3. ผลการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอน (Two Step Clustering) โดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical อัลกอริทึม

การจัดกลุ่มข้อมูลแบบ 2 ขั้นตอน โดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical นั้น จะทำการจัดกลุ่มโดยใช้โปรแกรมทางสถิติ โดยกำหนดให้โปรแกรมทำการจัดกลุ่มอัตโนมัติ โดยระบุจำนวนกลุ่มมากสุด คือ 10 กลุ่ม และ กำหนดให้โปรแกรมทำการจัดกลุ่มข้อมูลเท่ากับ 8 กลุ่ม ซึ่งเป็นผลลัพธ์ที่ได้จากเทคนิคการจัดกลุ่มแบบ SOM อัลกอริทึม เพื่อทำการเปรียบเทียบประสิทธิภาพการจัดกลุ่ม

ตารางที่ 8 ผลการวัดประสิทธิภาพการจัดกลุ่มลูกค้าโดยใช้เทคนิคการจัดกลุ่ม 2 ขั้นตอน Clustering Feature Tree และ Agglomerative Hierarchical algorithm

วิธีการแบ่งกลุ่ม	จำนวนกลุ่ม (Cluster)	ความแตกต่างภายในกลุ่ม	ความแตกต่างระหว่างกลุ่ม
จัดกลุ่มอัตโนมัติ	4	0.773	0.342
ระบุจำนวนกลุ่ม	8	0.775	0.471

ตารางที่ 8 แสดงผลการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical โดยโปรแกรมทางสถิติ จะทำการคำนวณจำนวนกลุ่มอัตโนมัติได้ 4 กลุ่ม ซึ่งผลลัพธ์เมื่อเทียบกับจำนวน 8 กลุ่มแล้วจะเห็นได้ว่าค่าความแตกต่างภายในกลุ่ม จำนวน 4 กลุ่มจะได้ผลลัพธ์ที่ดีกว่า คือมีค่าความแตกต่างกันเพียง 0.773 ส่วน 8 กลุ่มมีค่าความแตกต่างกัน 0.775 ส่วนค่าความแตกต่างระหว่างกลุ่ม จำนวน 8 กลุ่มได้ผลลัพธ์ที่ดีกว่าคือมีค่าความต่างกัน 0.471 ส่วน 4 กลุ่มมีค่าความต่างกัน 0.342 ซึ่งจะเอาค่าที่ได้นี้ไปเปรียบเทียบกับการจัดกลุ่มโดยใช้ SOM กับ K-means ในขั้นตอนต่อไป

4. ผลการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนโดยใช้ SOM และ K-Means อัลกอริทึม และการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical

จากการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนทั้ง 2 วิธีการข้างต้นนั้น นำผลลัพธ์ที่ได้มาเปรียบเทียบประสิทธิภาพของการจัดกลุ่ม และทำการเลือกผลการจัดกลุ่มที่ดีที่สุด เพื่อนำไปหาความสัมพันธ์การใช้บริการภายในระบบต่อไป

ตารางที่ 9 ผลการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนโดยใช้ SOM และ K-Means อัลกอริทึม กับเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ Clustering Feature Tree และ Agglomerative Hierarchical

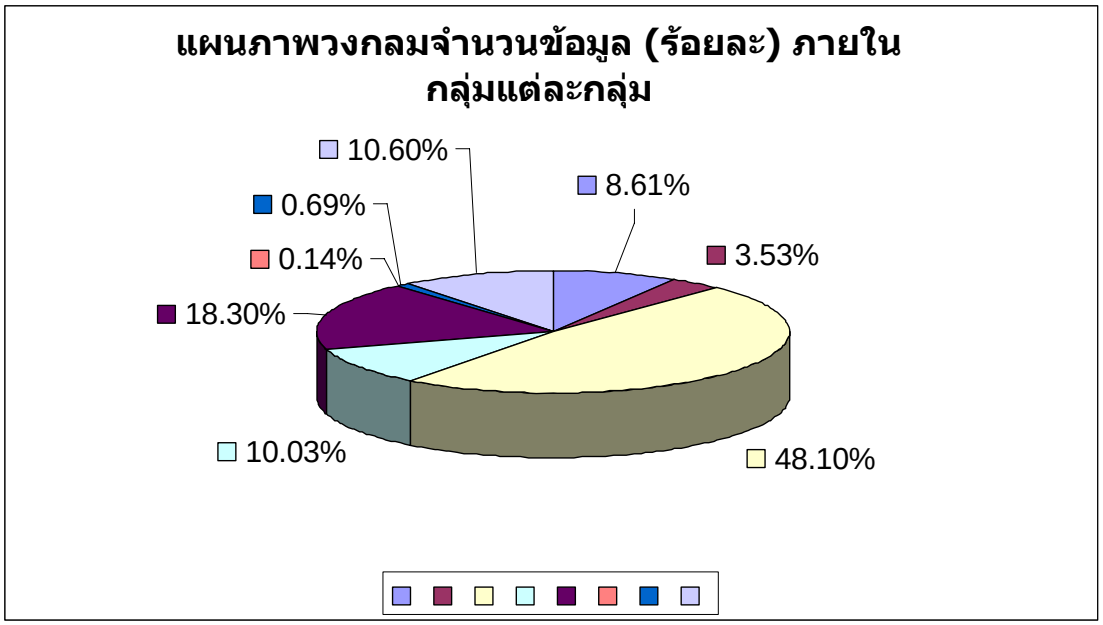
วิธีการแบ่งกลุ่ม	จำนวนกลุ่ม (Cluster)	ความแตกต่างภายในกลุ่ม	ความแตกต่างระหว่างกลุ่ม
SOM อัลกอริทึม	8	0.634	0.679
K-means อัลกอริทึม	8	0.616	0.697
Clustering Feature Tree และ	4	0.773	0.342
Agglomerative Hierarchical	8	0.775	0.471

จากตารางที่ 9 แสดงถึงผลการเปรียบเทียบการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนโดย SOM และ K-Means อัลกอริทึม กับเทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดย Clustering Feature Tree และ Agglomerative Hierarchical ซึ่งจะเห็นว่าการจัดกลุ่มโดย SOM และ K-means อัลกอริทึมนั้นจะสามารถจัดกลุ่มได้ดีกว่า คือได้ค่าความต่างภายในกลุ่มดีที่สุดเพียง 0.616 และค่าความต่างระหว่างกลุ่มดีที่สุด 0.697 จากผลการเปรียบเทียบข้างต้นจึงใช้เทคนิคการจัดกลุ่มแบบ 2 ขั้นตอนโดยใช้ SOM และ K-Means อัลกอริทึม จำนวน 8 กลุ่ม มาดำเนินการวิเคราะห์และสรุปผลต่อไป

5. ผลลัพธ์จากการจัดกลุ่มแบบ 2 ขั้นตอนของ SOM และ K-Means อัลกอริทึม

ตารางที่ 10 แสดงจำนวนลูกค้าและร้อยละของจำนวนลูกค้าทั้ง 8 กลุ่ม

กลุ่มที่	1	2	3	4	5	6	7	8	รวม
จำนวนลูกค้า (ราย)	551	226	3078	642	1171	9	44	678	6399
จำนวนลูกค้า (ร้อยละ)	8.61	3.53	48.10	10.03	18.30	0.14	0.69	10.60	100



ภาพที่ 20 แสดงข้อมูลจำนวน (ร้อยละ) ของลูกค้า ภายในกลุ่มแต่ละกลุ่ม

ภาพที่ 20 แสดงถึงผลการจัดกลุ่มข้อมูลของลูกค้า ซึ่งจำนวนสมาชิกในแต่ละกลุ่มมีความแตกต่างกันโดยสมาชิกที่มีมากที่สุดนั้น คือ กลุ่มที่ 3 ซึ่งมีจำนวนสมาชิกมากถึงร้อยละ 48.10 ของจำนวนข้อมูลทั้งหมด ดังตารางที่ 11 ที่ได้แจกแจงรายละเอียดของแต่ละกลุ่ม

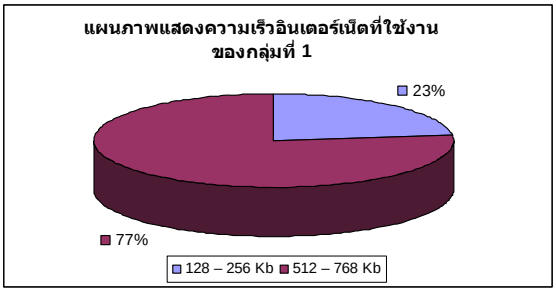
ตารางที่ 11 ผลการจัดกลุ่มข้อมูลโดยใช้เทคนิคการจัดกลุ่มข้อมูลแบบ 2 ขั้นตอนระหว่าง SOM และ K-Means อัลกอริทึม จำนวน 8 กลุ่ม

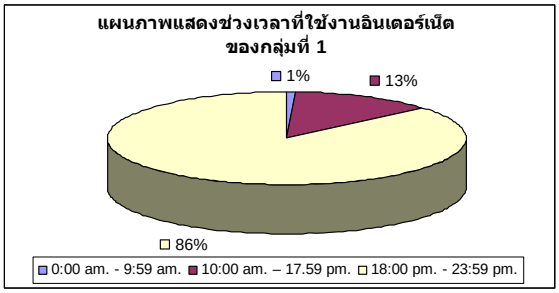
Factor		Cluster							
		1	2	3	4	5	6	7	8
Period of working	0:00 am. - 9:59 am.	5	13	0	24	283	0	3	0
	10:00 am. – 17.59 pm.	70	36	0	165	888	1	10	0
	18:00 pm. - 23:59 pm.	476	177	3078	453	0	8	31	678
Operation System	Windows XP	551	226	3068	639	1162	0	0	677
	Windows 2000	0	0	4	1	3	0	0	0
	Windows ME	0	0	0	0	0	3	27	0
	Windows 98	0	0	0	0	0	6	17	0
	Other	0	0	6	2	6	0	0	1
Brand name of Modem	Zyxel	0	0	0	202	334	0	10	678
	Billion	0	0	3078	440	805	0	34	0
	Huawei	418	95	0	0	32	7	0	0
	Other	133	131	0	0	0	2	0	0

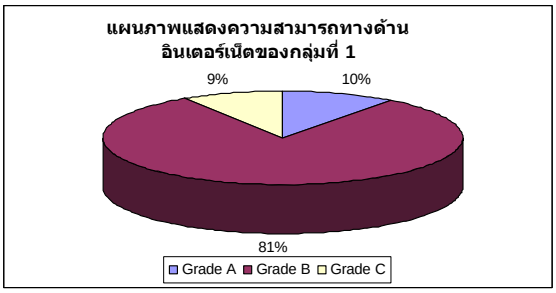
ตารางที่ 11 (ต่อ)

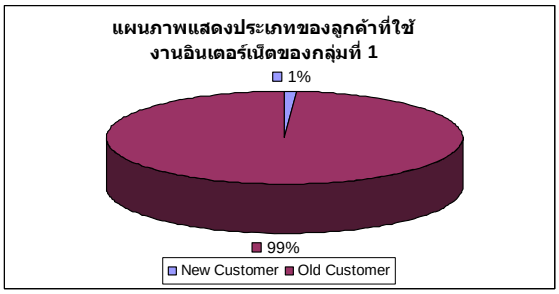
Factor		Cluster							
		1	2	3	4	5	6	7	8
IT literacy	Grade A	57	41	275	78	123	0	3	70
	Grade B	444	174	2555	529	938	7	36	543
	Grade C	50	11	248	35	110	2	5	65
Customer	New Customer	6	3	216	71	319	0	3	27
	Old Customer	545	223	2862	571	852	9	41	651
Speed of internet	128 – 256 Kb	127	0	377	0	308	5	9	133
	512 – 768 Kb	424	0	2701	0	845	3	33	545
	1 – 2 Mb	0	123	0	333	17	1	2	0
	> 2 Mb	0	103	0	309	0	0	0	0

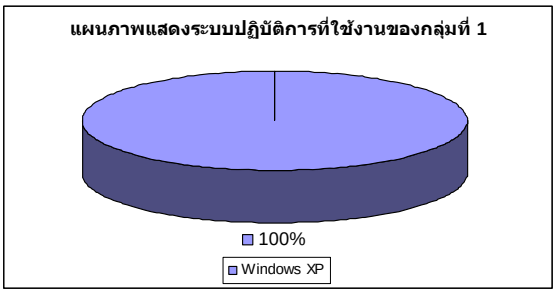
5.1 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 1

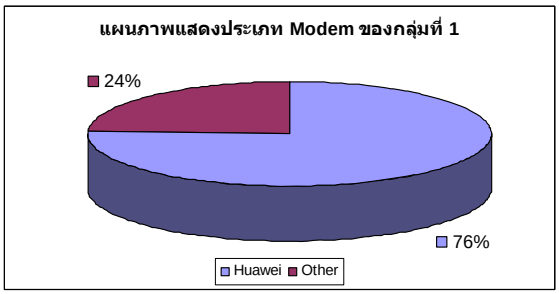






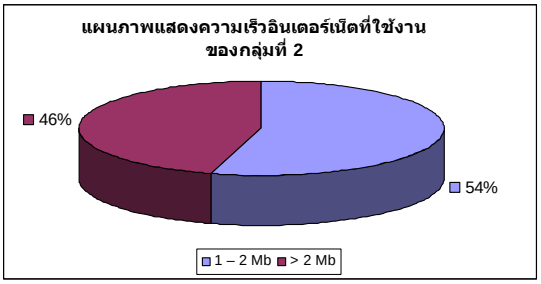


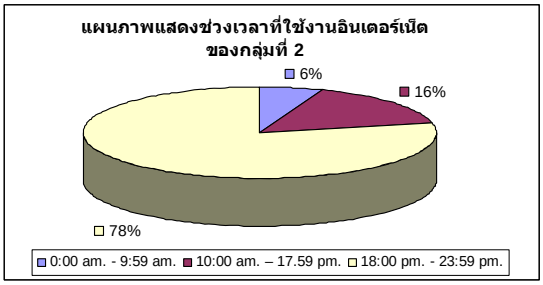


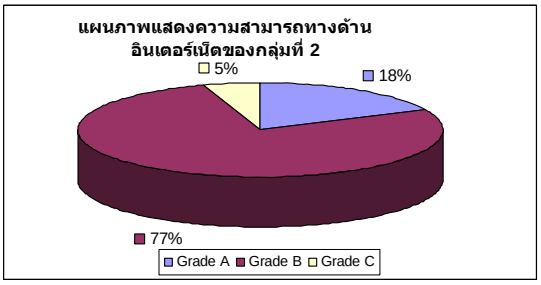


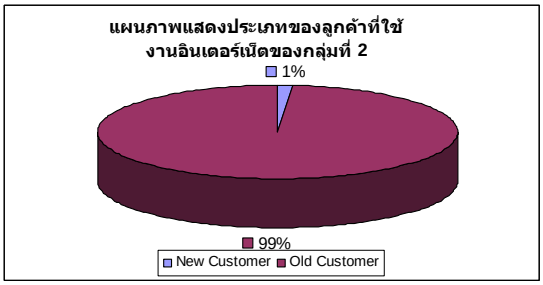
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 1 นี้ มีจำนวนทั้งสิ้น 551 รายการ หรือคิดเป็นร้อยละ 8.61 ของจำนวนข้อมูลทั้งหมด ลักษณะโดยทั่วไปของลูกค้าในกลุ่มนี้ จะใช้อินเตอร์เน็ตที่มีความเร็ว 514 - 768 Kb เป็นส่วนใหญ่ คิดเป็นร้อยละ 77 โดยร้อยละ 86 ของกลุ่มนี้จะใช้งานอินเตอร์เน็ตในช่วง เวลา 18.00 pm. – 23.59 pm. ความรู้ความสามารถด้านอินเตอร์เน็ตอยู่ในระดับ Grade B ร้อยละ 81 ลูกค้าเกือบทั้งหมดในกลุ่มจะเป็นลูกค้าที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน คิดเป็นร้อยละ 99 ยี่ห้อ Modem ที่ใช้งานในกลุ่มนี้เป็นยี่ห้อ Huawei ร้อยละ 76 โดยทุกคนในกลุ่มนี้ใช้ระบบปฏิบัติการ Windows XP

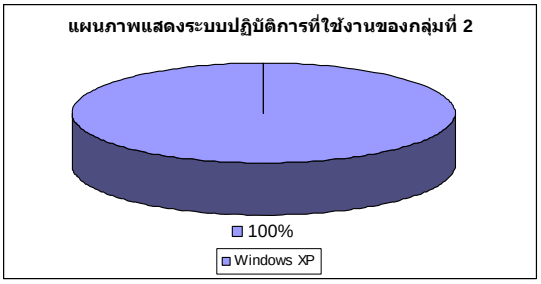
5.2 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 2

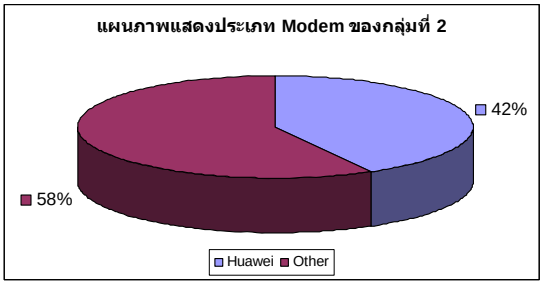






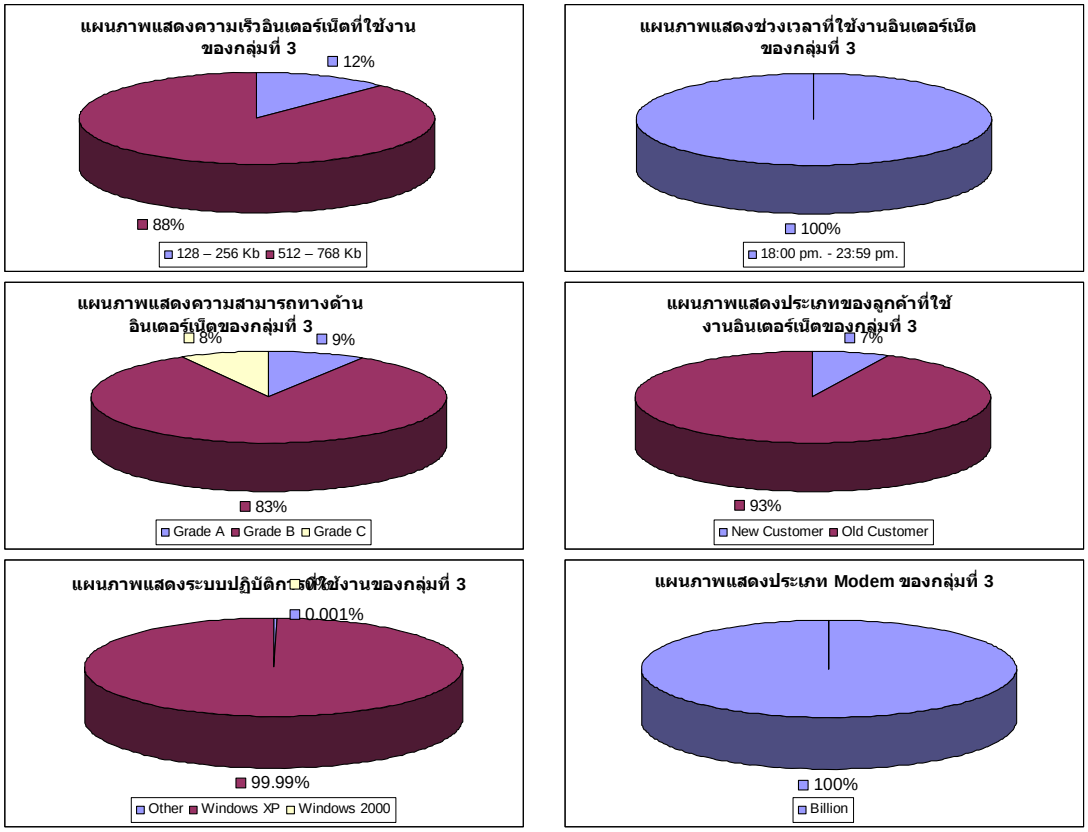






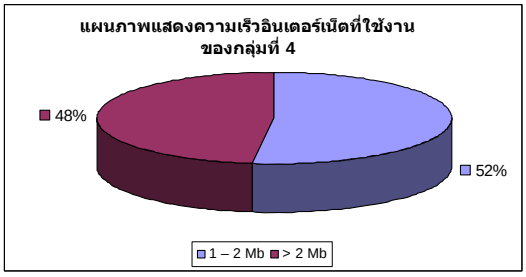
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 2 นี้ มีจำนวนทั้งสิ้น 226 รายการคิดเป็นร้อยละ 3.53 ของจำนวนข้อมูลทั้งหมด ลักษณะโดยทั่วไปของลูกค้าในกลุ่มนี้ จะแบ่งการใช้งาน Modem เป็น 2 ประเภท คือ ยี่ห้อ Huawei ร้อยละ 42 และยี่ห้ออื่น ๆ ในตลาดคิดเป็นร้อยละ 58 ความเร็วของอินเตอร์เน็ตที่ใช้งานสามารถแบ่งได้ 2 ประเภท อยู่ระหว่าง 1-2 Mb ร้อยละ 54 และใช้ความเร็วอินเตอร์เน็ตมากกว่า 2 Mb ร้อยละ 46 โดยลูกค้าส่วนใหญ่ใช้งานอินเตอร์เน็ตในช่วงเวลา 18:00 pm. – 23.59 pm. และความรู้ทางด้านอินเตอร์เน็ตของลูกค้ากลุ่มนี้อยู่ในระดับความสามารถ Grade B (คิดเป็นร้อยละ 77) ลูกค้าเกือบทั้งหมดในกลุ่มเป็นลูกค้าที่ใช้อินเตอร์เน็ตมากกว่า 3 เดือนและทั้งหมดในกลุ่มใช้ระบบปฏิบัติการ Windows XP

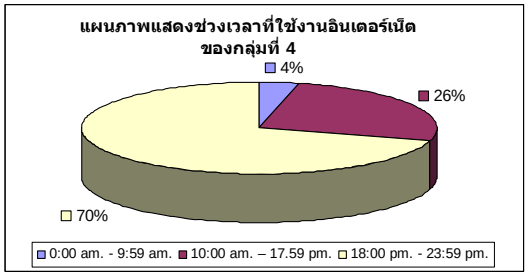
5.3 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 3

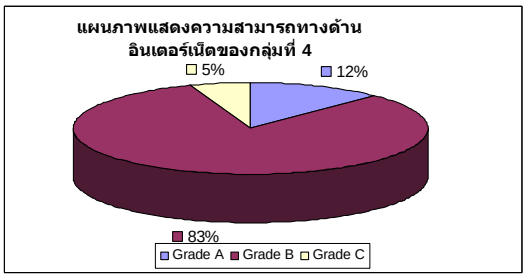


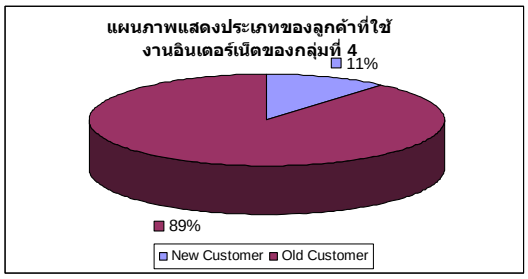
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 3 นี้ มีจำนวนทั้งสิ้น 3078 รายการ ซึ่งถือว่ามีจำนวนสมาชิกมากที่สุด หรือคิดเป็นร้อยละ 48.1 ของจำนวนข้อมูลทั้งหมด ลักษณะโดยทั่วไปของลูกค้าในกลุ่มนี้เป็นลูกค้าที่ใช้งานอินเตอร์เน็ตในช่วงเวลา 18:00 pm. – 23.59 pm. และใช้ Modem ยี่ห้อ Billion ทุกคน เกือบทั้งหมดของลูกค้าในกลุ่มนี้ใช้ระบบปฏิบัติการเป็น Windows XP ระดับความรู้ด้านอินเตอร์เน็ตของลูกค้ากลุ่มนี้อยู่ใน Grade B (คิดเป็นร้อยละ 83) ความเร็วอินเตอร์เน็ตที่ใช้งานอยู่ที่ 512 – 768 Kb ร้อยละ 88 และร้อยละ 93 ของลูกค้าในกลุ่มเป็นลูกค้าที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน

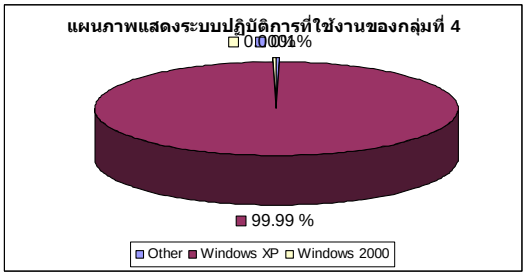
5.4 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 4

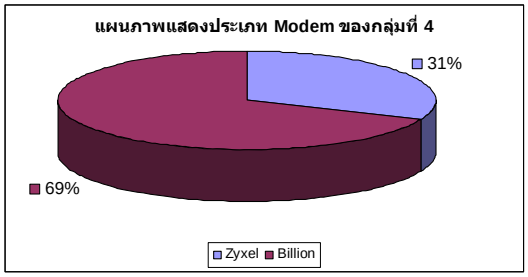






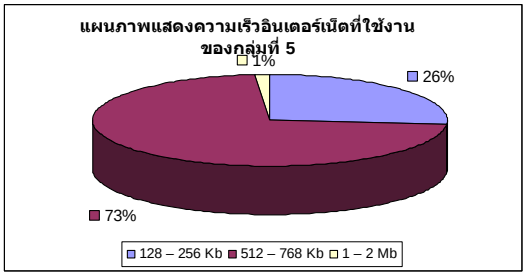


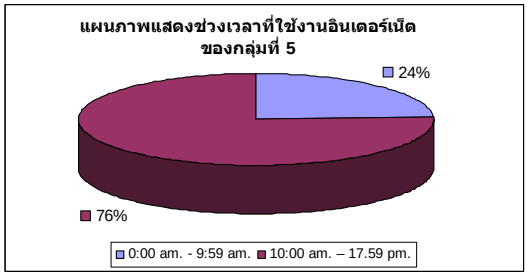


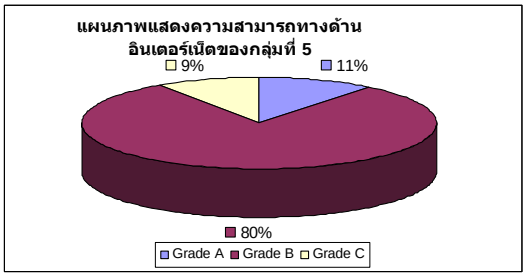


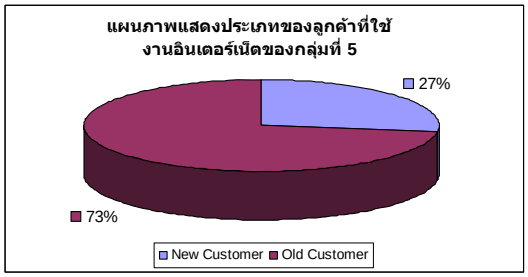
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 4 นี้ มีจำนวนทั้งสิ้น 642 รายการหรือคิดเป็นร้อยละ 10.03 ของจำนวนข้อมูลทั้งหมด ลักษณะโดยทั่วไปของลูกค้าในกลุ่มนี้ เกือบทุกคนในกลุ่มใช้ระบบปฏิบัติการ Windows XP ส่วนมากใช้งานอินเตอร์เน็ตในช่วง 18:00 pm. – 23.59 pm. (คิดเป็นร้อยละ 70) โดยร้อยละ 83 ของลูกค้ากลุ่มนี้ ความรู้ทางด้านอินเตอร์เน็ตอยู่ในระดับ Grade B ลูกค้าเป็นกลุ่มลูกค้าเดิมที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน ร้อยละ 89 โดย Modem ที่ใช้งานส่วนมากเป็นยี่ห้อ Billion ความเร็วอินเตอร์เน็ตของลูกค้ากลุ่มนี้แบ่งได้ 2 ส่วนคือ ความเร็วระหว่าง 1-2 Mb ร้อยละ 52 และร้อยละ 48 ใช้งานความเร็วมากกว่า 2 Mb

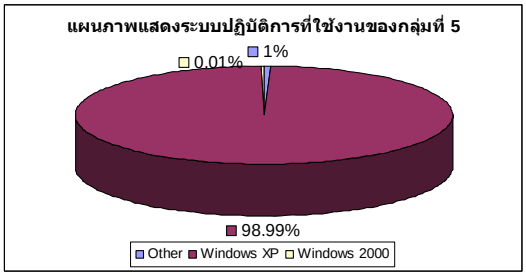
5.5 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 5

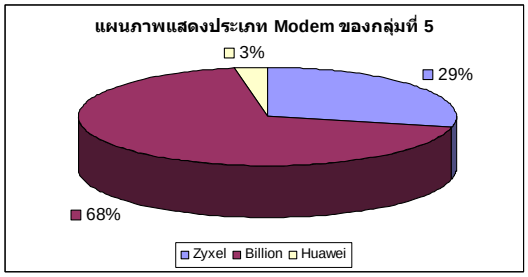






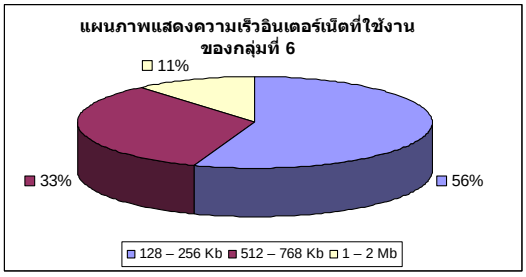


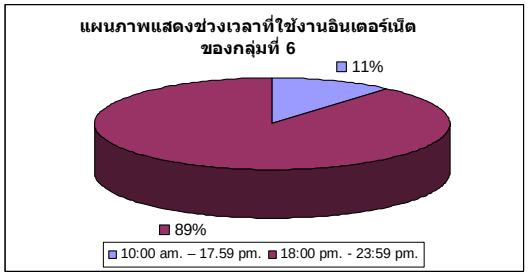


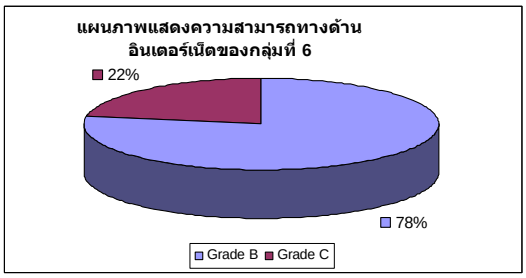


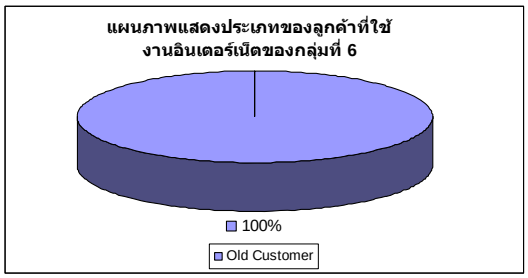
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 5 นี้ มีจำนวนทั้งสิ้น 1171 รายการ หรือคิดเป็นร้อยละ 18.3 ของจำนวนข้อมูลทั้งหมด โดยลักษณะทั่วไปของลูกค้ากลุ่มนี้เป็นลูกค้าเดิมที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน โดยใช้งานอินเตอร์เน็ตที่ความเร็ว 512 – 768 Kb และใช้งานในช่วง 10:00 am. – 17:59 pm. เป็นส่วนใหญ่ (คิดเป็นร้อยละ 76) ความรู้ทางด้านอินเตอร์เน็ตของลูกค้ากลุ่มนี้ร้อยละ 80 อยู่ในระดับ Grade B ส่วนใหญ่ใช้งาน Modem ยี่ห้อ Billion (คิดเป็นร้อยละ 68) เกือบทั้งหมดของกลุ่มใช้งานระบบปฏิบัติการ Windows XP (คิดเป็นร้อยละ 98.99)

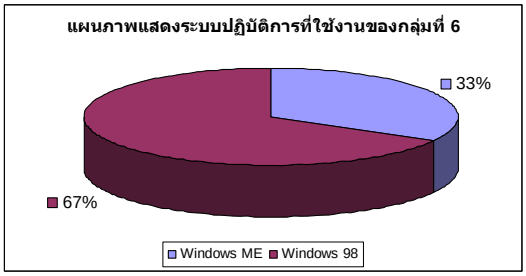
5.6 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 6

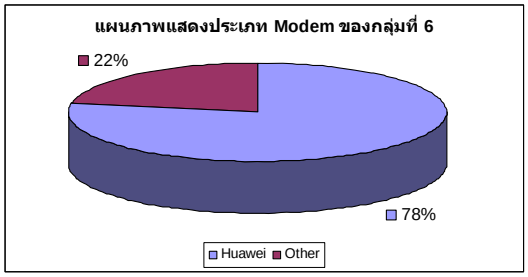






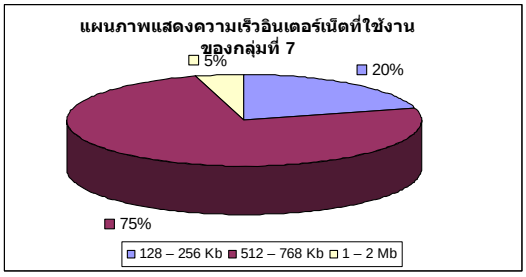


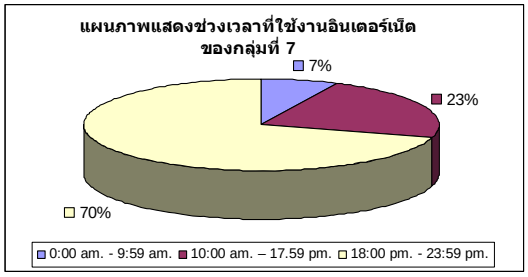


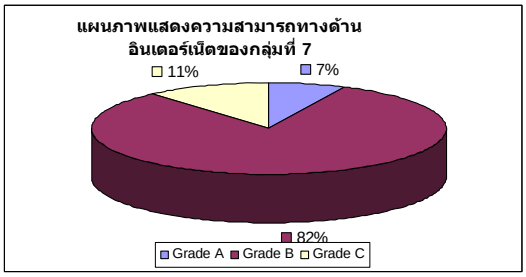


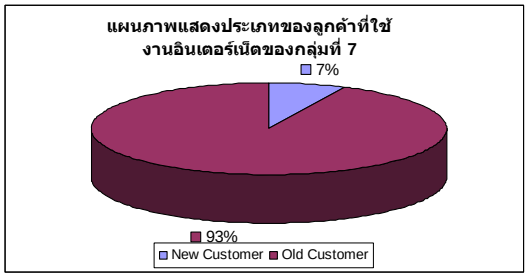
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 6 นี้ มีจำนวนทั้งสิ้น 9 รายการ หรือคิดเป็นร้อยละ 0.14 ของจำนวนข้อมูลทั้งหมด โดยลักษณะทั่วไปของลูกค้ากลุ่มนี้ ส่วนใหญ่ใช้ Modem ยี่ห้อ Huawei และความรู้ความสามารถทางด้านอินเตอร์เน็ตอยู่ในระดับ Grade B (คิดเป็นร้อยละ 78) ร้อยละ 56 ของกลุ่มใช้งานอินเตอร์เน็ตความเร็ว 128 – 256 Kb ส่วนใหญ่ในกลุ่มนี้ใช้งานอินเตอร์เน็ตใน ช่วงเวลา 18:00 pm. – 23:59 pm. (คิดเป็นร้อยละ 89) ในกลุ่มนี้ใช้งานระบบปฏิบัติการ Windows 98 ร้อยละ 67 และทั้งหมดของลูกค้ากลุ่มนี้เป็นลูกค้าเก่าที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน

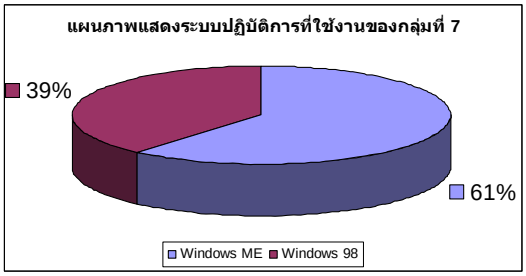
5.7 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 7

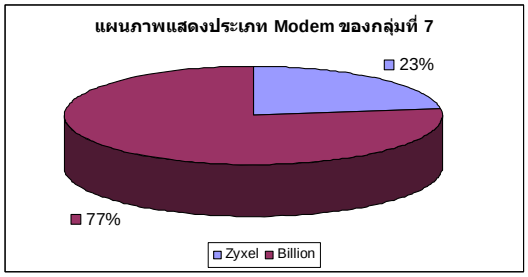






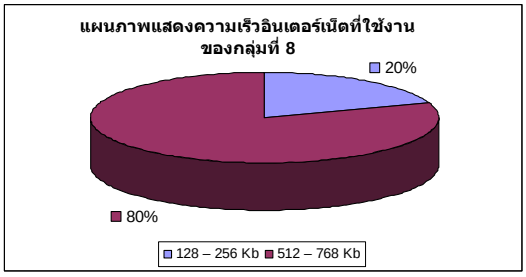


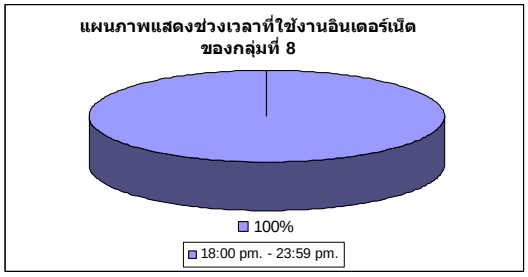


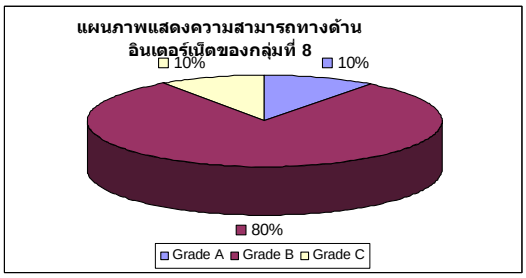


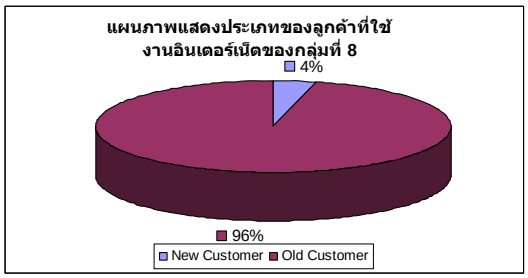
จำนวนลูกค้าทั้งหมดในกลุ่มที่ 7 นี้ มีจำนวนทั้งสิ้น 44 รายการ หรือคิดเป็นร้อยละ 0.69 ของจำนวนข้อมูลทั้งหมด โดยลักษณะทั่วไปของลูกค้ากลุ่มนี้จะใช้อินเตอร์เน็ตที่ความเร็ว 512 - 768 kb และใช้งาน Modem ยี่ห้อ Billion เป็นส่วนใหญ่ (คิดเป็นร้อยละ 77) ร้อยละ 82 ของกลุ่มนี้มีระดับความรู้ความสามารถด้านอินเตอร์เน็ตอยู่ในระดับ Grade B ร้อยละ 70 ใช้อินเตอร์เน็ตในช่วงเวลา 18:00 pm. – 23:59 pm. ระบบปฏิบัติการที่ใช้ในกลุ่มนี้ ร้อยละ 61 ใช้ระบบปฏิบัติการ Windows ME กลุ่มลูกค้าในกลุ่มส่วนใหญ่เป็นกลุ่มลูกค้าที่ใช้งานอินเตอร์เน็ตมากกว่า 3 เดือน (คิดเป็นร้อยละ 93)

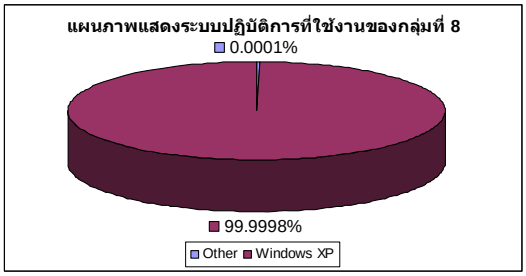
5.8 ลักษณะพฤติกรรมการใช้งานของลูกค้าบุคคลกลุ่มที่ 8

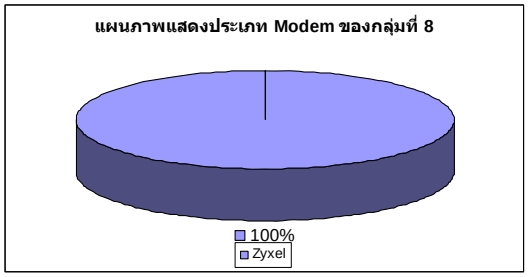












จำนวนลูกค้าทั้งหมดในกลุ่มที่ 8 นี้ มีจำนวนทั้งสิ้น 678 รายการ หรือคิดเป็นร้อยละ 10.6 ของจำนวนข้อมูลทั้งหมด โดยลักษณะทั่วไปของลูกค้ากลุ่มนี้จะใช้งานอินเตอร์เน็ตที่ความเร็ว 512 – 768 Kb โดยมีความรู้ความสามารถด้านอินเตอร์เน็ตในระดับ Grade B เป็นส่วนใหญ่ (คิดเป็นร้อยละ 80) ร้อยละ 96 ของลูกค้าเป็นลูกค้าเดิมที่ใช้งานอินเตอร์เน็ตนานมากกว่า 3 เดือน ระบบปฏิบัติการที่ใช้งานเป็น Windows XP โดยลูกค้าทุกคนในกลุ่มนี้ใช้งานอินเตอร์เน็ตในช่วงเวลา 18:00 pm. – 23:59 pm. และใช้ modem ยี่ห้อ Zyxel ทั้งหมด

6. ผลการวัดประสิทธิภาพการค้นคืนสาเหตุการเกิดปัญหากรณีเชื่อมต่ออินเตอร์เน็ตได้ โดยใช้เทคนิค OLAP Cube

การวัดประสิทธิภาพในส่วนนี้จะใช้เทคนิค Precision จากสูตร

$$\text{Precision} = \frac{\text{จำนวนคำตอบที่ถูกค้นคืนถูกต้อง}}{\text{จำนวนข้อมูลทั้งหมดที่ถูกค้นคืน}}$$

จากข้อมูลที่ใช้ทดลองจำนวน 7,110 รายการ โดยแบ่งข้อมูลจำนวน 6,399 รายการ (90 %) สำหรับ Training และอีก 711 รายการ (10 %) สำหรับการทดลอง พบว่ามีความถูกต้องในการค้นคืนถึง 504 รายการ ดังนั้นค่า Precision ที่วัดได้จากการทดลองนี้เท่ากับ 70.46 % ซึ่งได้ผลลัพธ์ออกมาดี

ตารางที่ 12 ตัวอย่างการวิเคราะห์สาเหตุของอาการผิดปกติโดยใช้เทคนิค OLAP Cube

OLAP Cubes (Group 1)			
Problem	Cause	N	% of Total N
Frequently disconnect	Splitter not install	2	3.39 %
	Many connecting points	38	64.41%
	Virus	4	6.78 %
	Line deteriorated	15	25.42 %
Total		59	100.0%

จากตารางที่ 12 แสดงสาเหตุของอินเตอร์เน็ตหลุดบ่อยของกลุ่มตัวอย่างกลุ่มที่ 1 มีทั้งสิ้น 59 รายการ โดยสาเหตุที่น่าจะเป็นปัญหามากที่สุด คือ มีจุดพ่วงหลายจุด โดยมีทั้งสิ้น 38 รายการ หรือคิดเป็นร้อยละ 64.41

วิจารณ์

งานวิจัยที่นำเสนอนี้ เป็นการนำเทคนิคการทำเหมืองข้อมูลมาประยุกต์ใช้ในการทำนายสาเหตุการเกิดปัญหาการใช้งานอินเตอร์เน็ต โดยได้แบ่งการแก้ปัญหาออกเป็น 2 ส่วน คือ ส่วนคำถามกรณีเชื่อมต่ออินเตอร์เน็ตไม่ได้ และ คำถามเมื่อเชื่อมต่ออินเตอร์เน็ตได้ ซึ่งที่ต้องแบ่งปัญหาออกเป็น 2 ส่วนเพื่อให้ระบบสามารถวิเคราะห์สาเหตุได้ตรงกับปัญหา

ส่วนที่ตอบปัญหาในกรณีที่ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้นั้น การสร้าง Software agent ตรวจสอบค่าต่าง ๆ ของอุปกรณ์ ได้ผลที่แม่นยำถึง 87.56 % เพราะปัญหาส่วนใหญ่ที่เกิดขึ้นที่ทำให้ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้นั้นสาเหตุส่วนมากมาจากการติดตั้งอุปกรณ์หรือการตั้งค่าของอุปกรณ์ไม่สมบูรณ์พร้อมให้บริการ การสร้างตัว Software agent นี้เปรียบเสมือนการสร้างตัวตรวจสอบอุปกรณ์ต่าง ๆ ก่อนว่าพร้อมสำหรับการใช้งานหรือไม่ เมื่อประยุกต์กับงานวิจัยนี้แล้วพบว่าการค้นคืนข้อมูลสาเหตุจะมีความแม่นยำแก้ปัญหาได้รวดและเมื่อเปรียบเทียบกับงานวิจัยของกฤษณะ ไวยมัย และคณะ (2001) ที่ได้ใช้ Decision tree กับระบบพัฒนาคุณภาพการศึกษาที่ได้ผลลัพธ์แม่นยำ 84.58 จะสรุปได้ว่าการประยุกต์ Decision tree กับ เทคนิคการตรวจสอบระบบอัตโนมัติ ทำให้มีผลลัพธ์ที่แม่นยำขึ้น

ส่วนการวิเคราะห์ปัญหากรณีที่สามารถเชื่อมต่ออินเตอร์เน็ตได้นั้น ได้ใช้การจัดกลุ่มเข้ามาช่วย ซึ่งผลของการจัดกลุ่มนั้น ประสิทธิภาพของการใช้ K-means ในการจัดกลุ่มได้ผลลัพธ์ดีที่สุดโดยดูจากค่า ความแตกต่างของข้อมูลภายในกลุ่ม (Root Mean Square Standard Deviation: RMSSTD) ซึ่งค่าความแตกต่างภายในกลุ่มน้อยที่สุดเพียง 0.616 และค่าความแตกต่างของข้อมูลระหว่างกลุ่ม (R Squared: RS) ซึ่งความแตกต่างมีค่ามากถึง 0.697 แสดงว่ามีการแบ่งกลุ่มที่ดีจึงทำให้ความแตกต่างภายในกลุ่มมีค่าน้อย และความแตกต่างระหว่างกลุ่มมีค่าสูง

ผลที่ได้จากขั้นตอนการวิเคราะห์ปัญหากรณีที่เชื่อมต่อไม่ได้ (87.56 %) มีผลลัพธ์ดีกว่าขั้นตอนการวิเคราะห์ปัญหากรณีที่เชื่อมต่อได้ (70.46 %) นั้นสาเหตุเพราะการสร้าง Software Agent ตรวจสอบอุปกรณ์ต่าง ๆ โดยตรงและบอกถึงค่าที่ผิดปกติจึงมีความแม่นยำมาก ส่วนการวิเคราะห์ปัญหาเชื่อมต่อได้นั้นจะเป็นการพยากรณ์คาดเดาถึงสาเหตุของปัญหาที่เกิดขึ้นให้ตรงกับลักษณะของลูกค้าแต่เมื่อเทียบกับงานของ Li และคณะ (2007) ที่นำเสนอข้อมูลให้ผู้ใช้ M-Commerce นั้นสามารถวัดความถูกต้องได้ 69 % สำหรับงานวิจัยปัจจุบันมีผลลัพธ์ที่ดีกว่า เนื่องจากในงานวิจัยนี้มี

การดำเนินงานแบบ 2 ขั้นตอนคือ การจัดกลุ่มข้อมูลของผู้ใช้ และการวิเคราะห์โดยใช้ OLAPจัดกลุ่มข้อมูลก่อนนำมาวิเคราะห์ ทำให้ได้ผลลัพธ์ที่ดีกว่า

สรุปและข้อเสนอแนะ

สรุป

งานวิจัยนี้มีจุดประสงค์ในการวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ตของลูกค้า เพื่อค้นคืนสาเหตุที่ทำให้เกิดปัญหา และนำเสนอการแก้ไขได้แม่นยำตรงกับสาเหตุที่เกิดขึ้นจริง อันดับแรกลูกค้าถามปัญหาเข้ามาในระบบ ต่อมาฟังก์ชั่นตัดคำจะดำเนินการสกัดคำสำคัญจากคำถามของลูกค้า หลังจากนั้นมีการตรวจสอบว่าคำถามนั้นเกิดจากการทำงานประเภทใด เมื่อวิเคราะห์ถึงปัญหาที่เกิดขึ้นกับการใช้งานอินเตอร์เน็ต พบว่าปัญหาที่ถามเข้ามาแบ่งเป็น สองชนิด คือ การสอบถามกรณีที่ไม่สามารถเชื่อมต่ออินเตอร์เน็ตได้ (offline) และกรณีที่เชื่อมต่อได้แล้วพบปัญหา (online)

การวิเคราะห์คำถามกรณี Offline ทำได้โดยใช้ Software agent ที่มีอยู่ในระบบทำงานตรวจสอบอุปกรณ์ที่ใช้ในการเชื่อมโยงกับระบบเครือข่ายอินเตอร์เน็ตว่าสามารถทำงานได้ปกติหรือไม่ หลังจากนั้นนำข้อมูลของปัญหาที่ได้ตรวจสอบกับ Decision tree ที่เก็บข้อมูลปัญหาการใช้งานของระบบอินเตอร์เน็ตของลูกค้าในอดีต และ ระบบรายงานสาเหตุของปัญหาที่เกิดขึ้นให้ลูกค้า ผลที่ได้จากการทดลอง ใช้ข้อมูลทดสอบ 10,000 รายการ พบว่าการทำงานของ Software agent มีความถูกต้อง 87.56 % ซึ่งผลลัพธ์น่าพอใจ ส่วนความผิดพลาดในกรณีนี้อาจเกิดจาก Decision tree ไม่ได้เก็บข้อมูลครอบคลุมปัญหาที่เกิดขึ้นทุกกรณีเพราะอุปกรณ์คอมพิวเตอร์ที่มีมากมายหลายประเภทและมีการพัฒนาฮาร์ดแวร์รุ่นใหม่สู่ท้องตลาดอย่างรวดเร็ว กรณีของปัญหาในการทดลอง เช่น การใช้อุปกรณ์เครือข่าย ของคอมพิวเตอร์ที่ผลิตขึ้นในรุ่นแตกต่างกัน อาจได้ผลลัพธ์ในการตรวจสอบความผิดพลาดไม่ตรงกันกัน

การวิเคราะห์คำถามกรณี Onlineใช้เทคนิค OLAP Cube ช่วยในการวิเคราะห์ปัญหาโดยก่อนการวิเคราะห์นั้นจะทำการจัดกลุ่มข้อมูลส่วนตัวลูกค้า โดยใช้คุณลักษณะ ดังนี้ คือ (1) ช่วงเวลาที่ใช้งาน (2) ระบบปฏิบัติการที่ใช้ (3) ประเภทของลูกค้า (4) ความสามารถด้าน IT (5) Modem ที่ใช้งาน และ (6) ความเร็วอินเตอร์เน็ตที่ขอใช้งาน ต่อมานำเทคนิคการจัดกลุ่ม 2 ขั้นตอน (SOM และ K-Means) เพื่อจัดกลุ่มลักษณะของลูกค้าที่มีคุณสมบัติเหมือนกันไว้กลุ่มเดียวกัน การวิเคราะห์ด้วยเทคนิค OLAP Cube มีคุณลักษณะของข้อมูลที่ใช้วิเคราะห์คือ (1) ข้อมูลลูกค้าที่ได้จากการจัดกลุ่มแบบ 2 ขั้นตอน, (2) ปัญหาที่เกี่ยวข้องกับการใช้งาน และ (3) สาเหตุที่ทำให้เกิดปัญหา ผลลัพธ์ของการทดลองนี้มีความแม่นยำในการค้นคืนข้อมูลของสาเหตุได้ถูกต้องร้อยละ 70.46 และ

การทดลองครั้งนี้จำนวนข้อมูลที่นำมาใช้มีเพียง 7,110 รายการถ้ามีการเพิ่มขนาดข้อมูลให้มีจำนวนมากขึ้นอาจจะมีความเที่ยงตรงสูงขึ้น

ข้อเสนอแนะ

การแก้ปัญหากรณีวิเคราะห์สาเหตุจากกลุ่มลูกค้านั้น ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลที่ได้จากระบบเมื่อนำมาใช้ในการแบ่งกลุ่มอาจมีน้อยเกินไป ดังนั้นการเพิ่มประสิทธิ์ภาพของงานวิจัยนี้คือ (1) เพิ่มปริมาณปัญหาการใช้งานอินเทอร์เน็ตที่นำมาใช้ในฐานข้อมูลของระบบ เมื่อใช้ข้อมูลปริมาณมากขึ้นทำให้ระบบมีโอกาสทำนายสาเหตุของปัญหาได้ถูกต้องมากขึ้น (2) เพิ่มคุณลักษณะในการแบ่งกลุ่ม เช่นปัจจัยภายนอกที่มีผลต่อปัญหาการใช้งานอินเตอร์เน็ตอีก ได้แก่ ฤดู หรือ พื้นที่ใช้งาน เพราะบางครั้งฤดูกาลมีผลต่อปัญหาที่ตามมา เช่นในฤดูฝนจะมีปัญหาอินเตอร์เน็ตหลุดบ่อยตามมาเพราะสาเหตุสายที่ใช้งานอินเตอร์เน็ตชำรุดเนื่องจากพายุฝนฟ้าคะนอง หรือพื้นที่ใช้บริการในพื้นที่นั้นอาจจะมีการปรับปรุงบ่อยครั้ง ดังนั้นการเพิ่มตัวแปรที่ใช้ในการจัดกลุ่มจะสามารถเพิ่มประสิทธิภาพในการจัดกลุ่มให้เด่นชัดมากยิ่งขึ้น ผลการทำนายสาเหตุของปัญหาช่วยให้แม่นยำขึ้น นอกจากการเพิ่มตัวแปรในการจัดกลุ่มข้อมูลแล้วการวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ตนั้นอาจใช้ข้อมูลปัญหาในอดีตมาช่วยพิจารณาเช่น อาจดูความน่าจะเป็นของปัญหาว่าวิธีแก้ไขหรือปัญหาที่เคยเกิดในอดีตมีผลส่งถึงปัญหาที่พบในปัจจุบันหรือไม่ เป็นต้น จะได้ช่วยวิเคราะห์สาเหตุได้แม่นยำมากขึ้น

เอกสารและสิ่งอ้างอิง

กัลยา วานิชย์บัญชา. 2546. **การวิเคราะห์สถิติขั้นสูงสุดด้วย SPSSโปรแกรมทางสถิติ for Windows.** ครั้งที่ 3. โรงพิมพ์จุฬาลงกรณ์มหาวิทยาลัย.

กฤษณะ ไวยมัย, ชิดชนก ส่งศิริ และ ธนาวินท์ รักธรรมานนท์. 2001. การใช้เทคนิคดาต้าไมน์นิงเพื่อพัฒนาคุณภาพการศึกษาคณะวิศวกรรมศาสตร์, น. 134-142. *ใน* **NECTEC Technical Journal (NTJ) 2. ed.**

ไพศาล เจริญพรสวัสดิ์. **Swath.** http://www.links.nectec.or.th/~yai/. 12 สิงหาคม 2548.

นภดล หมื่นไพ และ อัศนีย์ ก่อตระกูล. 2003. เทคนิคการประมวลผลภาษาสำหรับระบบบริหารจัดการความรู้ในองค์กรอิเล็กทรอนิกส์, *ใน* **National Conference on Electronic Business (NCEB) 2. ed.** Bangkok.

Bradley, P. S. and U. M. Fayyad. 1998. Refining initial points for K-Means clustering, p. 91–99. *In* **Proceedings of the Fifteenth International Conference on Machine Learning.** San Francisco, CA.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. **Knowledge Acquisition archive** 5 (2): 199 - 220.

Jain, A.K. and R.C. Dubes. 1988. **Algorithms for Clustering Data.** Jain, A. K. and Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Li, Q., C. Wang, G. Geng and R. Dai. 2007. A Novel Collaborative Filtering-Based Framework for Personalized Services in M-Commerce, pp. 1251-1252. *In* **WWW 2007** . Banff, Alberta, Canada.

Miller, G.A., R. Beckwith and C. Fellbaum. 1990. Introduction to WordNet: an on-line lexical database. *In* **International Journal of Lexicography** 3 (4): 235 - 244.

Niyagas, W., A. Srivihok and S. Kitisin. 2006. Clustering e-Banking Customer using Data Mining and Marketing Segmentation, *In* **Proceedings of the 2006 Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI) International Conference** . Ubon Ratchatani, Thailand.

Ragngam, N. and A. Srivihok. 2005. Question analysis and answering system for the internet usages by using Ontology and Concept Weighting, *In* **Proceedings of The Fifth International Conference on Electronic Business ICEB2005** . Hong Kong.

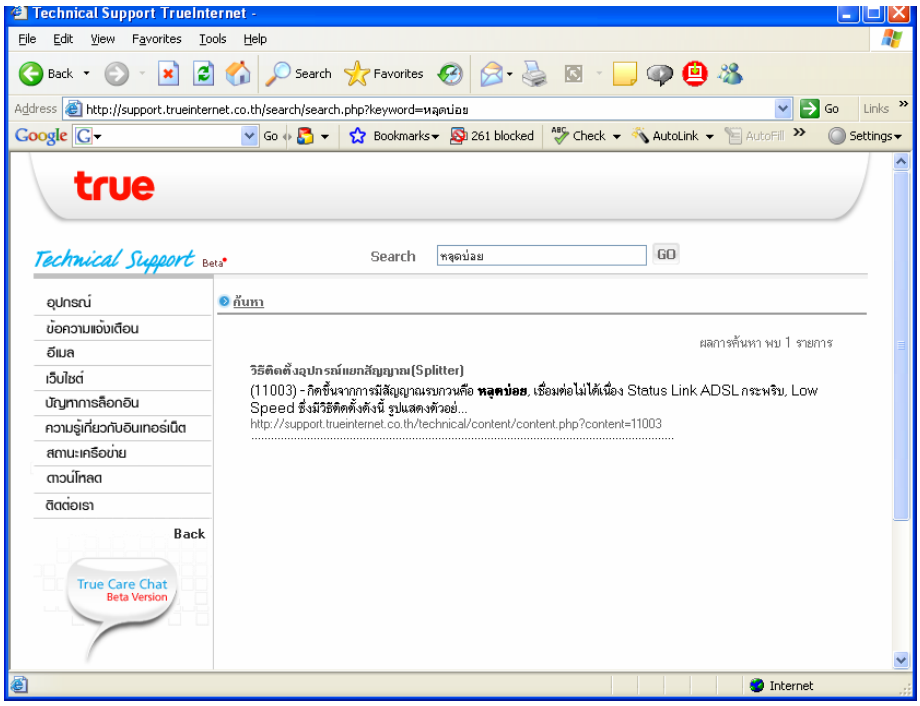
Ward, J.H. 1963. Hierarchical Grouping to Optimize an Objective Function. **Journal of the American Statistical Association** 58: 236-244.

Weng, S.S. and M.J. Liu. 2004. Feature-based recommendations for one-to-one marketing. **Expert Systems with Applications** 26 (4): 493-508.

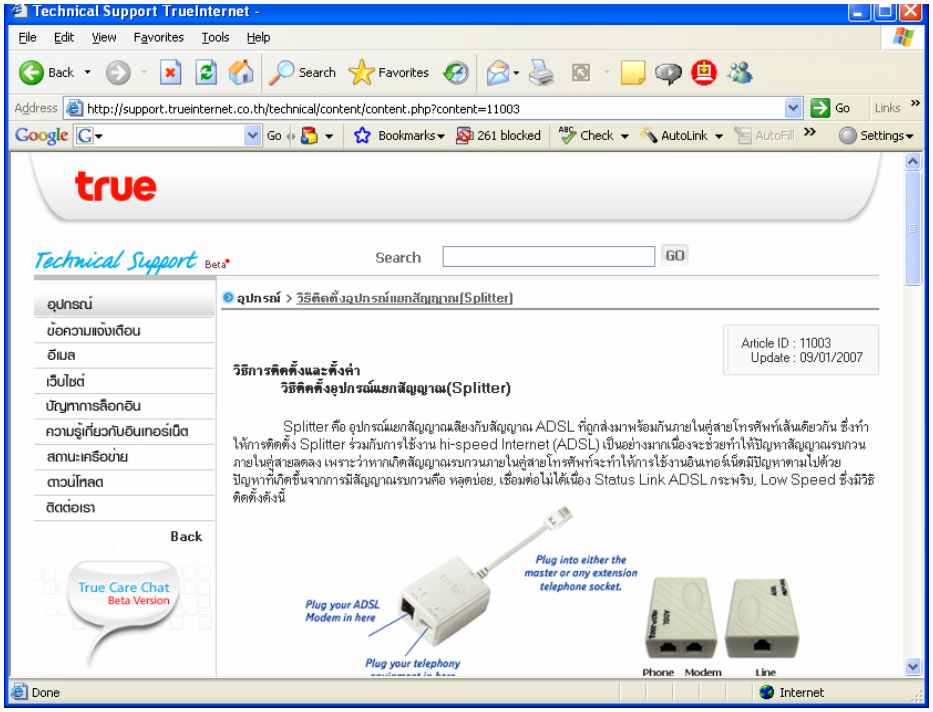
Yeh, J. F., M. J. Chen and C. H. Wu. 2003. Semantic inference based on ontology for medical FAQ mining, pp. 710- 715. *In* **International Conference on Natural Language Processing and Knowledge Engineering** . Taiwan.

Zhang, T., R. Ramakrishnan and M. Livny. 1996. BIRCH: An efficient data clustering method for very large databases, pp. 103–114. *In* **Proceedings of ACM SIGMOD Conference**. Montreal, Canada.

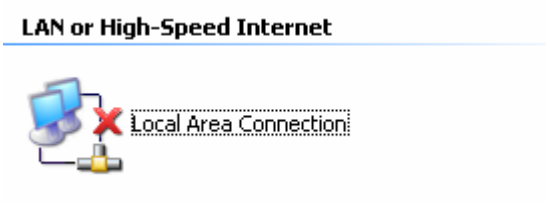
ภาคผนวก



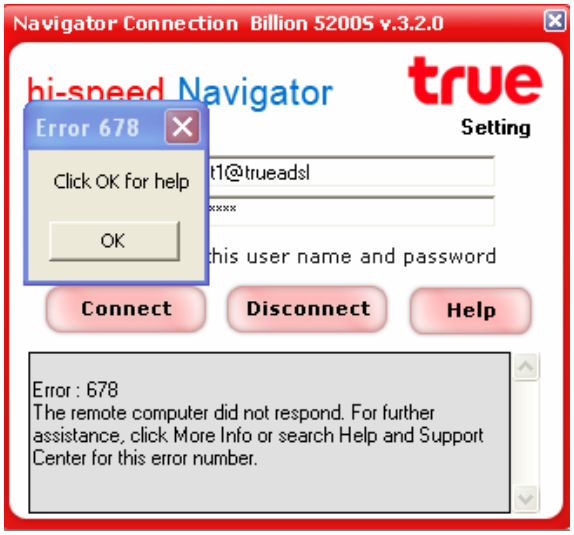
ภาพผนวกที่ 1 ตัวอย่างการวิเคราะห์สาเหตุและการแก้ปัญหากรณี Online (อินเตอร์หลุดบ่อย)



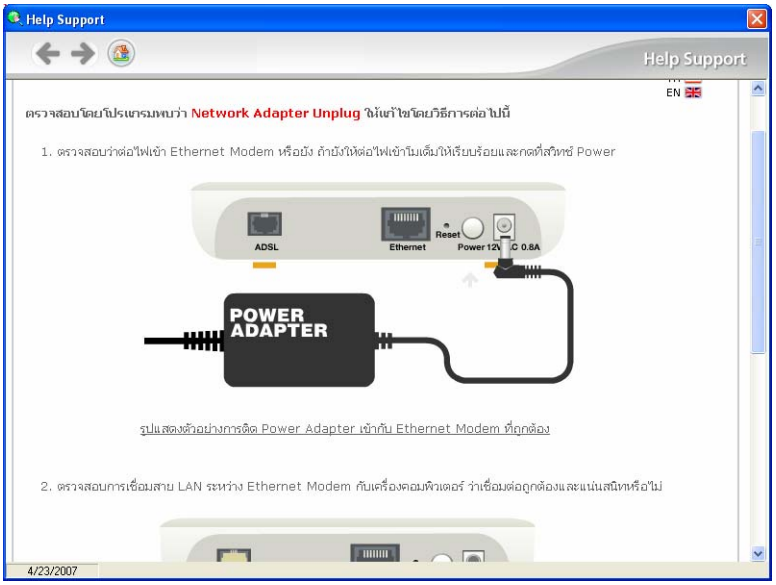
ภาพผนวกที่ 2 แสดงการนำเสนอวิธีการแก้ปัญหากรณี Online (การติดตั้ง Splitter)



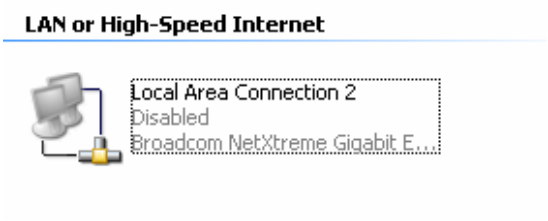
ภาพผนวกที่ 3 ตัวอย่างสาเหตุของอาการผิดปกติ (สาย LAN ไม่ได้เสียบ)



ภาพผนวกที่ 4 ตัวอย่างแสดงการวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ต (Error 678) โดยใช้ Software agent

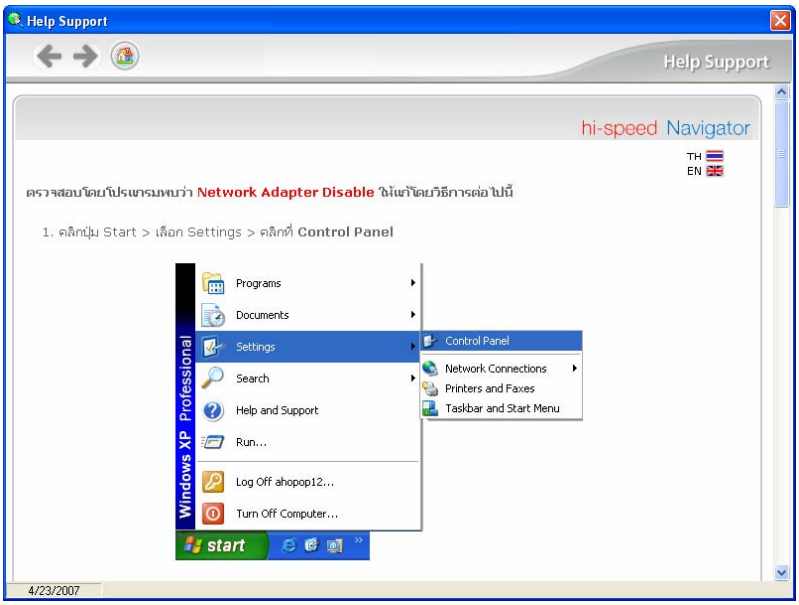


ภาพผนวกที่ 5 ตัวอย่างแสดงการนำเสนอวิธีแก้ปัญหากรณี Offline (สาย LAN ไม่ได้เสียบ)



ภาพผนวกที่ 6 ตัวอย่างสาเหตุของอาการผิดปกติ (LAN Disable)

ภาพผนวกที่ 7 ตัวอย่างแสดงการวิเคราะห์ปัญหาการใช้งานอินเตอร์เน็ต (Error 769) โดยใช้ Software agent



ภาพผนวกที่ 8 ตัวอย่างแสดงการนำเสนอวิธีแก้ปัญหากรณี Offline (LAN Disable)

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายนฤป รักงาม
วัน เดือน ปี ที่เกิด	วันที่ 3 กรกฎาคม 2525
สถานที่เกิด	ยะลา
ประวัติการศึกษา	วท.บ. (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยกรุงเทพ (พ.ศ. 2547)
ตำแหน่งหน้าที่การงานปัจจุบัน	System Analysis
สถานที่ทำงานปัจจุบัน	True corporation
ผลงานดีเด่นและรางวัลทางวิชาการ	Question analysis and answering system for the internet usages by using Ontology and Concept Weighting (ICEB 2005) Hong Kong 5-9 December 2005