

บทที่ 1

บทนำ

1.1 ปัญหาและที่มาของการศึกษา

ปัจจุบันการดำเนินธุรกิจไม่ว่าจะเป็นธุรกิจขนาดเล็กหรือว่าขนาดใหญ่ ต่างก็ให้ความสำคัญกับการจัดเก็บข้อมูลลงในฐานข้อมูล เพื่อใช้ในการวิเคราะห์ข้อมูลการดำเนินงานเกี่ยวกับภาคธุรกิจในอนาคต ซึ่งการสืบค้นความรู้ที่เป็นประโยชน์และน่าสนใจบนฐานข้อมูลขนาดใหญ่ (Knowledge Discovery from very large Databases หรือ KDD) หรือที่เรียกกันว่าการทำเหมืองข้อมูล (Data Mining) เป็นเทคนิคที่ใช้จัดการกับข้อมูลขนาดใหญ่ โดยจะนำข้อมูลที่มีอยู่มาวิเคราะห์แล้วนำความรู้เหล่านั้นออกมาใช้ในการวิเคราะห์หรือทำนายสิ่งต่าง ๆ ที่อาจจะเกิดขึ้น

โดยการสืบค้นความรู้นี้ สามารถวิเคราะห์ข้อมูลขนาดใหญ่ออกมาเป็นข้อมูลสรุปหรือความรู้ที่สำคัญซึ่งเป็นคุณสมบัติหรือลักษณะเฉพาะของข้อมูลในฐานข้อมูล (Descriptive Mining Tasks) เช่น การสืบค้นกฎความสัมพันธ์ (Association Rules Mining) การทำเหมืองข้อมูลรูปแบบลำดับ (Sequential Patterns Mining) เป็นต้น [1]

การสืบค้นกฎความสัมพันธ์ เป็นการหาความสัมพันธ์ระหว่างข้อมูลภายในรายการเปลี่ยนแปลง (Transaction) เดียวกัน โดยมักจะถูกนำไปประยุกต์ใช้สำหรับการวิเคราะห์การซื้อสินค้าของลูกค้า (Market Basket Analysis) กฎที่ได้แสดงว่าถ้าหากพบข้อมูลสินค้าชุดใดแล้ว มีแนวโน้มที่จะพบข้อมูลสินค้าชุดใดด้วย ซึ่งความเป็นไปได้ที่จะเกิดข้อมูลสินค้าทั้งสองชุดในรายการเปลี่ยนแปลงเดียวกันเรียกว่า ค่าสนับสนุน (Support) และสามารถมั่นใจว่ากฎจะเป็นจริงได้ด้วยค่าความเชื่อมั่น (Confidence) เช่น ถ้าสามารถสืบค้นกฎความสัมพันธ์จากฐานข้อมูลการซื้อสินค้าเป็น “{ขนมปัง, เนยสด} $\rightarrow$ {นมสด} ค่าสนับสนุนเท่ากับ 50 และค่าความเชื่อมั่น (Confidence) เท่ากับ 80%” หมายความว่า จะมีการซื้อขนมปัง, เนยสด และ นมสด พร้อมกัน 50 รายการเปลี่ยนแปลง (Transaction) และ 80 เปอร์เซ็นต์ของลูกค้าที่ซื้อขนมปังพร้อมกับเนยสด จะซื้อนมสด ไปด้วย

แต่ถ้าต้องการหาความสัมพันธ์ระหว่างข้อมูลโดยสนใจลำดับการเกิดของข้อมูลด้วยสามารถใช้เทคนิคการสืบค้นรูปแบบลำดับเหตุการณ์ได้ โดยเทคนิคนี้เป็นความสัมพันธ์ของข้อมูลระหว่างรายการเปลี่ยนแปลง ซึ่งทำให้มีเวลาเข้ามาเกี่ยวข้องด้วย รูปแบบลำดับ (Sequence Patterns) ที่ได้จะแสดงว่า ถ้าเกิดเหตุการณ์อะไรหรือพบกลุ่มของข้อมูลใดตามมาภายหลัง ความเป็นไปได้ที่จะเกิดรูปแบบลำดับนั้นๆ มีค่าเท่ากับค่าสนับสนุน (Support) เช่น ถ้ารูปแบบลำดับที่ได้เป็น “<(เบียร์, ผ้าอ้อมเด็ก) (นมสด)> ค่าสนับสนุนเท่ากับ 75” หมายความว่า ลูกค้าจำนวน 75 คนซื้อเบียร์ไปพร้อมกับผ้าอ้อมเด็ก ต่อมาภายหลังจะกลับมาซื้อนมสด โดยที่ลูกค้าเหล่านั้นอาจจะซื้อสินค้าอื่นๆ ก่อนนมสดก็ได้เพียงแต่รูปแบบลำดับนั้นอาจมีค่าสนับสนุนขั้นต่ำน้อยกว่าที่กำหนดไว้ โดยเทคนิคการสืบค้นรูปแบบลำดับ สามารถประยุกต์ใช้ในหลายด้าน เช่น การวิเคราะห์พฤติกรรม การซื้อสินค้า การวิเคราะห์ลำดับดีเอ็นเอ (Deoxyribonucleic Acid หรือ DNA) การวิเคราะห์ลำดับการเข้าไปยังหน้าต่างๆ ในเว็บไซต์ เป็นต้น

การสืบค้นรูปแบบลำดับเป็นงานที่ท้าทาย เนื่องจากข้อมูลขนาดใหญ่ในฐานข้อมูล ทำให้มีรูปแบบลำดับที่เป็นไปได้จำนวนมาก การหารูปแบบลำดับที่น่าสนใจคือ ค่าสนับสนุน (Support) มากกว่าหรือเท่ากับค่าสนับสนุนขั้นต่ำ (Minimum support) จากรูปแบบลำดับที่เป็นไปได้ทั้งหมดมีความซับซ้อนเป็น $O(nm)$ เมื่อ n คือจำนวนของรายการเปลี่ยนแปลง และ m คือความยาวของรูปแบบลำดับ ถ้าฐานข้อมูลมีข้อมูลขนาดใหญ่อาจทำให้ต้องใช้เวลาและเนื้อที่ของหน่วยความจำในคอมพิวเตอร์มหาศาลในการประมวลผล[1][2]

ขั้นตอนวิธี (Algorithm) กลุ่มแรกที่ถูกเสนอขึ้นเพื่อแก้ปัญหาการสืบค้นรูปแบบลำดับคือ AprioriSome และ AprioriAll ซึ่งถูกพัฒนามาจากขั้นตอนวิธี Apriori [62] ซึ่งใช้ในการหาความสัมพันธ์โดยที่ขั้นตอนวิธี AprioriAll ทำงานได้ดีที่สุดสำหรับกรณีทั่วไป ขั้นตอนวิธีนี้ได้ถูกพัฒนาต่อเป็นขั้นตอนวิธี GSP ซึ่งทำงานได้เร็วกว่า ขั้นตอนวิธี AprioriAll โดยขั้นตอนวิธี GSP จะมีการพิจารณาค่าควบคุมของเวลา คือ ค่าความแตกต่างของเวลาสูงสุด (Maxgap) ระหว่างรายการเปลี่ยนแปลง และค่าความแตกต่างของเวลาดำสุด (Mingap) ระหว่างรายการเปลี่ยนแปลง พร้อมทั้งพิจารณาอนุกรมวิธาน (Taxonomies) เพื่อให้ได้รูปแบบลำดับที่มากขึ้น [3]

ดังที่กล่าวไปข้างต้นเทคนิคการสืบค้นรูปแบบลำดับ เป็นเทคนิคที่ใช้สืบค้นหาความสัมพันธ์ของข้อมูลระหว่างรายการเปลี่ยนแปลงซึ่งทำให้มีเวลาเข้ามาเกี่ยวข้อง และเนื่องจาก

ปัจจุบัน การจัดเก็บข้อมูลของธุรกิจบางประเภท ถูกจัดเก็บในรูปแบบของฐานข้อมูลที่เก็บลำดับค่าคุณลักษณะที่แปรเปลี่ยนไปตามเวลา (Temporal Database) ซึ่งจะช่วยในการค้นพบลักษณะวิวัฒนาการ หรือแนวโน้มการเปลี่ยนแปลงของสิ่งที่สนใจจากข้อมูลที่จัดเก็บอยู่ในฐานข้อมูล หรืออาจนำความรู้ที่ได้ไปช่วยวางแผนการดำเนินงานทางธุรกิจได้

ตารางที่ 1.1 ข้อมูลประวัติการซื้อสินค้าของลูกค้า

Customer ID	Transaction Time	Items Bought
1	2011-06-25	30
1	2011-06-30	90
2	2011-06-14	10,20
2	2011-06-15	10,20
2	2011-06-16	10,20
2	2011-06-17	10,20
2	2011-06-20	30
3	2011-06-25	30,50,70
4	2011-06-25	30
4	2011-06-30	40,70
4	2011-07-25	90
5	2011-06-25	90

จากตัวอย่างของฐานข้อมูลเชิงเวลา ในตารางที่ 1 ที่แสดงข้อมูลเกี่ยวกับประวัติการซื้อสินค้าของลูกค้า ซึ่งหากต้องการสอบถาม (Query) ข้อมูลเพื่อนำมาใช้งาน ตัวอย่างเช่น หากผู้ใช้ต้องการสอบถามข้อมูลการซื้อสินค้าของลูกค้า รหัส “2” โดยใช้ภาษาสำหรับการสอบถามเอสคิวเควล (SQL Query) เพื่อสอบถามข้อมูล จะใช้คำสั่งต่อไปนี้

```

SELECT Customer_id, Transaction_time, Items_Bought
FROM ITEMS_BOUGHT
WHERE EMPNO = 2
GROUP BY Customer_id, Transaction_time, Item_Bought

```

โดยผลลัพธ์ที่ได้จากสอบถามข้อมูล จากคำสั่งข้างต้นจะทำให้ผู้ใช้งานได้ข้อมูลดัง
ตารางที่ 2

ตารางที่ 1.2 ผลลัพธ์จากการสอบถามข้อมูลการซื้อสินค้าของลูกค้า รหัส 2

Customer ID	Transaction Time	Items Bought
2	2011-06-14	10,20
2	2011-06-15	10,20
2	2011-06-16	10,20
2	2011-06-17	10,20
2	2011-06-20	30

ซึ่งในมุมมองของผู้ใช้งาน ผลลัพธ์ของข้อมูลที่ได้จากการสอบถามดังตารางที่ 2 ทำให้ผู้
งานเกิดความสับสนและยากต่อการนำใช้งาน ซึ่งโดยปกติข้อมูลที่ได้จากการสอบถามบนฐานข้อมูล
เชิงเวลาควรมีการรวมตัว (Coalescing) ของแถวข้อมูลด้วย เพื่อให้ผู้ใช้งานได้ข้อมูลที่ง่ายต่อ การใช้
งาน

ทั้งนี้สำหรับการสอบถามข้อมูลการซื้อสินค้าของลูกค้า รหัส “2” นั้น ควรมีผลลัพธ์ดัง
แสดงในตารางที่ 3 ซึ่งจะเห็นได้ว่าการรวมตัวของแถวข้อมูลตามช่วงเวลา บนแอตทริบิวต์
(Attribute) CustomerID และ ItemsBought เมื่อ CustomerID เป็นคีย์หลัก (Primary key) ของ
ฐานข้อมูลเชิงเวลา และแอตทริบิวต์ SALARY ซึ่งเป็นแอตทริบิวต์ที่อาจมีค่าแตกต่างกันในเวลา
ที่ต่างกัน และสามารถสังเกตได้ว่าช่วงเวลาที่ลูกค้าคนดังกล่าวซื้อสินค้ารหัสเดียวกัน จะถูก

รวมกลุ่มกันเป็นค่าเดียว เช่น การซื้อสินค้ารหัส “10,20” ซึ่งอยู่ในช่วงเวลาตั้งแต่ 2011-06-14 – 2011-06-17

ตารางที่ 1.3 ผลลัพธ์การสอบถามข้อมูลการซื้อสินค้าของลูกค้า รหัส 2 ในกรณีมีการรวมตัวของแถวข้อมูล

Customer ID	TStart	TEnd	Items Bought
2	2011-06-14	2011-06-17	10,20
2	2011-06-20	2011-06-20	30

จากการสอบถามข้อมูลที่ผ่านมา หากต้องการสอบถามข้อมูลผ่านภาษาสำหรับการสอบถามเอสคิวแอล เพื่อให้ได้ข้อมูลดังตารางที่ 3 สามารถสอบถามข้อมูลผ่านภาษาสำหรับการสอบถามเอสคิวแอล จากคำสั่งดังต่อไปนี้

```
WITH Temp(ITEMS_BOUGHT, TRANSACTION_TIME AS TSTART
, TRANSACTION_TIME AS TEND) AS
(SELECT SALARY, TSTART, TEND FROM ITEMS_BOUGHT WHERE CUSTOMER
= 2 )
SELECT DISTINCT F.ITEMS_BOUGHT, F. TRANSACTION_TIME AS TSTART
, F. TRANSACTION_TIME AS TEND FROM Temp AS F, Temp AS L
WHERE F. TSTART < L. TEND AND F.Items_bought =L.Items_bought
AND NOT EXISTS
(SELECT * FROM Temp AS M WHERE M.Items_bought = F.Item_bought
AND F. TRANSACTION_TIME < M. TRANSACTION_TIME AND M.TSTART
< L.TEND
AND NOT EXISTS
```

```

(SELECT * FROM Temp AS T1
WHERE T1. Item_bought = F. Item_bought
AND T1. TSTART < M. TSTART AND M.TSTART <= T1.TEND)
)
AND NOT EXISTS
(SELECT * FROM Temp AS T2
WHERE T2. Item_bought = F. Item_bought
AND
((T2. TSTART < F. TSTART AND F.TSTART <= T2.TEND)
OR
(T2.TSTART < L.TEND AND L.TEND < T2.TEND))

```

จากการสอบถามที่ยกตัวอย่างไปแล้วนั้น แสดงให้เห็นถึงความซับซ้อนและความยุ่งยากในการสอบถามข้อมูล จากฐานข้อมูลเชิงเวลาอีกทั้งในการรวมแถวของข้อมูลจะต้องทำประมวลผลแบบรีเคอร์ซีฟ (Recursive) เพื่อให้ได้ผลลัพธ์ที่ถูกต้อง ซึ่งอาจทำให้ต้องใช้เวลานานในการประมวลผล [4]

โดยธรรมชาติของการสืบค้นรูปแบบลำดับบนขั้นตอนวิธี GSP นั้นจะเป็นการสืบค้นรูปแบบลำดับ บนช่วงของเวลาโดยจะมีการพิจารณาค่าควบคุมของเวลา เช่น ค่าความแตกต่างของเวลาสูงสุด หรือค่าความแตกต่างของเวลาขั้นต่ำ เพื่อนำมาใช้ในการดำเนินการสืบค้นรูปแบบลำดับ ดังนั้นการสืบค้นรูปแบบลำดับ ขั้นตอนวิธี GSP จึงเหมาะสำหรับการดำเนินการบนฐานข้อมูลเชิงเวลา จากการศึกษพบว่ายังไม่มีผู้ใด นำการสืบค้นรูปแบบลำดับบนขั้นตอนวิธี GSP มาอิมพลีเมนต์ (Implement) การสืบค้นข้อมูลบนฐานข้อมูลเชิงเวลา

ทางเลือกหนึ่งที่สามารถทำได้ ในการดำเนินการสืบค้นรูปแบบลำดับบนขั้นตอนวิธี GSP จากข้อมูลบนฐานข้อมูลเชิงเวลา คือการใช้ภาษาสำหรับการสอบถามเอสคิวแอล ซึ่งจากที่ได้แสดงไปแล้วนั้น ค่อนข้างยุ่งยากและต้องใช้ทักษะในการเขียนภาษาสำหรับการสอบถามเอสคิวแอล เพื่อให้ได้ผลลัพธ์ของข้อมูลที่ได้จากการสอบถามที่ถูกต้อง

1.2 แนวทางในการแก้ปัญหา

จากปัญหาทั้งหมดที่ได้กล่าวมานั้น ทำให้ผู้ศึกษาเกิดแนวคิดในการดำเนินการสืบค้นรูปแบบลำดับ บนขั้นตอนวิธี GSP บนฐานข้อมูลเชิงเวลา โดยฐานข้อมูลเชิงเวลาจะถูกอิมพลีเมนต์ด้วยฐานข้อมูลเอ็กซ์เอ็มแอล (XML) เนื่องจากโดยธรรมชาติของข้อมูลเอ็กซ์เอ็มแอลมีประสิทธิภาพในการแสดงผลข้อมูลเชิงเวลา และสามารถนำภาษาสำหรับการสอบถามเอ็กซ์คิวรี (XQuery) มาช่วยในการสอบถามข้อมูลได้อย่างมีประสิทธิภาพ [5]

ซึ่งในปัจจุบันมีผู้ผลิตระบบจัดการฐานข้อมูล ที่สามารถรองรับการจัดเก็บข้อมูลในรูปแบบเอ็กซ์เอ็มแอล เช่น IBM DB2 pureXML และสามารถใช้ภาษาสำหรับการสอบถามเอ็กซ์คิวรีสำหรับสอบถามข้อมูลที่อยู่ในรูปแบบของเอ็กซ์เอ็มแอลได้

1.3 สรุปสาระสำคัญจากเอกสารที่เกี่ยวข้อง

Ramakrishnan Srikant และ Rakesh Agrawal [3] ได้พัฒนาขั้นตอนวิธี GSP ซึ่งเป็นการสืบค้นรูปแบบลำดับ โดยได้ทำการประเมินผลจากการใช้ข้อมูลจริงในการสังเคราะห์และแสดงให้เห็นว่าขั้นตอนวิธี GSP สามารถทำงานได้เร็วกว่าขั้นตอนวิธี AprioriAll โดยมีการนำเสนอการพิจารณาค่าควบคุมของเวลา คือ ค่าความแตกต่างของเวลาสูงสุดระหว่างรายการปรับปรุง (Maxgap) และค่าความแตกต่างของเวลาต่ำสุดระหว่างรายการปรับปรุง (Mingap) พร้อมทั้งพิจารณาอนุกรมวิธาน (Taxonomies) ด้วย

Fusheng Wang, Carlo Zaniolo และ Xin Zhou [5] ได้แสดงระบบ ArchIS ซึ่งมีวัตถุประสงค์ในการรองรับการรวมกลุ่มของข้อมูลบนระบบจัดการฐานข้อมูล โครงสร้างของ ArchIS ใช้ 1) เอ็กซ์เอ็มแอลที่สนับสนุนรวมกลุ่มของเวลาสำหรับแสดงประวัติฐานข้อมูล 2) ภาษาสำหรับการสอบถามเอ็กซ์คิวรีเพื่อแสดงประโยชน์ของการสอบถามข้อมูลบนฐานข้อมูลเชิงเวลา 3) เทคนิคการทำดัชนีสำหรับการจัดการข้อมูลในฐานข้อมูลเชิงสัมพันธ์ 4) ภาษาสำหรับการสอบถามเอสคิวแอล/เอ็กซ์เอ็มแอล สำหรับสอบถามข้อมูลบนฐานข้อมูลเอ็กซ์เอ็มแอล ที่มีค่าเท่ากับการสอบถามข้อมูลบนฐานข้อมูลเชิงสัมพันธ์ และได้แสดงเกี่ยวกับประสิทธิภาพในการเรียงลำดับ และการค้นคืน ภายใต้เงื่อนไข และความยุ่งยากของการสอบถามข้อมูลรายการเปลี่ยนแปลงบนฐานข้อมูลเชิงสัมพันธ์ และสามารถรับรองประสิทธิภาพในการจัดเก็บข้อมูลโดยการบีบอัดข้อมูล

1.4 วัตถุประสงค์ของการศึกษา

เพื่อวิเคราะห์ประสิทธิภาพการสืบค้นรูปแบบลำดับโดยขั้นตอนวิธีจีเอสพีบนฐานข้อมูลเชิงเวลาซึ่งอิมพลีเมนต์บนฐานข้อมูลเชิงสัมพันธ์ และฐานข้อมูลเอ็กซ์เอ็มแอล

1.5 ประโยชน์ที่จะได้รับจากการศึกษา

เพื่อเป็นทางเลือกให้นักพัฒนาโปรแกรมสามารถนำผลที่ได้จากการทดลองไปประยุกต์ใช้ในเลือกอิมพลีเมนต์การสืบค้นรูปแบบลำดับ หรือนำไปใช้ในการอิมพลีเมนต์การประมวลผลข้อมูลบนฐานข้อมูลเชิงเวลา ในงานลักษณะอื่นที่คล้ายคลึงกัน

1.6 ขอบเขตของการวิจัย

1. ฐานข้อมูลเอ็กซ์เอ็มแอลใช้ระบบ IBM DB2 pureXML
2. ใช้ข้อมูลสังเคราะห์ในการวิเคราะห์ประสิทธิภาพ
3. วิเคราะห์ประสิทธิภาพของขั้นตอนวิธีโดยใช้เวลาในการประมวลผล

1.7 ระเบียบวิธีการทำวิจัย

1. ศึกษางานวิจัย รวบรวมข้อมูลที่เกี่ยวข้อง
2. ศึกษาทฤษฎี การสืบค้นรูปแบบลำดับ ขั้นตอนวิธี GSP
3. กำหนดกรณีศึกษาที่ใช้ในการสอบถามที่มีความซับซ้อนเพื่อใช้ในการทดลอง
4. พัฒนาโปรแกรมที่ใช้ในการทดลองการทำการสืบค้นรูปแบบลำดับ โดยใช้ภาษาสำหรับการสอบถามเอสคิวแอล บนระบบจัดการฐานข้อมูลและ การใช้ภาษาสำหรับการสอบถามเอ็กซ์ คิวรี บนฐานข้อมูลเอ็กซ์เอ็มแอล
5. ตรวจสอบผลลัพธ์ที่ได้ และเปรียบเทียบความถูกต้องในการทำงาน และความเร็วในการประมวลผลข้อมูล ที่ได้จากการทดลอง

1.8 วิธีการทำวิจัย

1. ศึกษางานวิจัย รวบรวมข้อมูลที่เกี่ยวข้อง
2. ศึกษาทฤษฎี การสืบค้นรูปแบบลำดับ –ขั้นตอนวิธีGSP
3. กำหนดกรณีศึกษาใช้ในการสอบถามที่มีความซับซ้อนเพื่อใช้ในการทดลอง
4. พัฒนาโปรแกรมที่ใช้ในการทดลองการทำการสืบค้นรูปแบบลำดับเหตุการณ์ โดยใช้ภาษาสำหรับการสอบถามเอสคิวแอลบนระบบจัดการฐานข้อมูล และการสอบถามข้อมูลโดยใช้ภาษาสำหรับการสอบถามข้อมูลเอ็กซ์คิวรี บนฐานข้อมูลเอ็กซ์เอ็มแอล
5. ตรวจสอบผลลัพธ์ที่ได้ และเปรียบเทียบความถูกต้องในการทำงาน และความเร็วในการประมวลผลข้อมูล ที่ได้จากการทดลอง
6. วิเคราะห์ วิจารณ์ สรุปผลการวิจัย
7. จัดทำและเสนอรายงานวิทยานิพนธ์

1.9 สถานที่ทำการวิจัย

ห้องวิจัยวิศวกรรมข้อมูลและเทคโนโลยีเครือข่าย ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเชียงใหม่

1.10 เครื่องมือในการพัฒนา

ในวิทยานิพนธ์นี้การพัฒนาขั้นตอนวิธีทั้งหมดจะใช้เครื่องมือในการพัฒนาดังนี้

1. ฮาร์ดแวร์
 - 2.4 Ghz Intel core i5
 - 4 Gigabytes main memory
 - 500 GB SATA hard drive
2. ซอฟต์แวร์
 - Microsoft Windows 7
 - Microsoft Visual C#.Net 2008
 - DB2 Express C
 - IBM Quest Market-Basket Synthetic Data Generator