สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง



T148615

รายงานสหกิจศึกษาฉบับสมบูรณ์

การวิเคราะห์ข้อมูลและสำรวจเชิงลึกเพื่อหาพฤติกรรมการใช้งานโมบายแอปพลิเคชัน

Data Analysis & Pattern Exploration of Mobile Application Usage

นางสาวปิยาภรณ์ เสาให้สกุณีย์

วท.
ว162ป
2559
สขหมู่
เลขทะเบียน 148615
วัน.เดือน.ปี 6 พ.ย. 2560

b 10897266O
1

ภาควิชาวิศวกรรมคอมพิวเตอร์ สาขาวิชาวิศวกรรมสารสนเทศ
คณะวิศวกรรมศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา 2559

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชื่อโครงงานสหกิจศึกษา การวิเคราะห์ข้อมูลและสำรวจเชิงลึกเพื่อหาพฤติกรรมการใช้งานโมบายแอปพลิเคชัน

ชื่อ-สกุล นักศึกษา นางสาวปิยาภรณ์ เสาให้สกุณีย์

คณะ วิศวกรรมศาสตร์ **ภาควิชา** วิศวกรรมคอมพิวเตอร์ **สาขาวิชา** วิศวกรรมสารสนเทศ

ชื่อ-สกุล อาจารย์นิเทศ ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์

ชื่อ-สกุล ผู้นิเทศงาน 1.คุณธนโชค เมฆาวงศ์ธนา 2.คุณนพวงษ์ โพธิ์เตียน

สถานประกอบการ บริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน)

บทคัดย่อ

องค์กร และบริษัทต่าง ๆ ดำเนินธุรกิจภายใต้สภาวะการแข่งขัน และเผชิญกับสถานการณ์ที่ต้องตัดสินใจว่าจะยังคงใช้แผนงานธุรกิจแบบเดิม ๆ ที่มีอยู่ หรือจะหาแผนงานธุรกิจแบบใหม่ๆ เพื่อตอบสนองกับความต้องการของลูกค้าในปัจจุบัน ซึ่งการตัดสินใจต่าง ๆ เหล่านี้นั้นขึ้นกับปัจจัยอื่น ๆ มากมายหลายประการ เช่น การเปลี่ยนแปลงทางเทคโนโลยีที่รวดเร็ว และการคาดหวังของลูกค้า รวมถึงการคาดการณ์ที่แม่นยำถึงความต้องการของลูกค้าในอนาคต ซึ่งในโครงงานนี้ได้เล็งเห็นว่า ข้อมูลด้านการบันเทิงเป็นอีกสาขาหนึ่งที่มีผลต่อการแข่งขันในภาคธุรกิจเป็นอย่างมาก จึงได้มีการนำข้อมูลต่าง ๆ ด้านความบันเทิง มา ทำการวิเคราะห์เพื่อเป็นการเพิ่มแนวทางในการเข้าถึง ความสนใจของลูกค้าเกี่ยวกับความบันเทิง เพื่อ นำมาใช้ในการวิเคราะห์แนวโมในอนาคตของการบริหารธุรกิจบันเทิงได้

คำสำคัญ: ฐานข้อมูล เหมืองข้อมูล การวิเคราะห์ข้อมูล โมบายแอปพลิเคชัน แทบโบลว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Co-operative Title:	Data Analysis & Pattern Exploration of Mobile Application Usage
Student Intern Name:	Miss Piyaporn Saohisakunee
Faculty: Engineering	**Department:** Computer Engineering
Program:	Information Engineering
Advisor Name:	Asst.Prof.Dr.Sutheera Puntheeranurak
Mentor Name:	1. Mr. Tanachok Mekawongtana 2. Mr. Noppawong Potian
Company:	Advanced Info Service Public Company Limited (AIS)

ABSTRACT

Firms and organizations, operating in the competitive environment, are constantly faced with the situations whether to continue using existing plans or to replace the new plans with more economical ones in order to accommodate with current and future needs. These decisions will depend on the increasing pressure from rapid changes in technology and customer expectations. Including the constant prediction trends of future customer needs. The primary purpose of entertainment analysis in order to increase enjoyment for look at attention of customers, including what can be trends in future.

Keyword: Database, Data Mining, Data Analysis, Mobile Application, Tableau

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ข้าพเจ้าได้รับผิดชอบและปฏิบัติหน้าที่ในบริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน)ระหว่างวันที่ 8 สิงหาคม ถึงวันที่ 30 พฤศจิกายน พ.ศ.2559 ในโครงการวิชาสหกิจศึกษาที่ทางคณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง และบริษัทฯร่วมมือกันจัดตั้งขึ้นในหัวข้อโครงการ การวิเคราะห์ข้อมูลและสำรวจข้อมูลเชิงลึกด้านความบันเทิง ซึ่งข้าพเจ้าได้รับความรู้ ความเข้าใจและประสบการณ์ในการทำงานที่เป็นประโยชน์อย่างมาก อีกทั้งการดูแลและการช่วยเหลือต่าง ๆ ตลอดเวลาการทำงาน โดยการปฏิบัติงานสหกิจศึกษาในครั้งนี้สำเร็จลุล่วงได้ เพราะมีการชี้แนะและได้รับความร่วมมือจากบุคคลต่าง ๆ ดังต่อไปนี้

พนักงานแผนกวิเคราะห์และสำรวจข้อมูล (Data Analytic & Explorations หรือ DAE)

1. คุณธนโชค เมฆาวงศ์ธนา
2. คุณนพวงษ์ โพธิ์เตียน
3. คุณเอกอนันต์ แสงวัฒนาโรจน์
4. คุณเมธินี เมธาวนิช
5. คุณสมบัติ ยิ้มไพบูลย์
6. คุณมนัญญา วัฒนาชัยยงกุล

และข้าพเจ้าขอขอบคุณอาจารย์ที่ปรึกษา ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์ ที่คอยให้คำแนะนำ คำปรึกษาและคอยรับฟังและช่วยเหลือปัญหาต่าง ๆ ในการทำโครงการครั้งนี้ และท้ายที่สุดข้าพเจ้าขอขอบคุณครอบครัวที่คอยให้กำลังใจที่ดีแก่ข้าพเจ้าเสมอมาทำให้ปริญญานิพนธ์ฉบับนี้ให้สำเร็จลุล่วงไปได้ด้วยดี

ปิยาภรณ์ เสาให้สกุณีย์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

บทคัดย่อภาษาไทย I
บทคัดย่อภาษาอังกฤษ II
กิตติกรรมประกาศ III
สารบัญ IV
สารบัญตาราง VII
สารบัญภาพ VIII
บทที่ 1 บทนำ 1
1.1 ความเป็นมาและความสำคัญ 1
1.2 วัตถุประสงค์ของการปฏิบัติงาน 2
1.3 ขอบเขตของโครงงาน 2
1.4 วิธีการดำเนินงาน 2
1.5 ประโยชน์ที่คาดว่าจะได้รับ 5
1.5.1 ประโยชน์ต่อตนเอง 5
1.5.2 ประโยชน์ต่อองค์กรตนเอง 5
1.6 อุปกรณ์ที่ใช้ในการพัฒนาโครงงาน 6
1.6.1 ฮาร์ดแวร์ 6
1.6.2 ซอฟต์แวร์ 6
1.7 ภาษาที่ใช้ในการทำ 6
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 7
2.1 เหมืองข้อมูล (Data Mining) 7
2.1.1 ประเภทข้อมูลที่สามารถทำเหมืองข้อมูล 7
2.1.2 ลักษณะเฉพาะของข้อมูลที่สามารถทำเหมืองข้อมูล 7
2.2 การจัดกลุ่ม (Clustering) 8
2.2.1 เทคนิคการจัดกลุ่ม 8
2.2.2 ความหมายของการจัดกลุ่ม 8
2.2.3 วัตถุประสงค์ของการจัดกลุ่ม 9
2.3 เทคนิคการจัดกลุ่มแบบเคมีน (K-means Clustering) 9
2.3.1 การจัดกลุ่มแบบเคมีน 9
2.3.2 อัลกอริทึมการจัดกลุ่มโดยเคมีน 10

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

2.4 เทคนิคการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) 11
2.5 หลักเกณฑ์ในการรวมกลุ่ม 12
2.5.1 การจัดกลุ่มโดยการหาค่าระยะห่างเฉลี่ยระหว่างกลุ่ม 12
2.5.2 การจัดกลุ่มโดยใช้ค่าระยะห่างที่น้อยที่สุด 13
2.5.3 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างสั้นสุด 13
2.5.4 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างมากสุด 14
2.5.5 การจัดกลุ่มโดยใช้จุดศูนย์กลาง 14
2.5.6 การจัดกลุ่มโดยใช้ค่ากลาง 15
2.5.7 การจัดกลุ่มโดยหาความเหมาะสมเอง 15
2.6 เทคนิคต่าง ๆ ของการทำเหมืองข้อมูล 15
2.6.1 การหากฎความสัมพันธ์ (Association Rule Discovery) 15
2.6.2 การจำแนกข้อมูล และการทำนาย (Classification & Prediction) 17
2.6.3 การแบ่งส่วน (Segmentation) 22
2.6.4 ดิเวียชันดีเทคชันลิงค์อนาไลซิส (Deviation Detection Link Analysis) 22
2.6.5 ลิงค์อนาไลซิส (Link Analysis) 23
2.7 ข้อมูลขนาดใหญ่ (Big Data) 23
2.7.1 คุณลักษณะ 23
2.7.2 การจัดการข้อมูลขนาดใหญ่โดยรูปแบบการกระจายข้อมูล 24
2.8 สวิตช์ (Switch) 25
บทที่ 3 วิธีดำเนินงานวิจัย 26
3.1 การเตรียมโครงสร้างพื้นฐาน (Infrastructure) 26
3.1.1 การเข้าถึงข้อมูล 27
3.1.2 การเตรียมสภาพแวดล้อมการทำงาน (Environment) 28
3.1.3 การตั้งค่าอีเมล์ (E-mail) 28
3.2 การเตรียมข้อมูลในที่สเตจจิ้งดาต้าเบส (Staging Database) 35
3.3 การเลือกกลุ่มลูกค้า หรือกลุ่มที่สนใจจะนำมาวิเคราะห์ 35
3.4 การเตรียมข้อมูลให้ระบบ (Data Preparation) 35
3.5 การนำข้อมูลที่สนใจมาวิเคราะห์ (Data Exploration) 36

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

3.6 นำเสนอผลลัพธ์ออกมาในรูปของการนำเสนอ (Visualization) 36
บทที่ 4 ผลการวิจัย 37
4.1 กลุ่มลูกค้าที่มีการใช้งานแอปพลิเคชันที่เกี่ยวกับความบันเทิงและมีการใช้ *ABC 37
4.2 กลุ่มลูกค้าที่มีการใช้งานรายการสิทธิเศษ (Privilege) 38
4.3 ดูข้อมูลอาปู (ARPU) เพื่อใช้เป็นข้อกำหนดในการเลือกกลุ่มลูกค้า.....39
4.4 การใช้งานความบันเทิงรวมกับอาปู (Test_Cu_Entertain_Arpu_Total_Mth10) 39
4.5 สรุปการใช้งานของลูกค้า (Test_Cu_Customer_Appsum) 39
4.6 ผลลัพธ์จากการวิเคราะห์ความสนใจในการใช้โปรโมชั่นด้านเอ็นเตอร์เทนเมนท์ 40
4.7 ผลลัพธ์เมื่อเอาท์พุทค่าเคเท่ากับห้า (K=5) ของการจัดกลุ่มโดยใช้ภาษาอาร์.....46
4.8 ผลลัพธ์เมื่อเอาท์พุทค่าเคเท่ากับสี่ (K=4) ของการจัดกลุ่มโดยใช้ภาษาไพธอน 47
4.9 ผลลัพธ์เมื่อใช้กดการหาความสัมพันธ์แบบอะพริโอริ (Apriori) โดยใช้ภาษาอาร์.....50
บทที่ 5 บทสรุปและวิจารณ์ผลการดำเนินงาน.....52
5.1 บทสรุปปริญญานิพนธ์ 52
5.2 ปัญหาที่พบในระหว่างการดำเนินงาน และแนวทางแก้ปัญหา 52
5.3 ข้อเสนอแนะ 52
เอกสารอ้างอิง.....53
ภาคผนวก 57
ภาคผนวก ก โปสเตอร์ 58

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่ **หน้า**

ตารางที่ 1.1 รายละเอียดตารางการทำงานตั้งแต่เดือนสิงหาคมถึงธันวาคม 2559 2
ตารางที่ 2.1 ข้อมูลเกี่ยวกับลูกค้า 19
ตารางที่ 2.2 ผลลัพธ์ของเอสคิลแอลสำหรับรากโหนด 20
ตารางที่ 2.3 ผลลัพธ์โหนดที่เป็นลูกทางขวาของราก 20
ตารางที่ 3.1 รายละเอียดข้อมูลตารางที่ใช้ในการวิเคราะห์ข้อมูล 36

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ

ภาพที่ **หน้า**

ภาพที่ 2.1 ขั้นตอนการทำเหมืองข้อมูล 8

ภาพที่ 2.2 การจัดกลุ่มโดยวิธีเคมีน 10

ภาพที่ 2.3 ลิงค์เกจเมธอดแต่ละวิธี 11

ภาพที่ 2.4 การจัดกลุ่มโดยการหาค่าระยะห่างเฉลี่ยระหว่างกลุ่ม 12

ภาพที่ 2.5 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างสั้นสุด 13

ภาพที่ 2.6 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างมากสุด 14

ภาพที่ 2.7 การจัดกลุ่มโดยใช้จุดศูนย์กลาง 14

ภาพที่ 2.8 ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง 16

ภาพที่ 2.9 การเขียนภาษาเอสคิลแอลสาหรับการเลือกข้อมูลตามเงื่อนไข 16

ภาพที่ 2.10 ขั้นตอนการทำการจำแนกข้อมูล 17

ภาพที่ 2.11 โครงสร้างแบบต้นไม้เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน 18

ภาพที่ 2.12 การเขียนภาษาเอสคิลแอลสาหรับรากโหนด 19

ภาพที่ 2.13 เอสคิลแอลสาหรับโหนดลูก 20

ภาพที่ 2.14 เครือข่ายประสาทเทียมเพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า 21

ภาพที่ 2.15 ตัวอย่างผลลัพธ์การทำคลัสเตอร์ 22

ภาพที่ 2.16 องค์ประกอบของข้อมูลขนาดใหญ่ 24

ภาพที่ 2.17 องค์ประกอบการจัดการข้อมูลที่เป็นข้อมูลขนาดใหญ่ 25

ภาพที่ 3.1 โครงสร้างพื้นฐานของการทาเหมืองข้อมูลรวมถึงอุปกรณ์และเงื่อนไขต่างๆ 26

ภาพที่ 3.2 รูปประกอบการอธิบายการเข้าถึงข้อมูล 27

ภาพที่ 3.3 การเข้าใช้งานเอาท์ลุคครั้งแรกหลังจากคลิกไอคอนเอาท์ลุค 29

ภาพที่ 3.4 ความต้องการที่จะเพิ่มบัญชีอีเมล 29

ภาพที่ 3.5 การกรอกข้อมูลการตั้งค่าบัญชีแบบอัตโนมัติ (Auto Account Setup) 30

ภาพที่ 3.6 การติดตั้ง 30

ภาพที่ 3.7 ความผิดพลาดในการติดตั้ง 31

ภาพที่ 3.8 การเปลี่ยนการตั้งค่าของผู้ใช้งาน 31

ภาพที่ 3.9 การเลือกใช้งานเซอร์วิสของอีเมล์ 32

ภาพที่ 3.10 การตั้งค่าบัญชีแบบป๊อปและบัญชีแบบไอแมพ 33

ภาพที่ 3.11 หน้าจอการทดสอบการตั้งค่าผู้ใช้งาน 33

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

ภาพที่ 3.12 หน้าต่างเขียนอีเมลของเอาท์ลุค....34

ภาพที่ 3.13 หน้าต่างของโปรแกรมขณะมีอีเมลเข้ามา....34

ภาพที่ 4.1 ข้อมูลในการใช้งานแอปพลิเคชันของข้อมูลในวันที่ 29/9/2016 37

ภาพที่ 4.2 เอสคิลแอลในการนับจานวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแบบรายเดือน 38

ภาพที่ 4.3 เอสคิลแอลในการนับจานวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแบบเติมเงิน 38

ภาพที่ 4.4 เอสคิลแอลในการนับจานวนเบอร์โทรที่ไม่ซ้ากันของอาปูวในเดือนตุลาคม....39

ภาพที่ 4.5 เอสคิลแอลในการนับจานวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแอปพลิเคชันตัวอย่าง....39

ภาพที่ 4.6 ผลลัพธ์จากการวิเคราะห์ความสนใจในการใช้โปรโมชั่นด้านความบันเทิง 41

ภาพที่ 4.7 ปริมาณการใช้งาน (ครั้ง) ในการใช้งาน *ABC....42

ภาพที่ 4.8 กราฟแสดงความสัมพันธ์ระหว่างปริมาณการใช้งานกับประเภทโปรโตคอล 43

ภาพที่ 4.9 ภาพแสดงการนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 16/12/2016 45

ภาพที่ 4.10 การนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 17/12/2016....45

ภาพที่ 4.11 การนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 18/12/2016....46

ภาพที่ 4.12 ผลลัพธ์การใช้งานปริมาณการใช้งานจากข้อมูลตัวอย่าง (ข้อมูล 3 วัน)....46

ภาพที่ 4.13 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 0....47

ภาพที่ 4.14 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 1....48

ภาพที่ 4.15 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 2 48

ภาพที่ 4.16 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 3....49

ภาพที่ 4.17 รูปภาพรวมของทั้ง 4 คลัสเตอร์....49

ภาพที่ 4.18 ผลลัพธ์เมื่อใช้กดการหาความสัมพันธ์แบบอะพริโอริ (Apriori) โดยใช้ภาษาอาร์ 50

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญ

บริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) ได้จัดโครงการสหกิจศึกษาระหว่าง บริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) กับคณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง โดยบริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) หรือ เอไอเอส (AIS) ซึ่งผลิตภัณฑ์และบริการหลักที่เอไอเอสให้บริการแก่ลูกค้า คือบริการโทรศัพท์เคลื่อนที่ทั้งในระบบเติมเงินและระบบชำระค่าบริการรายเดือน นอกจากนี้นอกจากนี้ เอไอเอสยังมีบริการเสริมเพื่ออำนวยความสะดวกในรูปแบบต่าง ๆ เช่น การจ่ายเงินออนไลน์ผ่านอีเซอร์วิส (E-service) การชำระค่าสินค้าหรือบริการผ่านเอ็มเปย์ (mPay) การออกแรพบิทซิมการ์ด (Rabbit Simcard) สำหรับคนใช้บริการรถไฟฟ้า แอปพลิเคชั่นเสริมสำหรับคนชอบดูหนังฟังเพลง เป็นต้น

และจากสถานการณ์ปัจจุบันที่ค่ายมือถือต่างมีการแข่งขัน ออกแคมเปญโฆษณาต่างๆ เพื่อเอาใจลูกค้าเก่า และหาลูกค้าใหม่ ดังนั้นเพื่อที่จะทำให้การออกโปรโมชั่นใหม่ๆ การให้สิทธิพิเศษแก่ลูกค้า หรือการออกบริการอะไรต่างๆ ล้วนแต่ต้องมีการคิดให้ดีก่อน เนื่องจากการลงทุนแต่ละครั้งล้วนใช้เงินปริมาณมาก ซึ่งถ้าผลตอบลัพธ์ไม่ดี ก็อาจจะทำให้บริษัทขาดทุนแทนที่จะเป็นการเพิ่มกำไร เพิ่มลูกค้าให้กับบริษัท และจากผลสำรวจในปี 2559 คนไทยมีการใช้อินเตอร์เน็ตผ่านสมาร์ทโฟนเฉลี่ย 6.2 ชั่วโมง/วัน ซึ่งเพิ่มขึ้นจากปี 2558 ที่ค่าเฉลี่ย 5.7 ชั่วโมง/วัน โดยที่แต่ละช่วงวัยมีการใช้อินเตอร์แตกต่างกัน โดยเจเนอเรชันแซด (Generation Z) และเจเนอเรชันวาย (Generation Y) มีการใช้งานยูทูป(Youtube) มากเป็นอันดับหนึ่ง ส่วนเจเนอเรชันเอ็กซ์ (Generation X) และ เบบีบูมเมอร์ (Baby Boomer) มีการใช้งานแอปพลิเคชันไลน์ มากเป็นอันดับหนึ่ง ซึ่งการที่ค่าเฉลี่ยการใช้งานอินเทอร์เน็ตที่เพิ่มมากขึ้นนี้ ทำให้สนใจที่ทราบว่าลูกค้าของสนใจบริการเอ็นเตอร์เทนเมนท์แบบใด เพื่อที่จะได้สร้างความรู้สึกที่ดี รวมถึงสร้างความภักดีเพื่อครองใจลูกค้า โดยในโครงงานนี้จำทำการจัดกลุ่มลูกค้าเป็นกลุ่มรายย่อย หรือ ไมโครเซกเมนต์โดยอาศัยการข้อมูลการใช้ข้อมูลการสนทนา (Call Detail Record) และข้อมูลการใช้งานดาต้าเพิ่มเติม (xDR)

- สำหรับข้อมูลการสนทนาสามารถวิเคราะห์ได้จากสถิติการโทรติดต่อกับหมายเลขที่เป็นธุรกิจหรือบริการ เช่น การโทรสอบถามสถานที่ขายบัตรคอนเสิร์ต การโทรสอบถามรอบหนัง การโทรจองทอล์คโชว์ (Talk Shows) ฯลฯ

- สำหรับ xDR สามารถนำมาวิเคราะห์ความสนใจจากการเข้าใช้งานเว็บไซต์ หรือแอปพลิเคชัน (Application) โดยในส่วนงานของแผนกวิเคราะห์และสำรวจข้อมูล (Data Analytic & Explorations หรือ DAE) จะเป็นแผนกที่เกี่ยวข้องกับงาน 3 ส่วน ได้แก่

1. คลังข้อมูล (Data Warehouse)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. การวิเคราะห์ข้อมูล (Data Analysis)

3. การนำเสนอ (Report)

และในสำหรับโครงงานครั้งนี้ จะเน้นหลัก ๆ ไปทางการวิเคราะห์ข้อมูลและนำเสนอผ่านกราฟ

1.2 วัตถุประสงค์ของการปฏิบัติงาน

1. เพื่อจัดกลุ่มลูกค้าตามความสนใจในเชิงลึก
2. สามารถนำโครงงานไปประกอบการตัดสินใจต่าง ๆ ในการออกโปรโมชั่น
3. เพื่ออัพเดทข่าวสารให้กับลูกค้าตามความสนใจ
4. ใช้ภาษาอาร์ ภาษาไพธอน และแทบโบลว (Tableau) ในการวิเคราะห์และสร้างแบบจำลองจากข้อมูลได้

1.3 ขอบเขตของโครงงาน

1. การวิเคราะห์ข้อมูลครั้งนี้ ได้ใช้ข้อมูลของกลุ่มตัวอย่างจากสถานประกอบการแห่งหนึ่ง เป็นกรณีศึกษา
2. มีการเข้ารหัสทางเดียว (Hashing) ข้อมูลของกลุ่มตัวอย่างไว้ก่อนนำข้อมูลมาวิเคราะห์ เพื่อป้องกันความปลอดภัยในการนำข้อมูลไปใช้ในทางผิดกฎหมาย
3. ใช้ภาษาอาร์และภาษาไพธอนในการวิเคราะห์ทั้งโครงการ
4. ใช้แทบโบลว (Tableau) ในการพล็อตกราฟ และแสดงผลข้อมูล

1.4 วิธีการดำเนินงาน

ในการที่จะวิเคราะห์ข้อมูลความสนใจของการใช้งานโมบายแอปพลิเคชันนั้น มีการวางแผนการทำงานในขั้นตอนต่าง ๆ ซึ่งสามารถจำแนกออกได้ตามแต่ละส่วนของขอบเขตงานวิจัยดังตารางที่ 1.1

ตารางที่ 1.1 รายละเอียดตารางการทำงานตั้งแต่เดือนสิงหาคม-ธันวาคม 2559

หัวข้อ	สิงหาคม				กันยายน				ตุลาคม				พฤศจิกายน				ธันวาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1. ศึกษาและทำความเข้าใจในเรื่องของข้อมูลและระบบ																				

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 1.1 รายละเอียดตารางการทำงานตั้งแต่เดือนสิงหาคม-ธันวาคม 2559 (ต่อ)

หัวข้อ	สิงหาคม				กันยายน				ตุลาคม				พฤศจิกายน				ธันวาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
2. ศึกษาการใช้งาน Vistualization (Tableau)																				
3. ศึกษาและรวบรวมข้อมูลในส่วนของการใช้งานแอปพลิเคชันของผู้ใช้																				
4. ศึกษาภาษา R และโปรแกรมฮาร์สตูดิโอและศึกษาภาษาไพทอน																				
5. เตรียมอินฟราสตรัคเจอร์																				
6. เตรียมข้อมูลในดาต้าเบส																				
7. เลือกกลุ่มลูกลูกค้าที่จะวิเคราะห์จากปัจจัยต่างๆ																				
8. การเตรียมข้อมูล																				

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 1.1 รายละเอียดตารางการทำงานตั้งแต่เดือนสิงหาคม-ธันวาคม 2559 (ต่อ)

หัวข้อ	สิงหาคม				กันยายน				ตุลาคม				พฤศจิกายน				ธันวาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
9. Explore Data(นำข้อมูลมาวิเคราะห์หาสิ่งที่สนใจ)																				
10. Feature Extraction (ดึงจุดเด่น หรือส่วนสำคัญจากผลการวิเคราะห์ที่ได้)																				
11. เช็คผลการวิเคราะห์และรวบรวมข้อมูลที่ทำมาทั้งหมด																				

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 1.1 รายละเอียดตารางการทำงานตั้งแต่เดือนสิงหาคม-ธันวาคม 2559 (ต่อ)

หัวข้อ	สิงหาคม				กันยายน				ตุลาคม				พฤศจิกายน				ธันวาคม			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
12. จัดทำรูปเล่มรายงาน โปสเตอร์ และทำสื่อการนำเสนอ																				
13. นำเสนอข้อมูลในเรื่องโมบายแอปพลิเคชัน																				

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่ได้รับในการพัฒนาระบบที่ได้รับมอบหมายในการเข้าร่วมโครงการสหกิจศึกษากับทางบริษัท แอดวานซ์ อินโฟร์ เซอร์วิส จำกัด (มหาชน) สามารถจำแนกได้ออกเป็น 2 ส่วน ดังนี้

1.5.1 ประโยชน์ต่อตนเอง

ได้เรียนรู้สิ่งใหม่ ๆ เห็นการทำงานที่แท้จริงและนำความรู้จากที่เรียนมาใช้ให้เกิดประโยชน์ ฝึกการทำงานที่เป็นระบบ ปรับตัวให้เข้ากับสภาพแวดล้อมในองค์กรได้ มีวินัยในการดำเนินชีวิต มีความตรงต่อเวลา มีความรับผิดชอบและมีความขวนขวายในการหาความรู้มากขึ้น ได้พูดคุย ได้ฝึกการสื่อสาร ทำให้มีมนุษย์สัมพันธ์ที่ดีขึ้น และยังได้รู้จักกับบุคคลในสายงานของตนเอง

1.5.2 ประโยชน์ต่อบริษัท

1. สามารถนำไปช่วยใช้ในการประกอบการตัดสินใจในการออกโปรโมชั่นใหม่ ๆ หรือบริการต่าง ๆ ได้

2. สามารถจัดกลุ่มลูกค้าตามความสนใจอย่างถูกต้องได้

3. สามารถตอบโจทย์ความต้องการของลูกค้าอย่างตรงตามประเด็นได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.6 อุปกรณ์ที่ใช้ในการพัฒนาโครงงาน

1.6.1 ฮาร์ดแวร์

1. โน้ตบุ๊ก (Notebook) สำหรับพัฒนาโปรแกรมจำนวน 1 เครื่อง
2. สวิตช์ (Switch) สำหรับเชื่อมต่อพอร์ท (Port) ของโน้ตบุ๊กเข้ากับสายแลน (LAN) เพื่อปิดกั้นไม่ให้ข้อมูลถูกส่งไปข้างนอกได้

1.6.2 ซอฟต์แวร์

1. อิมพาลาฮาดูป (Impala Hadoop)
2. จูปีเตอร์จากอนาคอนด้า (Jupyter from Anaconda)
3. อาร์สตูดิโอ (Rstudio)
4. แทบโบลว (Tableau)

1.7 ภาษาที่ใช้ในการทำ

1. ภาษาไพธอน (Python)
2. ภาษาอาร์ (R)
3. ภาษาเอสคิวแอล (SQL)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เหมืองข้อมูล (การทำเหมืองข้อมูล) [1]

การทำเหมืองข้อมูล คือ การค้นหาความสัมพันธ์และรูปแบบ (Pattern) ทั้งหมด ซึ่งมีอยู่จริงในฐานข้อมูล แต่ได้ถูกซ่อนไว้ภายในข้อมูลจำนวนมาก การทำเหมืองข้อมูลจะทำการสำรวจและวิเคราะห์อย่างอัตโนมัติหรือกึ่งอัตโนมัติ ในปริมาณข้อมูลจำนวนมากให้อยู่ในรูปแบบที่เต็มไปด้วยความหมายและอยู่ในรูปของกฎ (Rule) โดยความสัมพันธ์เหล่านี้แสดงให้เห็นถึงความรู้ต่างๆที่มีประโยชน์ในฐานข้อมูล

2.1.1 ประเภทข้อมูลที่สามารถทำเหมืองข้อมูล

1. ฐานข้อมูลที่จัดเก็บอยู่ในรูปแบบของตาราง (Relational Database) โดยในแต่ละตารางประกอบไปด้วยแถวและคอลัมน์ ความสัมพันธ์ของข้อมูลทั้งหมดสามารถแสดงได้โดย แบบจำลองข้อมูลเชิงสัมพันธ์ (Entity-Relationship Model)

2. คลังข้อมูล (Data Warehouses) เป็นการเก็บรวบรวมข้อมูลจากหลายแหล่งมาเก็บไว้ในรูปแบบเดียวกันและรวบรวมไว้ในที่ ๆ เดียวกัน หรือ จะเก็บเป็น ฐานข้อมูล ก็ได้

3. ฐานข้อมูลรายการเปลี่ยนแปลง (Transaction Database) ประกอบด้วยข้อมูลที่ แต่ละรายการแทนด้วยเหตุการณ์ในขณะใดขณะหนึ่ง เช่น ใบเสร็จรับเงินซึ่งเก็บข้อมูลชื่อลูกค้าและรายการสินค้าที่ลูกค้ารายนั้นซื้อ เป็นต้น

4. ฐานข้อมูลที่จัดเก็บในรูปแบบอื่น (Advanced Database) เช่น ข้อมูลเท็กซ์ไฟล์ ข้อมูลมัลติมีเดีย หรือข้อมูลในรูปของเว็บ เป็นต้น

2.1.2 ลักษณะเฉพาะของข้อมูลที่สามารถทำเหมืองข้อมูล

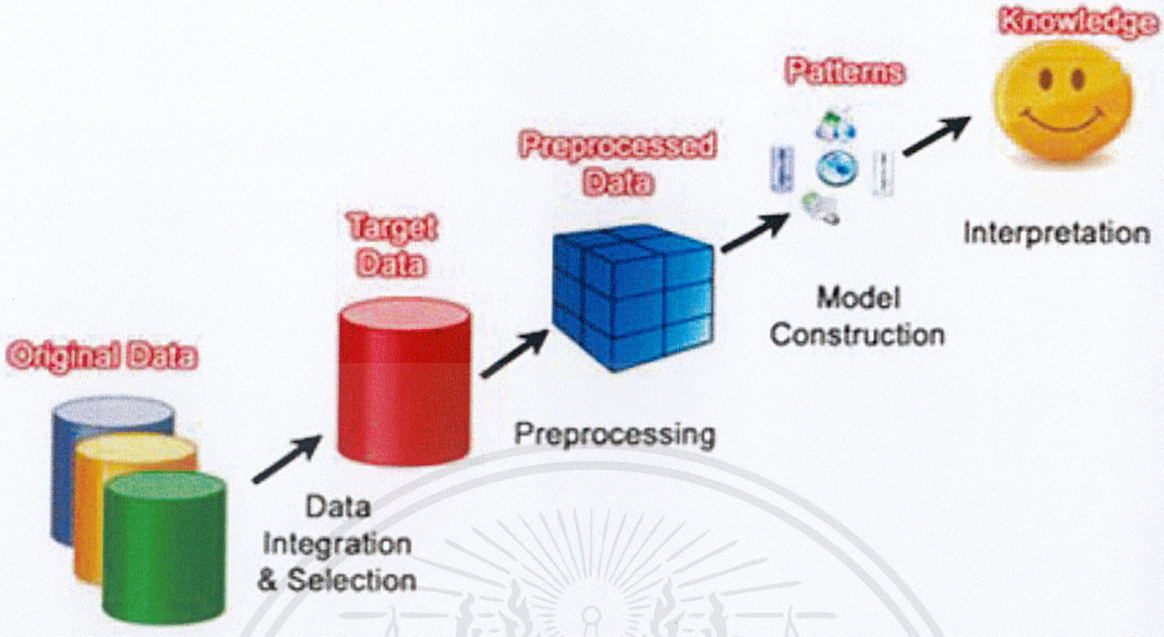
1. ข้อมูลที่มีขนาดใหญ่เกินกว่าจะพิจารณาความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลนั้นได้ด้วยตาเปล่า หรือด้วยการใช้ระบบจัดการฐานข้อมูล (Database Management System) ในการจัดการฐานข้อมูล

2. ข้อมูลที่มาจากหลายแหล่งโดยอาจรวบรวมมาจากหลายระบบปฏิบัติการหรือหลายระบบจัดการฐานข้อมูล

3. ข้อมูลที่ไม่มีการเปลี่ยนแปลงตลอดช่วงเวลาของการทำเหมืองข้อมูล หากข้อมูลที่มี อยู่นั้นเป็นข้อมูลที่เปลี่ยนแปลงตลอดเวลาจะต้องแก้ปัญหานี้ก่อน โดยบันทึกฐานข้อมูลนั้นไว้และนำฐานข้อมูลที่บันทึกไว้มาทำเหมืองข้อมูล แต่เนื่องจากข้อมูลนั้นมีการเปลี่ยนแปลงอยู่ตลอดเวลา จึงทำให้ผลลัพธ์ที่ได้จาการทำเหมืองข้อมูลสมเหตุสมผลในช่วงเวลาหนึ่งเท่านั้น ดังนั้นเพื่อให้ได้ผลลัพธ์ที่มี ความถูกต้องเหมาะสมอยู่ตลอดเวลาจึงต้องทำเหมืองข้อมูลใหม่ทุกครั้งในช่วงเวลาที่เหมาะสม

4. ข้อมูลที่มีโครงสร้างซับซ้อน เช่น ข้อมูลรูปภาพ ข้อมูลมัลติมีเดีย ข้อมูลเหล่านี้สามารถนำมาทำเหมืองข้อมูลได้เช่นกันแต่ต้องใช้เทคนิคการทำเหมืองข้อมูลขั้นสูง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 2.1 ขั้นตอนการทำเหมืองข้อมูล

จากภาพที่ 2.1 แสดงขั้นตอนการทำเหมืองข้อมูลพื้นฐาน ซึ่งมี 5 ขั้นตอน ได้แก่ การเตรียมข้อมูล การจำแนกข้อมูล การประมวลผล การหาแบบจำลอง หรือรูปแบบและแปลงออกมาเป็นองค์ความรู้

2.2 การจัดกลุ่ม (Clustering) [2]

2.2.1 เทคนิคการจัดกลุ่ม (Clustering)

ตามหลักการของการทำเหมืองข้อมูล (Data Mining) จะต่างกับการแบ่งกลุ่มในเชิงสถิติ เนื่องจากการแบ่งกลุ่มทางสถิติ จะหมายถึง การทำการสุ่มข้อมูล (Sampling) เช่น มีข้อมูลอยู่ 30 กลุ่มซึ่งมีลักษณะไม่แตกต่างกันมาก ก็จะแบ่งกลุ่มโดยวิธีการสุ่มใช้ข้อมูลเพียงบางกลุ่มเท่านั้น ส่วนการทำการแบ่งกลุ่มทางด้านการขุดค้นข้อมูลนั้น จะหมายถึง การแบ่งกลุ่มที่ต้องใช้ข้อมูลของทุกกลุ่มในการหาจุดเด่นของแต่ละกลุ่มออกมาให้เห็นอย่างชัดเจน เพื่อได้มาซึ่งลักษณะเฉพาะของกลุ่มนั้น ๆ

2.2.2 ความหมายของการจัดกลุ่ม

การจัดกลุ่มเป็นเทคนิคที่ใช้จำแนกหรือแบ่งเป็นกรณี (Case) หมายถึง คน สัตว์ สิ่งของ หรือองค์กร ฯลฯ หรือแบ่งตัวแปรออกเป็นกลุ่มย่อย ๆ ตั้งแต่ 2 กลุ่มขึ้นไป กรณีที่อยู่ในกลุ่มเดียวกันจะมีลักษณะที่เหมือนกันหรือคล้ายกัน ส่วนกรณีที่อยู่ต่างกลุ่มกัน จะมีลักษณะที่แตกต่างกัน ดังนั้น การพิจารณาเลือก

ลักษณะหรือตัวแปรที่จะนามาใช้ในการแบ่งกลุ่มกรณี จึงมีความสำคัญ นอกจากนั้นกรณีใดกรณีหนึ่งจะต้องอยู่ในกลุ่มหนึ่งเพียงกลุ่มเดียวถ้านำเทคนิคการจัดกลุ่ม มาใช้ในการแบ่งกลุ่มตัวแปร จะให้ตัวแปรอยู่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในกลุ่มเดียวกันมีความสัมพันธ์กันมากกว่าตัวแปรที่อยู่ต่างกลุ่มกัน ตัวแปรที่อยู่ต่างกลุ่มกันมีความสัมพันธ์กันน้อยหรือไม่มีความสัมพันธ์กันเลย ส่วนใหญ่การแบ่งกลุ่มตัวแปรจะใช้เทคนิคตัวแปร ส่วนการแบ่งกลุ่มกรณีคน สัตว์ สิ่งของ จะใช้เทคนิคการจัดกลุ่ม

2.2.3 วัตถุประสงค์ของการจัดกลุ่ม

เพื่อการลดขนาดข้อมูล (Data Reduction) ด้วยเทคนิคที่แม่นยำและมีประสิทธิภาพเป็นอีกแนวทางหนึ่งเพื่อจัดการกับปัญหาดังกล่าว หลักการที่สำคัญของการลดขนาดข้อมูลคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด เนื่องจากข้อมูลแต่ละตัวจะมีความสำคัญต่อการจัดกลุ่มข้อมูลไม่เท่ากัน ด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มีความสำคัญ และสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ ข้อมูลที่มีการรวมกลุ่มกันอย่างหนาแน่นจะเป็นข้อมูลที่มีความสำคัญต่อการจัดกลุ่มข้อมูลในอนาคต

2.2.4 คุณสมบัติของเทคนิควิธีการจัดกลุ่ม

1. ความต้องการทางด้านข้อมูล สำหรับการวิเคราะห์จัดกลุ่มหน่วยวิเคราะห์ ผู้วิจัยอาจใช้ข้อมูลที่ระบุหน่วยวิเคราะห์และตัวแปรตามที่จัดเก็บมาได้เลย เช่น การวิเคราะห์ที่ได้กล่าวมาแล้วข้างต้น ส่วนการวิเคราะห์จัดกลุ่มตัวแปร ผู้วิจัยไม่อาจจะใช้แฟ้มข้อมูลดังกล่าวได้โดยใช้เมตริกแสดงความสัมพันธ์ระหว่างตัวแปรแทนได้

2. แนวคิดพื้นฐาน สิ่งสำคัญที่สุดของการวิเคราะห์การจัดกลุ่มคือ ตัวแปรที่ใช้หากผู้วิจัยไม่ได้เก็บข้อมูลเกี่ยวกับตัวแปรที่สำคัญ ๆ ผลที่ได้ก็จะไม่ดีหรือทำให้ไขว้เขวได้ ทั้งนี้เพราะตัวแปรที่เลือกไว้ตั้งแต่แรกจะเป็นสิ่งที่กำหนดคุณสมบัติของ สิ่งที่ระบุความเป็นกลุ่มย่อย เช่นในการจัดกลุ่มโรงเรียนในเมือง หากผู้วิจัยไม่เก็บข้อมูลเกี่ยวกับจำนวนนักเรียนและครู ขนาดของโรงเรียนก็ไม่อาจเป็นเกณฑ์ในการแบ่งกลุ่มได้

3. ความคล้ายกันของหน่วย ความคิดเกี่ยวกับความคล้ายของหน่วยศึกษาเป็นเทคนิคของการวิเคราะห์ทางสถิติหลายวิธี โดยทั่วไปการวัดความคล้ายจะพิจารณาจากความห่างระหว่างวัตถุ หรือพิจารณาจากความคล้ายกัน

4. การวัดความห่าง วิธีการวัดความห่างสามารถวัดได้หลายวิธี วิธีการหนึ่งที่นิยมวัดกันมากก็คือวิธีที่เรียกว่า ระยะห่างเชิงยูคลิดยกกำลังสอง (Squared Euclidean Distance) คือ ผลรวมของผลต่างยกกำลังสองของทุกตัวแปร เช่น ต้องการดูความห่างกันของเบียร์ 2 ยี่ห้อ ซึ่งทราบราคา

2.3 เทคนิคการจัดกลุ่มแบบเคมีน (K-means Clustering) [3]

2.3.1การจัดกลุ่มแบบเคมีน

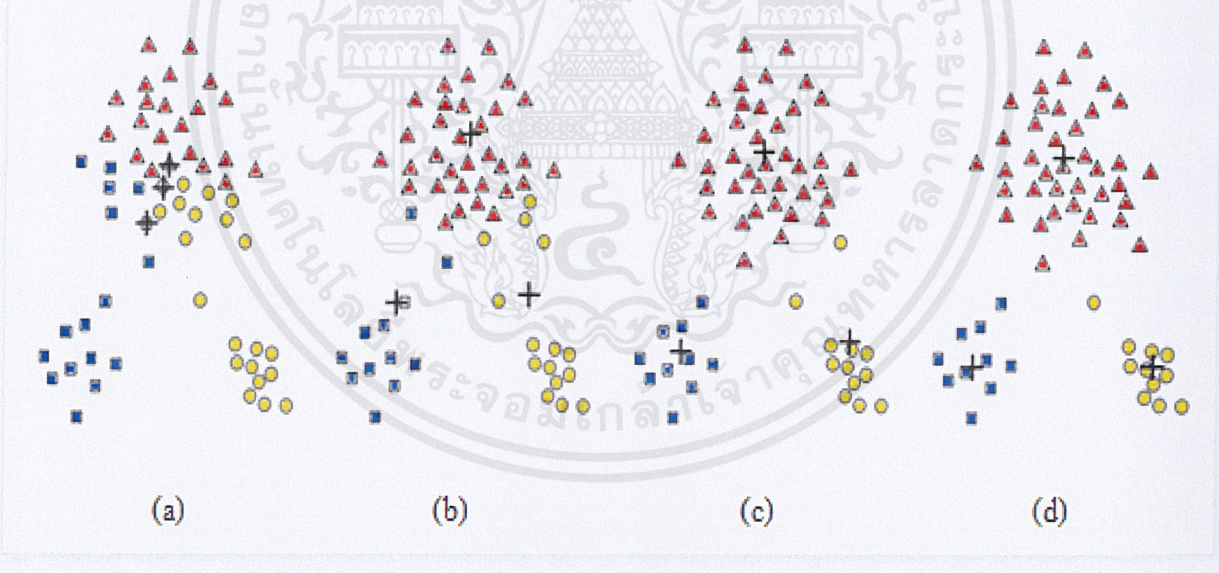
การจัดกลุ่มแบบเคมีน (K-means) คือ หนึ่งในอัลกอริทึมเทคนิคการเรียนรู้โดยไม่มีผู้สอนที่ง่ายที่สุด เพราะเป็นการแก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป โดยอัลกอริทึมเคมีนจะตัดแบ่ง (Partition) วัตถุออกเป็น K กลุ่ม โดยแทนแต่ละกลุ่มด้วยค่าเฉลี่ยของกลุ่ม ซึ่งใช้เป็นจุดศูนย์กลาง (Centroid) ของกลุ่มในการวัดระยะห่างของข้อมูลในกลุ่มเดียวกัน ในขั้นแรกของการจัดกลุ่มโดยการหาค่าเฉลี่ยแบบเคมีน ต้องกำหนดจำนวนกลุ่ม (K)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ที่ต้องการ และกำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด สิ่งสำคัญในการกำหนดจุดศูนย์กลางเริ่มต้นของแต่ละกลุ่มนี้ ควรจะถูกกำหนดด้วยวิธีที่เหมาะสม เพราะตำแหน่งจุดศูนย์กลางเริ่มต้นที่แตกต่างกันทำให้ได้ผลลัพธ์สุดท้ายแตก ต่างกัน ดังนั้นในทางที่ดีควรจะกำหนดจุดศูนย์กลางนี้ให้หางจากจุดศูนย์กลางอื่นๆ ขั้นตอนต่อไปคือสร้างกลุ่มข้อมูลและความสัมพันธ์กับจุดศูนย์กลางที่ใกล้มากที่สุด โดยแต่ละจุดจะถูกกำหนดไปยังจุดศูนย์กลางที่ใกล้เคียงที่สุดจนครบหมดทุกจุด และคำนวณจุดศูนย์กลางใหม่ โดยการหาค่าเฉลี่ยทุกวัตถุที่อยู่ในกลุ่ม หากจุดศูนย์กลางในแต่ละกลุ่มถูกเปลี่ยนตำแหน่ง จะได้จุดมีความสัมพันธ์กับกลุ่มใหม่และใกล้กับจุดศูนย์กลางใหม่ ทำซ้ำแบบนี้ไปเรื่อย ๆ จะสังเกตเห็นว่าผลลัพธ์จากการทำซ้ำแบบนี้ทำให้จุดศูนย์กลางเปลี่ยนตำแหน่งทุกรอบ จนกระทั่งจุดศูนย์กลางจำนวน K จุด ไม่มีการเปลี่ยนแปลงจึงจะสิ้นสุดกระบวนการ

2.3.2 อัลกอริทึมการจัดกลุ่มโดยเคมีน

1. กำหนดจำนวนกลุ่ม K กลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้นจำนวน K จุด
2. นำวัตถุทั้งหมดจัดเข้ากลุ่มที่มีจุดศูนย์กลางที่อยู่ใกล้วัตถุนั้นมากที่สุด โดยคำนวณจากการวัดระยะห่างระหว่างจุดที่น้อยที่สุด
3. คำนวณจุดศูนย์กลาง K จุดใหม่ โดยหาจากค่าเฉลี่ยทุกวัตถุที่อยู่ในกลุ่ม
4. ทำซ้ำในข้อที่ 3 จนกระทั่งจุดศูนย์กลางไม่เปลี่ยนแปลง



ภาพที่ 2.2 การจัดกลุ่มโดยวิธีเคมีน [3]

รูป (a) เป็นการจัดกลุ่มในขั้นตอนแรกโดยที่กำหนดจำนวนกลุ่ม 3 กลุ่ม และกำหนดจุดศูนย์กลางเริ่มต้น ซึ่งใช้สัญลักษณ์ + แทนจุดศูนย์กลางของแต่ละกลุ่มทั้ง 3 กลุ่ม จากนั้นวัตถุจะถูกกำหนดให้เข้ากลุ่มที่มีจุดศูนย์กลางอยู่ใกล้วัตถุนั้นมากที่สุดแสดงได้ดังรูป (b) จุดศูนย์กลางมีการเปลี่ยนแปลงและเกิด

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

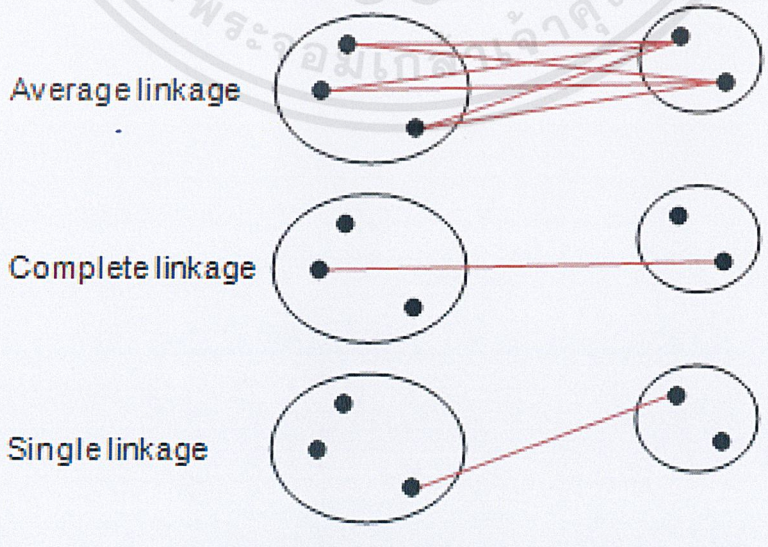
ความสัมพันธ์ระหว่างวัตถุกับจุดศูนย์กลางใหม่ และจัดวัตถุให้เข้ากลุ่มที่มีจุดศูนย์กลางอยู่ใกล้กับวัตถุนั้นมากที่สุดดังรูป (c) ทำซ้ำเช่นนี้ไปเรื่อย ๆ จนกระทั่งจุดศูนย์กลางไม่เปลี่ยนแปลงจึงจะได้ผลลัพธ์สุดท้ายดังรูป (d)

2.4 เทคนิคการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) [4]

การจัดกลุ่มแบบลำดับชั้นแอกกอลเมราทีฟ (Agglomerative Hierarchical) วิธีนี้เริ่มต้นจะสมมติว่ามี n กลุ่มย่อย สิ่งของ (item) ที่มีระยะสั้นที่สุด หรือคล้ายกันมากที่สุดจะรวมเข้าด้วยกันเป็นกลุ่มก่อน จึงเหลือ n-1 กลุ่มย่อย จากนั้นหาระยะทางหรือความคล้ายจาก n – 1 กลุ่มย่อยใหม่ แล้วดูว่ากลุ่มย่อยใดมีระยะทางสั้นที่สุด หรือคล้ายกันมากที่สุดก็รวมกลุ่มย่อยนั้นเข้าด้วยกัน ทำเช่นนี้ต่อ ๆ ไป ในท้ายที่สุดแล้วจะมีเหลือเพียง 1 กลุ่ม ข้อจำกัดของวิธีการจัดกลุ่มแบบลำดับชั้นแอกกอลเมราทีฟ เนื่องจากวิธีการจัดกลุ่มแบบลำดับชั้นแอกกอลเมราทีฟการจัดกลุ่ม จะเริ่มต้นให้จำนวนกรณี = จำนวนคลัสเตอร์ เช่น มี n กรณี = มี n คลัสเตอร์แล้วค่อย ๆ ลดจำนวนคลัสเตอร์ทีละ 1 โดยรวมกลุ่ม 2 คลัสเตอร์ ที่คล้ายกันมากที่สุด หรือต่างกันน้อยที่สุดเข้าด้วยกัน จึงค่อย ๆ ลดจำนวนคลัสเตอร์ครั้งละ 1 ดังนั้น ถ้ามี n มาก เช่น n = 1,000 คน จะต้องทำการรวมกลุ่ม 999 ครั้งโดยเริ่มจากมี 1,000 คลัสเตอร์แล้วลดเหลือ 999 คลัสเตอร์ 998 คลัสเตอร์เป็นเช่นนี้ไปเรื่อย ๆ จน เหลือ 1 คลัสเตอร์ ซึ่งจะทำให้เสียเวลามาก ดังนั้นโดยทั่วไปถ้ามี จำนวนกรณีมากว่า 200 กรณี จึงไม่นิยมใช้เทคนิคจัดกลุ่มแบบลำดับชั้น

ในการจัดกลุ่มแบบลำดับชั้นแอกกอลเมราทีฟจะมีวิธีที่เรียกว่า ลิงค์เกจเมธอด (Linkage method) ที่เป็นที่รู้จักกันโดยทั่วไปมี 3 วิธีคือ

1. ระยะทางที่สั้นที่สุดระหว่างสิ่งของ เรียกว่า ซิงเกิลลิงค์เกจ (Single Linkage)
2. ระยะทางที่ยาวที่สุดระหว่างสิ่งของ เรียกว่า คอมพลีทลิงค์เกจ (Complete Linkage)
3. ระยะทางที่ได้จากระยะทางเฉลี่ยระหว่างสิ่งของ เรียกว่า เอเวอเรจลิงค์เกจ (Average Linkage)



ภาพที่ 2.3 ลิงค์เกจเมธอดแต่ละวิธี [5]

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชนิดของข้อมูลหรือตัวแปรที่สามารถใช้ เทคนิคการจัดกลุ่มแบบลำดับชั้นได้ มี 3 ประเภท คือ

1. ข้อมูลเป็นสเกลอันตรภาค (Interval scale) หรือสเกล อัตราส่วน (Ratio scale)
2. ข้อมูลที่อยู่ในรูปความถี่ (Count Data)
3. ข้อมูลอยู่ในรูปไบนารี่ (Binary) นั่นคือ มีได้ 2 ค่า คือ 0 กับ 1 หรือ คุณภาพ

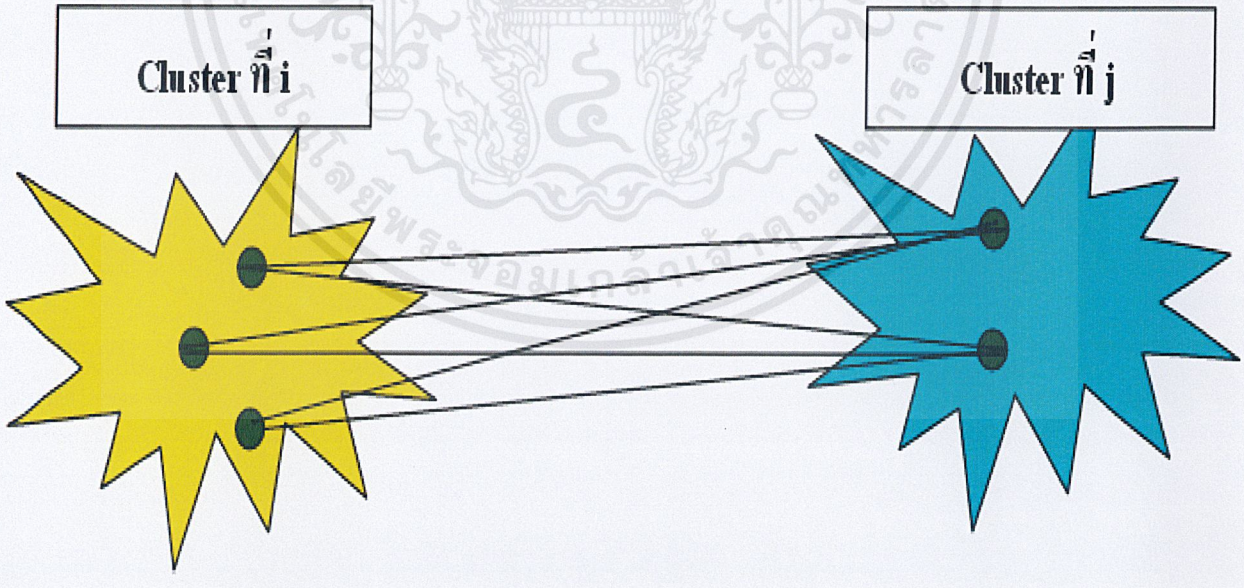
กรณีที่วัดความคล้ายด้วยระยะห่าง ถ้าระยะห่างระหว่างกรณีคู่ใดต่ำ แสดงว่ากรณีคู่นั้นอยู่ใกล้กัน หรือมีความคล้ายกัน ควรจะจัดให้อยู่ในกลุ่ม หรือคลัสเตอร์เดียวกัน สำหรับวิธีการคำนวณจะขึ้นอยู่กับชนิดของ ข้อมูลทั้ง 3 ชนิดข้างต้น คือ ข้อมูลเป็นสเกลอันตรภาค ข้อมูลที่อยู่ในรูปความถี่ ข้อมูลอยู่ในรูปไบนารี่

กรณีที่วัดความคล้ายด้วยของกรณีถ้าค่าความคล้ายของกรณีคู่ใดมีค่ามาก แสดงว่ากรณีคู่นั้นคล้ายกันมาก จึงควรจัด ให้อยู่ในกลุ่มเดียวกัน การคำนวณค่าความ คล้ายจะแตกต่างกัน ถ้าชนิดของข้อมูล แตกต่างกัน

กรณีที่วัดความคล้ายของตัวแปรด้วยค่า สัมประสิทธิ์สหสัมพันธ์ ถ้าตัวแปรคู่ใด มีค่าสัมประสิทธิ์สหสัมพันธ์ มาก แสดงว่าคู่นั้นสัมพันธ์กันมากควรจัดไว้ ในกลุ่มเดียวกัน

2.5 หลักเกณฑ์ในการรวมกลุ่ม [6]

2.5.1. การจัดกลุ่มโดยการหาค่าระยะห่างเฉลี่ยระหว่างกลุ่ม (Average Linkage Between Groups)



ภาพที่ 2.4 การจัดกลุ่มโดยการหาค่าระยะห่างเฉลี่ยระหว่างกลุ่ม

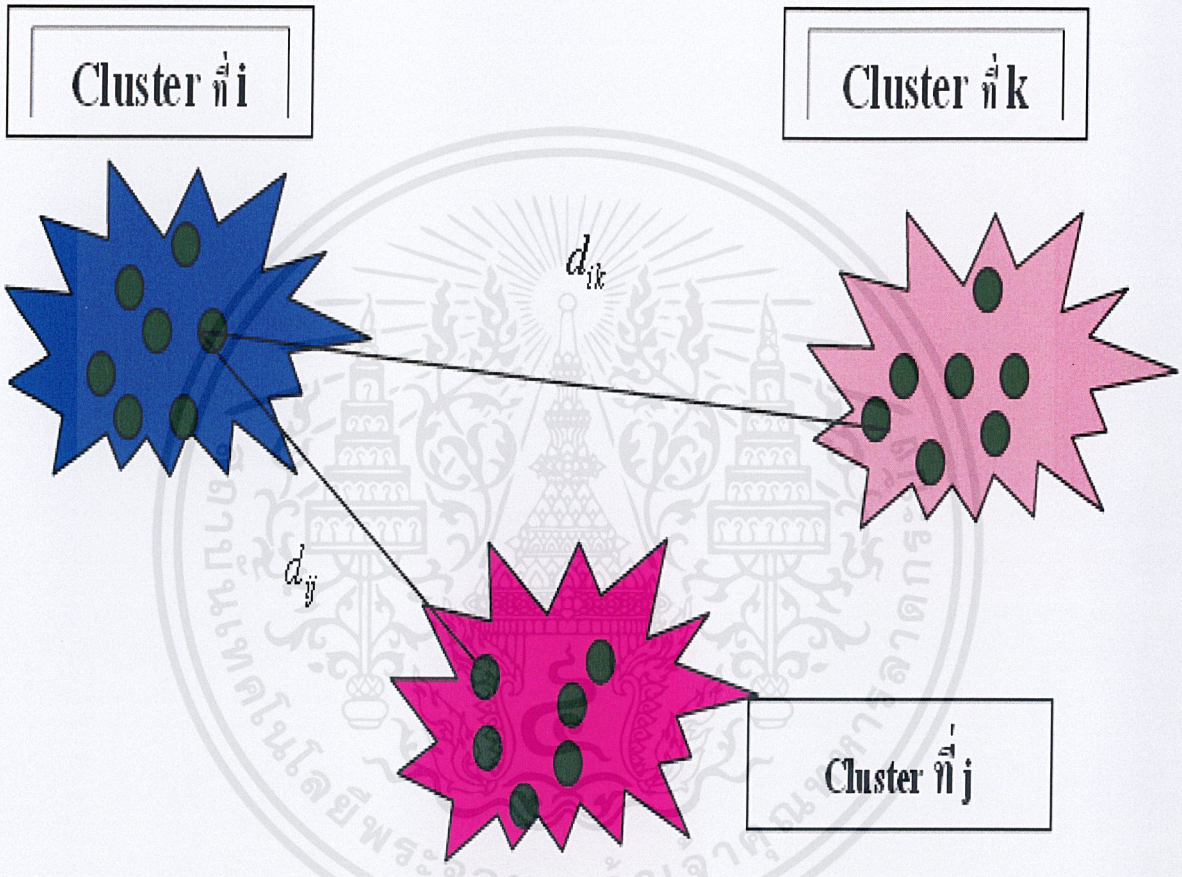
จากภาพที่ 2.4 วิธีนี้จะคำนวณหาระยะห่างเฉลี่ยของทุก คู่ของกรณีโดยที่กรณีหนึ่งอยู่ในคลัสเตอร์ที่ i ส่วนอีก กรณีหนึ่งอยู่ในคลัสเตอร์ ที่ j ถ้าคลัสเตอร์ ที่ i มีระยะห่างเฉลี่ยจากคลัสเตอร์ ที่ j สั้นกว่าระยะห่างจากคลัสเตอร์อื่น จะนำคลัสเตอร์ ที่ i และ j รวมกันเป็นคลัสเตอร์เดียวกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5.2. การจัดกลุ่มโดยใช้ค่าระยะห่างที่น้อยที่สุด (Within-group Linkage Technique)

วิธีนี้จะเป็นการรวมคลัสเตอร์เข้าด้วยกัน โดยใช้ระยะห่างเฉลี่ยระหว่างทุกกรณี ในคลัสเตอร์นั้น ๆ มีค่าน้อยที่สุด

2.5.3. การจัดกลุ่ม 2 กลุ่มแบบระยะห่างสั้นสุด (Single Linkage)

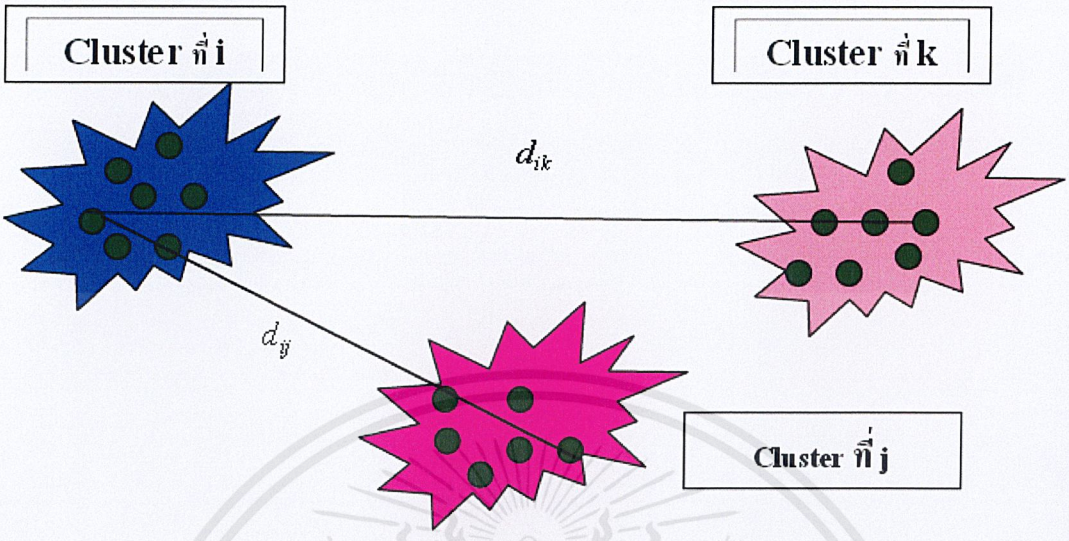


ภาพที่ 2.5 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างสั้นสุด

จากภาพที่ 2.5 วิธีนี้จะรวมคลัสเตอร์ 2 คลัสเตอร์เข้าด้วยกันโดยพิจารณาจากระยะห่างที่สั้นที่สุด โดยที่ d_{ik} เป็นระยะห่างที่สั้นที่สุดระหว่างคลัสเตอร์ i และ j ในรูปที่ 5.2 จะรวมคลัสเตอร์ i และ j เข้าด้วยกันเพราะ $d_{ij} < d_{ik}$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

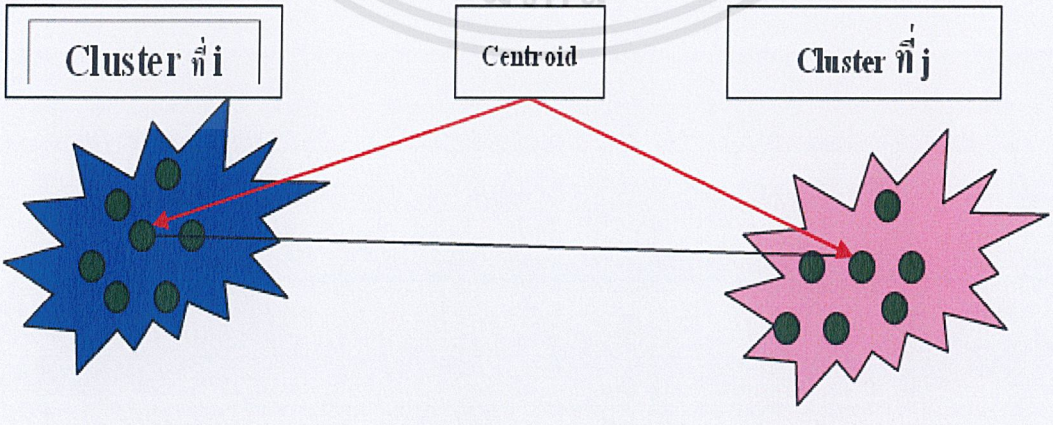
2.5.4 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างมากสุด (Complete Linkage)



ภาพที่ 2.6 การจัดกลุ่ม 2 กลุ่มแบบระยะห่างมากสุด

จากภาพที่ 2.6 วิธีนี้จะรวมคลัสเตอร์ 2 คลัสเตอร์เข้าด้วยกันโดยพิจารณาจากระยะห่างที่ยาวที่สุด d_{ik} = ระยะห่างที่ยาวที่สุดของคลัสเตอร์ ที่ i และ k d_{ij} = ระยะห่างที่ยาวที่สุดของคลัสเตอร์ ที่ i และ j ในที่นี้ $d_{ij} < d_{ik}$ จึงรวม คลัสเตอร์ ที่ i และ j เข้าเป็น คลัสเตอร์ เดียวกัน

2.5.5. การจัดกลุ่มโดยใช้จุดศูนย์กลาง (Centroid Clustering)



ภาพที่ 2.7 การจัดกลุ่มโดยใช้จุดศูนย์กลาง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากภาพที่ 2.7 วิธีการนี้จะคำนวณหาระยะห่างระหว่างจุดศูนย์กลางของคลัสเตอร์ทีละคู่ ในที่นี้ จะเรียกค่าเฉลี่ย หรือค่ากลางของแต่ละคลัสเตอร์ ว่าระยะห่างระหว่างจุดศูนย์กลางของคลัสเตอร์ เนื่องจากการจัดกลุ่มกรณีจะพิจารณาจากตัวแปรหลาย ๆ ตัวพร้อม ๆ กัน จึงเรียกค่ากลางหรือค่าเฉลี่ยว่าศูนย์กลาง ถ้าระยะห่างระหว่างจุดศูนย์กลางของคลัสเตอร์ คู่ใดต่ำจะรวมคลัสเตอร์คู่นั้นเข้าเป็นคลัสเตอร์เดียวกัน

2.5.6. การจัดกลุ่มโดยใช้ค่ากลาง (Median Clustering)

วิธีนี้จะรวมคลัสเตอร์ 2 คลัสเตอร์ เข้าด้วยกัน โดยให้แต่ละคลัสเตอร์สำคัญเท่ากัน (ให้น้ำหนักเท่ากัน) ในขณะที่วิธีของการจัดกลุ่มโดยใช้จุดศูนย์กลาง จะให้ความสำคัญแก่คลัสเตอร์ ที่มีขนาดใหญ่มากกว่าคลัสเตอร์ที่มีขนาดเล็ก (ให้น้ำหนักไม่เท่ากัน) การจัดกลุ่มโดยใช้ค่ากลาง จะใช้ค่าระหว่างค่ากลางเป็นค่ากลางของจุดศูนย์กลาง ถ้าระยะห่างระหว่างค่ากลาง (Median) ของคลัสเตอร์จะใช้ค่าระหว่างค่ากลาง เป็นค่ากลางของของจุดศูนย์กลาง ถ้าระยะห่างระหว่างค่ากลางของคลัสเตอร์คู่ใดต่ำจะรวม คลัสเตอร์ คู่นั้นเข้าด้วยกัน

2.5.7. การจัดกลุ่มโดยหาความเหมาะสมเอง (Ward's Method)

วิธีนี้จะพิจารณาจากค่าซัมสแควออฟเดอะสแควิทอินคลัสเตอร์ดิสแทนท์ (Sum Of The Squared Within- Cluster Distance) โดยจะรวมคลัสเตอร์ที่ทำให้ซัมสแควออฟเดอะสแควิทอินคลัสเตอร์ดิสแทนท์ของคลัสเตอร์เพิ่มขึ้นน้อยที่สุด

การพิจารณาเลือกจำนวนกลุ่มที่เหมาะสม ผลลัพธ์ของเทคนิคคลัสเตอร์ไม่ได้ให้ค่าสถิติ หรือผลการทดสอบ สมมติฐานเพื่อให้ตัดสินใจหาจำนวนกลุ่มที่เหมาะสมต้องพิจารณาความเหมาะสมเอง โดยอาจใช้ระยะห่างหรือความคล้ายโดยใช้

1. การใช้เดนโดแกรม (Dendogram) สำหรับ Dendogram ถ้ากำหนดระยะห่าง ระหว่างกลุ่ม เป็นหน่วยที่แตกต่างกันไปก็จะได้จำนวนคลัสเตอร์ ที่แตกต่างกันไป คือยิ่งระยะห่างยิ่งมาก จำนวนคลัสเตอร์ ก็จะเพิ่มขึ้น

2. มัลติไดเมนซันสเกลิ่ง (Multidimension Scaling)

3. ดิสครีมิแนนท์ (Discriminant)

2.6 เทคนิคต่าง ๆ ของ การทำเหมืองข้อมูล [7]

2.6.1. การหากฎความสัมพันธ์ (Association Rule Discovery)

เป็นเทคนิคหนึ่งของการทำเหมืองข้อมูลที่สำคัญ และสามารถนำไปประยุกต์ใช้ได้จริงกับงานต่าง ๆ หลักการทำงานของวิธีนี้ คือ การค้นหาความสัมพันธ์ของข้อมูลจากข้อมูลขนาดใหญ่ที่มีอยู่เพื่อนำไปใช้ในการวิเคราะห์ หรือทำนายปรากฏการณ์ต่าง ๆ หรือมากจากการวิเคราะห์การซื้อสินค้าของลูกค้า (Market Basket Analysis) ซึ่งประเมินจากข้อมูลในตารางที่รวบรวมไว้ ผลการวิเคราะห์ที่ได้จะเป็นคำตอบของปัญหาซึ่งการวิเคราะห์แบบนี้เป็นการใช้ กฎความสัมพันธ์ (Association Rule) เพื่อหาความสัมพันธ์ของข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง ได้แก่ ระบบแนะนำหนังสือให้กับลูกค้าแบบอัตโนมัติของอเมซอน (Amazon) ข้อมูลการสั่งซื้อทั้งหมดของอเมซอน ซึ่งมีขนาดใหญ่มากจะถูกนำมาประมวลผลเพื่อหาความสัมพันธ์ของข้อมูล คือลูกค้าที่ซื้อหนังสือเล่มหนึ่ง ๆ มักจะซื้อหนังสือเล่มใดพร้อมกันด้วยเสมอ ความสัมพันธ์ที่ได้จากกระบวนการนี้จะสามารถนำไปใช้คาดเดาได้ว่าควรแนะนำหนังสือเล่มใดเพิ่มเติมให้กับลูกค้าที่เพิ่งซื้อหนังสือจากร้าน ตัวอย่างเช่น

buys (x , database) -> buys (x , การทำเหมืองข้อมูล) [80% , 60%]

ภาพที่ 2.8 ตัวอย่างการนำเทคนิคนี้ไปประยุกต์ใช้กับงานจริง

จากภาพที่ 2.8 หมายความว่า เมื่อซื้อหนังสือฐานข้อมูล (database) แล้วมีโอกาสที่จะซื้อหนังสือ การทำเหมืองข้อมูล ด้วย 60 % และมีการซื้อทั้งหนังสือ database และหนังสือ การทำเหมืองข้อมูล พร้อม ๆ กัน 80 %

อีกตัวอย่างคือ ในการซื้อสินค้าของลูกค้า 1 ครั้ง โดยไม่ต้องจำกัดว่าจะซื้อสินค้าในห้างร้าน หรือสั่งผ่านทางไปรษณีย์ หรือการซื้อสินค้าจากร้านค้าเสมือนจริง (Virtual store) บนเว็บไซต์โดยปกติจะต้องการทราบว่าสินค้าใดบ้างที่ลูกค้ามักซื้อด้วยกัน เพื่อนำไปพิจารณาปรับปรุงการจัดวางสินค้าในร้าน หรือใช้เพื่อหาวิธีวางรูปคู่กันในใบโฆษณาสินค้า ก่อนอื่นขอกำหนดคำว่า

กลุ่มรายการ (Itemset) หมายถึง กลุ่มสินค้าที่ปรากฏร่วมกัน เช่น {รองเท้า , ถุงเท้า}, {ปากกา, หมึก} หรือ {นม , น้ำผลไม้} โดยกลุ่มรายการดังกล่าวนี้ อาจจะจับคู่กลุ่มลูกค้ากับสินค้าก็ได้เช่น

วิเคราะห์หา ลูกค้าที่ซื้อสินค้าบางชนิดซ้ำ ๆ กัน อย่างน้อย 5 ครั้งแล้ว กรณีนี้ฐานข้อมูลมีการเก็บรายการซื้อขายเป็นจำนวนมาก และคำถามข้างต้น (Query) นี้จำเป็นต้องค้นหาทุก ๆ คู่ของลูกค้ากับสินค้า เช่น { คุณ ก , สินค้า A} , { คุณ ก , สินค้าB} , {คุณ ก , สินค้า C } , { คุณ ข , สินค้า B} เป็นต้น นับเป็นงานที่หนักพอสมควรสำหรับ ระบบการจัดการฐานข้อมูล และถ้าจะเขียนภาษาเอสคิวแอล สำหรับการเลือกข้อมูลตามเงื่อนไขข้างบน(query) จะได้ว่าตามภาพที่ 2.9

```
SELECT    P.custid , P.item , SUM( P.qty )
FROM      Purchases P
GROUP  BY P.custid , P.item
HAVING    SUM(P.qty ) > 5
```

ภาพที่ 2.9 การเขียนภาษาเอสคิลแอลสำหรับการเลือกข้อมูลตามเงื่อนไข

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

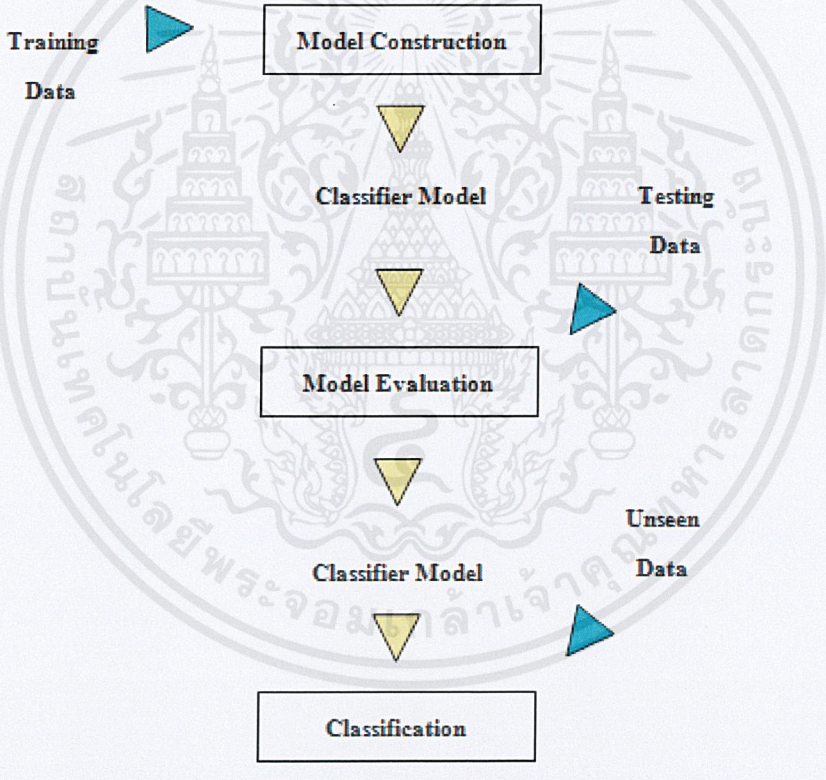
สำนักหอสมุดกลาง พระจอมเกล้าลาดกระบัง

หลังจากที่ ระบบการจัดการฐานข้อมูล ประมวลผลเอสคิลแอลนี้ เนื่องจากมีข้อมูลที่ต้องตรวจสอบมากมายหลายคู่และแต่ละคู่ต้องค้นหามาจากฐานข้อมูลเลย แต่ผลลัพธ์ของคิวรี่แบบนี้ มักจะมีจำนวนน้อยมาก เลยเรียกคิวรี่ชนิดนี้ว่าเป็น ไอซ์เบิร์กคิวรี่ (Iceberg Query) ซึ่งเปรียบกับสำนวนไทย คืองมเข็มในมหาสมุทร

2.6.2. การจำแนกข้อมูล และการทำนาย (Classification & Prediction)

การจำแนกข้อมูล (Classification) เป็นกระบวนการสร้างแบบจำลองข้อมูล โดยจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ ตัวอย่างเช่น

จัดกลุ่มนักเรียนออกเป็น 4 กลุ่ม คือ ดีมาก ดี ปานกลาง และไม่ดี โดยพิจารณาจากประวัติและผลการเรียน หรืออาจจะเป็นการแบ่งประเภทของลูกค้าว่าเชื่อถือได้ หรือไม่โดยพิจารณาจากข้อมูลที่มีอยู่ กระบวนการจำแนกข้อมูลนี้แบ่งออกเป็น 3 ขั้นตอน ดังภาพที่ 2.10



ภาพที่ 2.10 ขั้นตอนการทำการจำแนกข้อมูล [8]

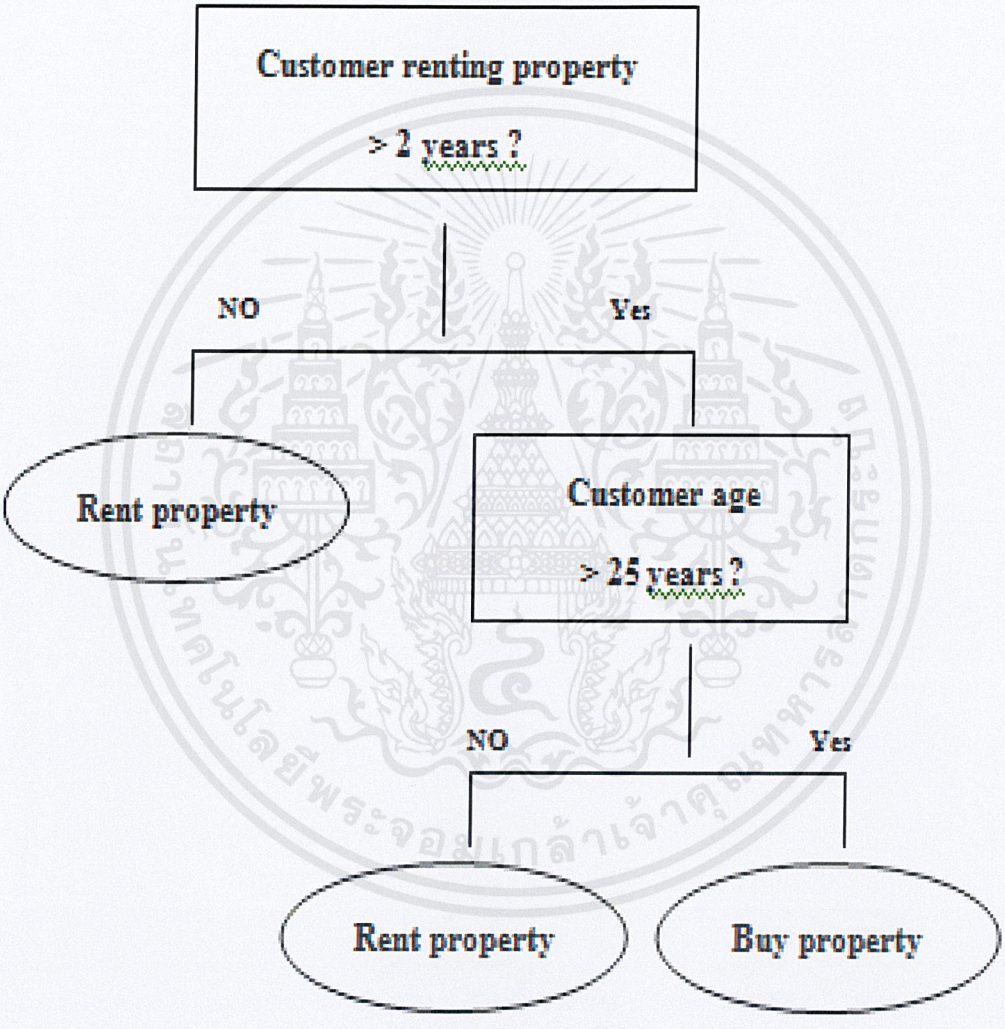
1. โครงสร้างแบบต้นไม้ (Decision Tree) เป็นที่นิยมกันมากเนื่องจากเป็นลักษณะที่คนจำนวนมากคุ้นเคย ทำให้เข้าใจได้ง่าย มีลักษณะเหมือนแผนภูมิองค์กร โดยที่แต่ละโหนดแสดงตัวแปร (Attribute) แต่ละกิ่งแสดงผลในการทดสอบ และลีฟโหนดแสดงคลาสที่กำหนดไว้

สมมติว่าบริษัทขนาดใหญ่แห่งหนึ่ง ทำธุรกิจอสังหาริมทรัพย์มีสำนักงานสาขาอยู่ประมาณ 50 แห่ง แต่ละสาขามีพนักงานประจำ เป็นผู้จัดการและพนักงานขาย พนักงานเหล่านี้แต่ละคนจะ ดูแลอาคารต่าง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

148615

ๆ หลายแห่งรวมทั้งลูกค้าจำนวนมาก บริษัทจำเป็นต้องใช้ระบบฐานข้อมูลที่กำหนดความสัมพันธ์ระหว่างองค์ประกอบเหล่านี้ เมื่อรวบรวมข้อมูลแบ่งเป็นตารางพื้นฐานต่าง ๆ เช่น ข้อมูลสำนักงานสาขา (Branch) ข้อมูลพนักงาน (Staff) ข้อมูลทรัพย์สิน (Property) และข้อมูลลูกค้า (Client) พร้อมทั้งกำหนดความสัมพันธ์ (Relationship) ของข้อมูลเหล่านี้ เช่น ประวัติการเช่าบ้านของลูกค้า (Customer Rental) รายการให้เช่า (Rentals) รายการขายสินทรัพย์ (Sales) เป็นต้น ต่อมาเมื่อมีประชุมกรรมการผู้บริหารของบริษัท ส่วนหนึ่งของรายงานจากฐานข้อมูลสรุปว่า “40 % ของลูกค้าที่เช่าบ้านนานกว่าสองปี และมีอายุเกิน 25 ปี จะซื้อบ้านเป็นของตนเอง โดยกรณีเช่นนี้เกิดขึ้น 35 % ของลูกค้าผู้เช่าบ้านของบริษัท”



ภาพที่ 2.11 โครงสร้างแบบต้นไม้เพื่อวิเคราะห์โอกาสที่ลูกค้าบ้านเช่าจะซื้อบ้าน [9]

จากภาพที่ 2.11 แสดงให้เห็นถึงโครงสร้างแบบต้นไม้ สำหรับการวิเคราะห์ว่าลูกค้าบ้านเช่าจะมีความสนใจที่จะซื้อบ้านเป็นของตนเองหรือไม่ โดยใช้ปัจจัยในการวิเคราะห์คือ ระยะเวลาที่ลูกค้าได้เช่าบ้านมาและอายุของลูกค้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.1 แสดงข้อมูลเกี่ยวกับลูกค้า

Age	Rent Period	Buy
23	3	No
36	1.5	No
20	1.5	No
27	2	Yes
20	1	No
50	2.5	Yes
36	1	No
36	2	Yes
22	2.5	No

จากตารางที่ 2.1 แสดงถึงรายการทั้งหมดเกี่ยวกับลูกค้าบ้านเช่าของบริษัทโดยมีรายละเอียดเกี่ยวกับอายุ และระยะเวลาเช่า รวมทั้งการซื้อบ้านของลูกค้าแต่ละราย

- เอสคิวแอล (SQL) สำหรับโครงสร้างแบบต้นไม้ ของตัวอย่างนี้แบ่งเป็น 2 ชุด สำหรับปัจจัยแต่ละอย่าง เอสคิลแอลสำหรับรากโหนด (Root Node) แสดงออกมาได้ดังภาพที่ 2.13 ดังนี้

```
SELECT  B.Rent_Period , B.Buy , COUNT(*)

FROM      Business_Info B

WHERE   B.Rent_Period > 2

GROUP BY  B.Rent_Period , B.Buy
```

ภาพที่ 2.12 การเขียนภาษาเอสคิลแอลสำหรับรากโหนด

ผลลัพธ์ของเอสคิลแอล นี้แสดงได้ดังตารางที่ 2.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 2.2 ผลลัพธ์ของเอสคิลแอลสำหรับรากโหนด

Rent Period	Buy	Yes	No
1	0	2	
1.5	0	2	
2	2	0	
2.5	1	1	
3	0	1	

- เอสคิลแอลสำหรับโหนด (Node) ที่เป็นลูก (Child) ทางขวาของราก (Root) แสดงได้ดังภาพที่ 2.15

```
SELECT   B.Age , B.Buy , COUNT(*)
FROM     Business_Info B
WHERE    B.Age > 25
GROUP  BY  B.Age , B.Buy
```

ภาพที่ 2.13 เอสคิลแอลสำหรับโหนดลูก

ผลลัพธ์ของเอสคิลแอลนี้แสดงได้ดังตารางที่ 2.3

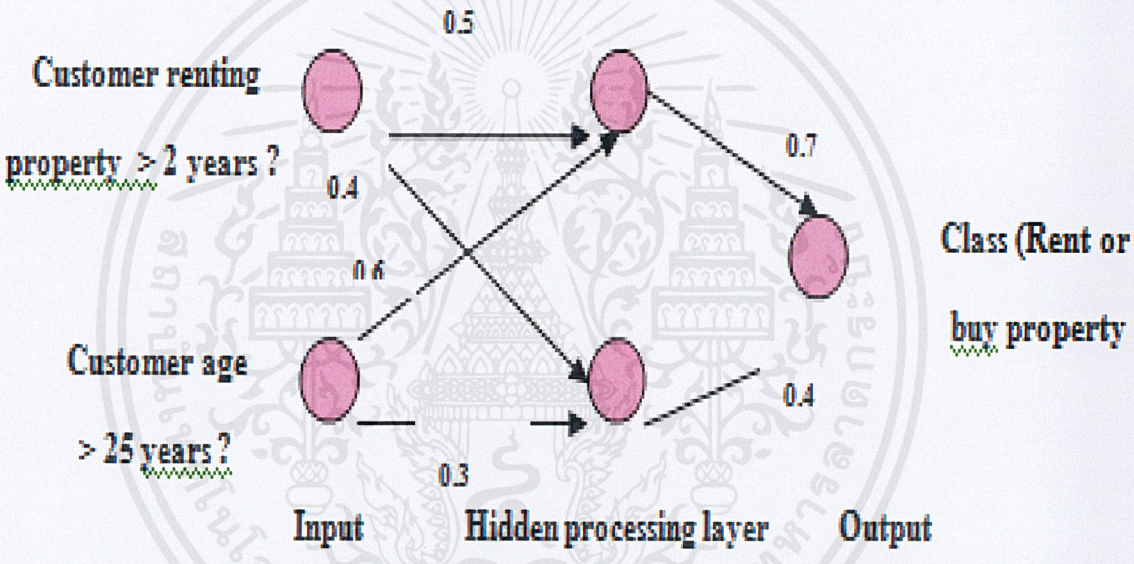
ตารางที่ 2.3 ตารางแสดงผลลัพธ์โหนดที่เป็นลูกทางขวาของราก

Rent Period	Buy	Yes	No
20	0	2	
22	0	1	
23	2	1	
27	1	0	
36	1	2	
50	1	0	

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผลลัพธ์ที่ได้จากแต่ละโหนดของ Decision Tree เรียกว่า เอทีวีเซท(AVC set) จากตัวอย่างข้างต้นจะเห็นว่ามี 2 AVC sets เพื่อใช้ในการจัดกลุ่มลูกค้า

2. เครือข่ายประสาทเทียม (Neural Net) เป็นเทคโนโลยีที่มีที่มาจากงานวิจัยด้านปัญญาประดิษฐ์ (Artificial Intelligence:AI) เพื่อใช้ในการคำนวณค่าฟังก์ชันจากกลุ่มข้อมูล วิธีการของครือข่ายประสาทเทียม เป็นวิธีการที่ให้เครื่องเรียนรู้จากตัวอย่างต้นแบบแล้วฝึก (Train) ให้ระบบได้รู้จักที่จะคิดแก้ปัญหาที่กว้างขึ้นได้ ในโครงสร้างของครือข่ายประสาทเทียม จะประกอบด้วยโหนด (Node) สำหรับอินพุต-เอาท์พุตและการประมวลผล กระจายอยู่ในโครงสร้างเป็นชั้น ๆ ได้แก่ อินพุทเลเยอร์ เอาท์พุทเลเยอร์ และ ฮิดเดนเลเยอร์ การประมวลผลของเครือข่ายประสาทเทียมจะอาศัยการส่งการทำงานผ่านโหนดต่าง ๆ ในเลเยอร์เหล่านี้ สำหรับภาพที่ 2.10 เป็นการวิเคราะห์การเช่าซื้อบ้านของลูกค้าในโครงสร้างแบบครือข่ายประสาทเทียม



ภาพที่ 2.14 ครือข่ายประสาทเทียมเพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า [9]

3. การประมาณความถูกต้อง (Accuracy) เป็นขั้นการประมาณความถูกต้องโดยอาศัยข้อมูลที่ใช้ทดสอบ (Testing Data) ซึ่งคลาสที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับคลาสที่หามาได้จาก โมเดล เพื่อทดสอบความถูกต้อง

4. การจำแนกข้อมูล (Classification) เป็นโมเดลสำหรับใช้ข้อมูลที่ไม่เคยเห็นมาก่อน (Unseen Data) โดยจะทำการกำหนดคลาสให้กับออบเจ็กต์ใหม่ที่ได้มา หรือทำนายค่าออกมาตามที่ต้องการ

5. การทำนาย (Prediction) เป็นการทำนายหาค่าที่ต้องการจากข้อมูลที่มีอยู่ ตัวอย่างเช่น หายอดขายของเดือนถัดไปจากข้อมูลที่มีอยู่ หรือทำนายโรคจากอาการของคนไข้ในอดีต เป็นต้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.6.3. การแบ่งส่วน (Segmentation)

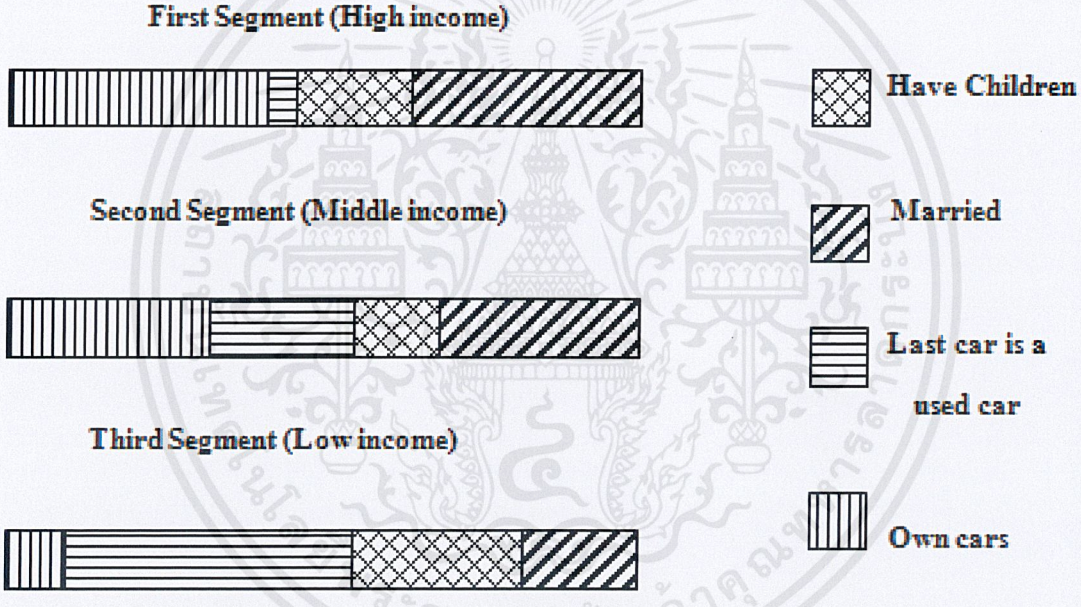
เป็นเทคนิคการลดขนาดของข้อมูลด้วยการรวมกลุ่มตัวแปรที่มีลักษณะเดียวกันไว้ด้วย กัน ตัวอย่างเช่น บริษัทจำหน่ายรถยนต์ได้แยกกลุ่มลูกค้าออกเป็น 3 กลุ่ม คือ

1.กลุ่มผู้มีรายได้สูง (>$80,000)

2.กลุ่มผู้มีรายได้ปานกลาง ($25,000 to $ 80,000)

3.กลุ่มผู้มีรายได้ต่ำ (Less Than $25,000) และภายในแต่ละกลุ่มยังแยกออกเป็น

- มีครอบครัว
- แต่งงานแล้ว
- เคยมีรถส่วนตัว
- มีรถส่วนตัว



ภาพที่ 2.15 ตัวอย่างผลลัพธ์การทำคลัสเตอร์ [10]

จากข้อมูลข้างต้นทำให้ทางบริษัทรู้ว่าเมื่อมีลูกค้าเข้ามาที่บริษัทควรจะเสนอขายรถประเภทใด เช่น ถ้าเป็นกลุ่มผู้มีรายได้สูงควรจะเสนอรถใหม่ เป็นรถครอบครัวขนาดใหญ่พอสมควร แต่ถ้าเป็นผู้มีรายได้ค่อนข้างต่ำควรเสนอรถมือสอง ขนาดค่อนข้างเล็ก

2.6.4 ดิเวียชันดีเทคชันลิงค์อนาไลซิส (Deviation Detection Link Analysis)

เป็นกรรมวิธีในการหาค่าที่แตกต่างไปจากค่ามาตรฐาน หรือค่าที่คาดคิดไว้ว่าต่างไปมากน้อยเพียงใด โดยทั่วไปมักใช้วิธีการทางสถิติ หรือการแสดงให้เห็นภาพ (Visualization) สำหรับเทคนิคนี้ใช้ในการตรวจสอบลายเซ็นปลอม หรือบัตรเครดิตปลอม รวมทั้งการตรวจหาจุดบกพร่องของชิ้นงานในโรงงานอุตสาหกรรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.6.5 ลิงค์อนาไลซิส (Link Analysis)

จุดมุ่งหมายของการวิเคราะห์ลิงค์ (Link Analysis) คือการสร้างลิงค์ที่รียกว่า กฎความสัมพันธ์ (Associations) ระหว่างเรคคอร์ดเดียว หรือกลุ่มของเรคคอร์ดในฐานข้อมูล

2.7 ข้อมูลขนาดใหญ่ (Big Data) [11]

ข้อมูลขนาดใหญ่ หมายถึงปริมาณข้อมูลที่มีขนาดใหญ่มหาศาลเกินกว่าขีดความสามารถในการประมวลผลของระบบฐานข้อมูลธรรมดาที่จะรองรับได้ปริมาณข้อมูลที่มีขนาดใหญ่มาก ๆ จะมีอัตราการเพิ่มข้อมูลได้อย่างรวดเร็วมากและจะมีรูปแบบที่ไม่มีโครงสร้างหรือกึ่งโครงสร้าง ซึ่งไม่สามารถอยู่ในระบบฐานข้อมูลที่จะจัดเก็บข้อมูลได้

ข้อมูลขนาดใหญ่สามารถวิเคราะห์ได้ด้วยเครื่องมือที่มีการใช้งานกันอย่างแพร่หลาย เช่น ฮาดูป (Hadoop) โดยจะถือว่าเป็นส่วนหนึ่งของการวิเคราะห์เชิงลึก เช่น การวิเคราะห์เพื่อคาดการณ์ล่วงหน้า การทำเหมืองข้อมูล (Data Mining) การวิเคราะห์ข้อความ และการวิเคราะห์ทางสถิติ

อุปสรรคต่อองค์กรในการจัดตั้งส่วนการวิเคราะห์ข้อมูลแบบข้อมูลขนาดใหญ่ก็คือการขาดทักษะความเข้าใจในการวิเคราะห์ข้อมูลของบุคคลภายใน และค่าใช้จ่ายที่สูงในการจ้างผู้เชี่ยวชาญที่มีประสบการณ์ในการวิเคราะห์ข้อมูล ปริมาณของข้อมูลที่มีส่วนเกี่ยวข้องที่ต้องนำมาวิเคราะห์ก็อาจจะมีหลากหลายประเภทและหลากหลายแหล่งข้อมูลซะจนทำให้การจัดการข้อมูลเป็นไปอย่างยากลำบาก รวมไปถึงคุณภาพของข้อมูลที่นำไปสู่ประเด็นปัญหาเรื่องความเกี่ยวเนื่องของข้อมูลที่ถูกนำมาใช้ ยิ่งไปกว่านั้นการเชื่อมต่อระหว่างฮาดูปและระบบจัดเก็บข้อมูล (Data Warehouses) ก็ยังเป็นอะไรที่มีความท้าทาย แม้ว่าในปัจจุบันจะมีการคิดค้นซอฟต์แวร์ที่เป็นตัวกลางเชื่อมระหว่างฮาดูป และระบบฐานข้อมูลเชิงสัมพันธ์ รวมไปถึงเครื่องมือเชื่อมต่อเพื่อจัดทำระบบข้อมูลขนาดใหญ่อื่น ๆ เกิดขึ้นมามากมายก็ตาม

2.7.1 คุณลักษณะ

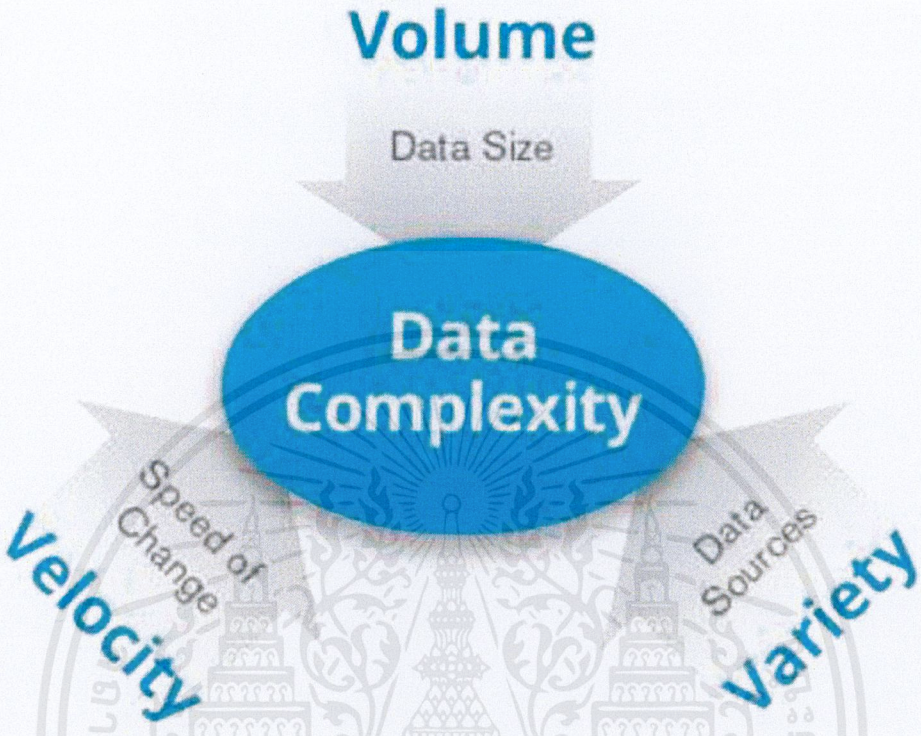
1. ปริมาตร (Volume) หมายถึงข้อมูลที่มีปริมาณมหาศาลซึ่งโครงสร้างข้อมูลของระบบฐานข้อมูลไม่สามารถจัดเก็บข้อมูลได้ปริมาณข้อมูลมหาศาลมีประโยชน์ เพื่อเป็นข้อมูลที่ใช้ในการตัดสินใจ หรือทำนายอนาคต หรือเพื่อเตรียมการวางแผนการทางานเชิงรุกทางธุรกิจ

2. ความเร็ว (Velocity) หมายถึงอัตราการเพิ่มขึ้นของข้อมูล ข้อมูลที่เข้าสู่ระบบฐานข้อมูลจะมีอัตราการเพิ่มขึ้นอย่างรวดเร็ว เช่น ข้อมูลที่เกิดขึ้นจากโทรศัพท์เคลื่อนที่ที่ผู้ใช้นาขึ้นเก็บเป็นข้อมูลภาพถ่าย ข้อมูลการพิมพ์การสนทนาข้อมูลการอัดภาพ วิดีโอหรือข้อมูลการอัดเสียง หรือแม้กระทั่งข้อมูลการสั่งซื้อสินค้า การขนส่ง และการบริการต่าง ๆ สามารถนาข้อมูลเหล่านั้นเข้าสู่ระบบฐานข้อมูลได้อย่างรวดเร็ว

3. รูปแบบที่หลากหลาย (Variety) หมายถึงรูปแบบมีความหลากหลายของรูปแบบข้อมูล ซึ่งอาจจะเป็นรูปแบบที่มีโครงสร้างไม่มีโครงสร้างและกึ่งมีโครงสร้างเป็นต้น รูปแบบที่ไม่มีโครงสร้างหรือกึ่งโครงสร้างจะไม่เหมือนข้อมูลที่จัดเก็บไว้ในระบบฐานข้อมูลเช่น ข้อความ อีเมล์รูปภาพ วิดีโอและเสียง เป็นต้น ซึ่งข้อมูลเหล่านี้มีความซับซ้อนและเชื่อมโยงกัน ข้อมูลมีรูปแบบที่หลากหลาย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

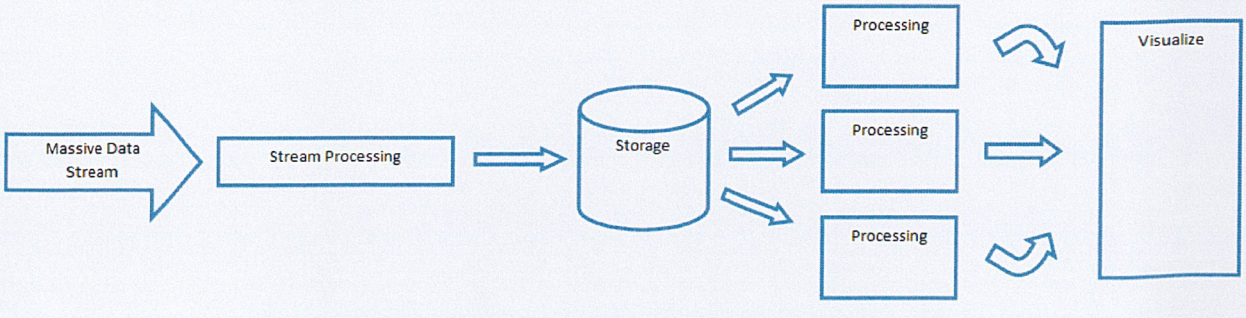
โดยลักษณะที่กล่าวมาข้างต้นนี้แสดงได้ดังภาพที่ 2.16



ภาพที่ 2.16 องค์ประกอบของข้อมูลขนาดใหญ่ [12]

2.7.2 การจัดการข้อมูลขนาดใหญ่โดยรูปแบบการกระจายข้อมูล

เมื่อขนาดของข้อมูลมีปริมาณที่ใหญ่โตมหาศาลที่เรียกว่า ข้อมูลขนาดใหญ่ จึงต้องหาวิธีที่สามารถประมวลผลข้อมูลที่มีปริมาณมหาศาลและข้อมูลนั้นไม่สามารถเคลื่อนย้ายได้
การจัดการข้อมูลที่เป็นข้อมูลขนาดใหญ่มีองค์ประกอบดังต่อไปนี้



ภาพที่ 2.17 องค์ประกอบการจัดการข้อมูลที่เป็นข้อมูลขนาดใหญ่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากภาพที่ 2.17 จะแสดงรูปแบบการแสดงการจัดการข้อมูลที่เป็นข้อมูลขนาดใหญ่ ได้แก่

1. การจัดเก็บ (Storage)
2. การประมวลผล (Processing)
3. การวิเคราะห์ (Analysis Algorithm)
4. การทำรายงานสรุป (Visualization)

จะมีการเคลื่อนย้ายกลุ่มของข้อมูลเข้าสู่ระบบฐานข้อมูลเพื่อการประมวลผล และนำไปเก็บในส่วนจัดเก็บข้อมูล ซึ่งจะเก็บปริมาณข้อมูลมหาศาลที่ไม่สามารถเคลื่อนย้ายได้ทั้งนี้ประมวลผลจะต้องเคลื่อนย้ายเข้ามาประมวลผลของข้อมูลให้อยู่ในมุมมองต่าง ๆ ที่ผู้ใช้ต้องการในปัจจุบันได้มีการพัฒนาซอฟต์แวร์อิลาสติกเสิร์ช (Elasticsearch) เข้ามาช่วยจัดการปริมาณข้อมูลมหาศาลนี้

2.8 สวิตช์ (Switch) [13]

เป็นอุปกรณ์เครือข่ายเช่นเดียวกันกับฮับ (Hub) และมีหน้าที่คล้ายกับฮับมาก แต่มีความแตกต่างที่ในแต่ละพอร์ต (Port) จะมีความสามารถในการส่งข้อมูลได้สูงกว่า เช่น สวิตช์ที่มีความเร็ว 10 Mbps นั้นจะหมายความว่าในแต่ละพอร์ตจะสามารถส่งข้อมูลได้ที่ความเร็ว 10 Mbps และนอกจากนั้นเครื่องทุกเครื่องที่ต่อมายังสวิตช์ยังไม่ได้อยู่ในคอลลิชันโดเมนเดียวกันด้วย (ซึ่งถ้าฮับจะอยู่ในคอลลิชันโดเมนเดียวกัน) นั่นหมายความว่าแต่ละเครื่องจะได้ครอบครองสายสัญญาณแต่เพียงผู้เดียว จะไม่เกิดปัญหาการแย่งสายสัญญาณและการชนกันของสัญญาณเกิดขึ้น

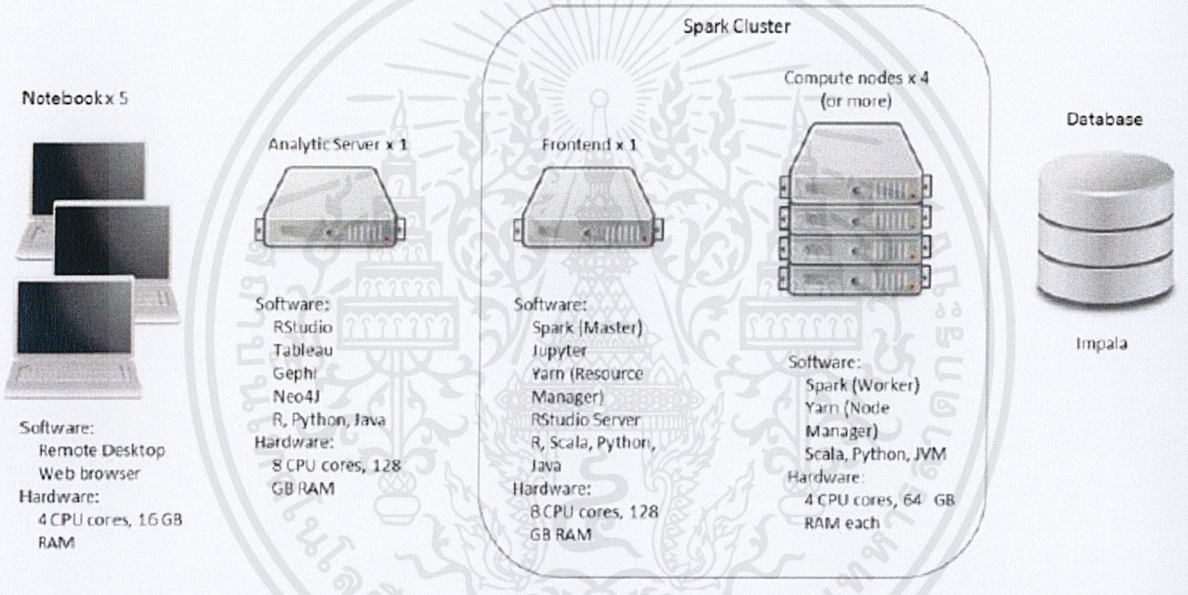
สวิตช์จะมีความสามารถมากกว่าฮับ แต่ยังมีการใช้งานอยู่ในวงจำกัดเพราะราคายังค่อนข้างสูงกว่าฮับอยู่มาก ดังนั้นจึงมีการนำสวิตช์มาใช้ในระบบเครือข่ายที่ต้องการแบ่งโดเมนเพื่อเพิ่มความเร็วในการติดต่อกับระบบ โดยอาจนำสวิตช์มาเป็นศูนย์กลาง และใช้ต่อเข้ากับเครื่องที่มีการเชื่อมต่อกับเครื่องอื่น ๆ เป็นจำนวนมาก ซึ่งส่วนใหญ่จะเป็นเครื่องเซิร์ฟเวอร์ เพื่อจะได้ส่งข้อมูลได้ทีละมาก ๆ และส่งด้วยความเร็วสูง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีการดำเนินการวิจัย

เมื่อได้ศึกษาสภาพแวดล้อมและสิ่งต่าง ๆ ภายในองค์กรจนมีความรู้ความเข้าใจในกระบวนการทำงานตลอดจนถึงข้อจำกัดและข้อพึงระวัง หลังจากนั้นจึงได้ริเริ่มการวางแผนการทำงานขึ้นมา โดยมีการปรับเปลี่ยนแผนงานไปเรื่อย ๆ อันเนื่องมาจากข้อจำกัดด้านความปลอดภัยภายในองค์กร เพราะการทำงานกับข้อมูลค่อนข้างต้องทำในระบบปิดและปลอดภัย ดังนั้นในส่วนแรกของการวางแผนการทำงานจะใช้เวลาส่วนใหญ่ไปกับเรื่องความปลอดภัยโดยเริ่มจากขั้นตอนต่าง ๆ ดังนี้

3.1 การเตรียมโครงสร้างพื้นฐาน (Infrastructure)



ภาพที่ 3.1 โครงสร้างพื้นฐานของการทำเหมืองข้อมูลรวมถึงอุปกรณ์และเงื่อนไขต่าง ๆ

เนื่องจากการทำงานกับข้อมูลจะเป็นต้องทำในระบบปิด ดังนั้นจึงต้องมีการเตรียม infrastructure เฉพาะสำหรับงานด้านนี้ เพื่อป้องกันข้อมูลรั่วไหลอันจะนำมาสู่ความเสียหายต่าง ๆ โดยในการเตรียมโครงสร้างพื้นฐาน สิ่งแรกที่คำนึงถึงคือเรื่องของความปลอดภัยดังนั้นในช่วงแรกจะเป็นการทำเรื่องขอใช้โน้ตบุ๊กจากส่วนกลางจำนวน 5 เครื่อง เนื่องจากโปรเจคนี้แบ่งออกเป็นหลายส่วน การวิเคราะห์โซเชียลมีเดียที่เกี่ยวกับเอ็นเตอร์เทนเมนต์นี้นั้นก็เป็นส่วนหนึ่งของโปรเจคที่ทางเอไอเอสได้วางแผนเอาไว้ ซึ่งส่วนอื่น ๆ จะเป็นการวิเคราะห์การใช้งานโทรศัพท์มือถือ (Mobility Analysis) วิเคราะห์ความสนใจของผู้ใช้งาน (Interest Analysis) และวิเคราะห์โซเชียลมีเดีย (Social Analysis) ซึ่งจะเป็นทางอาจารย์และนิสิตจากจุฬาลงกรณ์มหาวิทยาลัยเป็นผู้ได้รับมอบหมายในส่วนอื่น ๆ

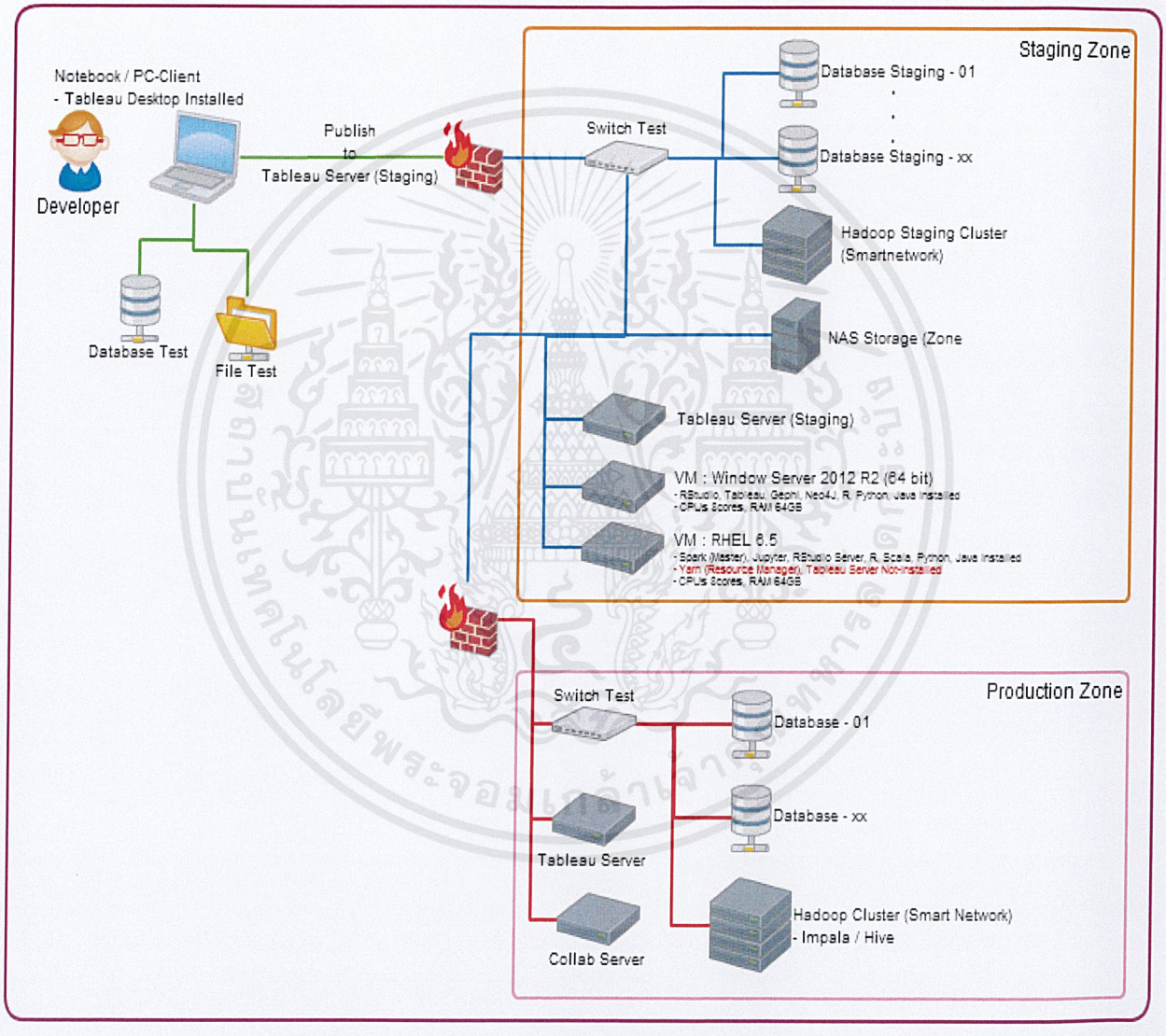
โดยการทำงานในส่วนเอ็นเตอร์เทนเมนต์มีรายละเอียดในการดำเนินงานดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1.1 การเข้าถึงข้อมูล

1. ปกติที่สเตจจิ้งไม่มีแทบโบลวเดสก์ทอป (Tableau Desktop) เพราะจะต้องติดตั้งแทบโบลวเดสก์ทอปบนเครื่องพีซีหรือโน้ตบุ๊ก (Notebook) ของผู้ใช้งานเอาเอง แล้วจึงทำการปล่อยขึ้นบนตัวแทบโบลวเซิร์ฟเวอร์สเตจจิ้ง (Tableau Server Staging)

2. ส่วนวีเอ็มเรดแฮช (VM RHEL) นั้นจะไม่ติดตั้งแทบโบลวเซิรฟเวอร์ (Tableau Server) เพราะเป็นเซอร์วิส (Service) ของเซิฟเวอร์ในระบบเดิมอยู่แล้ว



ภาพที่ 3.2 รูปประกอบการอธิบายการเข้าถึงข้อมูล

3.1.2 การเตรียมสภาพแวดล้อมการทำงาน (Environment)

1. เตรียมโน้ตบุ๊กสำหรับใช้งาน จำนวน 5 เครื่อง ตอนทำใช้แค่ 1 เครื่อง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. ทางไอทีซับพอร์ท (IT Support) จะเป็นผู้ตั้งค่าโน้ตบุ๊กให้ โดยในห้องมีแลน (LAN) พอร์ทเดียว ดังนั้นจะใช้เราเตอร์แยกออกเป็น 5 พอร์ทแทน เพื่อให้ทั้ง 5 เครื่องสามารถทำงานพร้อมกันได้

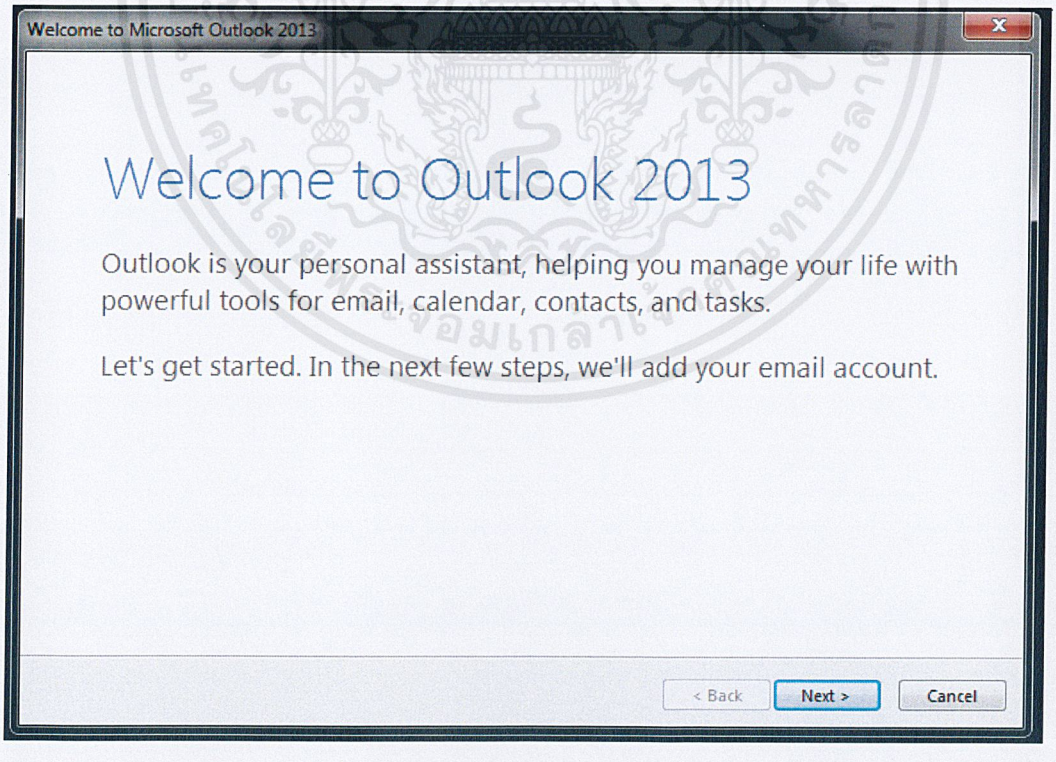
3. ไอทีซับพอร์ท (IT Support) จะทำการบล็อคพอร์ทต่าง ๆ ของโน้ตบุ๊กทุกตัว

4. ใช้แอลแดปยูสเซอร์ (LDAP User) ในการเข้าเครื่องโน้ตบุ๊กทำงานในระบบปิด คือไม่สามารถไปทำตรงสถานที่อื่นได้ และข้อมูลต่าง ๆ จะอยู่แค่ภายในห้องที่เตรียมไว้เท่านั้น

3.1.3 การตั้งค่าอีเมล์ (E-mail) ในอินทราเน็ต (Intranet) เพื่อติดต่อสื่อสารภายในองค์กร

ในที่นี้จะใช้เอาท์ลุค (Outlook) ซึ่งเป็นซอฟต์แวร์ที่ใช้ในบริษัท โดยในการส่งอีเมล์เพื่อติดต่อสื่อสารกันภายในองค์กร รวมถึงการส่งข้อมูลต่าง ๆ ออกข้างนอก เป็นโปรแกรมที่ใช้จัดการอีเมล์ในรูปแบบป๊อปเมล (Pop Mail) ที่มีรูปแบบการทำงานเร็วกว่าเว็บเมล (Web Mail) พร้อมทั้งความสามารถในการบันทึกรายชื่อผู้ที่เราติดต่อ เพื่อความสะดวกในการส่งเมล์ถึงกัน และอีกทั้งยังสนับสนุนการทำงานของผู้ใช้ในการที่จะบริหารจัดการเวลา นัดหมายต่าง ๆ จดบันทึกงานที่ต้องทำเอาไว้ และบริหารงานในแต่ละวันอย่างเป็นระบบอีกทั้งยังสนับสนุนการทำงานของผู้ใช้ในการที่จะบริหารจัดการเวลา นัดหมายต่าง ๆ จดบันทึกงานที่ต้องทำเอาไว้ และบริหารงานในแต่ละวันอย่างเป็นระบบซึ่งก่อนที่จะสามารถใช้งานอีเมล์ได้นั้น จะต้องมีการตั้งค่าให้อีเมล์สามารถรับและส่งข้อความได้โดยขั้นตอนการตั้งค่า มีดังนี้

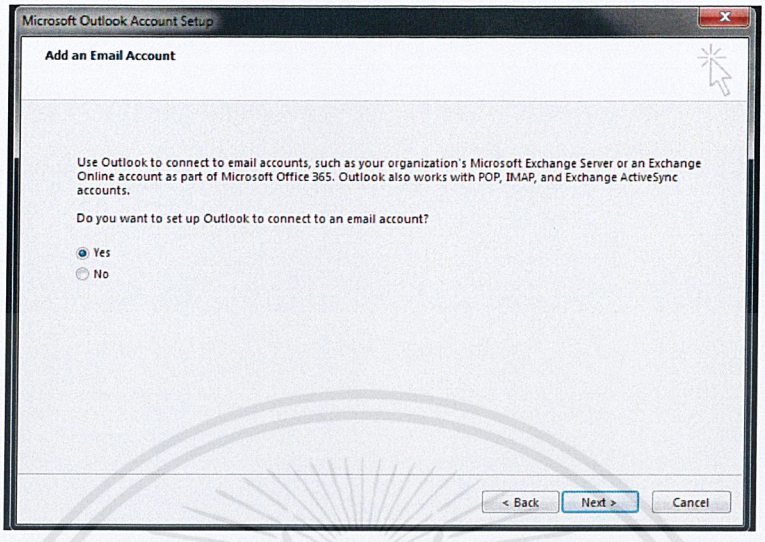
1. เปิดโปรแกรมเอาท์ลุค (Outlook 2013)



ภาพที่ 3.3 การเข้าใช้งานเอาท์ลุคครั้งแรกหลังจากคลิกไอคอนเอาท์ลุค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2. เลือกใช่ (Yes) ละคลิกถัดไป (Next) เพื่อที่จะไปหน้าเพิ่มผู้ใช้งาน

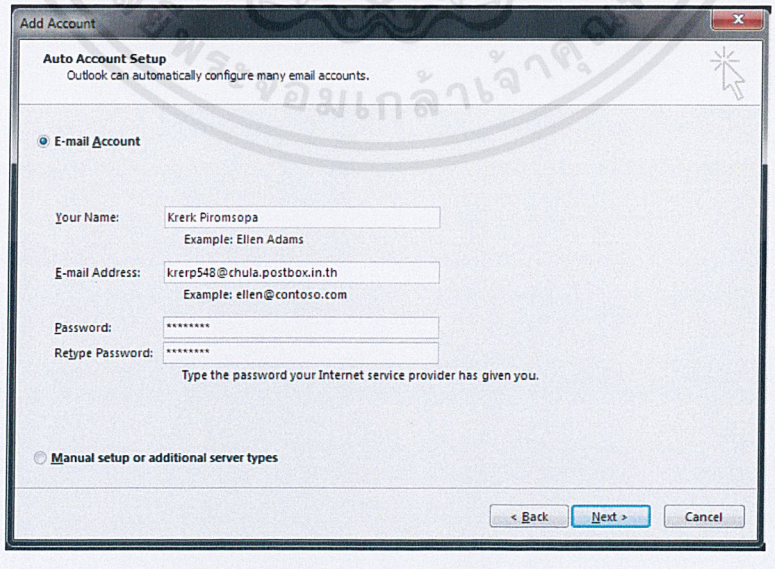


ภาพที่ 3.4 ความต้องการที่จะเพิ่มบัญชีอีเมล

3. เลือกบัญชีอีเมล (Email Account) ละกรอกข้อมูลตามช่องว่างให้ครบโดยที่

- ชื่อของผู้ใช้ (Your Name) คือชื่อที่จะปรากฏเวลาส่งอีเมลไปหาคนอื่นๆ
- ที่อยู่อีเมล (E-mail Address) คืออีเมลที่ต้องการใช้งานผ่านเอาท์ลุค (Outlook) โดยจากภาพที่ 3.5 จะเป็นชื่ออีเมลที่ทางฝ่ายวางแผนได้ขอมาไว้ให้สำหรับโปรเจคชิ้นนี้
- รหัส (Password) และรหัสอีกครั้ง (Retype Password) คือรหัสเข้าใช้อีเมล

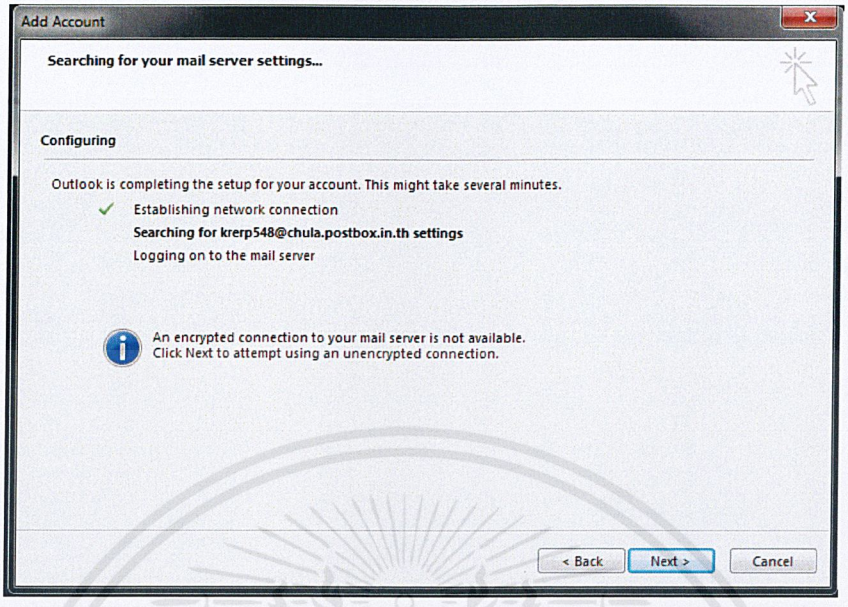
ซึ่งทั้งสองช่องจะต้องตรงกัน และเมื่อกรอกข้อมูลครบหมดแล้ว ก็กดถัดไป (Next) เพื่อไปขั้นตอนถัดไป



ภาพที่ 3.5 การกรอกข้อมูลการตั้งค่าบัญชีแบบอัตโนมัติ (Auto Account Setup)

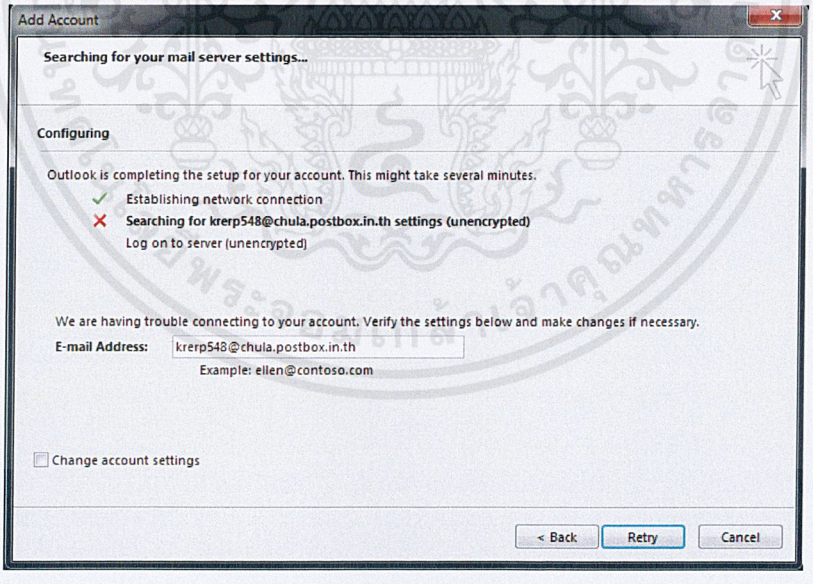
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4. หลังจากนั้นกดถัดไป (Next) ต่ออีกครั้ง



ภาพที่ 3.6 การติดตั้ง

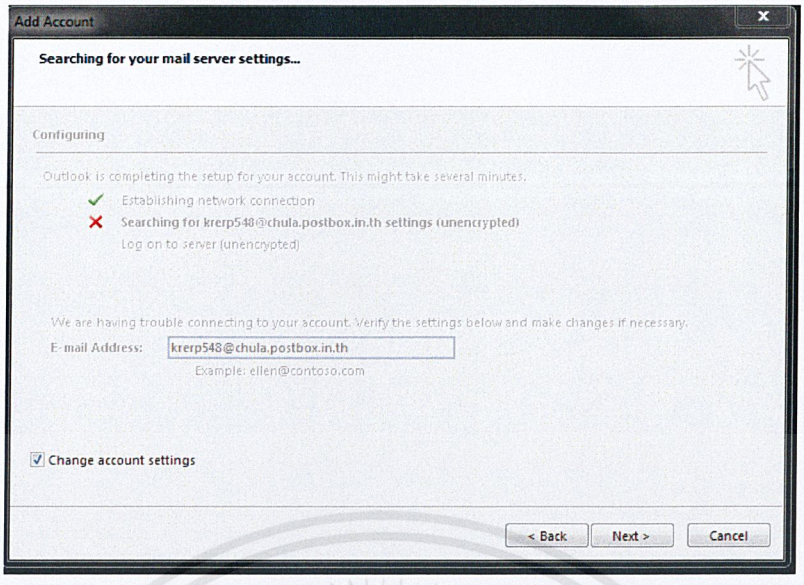
5. เมื่อมันเช็คเสร็จจะเห็นได้ว่าหัวข้อหาข้อมูลการตั้งค่าของอีเมล (Searching For Name_Email@Domain_Name_Email Setting) จะเกิดความผิดพลาดให้คลิกปุ่มลองอีกครั้ง (Retry)



ภาพที่ 3.7 ความผิดพลาดในการติดตั้ง

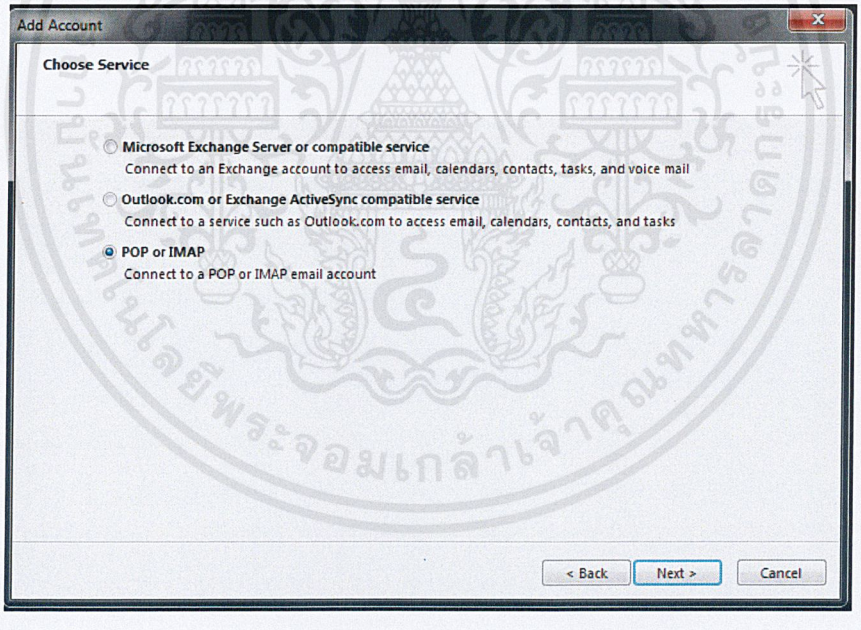
6. เมื่อกดลองอีกครั้ง (Retry) จะเห็นได้ว่าช่องเปลี่ยนการตั้งค่าบัญชี (Change Account Setting) มีเครื่องหมายถูกขึ้นมา หลังจากนั้นกดถัดไป (Next) เพื่อไปขั้นตอนถัดไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 3.8 การเปลี่ยนการตั้งค่าของผู้ใช้งาน

7. เลือกบัญชีแบบป๊อปหรือไอแมพ (POP or IMAP) เนื่องจากอีเมลที่ใช้สมัครอีเมลสำหรับเอาท์ซอส (Outsource) แล้วกดถัดไป (Next)



ภาพที่ 3.9 การเลือกใช้งานเซอร์วิสของอีเมล์

8. ตั้งค่าบัญชีแบบป๊อป (POP) และบัญชีแบบไอแมพ (IMAP Account ทั้งสองส่วน คือ

ส่วนที่ 1 คือ ข้อมูลทั่วไปของผู้ใช้ (User Information) ได้แก่

- ชื่อสำหรับอีเมล (Your Name)
- อีเมลที่จะใช้งาน (Email Address)

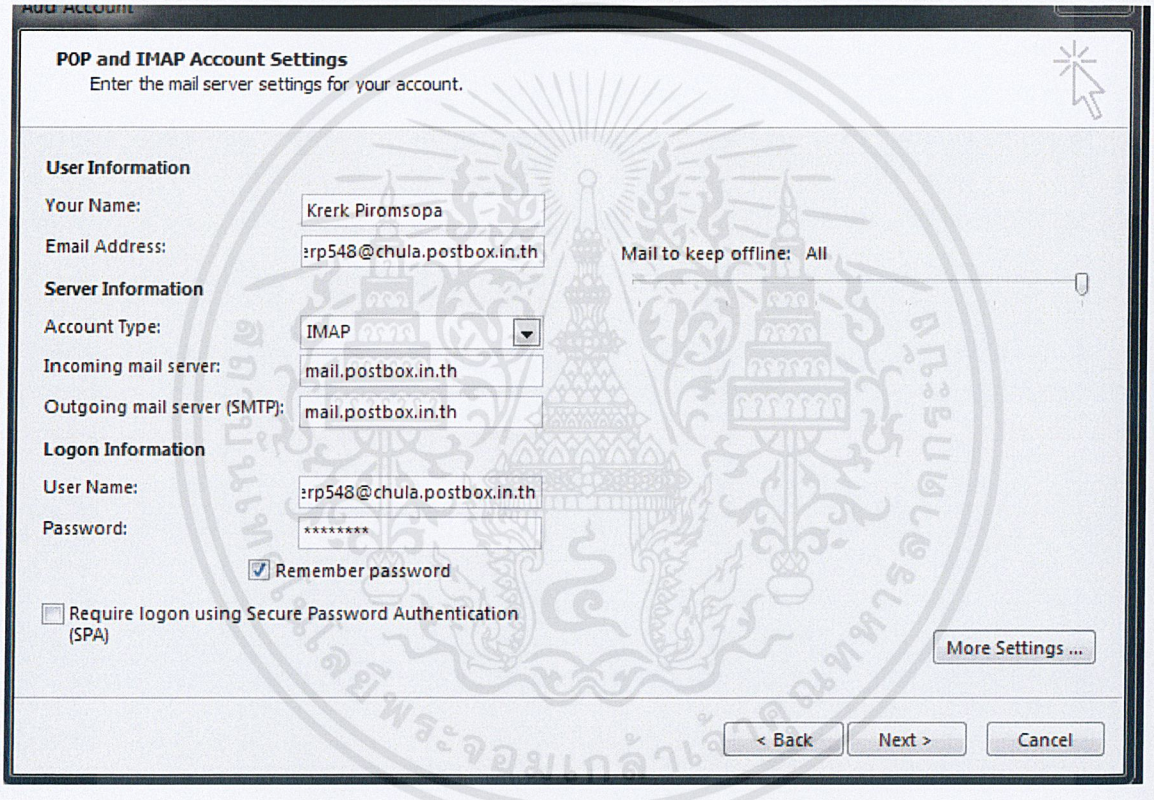
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ส่วนที่ 2 คือ ข้อมูลทั่วไปของเซิร์ฟเวอร์ (Server Information) ได้แก่

- ชนิดบัญชี (Account Type) มีให้เลือกเป็นไอแมพกับป๊อปทรี
- รูปแบบอีเมลที่รับเข้ามา (Incoming Mail Server)
- รูปแบบอีเมลที่ส่งออกไป (Outgoing Mail Server หรือ SMTP)

ส่วนที่ 3 คือ Logon Information ได้แก่

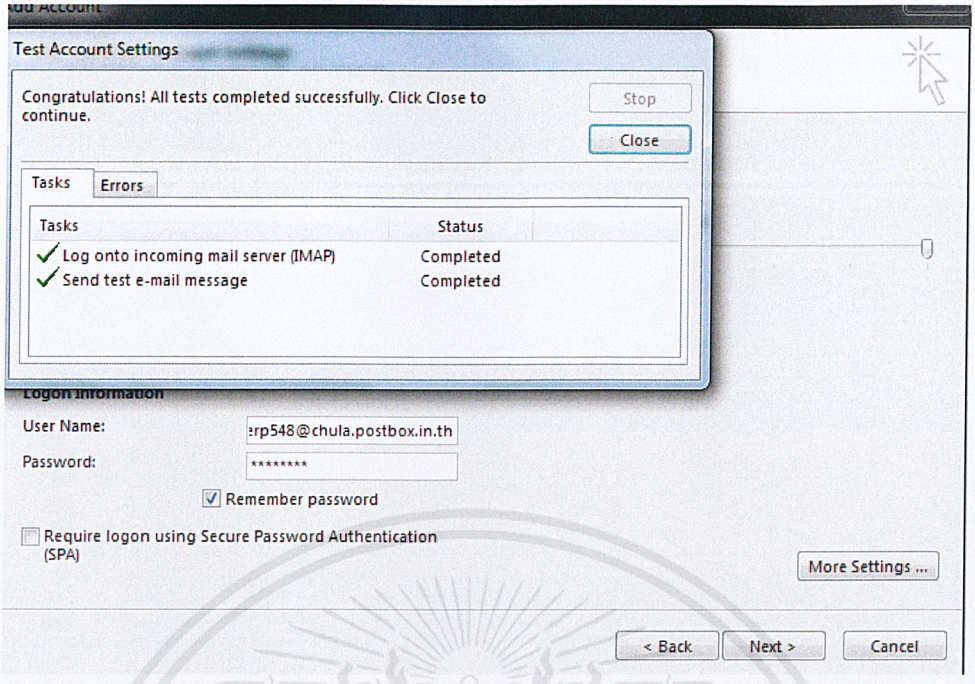
- ใช้ในการเข้าสู่ระบบ (Login) เป็นอันเดียวกับที่อยู่อีเมลที่ใช้งาน (Username)
- รหัสอันเดียวกับรหัสเข้าเครื่องคอมพิวเตอร์ (Password)



ภาพที่ 3.10 การตั้งค่าบัญชีแบบป๊อปและบัญชีแบบไอแมพ

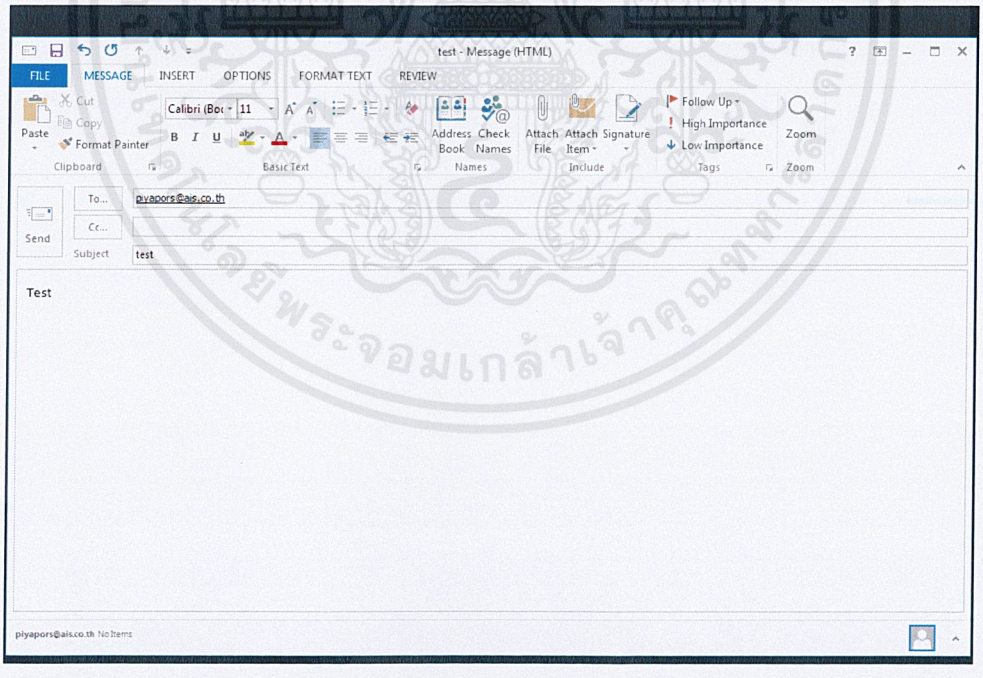
9. หลังจากนั้นจะขึ้นหน้าจอทดสอบการเข้าใช้งานบัญชี (Test Account Settings) และกดปิด (Close) เมื่อมันทดสอบในส่วนของรายละเอียดงาน (Tasks) เสร็จเรียบร้อย จะเห็นได้ว่าจะขึ้นลูกศรสีเขียวในส่วนของหัวข้อสามารถเข้าสู่อีเมลได้ (Log onto incoming mail server (IMAP) และมีสถานะ (Status) เป็นสมบูรณ์ (Completed) และรวมถึงหัวข้อทดสอบการส่งอีเมลได้ (Send test e-mail message) ก็ต้องมีสถานะ (Status) เป็นสมบูรณ์ (Completed) และมีลูกศรสีเขียวอยู่ข้างหน้าด้วยเช่นเดียวกัน หลังจากนั้นจึงกดปิดหน้าต่างได้ (Close)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 3.11 หน้าจอการทดสอบการตั้งค่าผู้ใช้งาน

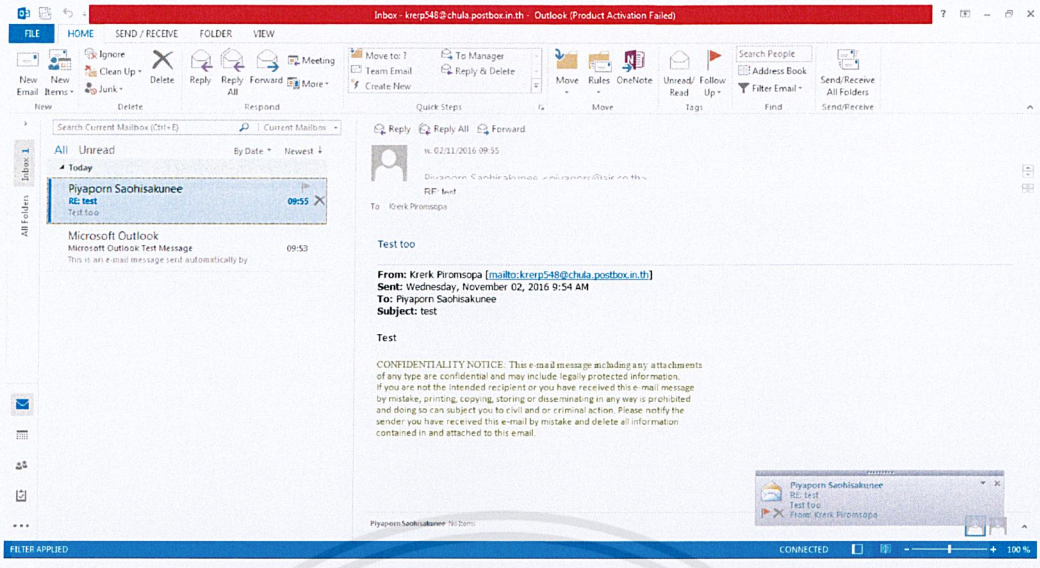
10. หลังจากกดเสร็จจากการเพิ่มบัญชี (Add Account) โปรแกรมจะกลับมาที่เอาท์ลุค และจะสามารถเช็คดูได้ว่าสามารถรับ-ส่งอีเมลได้ทั้งคู่หรือไม่



ภาพที่ 3.12 หน้าต่างเขียนอีเมลของเอาท์ลุค

11. หลังจากที่ส่งอีเมลไป ก็รออีกฝั่งตอบกลับมา เพื่อยืนยันว่าอีเมลที่ได้ทำการตั้งค่าแล้ว ได้ตั้งค่าเสร็จสมบูรณ์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 3.13 หน้าต่างของโปรแกรมขณะมีอีเมลส่งเข้ามา

3.2 การเตรียมข้อมูลในที่สเตจจิ้งดาต้าเบส (Staging Database)

- เลือกข้อมูลที่จะนำมาใช้ โดยเลือกคอลัมน์ที่จะใช้จากดาต้าดิกชันนารี (Data Dictionary) ของแต่ส่วนที่มีมีการเก็บข้อมูล (Source) ที่แตกต่างกันออกไปตามข้อกำหนดที่ต้องการ
- เลือกใช้ข้อมูล 3 เดือน คือช่วงสิงหาคม-พฤศจิกายน 2559
- เลือกลูกค้าที่มีการใช้งานอยู่ภายในกรุงเทพมากที่สุด (Most Usage)
- รายละเอียดข้อมูลที่ใช้ในการวิเคราะห์

3.3 การเลือกกลุ่มลูกค้า หรือกลุ่มที่สนใจจะนำมาวิเคราะห์

ในการเลือกกลุ่มลูกค้านั้นจะเลือกจากการอ้างอิงข้อมูลประชากรศาสตร์ (Demography) จากหลาย ๆ ที่ ซึ่งจากการวิจัยจะพบว่าคนกรุงเทพมีการใช้งานอินเตอร์เน็ตมากที่สุดจากทุก ๆ ภาคส่วน ทำให้สามารถจำกัดขอบเขตการเลือกกลุ่มเป้าหมายที่มีการใช้งานดาต้าในกรุงเทพมากที่สุด โดยดูจากโลเคชันที่กลุ่มเป้าหมายใช้ และเลือกช่วงเวลาก่อนการวิเคราะห์ข้อมูลที่มีล่าสุด 3 เดือนก็คือเดือนสิงหาคมถึงเดือนตุลาคม สำหรับการวิเคราะห์การใช้งานสิทธิพิเศษ (Privilege *ABC) ของลูกค้าว่ามีความสนใจในด้านไหน และใช้ข้อมูลล่าสุดในวันที่ 16-18 ธันวาคม 2559 สำหรับการวิเคราะห์ความสนใจจากแอปพลิเคชันที่สนใจ

3.4 การเตรียมข้อมูลให้ระบบ (Data Preparation)

การเตรียมข้อมูลให้ระบบเนื่องจากในบริษัทใหญ่ จะมีการแบ่งส่วนงานรับผิดชอบในหลาย ๆ ส่วน ซึ่งก่อนอื่นคือต้องคิดก่อนว่าจะใช้ข้อมูลส่วนไหนมาวิเคราะห์บ้าง เพื่อที่จะสอบถามไปยังพนักงานในแผนกว่ามีข้อมูลส่วนไหนต้องไปขอที่ใคร หรือมีการเก็บข้อมูลแบบนี้ไว้ไหม จะต้องมีการคัดเลือกข้อมูลโดยการคัดเลือก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อมูลหรือตาราง (Table) ที่จะนำมาใช้ จะคัดเลือกจากการดูดาต้าดิกชันนารี (Data Dictionary) เป็นหลักว่าตารางอะไรเก็บข้อมูลอะไร และในแต่ละตารางต้องการคอลัมม์อะไรบ้าง สาเหตุที่ต้องเลือกคอลัมม์ด้วย ก็เป็นผลมาจากการที่ระบบที่จะนำข้อมูลเข้าไปมีขนาดที่จำกัด ทำให้ถ้าเอาข้อมูลเข้าทั้งหมดจะทำให้เป็นการสิ้นเปลืองทรัพยากรเนื้อที่เกินจำเป็น ด้วยเหตุนี้จึงควรเลือกคอลัมม์แต่ที่สำคัญ ที่คิดว่าจะสามารถนำมาใช้กับโปรเจคนี้ได้อย่างเหมาะสม โดยตารางที่เลือกมาทั้งหมดนี้จะนำมาใช้ในหลายโปรเจค ไม่ว่าจะเป็นการหาผู้ที่มีความสนใจเกี่ยวกับอาหาร การหาผู้ที่สนใจในการเดินทาง การหาผู้ที่สนใจในการซื้อของ และการหาผู้ที่สนใจในสื่อมัลติมีเดียโดยในโปรเจคนี้จะพูดถึงหัวข้อเอ็นเตอร์เทนเมนท์

โดยข้อมูลที่จะทำการเลือกทั้งหมด จะมีรายละเอียดข้อมูลที่ใช้ในการวิเคราะห์แสดงเป็นตารางตามตารางที่ 3.1 ดังนี้

ตารางที่ 3.1 รายละเอียดข้อมูลตารางที่ใช้ในการวิเคราะห์ข้อมูล

Table name(ชื่อตาราง)	Remark(หมายเหตุ)
Source_Profile_Txt	ข้อมูลเกี่ยวกับประวัติ (Profile) ของลูกค้า
Privilege_Category_Txt	คือการกดรหัสพิเศษ ในที่นี้คือ *ABC# จะเก็บเกี่ยวกับโปรโมชั่นต่าง ๆ ที่ลูกค้ากดรับสิทธิ์ว่าเป็นรหัสเกี่ยวกับอะไร
Cutomer_Arpu	เกี่ยวกับค่าใช้จ่ายในการใช้งานดาต้า (Data) และการโทร (Voice)
Ext_App_Summary	การใช้งานแอปพลิเคชัน (Application) ในแบบสรุป
Privilege	รายละเอียดของลูกค้าว่ากดรับสิทธิ์อะไร

3.5 การนำข้อมูลที่สนใจมาวิเคราะห์ (Data Exploration)

ในการจะวิเคราะห์ข้อมูลจะมีการสร้างตารางหรือที่เรียกว่าตารางเกิดขึ้นมากมาย เพื่อให้สะดวกในการเรียกใช้ ไม่ต้องมาคอยเรียกข้อมูล (Query) ดูผลลัพธ์

และในขั้นตอนนี้จะมีการนำเครื่องมือ (Tools) ที่มีชื่อว่า Tableau เข้ามาช่วยในการวิเคราะห์ข้อมูล เนื่องจากดาต้ามีขนาดที่ใหญ่ และยากในการมองหาความสัมพันธ์ ดังนั้นจึงต้องมีการนำเครื่องมือ มาช่วยในการมอง โดยแทบโบลวจะช่วยทำให้มองเห็นข้อมูลออกมาเป็นภาพ ทำให้เข้าใจข้อมูลได้ง่ายมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6 นำเสนอผลลัพธ์ออกมาในรูปของการนำเสนอ (Visualization)

ในการนำเสนอข้อมูลที่วิเคราะห์ออกมา จะใช้เครื่องมือ (Tools) ที่มีชื่อว่าแทบโบลว (Tableau) เป็นตัวสื่อในการนำเสนอโดยตัวแทบโบลว จะแสดงผลการวิเคราะห์ออกมาเป็นกราฟที่สีสันสวยงาม สามารถอ่านแล้วเข้าใจได้ง่าย มองภาพได้ชัดเจน



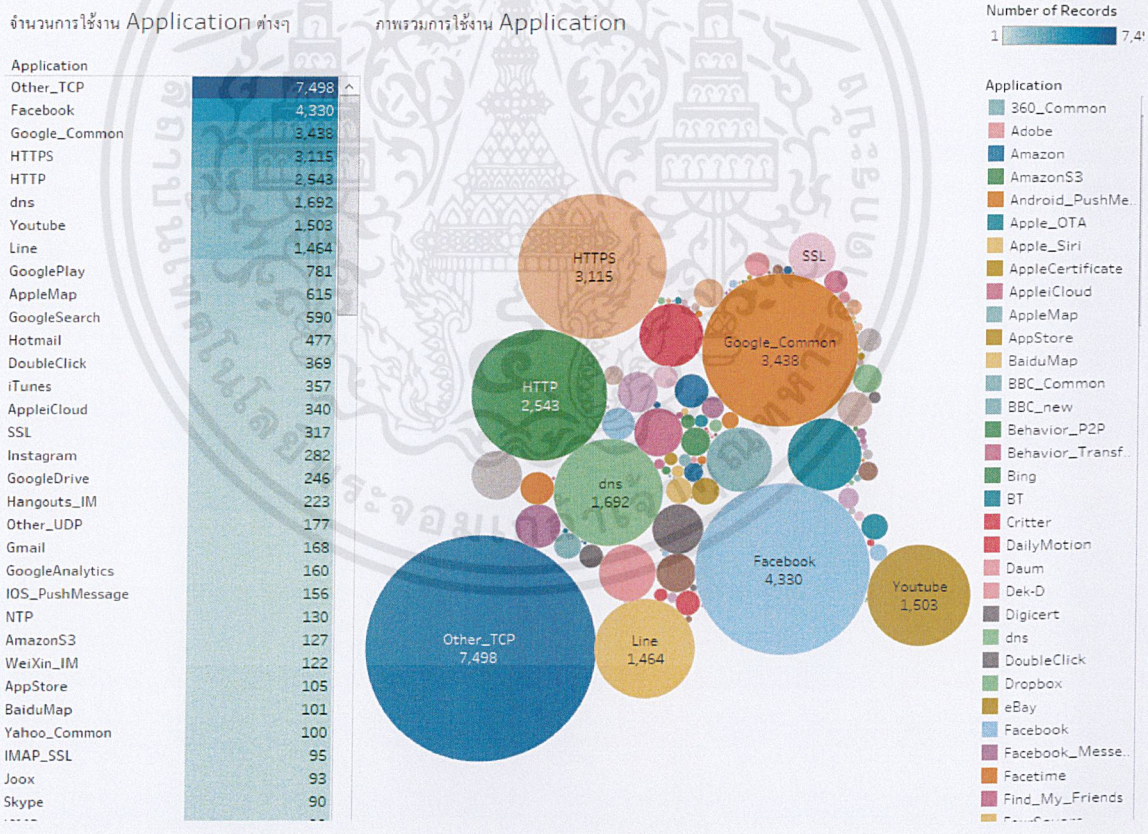
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัย

4.1 กลุ่มลูกค้าที่มีการใช้งานแอปพลิเคชันที่เกี่ยวกับความบันเทิง และมีการใช้ *ABC

ตารางการใช้งานความบันเทิง (Entertain App) และตารางการใช้งานสิทธิพิเศษ (Privilege) [ข้อมูลตัวอย่าง]

- นำข้อมูลของสองตารางมา รวมกัน (Join) กัน (ตารางการใช้งานแอปพลิเคชันความบันเทิง และตารางการใช้งานสิทธิพิเศษ)
- นับจำนวนแถว (Row) ของข้อมูลตัวอย่าง ได้ข้อมูลออกมา 33,124 แถว
- นับจำนวนเบอร์โทรที่ไม่ซ้ำกันได้ข้อมูลออกมา 227 เบอร์
- สร้างตาราง (Table) จากการคิวรี่ (Query) ข้างต้น เพื่อความสะดวกในการเรียกใช้ข้อมูล
- นำไปแสดงผลบนแทบโบลว (Tableau) เพื่อวิเคราะห์ข้อมูลในเบื้องต้น



ภาพที่ 4.1 ข้อมูลในการใช้งานแอปพลิเคชันของข้อมูลในวันที่ 29/9/2016

จากภาพที่ 4.1 จากการที่เลือก (Select) ข้อมูลการใช้งาน 33,124 ของลูกค้าจำนวน 227 เบอร์ จะเห็นว่า จำนวนการใช้งานในกลุ่มของโปรโตคอลทีซีพีอื่น ๆ (Other_TCP) มีการใช้งานมากที่สุดซึ่งเป็นเรื่อง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ปกติ เนื่องมาจากว่าการที่เข้าเว็ปไซต์ หรือการเข้าแอปพลิเคชันบางชนิดจะต้องมีการเรียกใช้โปรโตคอลในการเข้าถึงหน้าเว็บไซต์นั้น ๆ หรือหน้าแอปพลิเคชันนั้น ๆ แต่สิ่งที่จะเห็นรองลงมาคือจำนวนการใช้เฟสบุ๊ค (Facebook) ที่มีจำนวนการใช้งานมากกว่าแอปพลิเคชันต่าง ๆ และสูงกว่าการใช้งานพวกยูทูป (Youtube) หรือไลน์ (Line) ประมาณ 2-3 เท่าตัว

4.2 กลุ่มลูกค้าที่มีการใช้งานรายการสิทธิพิเศษ (Privilege)

หลังจากสร้างตารางข้อมูลสิทธิพิเศษรวมกับประวัติและมาสเตอร์ของสิทธิพิเศษ (Test_Cu_Ent_Profile_Privilege_Mstpri) โดยจะแยกเป็นการใช้งานตามรายเดือน (Postpaid) และเติมเงิน (Prepaid)

4.2.1 รายเดือน (Postpaid)

ได้ข้อมูลออกมา 1,017,569 Rows และเลือกข้อมูลการใช้งาน*ABC ในหมวดบันเทิง ได้ดังภาพที่ 4.2

```
SELECT COUNT(DISTINCT MOBILE_NO)
FROM AISCUGEN.TEST_CU_ENT_PROFILE_PRIVILEGE_MSTPRI_POSTPAID
```

ภาพที่ 4.2 เอสคิลแอลในการนับจำนวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแบบรายเดือน

จากภาพที่ 4.2 จะผลลัพธ์เป็นจำนวนเบอร์ทั้งหมดที่ไม่ซ้ำ 418,403 เบอร์ที่ไม่ซ้ำกันสำหรับตารางข้อมูลสิทธิพิเศษรวมกับประวัติและมาสเตอร์ของสิทธิพิเศษสำหรับข้อมูลรายเดือน

4.2.2 เติมเงิน (Prepaid)

ได้ข้อมูลออกมา 9,422,046 Rows และเลือกข้อมูลการใช้งาน*ABC ในหมวดบันเทิง ได้ดังภาพที่ 4.3

```
SELECT COUNT(DISTINCT MOBILE_NO)
FROM AISCUGEN.TEST_CU_ENT_PROFILE_PRIVILEGE_MSTPRI_PREPAID
```

ภาพที่ 4.3 เอสคิลแอลในการนับจำนวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแบบเติมเงิน

จากภาพที่ 4.3 จะผลลัพธ์เป็นจำนวนเบอร์ทั้งหมดที่ไม่ซ้ำ 490,992 เบอร์ที่ไม่ซ้ำกันสำหรับตารางข้อมูลสิทธิพิเศษรวมกับประวัติและมาสเตอร์ของสิทธิพิเศษสำหรับข้อมูลเติมเงิน

4.3 ดูข้อมูลอาปูว (ARPU) เพื่อใช้เป็นข้อกำหนดในการเลือกกลุ่มลูกค้า

เลือกลูกค้าจากตารางอาปูว โดยใช้ข้อมูลจากเดือน 10 คือเดือนตุลาคม โดยเลือกแต่ข้อมูลที่มียอดการ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ใช้งานมากกว่า 399 บาทขึ้นไปต่อเดือน และมียอดค่าใช้จ่ายการใช้ดาต้ามากกว่าการใช้การโทร

4.4 การใช้งานความบันเทิงรวมกับอาปู (Test_Cu_Entertain_Arpu_Total_Mth10)

เมื่อสร้างตารางการใช้งานความบันเทิงรวมกับตารางอาปูวสำหรับเดือนตุลาคม (Test_Cu_Entertain_Arpu_Total_Mth10) และเพิ่มข้อมูลเข้าไปในตารางแล้ว ก็มาดูจำนวนเบอร์ที่เลือกมาตามเงื่อนไขข้างต้น

- ดูจำนวนเบอร์ที่ไม่ซ้ำกันในตารางการใช้งานความบันเทิงรวมกับตารางอาปูวเดือนตุลาคม

```
SELECT COUNT(DISTINCT MOBILE_NO)
FROM AISCUGEN.TEST_CU_ENTERTAIN_ARPU_TOTAL_MTH10
```

ภาพที่ 4.4 เอสคิลแอลในการนับจำนวนเบอร์โทรที่ไม่ซ้ำกันของอาปูวในเดือนตุลาคม

จากภาพที่ 4.4 จะผลลัพธ์เป็นจำนวนเบอร์ทั้งหมดที่ไม่ซ้ำ 8,058,001 เบอร์ในเดือนตุลาคมที่ตรงตามเงื่อนไขเอสคิลแอล

4.5 สรุปการใช้งานของลูกค้า (Test_Cu_Customer_Appsum)

เมื่อสร้างตารางสรุปการใช้งานของลูกค้า (Test_Cu_Customer_Appsum) จากตารางการใช้งานความบันเทิงรวมกับตารางอาปูวสำหรับเดือนตุลาคม (Test_Cu_Entertain_Arpu_Total_Mth10) และเพิ่มข้อมูลเข้าไปในตารางแล้ว ก็มาดูจำนวนเบอร์ที่เลือกมาตามเงื่อนไขข้างต้น

```
SELECT COUNT(DISTINCT MOBILE_NO)
FROM AISCUGEN.TEST_CU_CUSTOMER_APPSUM
```

ภาพที่ 4.5 เอสคิลแอลในการนับจำนวนเบอร์โทรที่ไม่ซ้ำกันของการใช้งานแอปพลิเคชันตัวอย่าง

จากภาพที่ 4.5 จะสรุปได้ว่ามีเบอร์ที่จะนำไปใช้ช่วยในการดึงข้อมูลการใช้งานแอปพลิเคชันตัวอย่าง 65,644 เบอร์

โดยจะใช้เบอร์ที่เลือกตามเงื่อนไขเหล่านี้ ไปช่วยในการดึงข้อมูลการใช้งานแอปพลิเคชัน และยูอาร์แอล (Url) ของจริง เนื่องจากถ้าไม่มีการกำหนดเงื่อนไขในการเลือกข้อมูล ข้อมูลที่เข้ามาในแต่ละวันมีขนาดใหญ่มาก ทำให้มีปัญหาเวลาเอาเข้าตัวอิมพาลา (Impala) ทั้งหมดเพราะปริมาณที่เยอะเกินไป จึงจำเป็นต้องมี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การกรองข้อมูลเวลาจะนำเข้าตัวอิมพาลา

จากข้อมูลข้างต้นทำให้กำหนดเงื่อนไขกับข้อมูลชุดใหม่คือ

1. ใช้ข้อมูลเบอร์จากตารางสรุปการใช้งานของลูกค้า (Test_Cu_Customer_Appsum)

2. ใช้ข้อมูลเบอร์ที่นอกเหนือจากที่เลือกไว้โดยดูจากการใช้งานแอปพลิเคชัน โดยเลือกจากคอลัมม์ประเภทของโปรโตคอล (Protocol_Category) ที่มีการใช้งานประเภทของโปรโตคอลเหล่านี้

2.1 สำหรับกลุ่มโซเชียล (Social)

- โซเชียล (Social) ไอเอ็ม (IM) อีเมล (E-mail) พีทูพี (P2P) วีโอไอพี (VoIP) สตรีมมิ่ง (Streaming)

2.2 สำหรับกลุ่มใช้ค้นหาข้อมูล/อ่านข่าวสาร/การ์ตูนและนิตยสาร

- แผนที่และการนำทาง (Map & Navigation) การสืบค้นข้อมูล (Search Engine)

2.3 สำหรับกลุ่มที่ใช้อ่านข่าวสาร/การ์ตูนและนิตยสาร

- หนังสือและการ์ตูน(Book & Comics) ข่าวและนิตยสาร (News & Magazines)

2.4 สำหรับกลุ่มคนที่ชอบความบันเทิง

- มีเดียและวิดีโอ (Media & Video) เพลง ความบันเทิง

2.5 สำหรับกลุ่มที่ชอบเล่นเกม (Gaming)

- เกม (Game)

2.6 สำหรับกลุ่มที่ชอบด้านกีฬา

- กีฬา (Sports) สุขภาพและฟิตเนต (Health & Fitness)

4.6 ผลลัพธ์จากการวิเคราะห์ความสนใจในการใช้โปรโมชั่นด้านเอ็นเตอร์เทนเมนท์

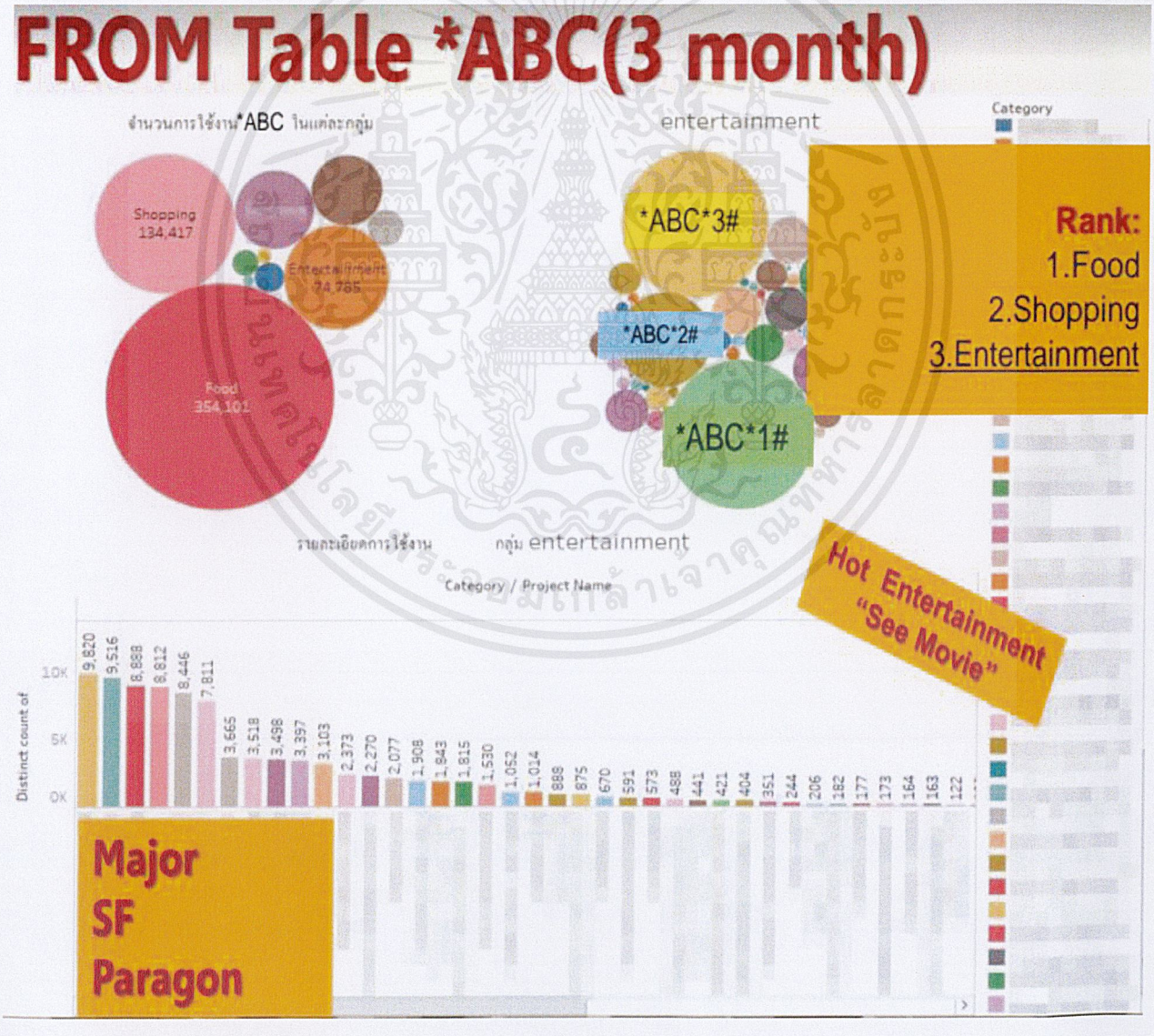
จากตาราง (Table) ของการใช้งานสิทธิพิเศษต่างๆ จาก *ABC ที่ผ่านการเลือก (Selection) โดยใช้เงื่อนไขในการสุ่มข้อมูลจากฐานข้อมูลโดยที่จะมีการตัดข้อมูลที่มีเงื่อนไขไม่ครบออก เช่นกรณีที่ เบอร์โทรศัพท์หนึ่งเบอร์ มีการใช้งานน้อย และไม่มีการระบุเพศไว้ตอนลงทะเบียน หรือเบอร์โทรศัพท์นั้น ไม่สามารถหาอายุได้ เนื่องจากอาจจะไม่ได้ให้วันเดือนปีเกิดไว้ ซึ่งในตัวโปรเจคนี้ จะมุ่งเน้นไปยังกลุ่มคนที่อยู่ในอายุ 18-60 ปี เนื่องจากเป็นวัยที่มีการใช้งานทางด้านความบันเทิงที่น่าสนใจ ซึ่งจากเงื่อนไขต่าง ๆ เหล่านี้ข้างต้น เมื่อได้ข้อมูลมาแล้ว จะนำข้อมูลที่ได้มาแสดงผลบนแทบโบลว ซึ่งบนแทบโบลวนั้นจะประกอบด้วยจำนวนการใช้งาน *ABC ในแต่ละกลุ่ม จำนวนการใช้งานรหัสสิทธิพิเศษว่า รหัสอะไรได้ถูกกดใช้งานมากน้อยเพียงใด ละส่วนสุดท้ายจะพูดถึงรายละเอียดการใช้งาน *ABC ในกลุ่มความบันเทิงว่ามีรายละเอียดการใช้งานอะไรบ้าง

ซึ่งจากการวิเคราะห์โดยใช้แทบโบลว สำหรับจำนวนการใช้งาน *ABC ในหมวดหมู่ของอาหาร (Food)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

มีมากที่สุด คือมีจำนวน 354,101 ลำดับที่สองคือการซื้อขาย (Shopping) คือมีจำนวน 134,417 และกลุ่มลำดับที่สามคือกลุ่มที่สนใจ ซึ่งเป็นในส่วนของความบันเทิง (Entertainment) มีปริมาณมากเป็นลำดับที่ 3 มีจำนวน 74,785 จากทั้งหมด 10 กลุ่ม

และจากการใช้งาน *ABC ในกลุ่มของความบันเทิงเราจะเห็นได้ว่า จะเป็น *ABC#1 มากที่สุด หลังจากนั้นจะเป็น *ABC#2 และ *ABC#3 ตามลำดับ โดยที่ทั้งสามสิทธิพิเศษนี้ จะเกี่ยวข้อกับการดูหนังทั้งหมด แตกต่างกันไปตามเครือของโรงหนัง โดย * ABC#1 จะเป็นในเครือของโรงภาพยนตร์เมเจอร์ซีเนเพล็กซ์ (Major Cineplex) สำหรับ *ABC#2 จะเป็นในเครือของโรงภาพยนตร์เอสเอฟซิเนม่า (SF Cinema) และสุดท้ายคือ *ABC#3 จะเป็นในส่วนของโรงภาพยนตร์ในเครือเมเจอร์ซีเนเพล็กซ์ (Major Cineplex) เช่นกัน แตกต่างตรงที่จะเป็นโรงภาพยนตร์ของสาขาสยามพารากอน หรือเรียกอีกชื่อว่าพารากอนซีเนเพล็กซ์ ซึ่งจากที่กล่าวมาข้างต้น แสดงผลได้ดังภาพที่ 4.6



ภาพที่ 4.6 ผลลัพธ์จากการวิเคราะห์ความสนใจในการใช้โปรโมชั่นด้านความบันเทิง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Entertainment

Project Name	
ดูหนังราคาพิเศษ @ โรงภาพยนตร์ในเครือ SF_Aug 16	9,516
AIS Movie Major_Sep16	8,888
AIS Movie Major_Oct 16	8,446
ดูหนังราคาพิเศษ @ โรงภาพยนตร์ในเครือ SF_Oct 16	7,811
Sea Life Bangkok Ocean World 2016	3,665
AIS Movie Paragon_Aug 16	3,518
AIS Movie Paragon_Sep 16	3,498
AIS Movie Paragon_Oct 16	3,397
Dinosaur Planet (25 Mar - 31 Dec 2017)	3,103
Major Bowling 2016	2,373
AIS Privilege @ Concession โรงภาพยนตร์ SF2016	2,270
Kidzania 2016	2,077
Ramayana Water Park	1,908
Funarium Discount 20%	1,843
AIS Dino Farm (25 Mar - 31 Dec 2017)	1,815
Major Karaoke 2016	1,530
LaZGam (Apr - Dec 2016)	1,052
สวนสัตว์ดุสิต 2016	1,014
Sub Zero 2016	888
Embassy (Mass) 2015 & 2016	875
พระที่นั่งอนันตสมาคม	670
FUN Planet & FUN Festa	591
พระราชวังสนามจันทร์	573
Embassy (Platinum) 2015 & 2016	488
Santorini Water Fantasy	441
SRN BD Embassy Diplomat Screen 2016	421
ดูหนังราคาพิเศษ @ First Class จ้า SFW / SFX 2016	404
Madame Tussauds 2016	351
Harbor Land	244
Payakhin Muaythai	206
Tiffany Show Pattaya 2016	182
Coliseum 2016	177
ลุ้นใจจิดเต็ม ฉลองวันแม่ @ KidZania Bangkok	173
Embassy Lounges	164
Arts in Paradise Bangkok (1 Dec 15 - 30 Nov 16)	163
สวนสัตว์เปิดเขาเขียว 2016	122
Scenical World สวนน้ำ Splash World	120
YMAX Real 4D 2016	115
สวนสัตว์เชียงใหม่ (1 May 2016 - 31 Apr 2017)	110
สวนสัตว์นครราชสีมา (1 May 2016 - 31 Apr 2017)	94
สวนสัตว์อุบลราชธานี (1 May 2016 - 31 Apr 2017)	92
สวนสัตว์เชียงใหม่ 2016	66
สวนสัตว์ขอนแก่น (1 May 2016 - 31 Apr 2017)	48
CAMP	27

ภาพที่ 4.7 ปริมาณการใช้งาน(ครั้ง)ในการใช้งาน *ABC

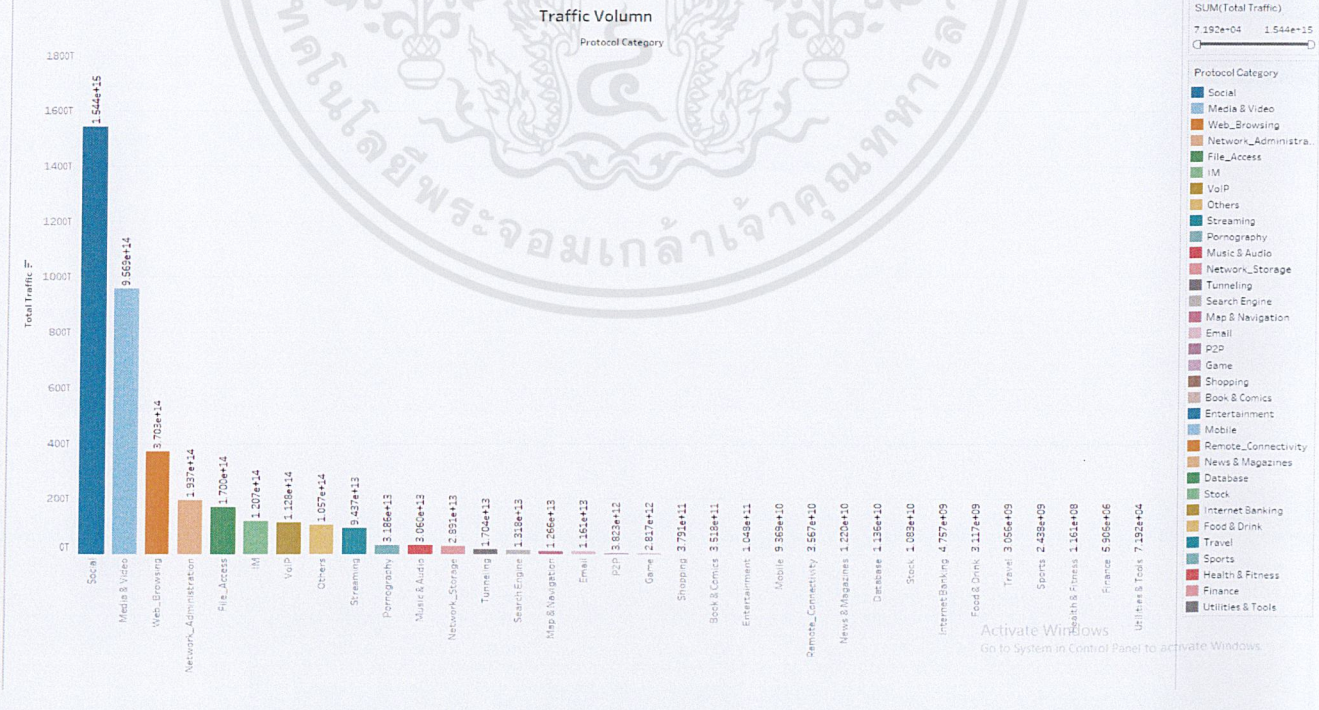
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

และจากภาพที่ 4.7 ในกลุ่มก้อนของเอ็นเตอร์เทนเมนท์ในจำนวนการใช้งานทั้งหมด 1,208,375 ครั้ง นั้นอันดับแรก ๆ ล้วนเป็นการใช้งานเกี่ยวกับการดูหนัง โดยจากข้อมูลการใช้งานสิทธิพิเศษ (Privilege) ในระยะเวลา 3 เดือน โดยเก็บข้อมูลตั้งแต่วันที่ 1 สิงหาคม 2559 ถึง 31 ตุลาคม 2559 รวมระยะเวลา 3 เดือน ได้ผลปรากฏว่าในเดือนสิงหาคมมีการใช้งานมากที่สุดคือ 9,820 ครั้งซึ่งล้วนแต่เกี่ยวข้องกับการดูหนัง และเดือนตุลาคมมีการใช้งานลดลงมาจนเหลือ 8,446 ครั้งและนอกจากการชมภาพยนตร์ก็จะเป็นพวกพิพิธภัณฑ์ หรือสถานที่ต่าง ๆ เช่น โอเชียนเวิลด์ (Ocean World) ดวงดาวใหม่แห่งไดโนเสาร์ (Dinosaur Planet) เป็นต้น

สำหรับการใช้ข้อมูลจริงจากในวันที่ 16/12/2016 สำหรับในส่วนการวิเคราะห์ Application หลังจากที่ได้ทดลองวิเคราะห์ข้อมูลในตอนต้นจากข้อมูลต่างๆแล้ว ก็นำเงื่อนไขและแนวโน้มที่ได้ไปช่วยในการวิเคราะห์และการใช้งานในสำหรับ 3 วัน โดยจาก Table Aiscugen.Ext_Cu_Ida_Mobile_App_Summary ปริมาณข้อมูลในแต่ละวันมีจำนวนดังนี้

- 2016-12-16 มี 477,639,604 Rows
- 2016-12-17 มี 480,211,078 Rows
- 2016-12-18 มี 472,631,833 Rows

ใช้แทบโบลวเพื่อแสดง ปริมาณ (Volumn) สำหรับข้อมูล 3 วัน ในแต่ละหมวดหมู่ของโปรโตคอล (Protocol_Category) ได้ผลลัพธ์ดังนี้



ภาพที่ 4.8 กราฟแสดงความสัมพันธ์ระหว่างปริมาณการใช้งานกับประเภทโปรโตคอล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

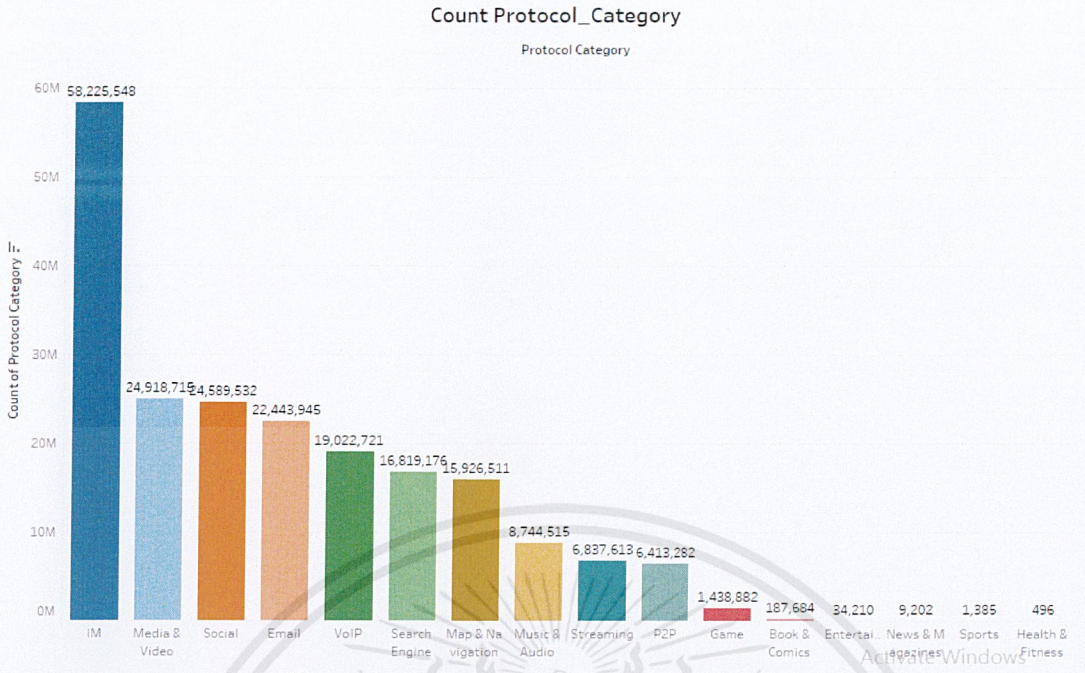
จากภาพที่ 4.8 เมื่อดูแนวโน้มจากทั้ง 3 วันแล้ว จะเห็นได้ว่าการใช้งานสำหรับแอปพลิเคชัน Application ประเภทโซเชียล (Social) มีปริมาณการใช้งาน (Traffic) มากอย่างเห็นได้ชัด โดยมากกว่ามีเดียและวีดิโอ (Media & Video)เกือบ 50% และมากกว่าการใช้งานพวกเว็บ (web browsing) เกือบ 75 % โดยจากกราฟจะสังเกตได้อีกว่า อันดับแรกๆล้วนเกี่ยวข้องกับเรื่องของความบันเทิง และยังรู้ว่าพฤติกรรมการใช้งานที่ใช้ Traffic มาก ๆ เกิดจากอะไร และนอกจากนั้นยังมีการใช้งานพวก การส่งข้อความทันที (IM) ซึ่งมีปริมาณการใช้งานพอ ๆ กับพวกการใช้งานการสนทนาวีดิโอ (VoIP)

- สร้างตารางสำหรับการใช้งานแอปพลิเคชันวันที่ 16/12/2016 โดยเลือกข้อมูลจากคอลัมม์ประเภทโปรโตคอล ดังนี้

  - การสนทนาวีดิโอ (VoIP)
  - สตรีมมิ่ง (Streaming)
  - กีฬา (Sports)
  - โซเชียล (Social)
  - การสืบค้นข้อมูล (Search Engine)
  - พีทูพี (P2P)
  - ข่าวและนิตยสาร (News & Magazines)
  - เพลง (Music & Audio)
  - มีเดียและวีดิโอ (Media & Video)
  - แผนที่และการนำทาง (Map & Navigation)
  - การส่งข้อความทันที (IM)
  - สุขภาพและฟิตเนต (Health & Fitness)
  - เกม (Game)
  - ความบันเทิง (Entertainment)
  - อีเมล (Email)
  - หนังสือและการ์ตูน (Book & Comics)

[OUTPUT] ได้ข้อมูล 205,613,417 Rows แสดงได้ดังภาพที่ 4.9

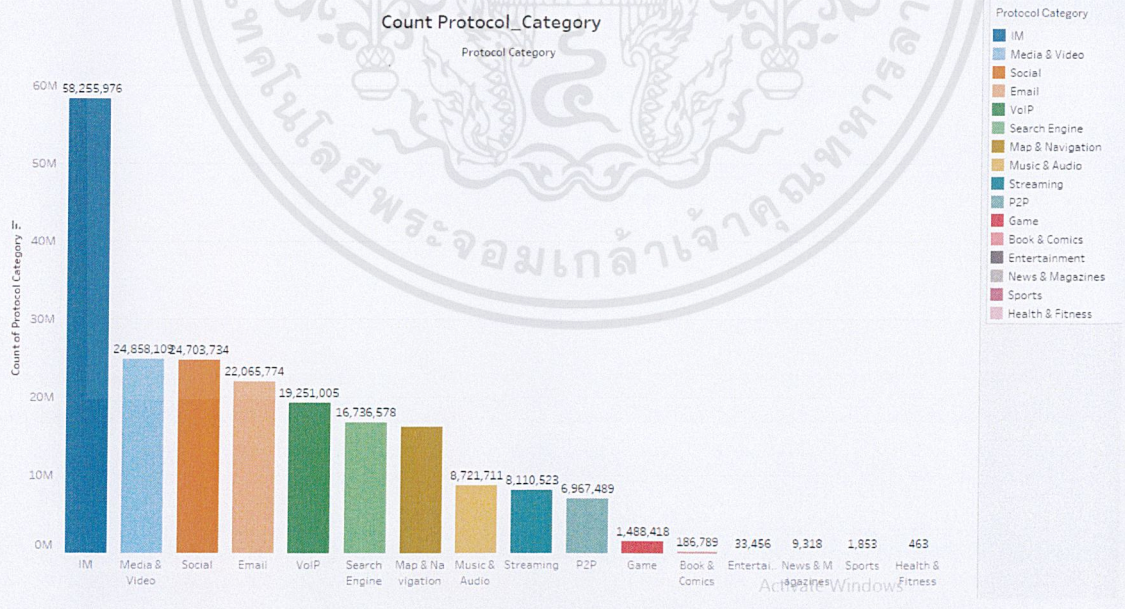
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 4.9 ภาพแสดงการนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 16/12/2016

- สร้าง Table สำหรับการใช้งาน App วันที่ 17/12/2016

[OUTPUT] ได้ข้อมูล 207,539,699 Rows แสดงได้ดังภาพที่ 4.10

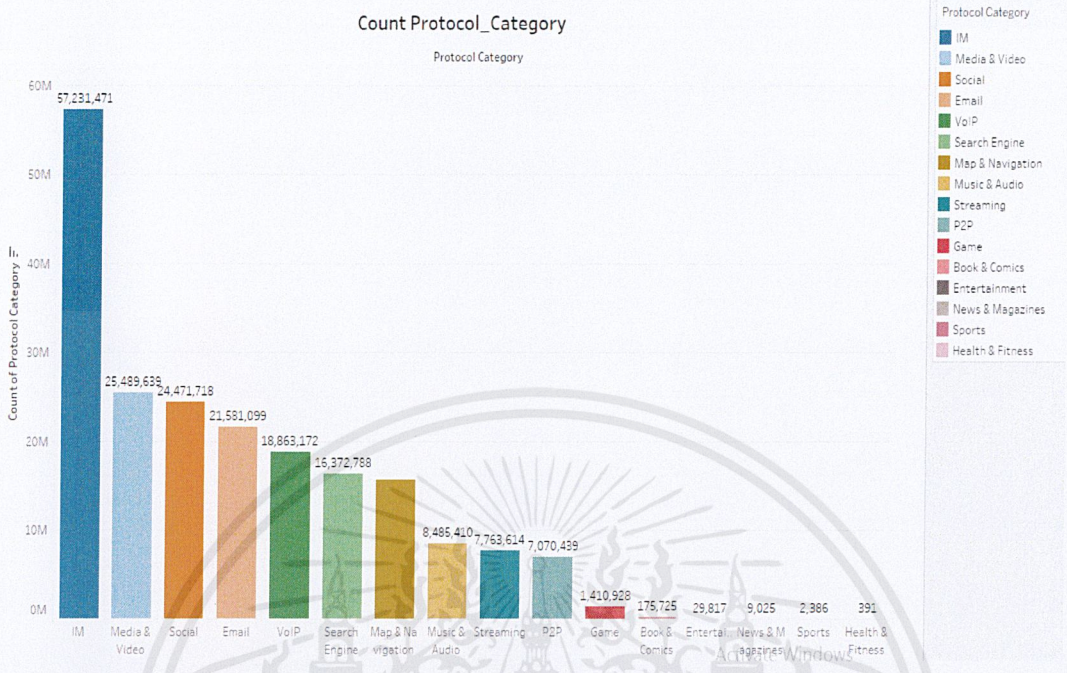


ภาพที่ 4.10 การนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 17/12/2016

- สร้าง Table สำหรับการใช้งาน App วันที่ 18/12/2016

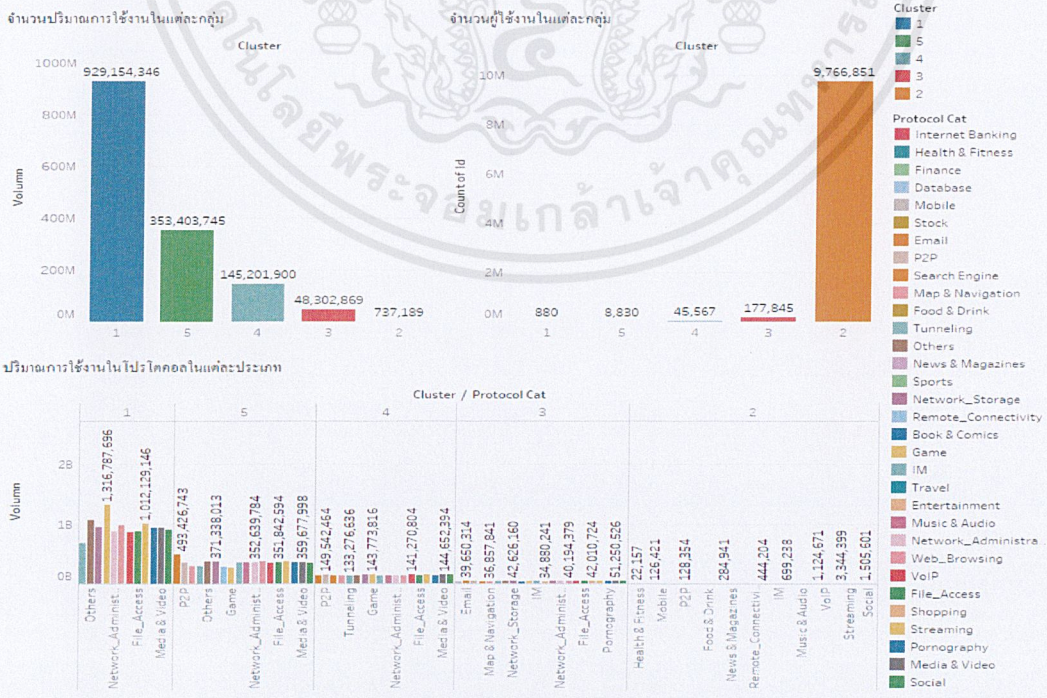
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

[OUTPUT] ได้ข้อมูล 204,574,301 Rows แสดงได้ดังภาพที่ 4.11



ภาพที่ 4.11 การนับจำนวนการใช้งานประเภทโปรโตคอลของวันที่ 18/12/2016

4.7 ผลลัพธ์เมื่อเอาท์พุทค่า K=5 ของการจัดกลุ่มโดยใช้ภาษาอาร์



ภาพที่ 4.12 ผลลัพธ์การใช้งานปริมาณการใช้งานจากข้อมูลตัวอย่าง (ข้อมูล 3 วัน)

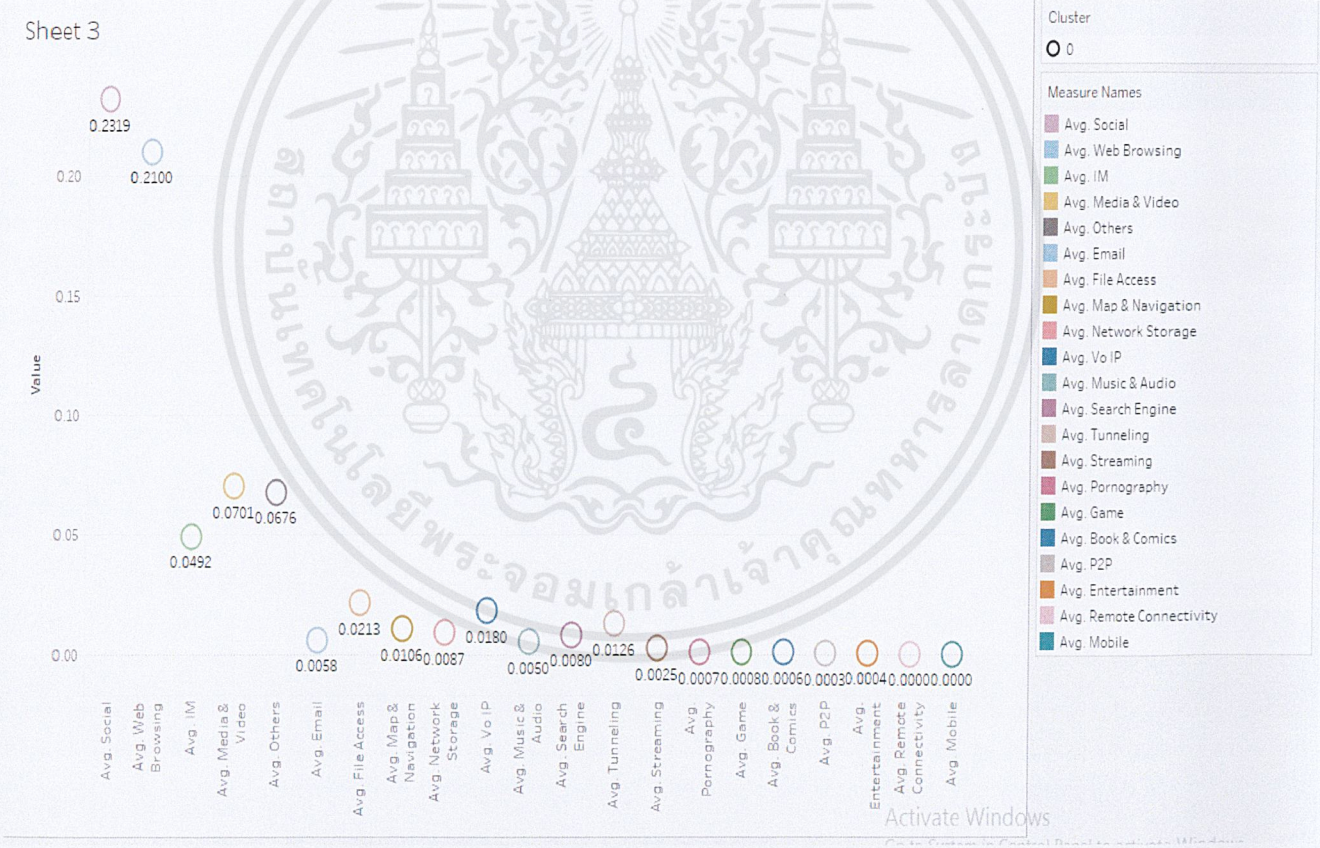
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากภาพที่ 4.12 จำนวนปริมาณการใช้งานในแต่ละกลุ่ม เทียบกับจำนวนผู้ใช้งานแต่ละกลุ่ม

คลัสเตอร์ 1 มีปริมาณการใช้งานมากสุด รองลงมาเป็น คลัสเตอร์ 5, 4, 3 และ 2 ตามลำดับซึ่งสอดคล้องกับจำนวนการใช้งาน คือ คลัสเตอร์ 1 มีปริมาณการใช้งานมากที่สุด แต่มีน้อยคนที่จะใช้งานปริมาณที่มากขนาดนี้ ซึ่งคนส่วนใหญ่จากกลุ่มตัวอย่างจะอยู่ในคลัสเตอร์ ที่ 3 และ 4

สำหรับปริมาณการใช้งานในโปรโตคอลในแต่ละประเภท จะเห็นได้ว่าทุก ๆ คลัสเตอร์ มีการใช้งานในระดับที่เท่าๆกันเกือบทุกแอปพลิเคชัน จะมีที่โดดเด่นออกมาคือ ในคลัสเตอร์ ที่ 1 และ 5 โดยที่ในคลัสเตอร์ 1 มีความโดดเด่นมากที่สุดเมื่อสังเกตจากกราฟ โดยในกลุ่มนี้จะเน้นไปในเรื่องของ game และรองลงมาเป็นพวกสตรีมมิ่ง (Streaming) ส่วนในคลัสเตอร์ 5 จะเน้นไปทางการสืบค้นข้อมูล (Search Engine) และรองลงมาเป็นสตรีมมิ่ง (Streaming)

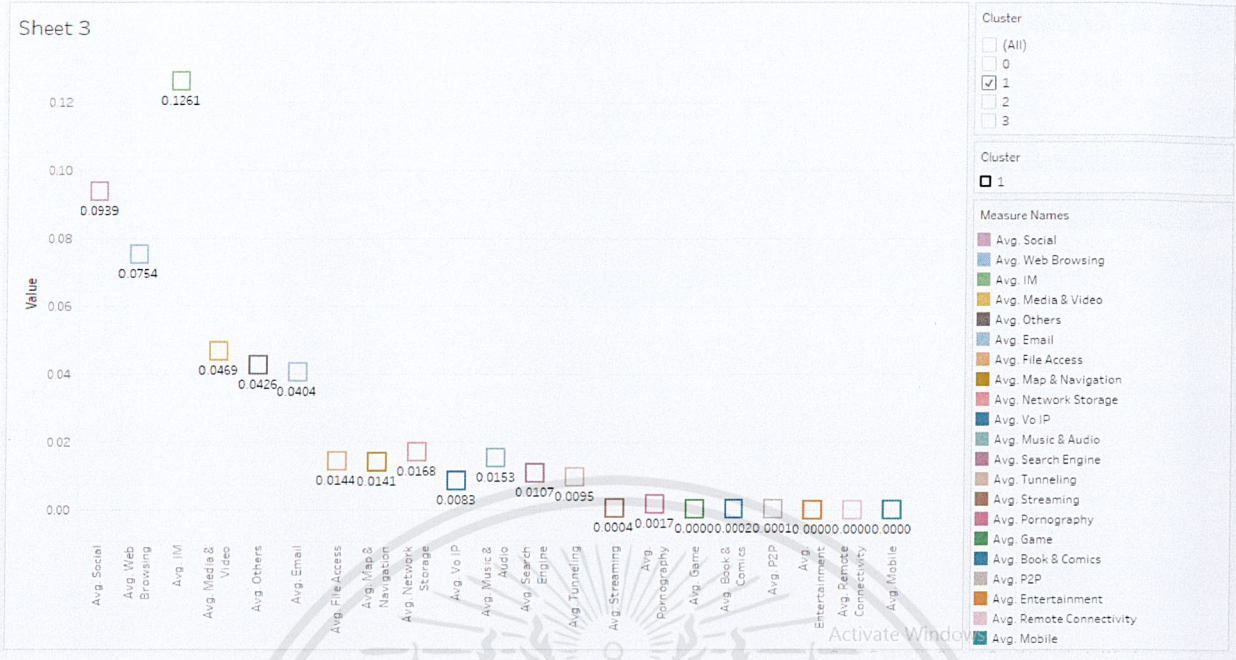
4.8 ผลลัพธ์เมื่อเอาท์พุทค่าเคเท่ากับ4 (K=4) ของการจัดกลุ่มโดยใช้ภาษาไพธอน



ภาพที่ 4.13 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 0

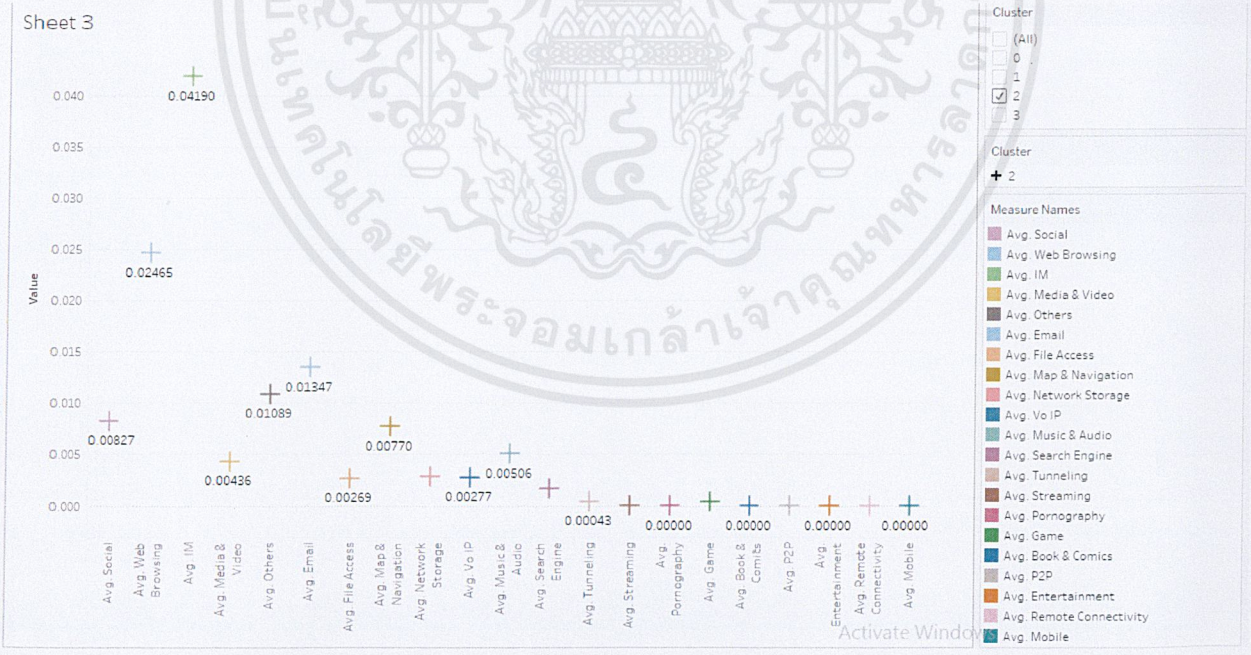
จากภาพจะเห็นได้ว่าสำหรับคลัสเตอร์ 0 สัดส่วนการใช้งานตามพฤติกรรมจะเน้นไปที่ด้านโซเชียล (Social) เป็นหลัก และมีการใช้งานเว็บบราวซิ่ง (Web Browsing) และมีเดียแอนด์วิดีโอ (Media & Video) รองลงมาเป็นหลัก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 4.14 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 1

จากภาพจะเห็นได้ว่าสำหรับคลัสเตอร์ 1 สัดส่วนการใช้งานตามพฤติกรรมจะเน้นไปที่ด้าน IM เป็นหลัก และมีการใช้งาน Social และ Web Browsing รองลงมาเป็นหลัก



ภาพที่ 4.15 ลักษณะการใช้งานแอปพลิเคชันของคลัสเตอร์ 2

จากภาพจะเห็นได้ว่าสำหรับคลัสเตอร์ 2 สัดส่วนการใช้งานตามพฤติกรรมจะเน้นไปที่ด้าน IM เป็นหลัก และมีการใช้งาน Web Browsing และ Social รองลงมาเป็นหลัก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



ภาพที่ 4.16 ลักษณะการใช้งาน Application ของ คลัสเตอร์ 3

จากภาพจะเห็นได้ว่าสำหรับคลัสเตอร์ 3 สัดส่วนการใช้งานตามพฤติกรรมจะเน้นไปที่ด้าน Social เป็นหลัก และมีการใช้งาน Web Browsing และ IM รองลงมาเป็นหลัก



ภาพที่ 4.17 รูปภาพรวมของทั้ง 4 คลัสเตอร์

และจากภาพ 4.17 จะแสดงถึงภาพรวม ซึ่งจะเห็นได้ว่า คลัสเตอร์ 3 จะมีการใช้งาน ด้าน Social มีความโดดเด่นกว่าคลัสเตอร์อื่นๆ เพราะมีลักษณะกราฟเป็นกราฟโดดๆ และคัลสเตอร์ 2 มีการใช้งาน Social น้อยกว่า คลัสเตอร์อื่นๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.9 ผลลัพธ์เมื่อใช้กดการหาความสัมพันธ์แบบอะพริโอริ (Apriori) โดยใช้ภาษาอาร์

```
inspect(sort(rules, by="lift"))
```

	lhs		rhs	support	confidence	lift
[1]	{value.Shopping}	=>	{gender=F}	0.023840299	0.71200980	1.2811603
[2]	{gender=F}	=>	{value.Shopping}	0.023840299	0.04289722	1.2811603
[3]	{gender=M}	=>	{value.Game}	0.433618514	0.97607722	1.0124167
[4]	{value.Game}	=>	{gender=M}	0.433618514	0.44976219	1.0124167
[5]	{}	=>	{value.Shopping}	0.033483105	0.03348310	1.0000000
[6]	{}	=>	{gender=M}	0.444246117	0.44424612	1.0000000
[7]	{}	=>	{gender=F}	0.555753883	0.55575388	1.0000000
[8]	{}	=>	{value.Game}	0.964106194	0.96410619	1.0000000
[9]	{gender=F}	=>	{value.Game}	0.530487680	0.95453706	0.9900746
[10]	{value.Game}	=>	{gender=F}	0.530487680	0.55023781	0.9900746
[11]	{value.Shopping}	=>	{gender=M}	0.009642806	0.28799020	0.6482672
[12]	{gender=M}	=>	{value.Shopping}	0.009642806	0.02170600	0.6482672

ภาพที่ 4.18 ผลลัพธ์เมื่อใช้กดการหาความสัมพันธ์แบบอะพริโอริ (Apriori) โดยใช้ภาษาอาร์

จากภาพที่ 4.18 จะเห็นได้ว่าผลลัพธ์ที่ได้ถ้าค่า ลิฟท์ (Lift) >> 1 มาก ๆ แล้วนั้น หมายถึงว่าถ้ามีแอลเอชเอส (LHS) แล้วจะมี อาเอชเอส (RHS) ด้วย ซึ่งจาก

โดยผลลัพธ์ใน 2 บรรทัดแรกนั้นจะเห็นว่า ถ้าพูดถึงเรื่องการซื้อของ (Shopping) ก็ไปโฟกัสที่เพศหญิงได้ หรือในอีกทาง ถ้าพูดถึงผู้หญิง ก็จะนึกถึงการซื้อของได้ โดยที่ถ้าค่าลิฟท์ หรือเรียกอีกอย่างว่าค่าความสัมพันธ์ (Correlation) ซึ่งถ้ายิ่งได้เกินค่า 1 เยอะ ๆ นั่นหมายถึงมีโอกาสที่เกิดแอลเอชเอส (LHS) แล้วจะเกิดอาเอชเอส (RHS) สูงมาก

กรณีค่า Lift == 1 หมายถึงว่าแอลเอชเอส (LHS) และอาเอชเอส (RHS) ไม่มีความสัมพันธ์กันเลย คือเป็นอิสระต่อกัน จากผลลัพธ์ก็จะเห็นว่า ช้อปปิ้ง (Shopping) เกม (Game) เพศชาย (Gender =M) เพศหญิง (Gender = F) ที่ช่องแอลเอชเอส (LHS) ไม่มีค่าอะไรออกมา ทำให้ทุกตัวปรากฏแค่ด้านอาเอชเอส (RHS) จึงทำให้มันเป็นอิสระไม่ขึ้นกับการเกิดอะไร

และสำหรับค่าถ้าค่าลิฟท์ (Lift) << 1 หมายถึงว่าแอลเอชเอส (LHS) ไม่มีผลต่อการเกิดอาเอชเอส (RHS) หรือในอีกนัยนึง คือผลลัพธ์ที่ได้มีความสัมพันธ์ที่ตรงข้ามกัน จากผลลัพธ์ที่ได้ในบรรทัดที่ 11-12 จะเห็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ว่า ถ้าพูดถึงการช้อปปิ้ง (Shopping) แล้ว เพศชายจะไม่ค่อยมีความสัมพันธ์ หรือถ้าพูดถึงเพศชายเรื่องช้อปปิ้ง (Shopping) ไม่มีผลลัพท์ความสัมพันธ์เช่นกัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
บทสรุปและวิจารณ์ผลการดำเนินงาน

5.1 บทสรุปปริญญานิพนธ์

จากการแข่งขันทางการตลาดของผู้ให้บริการสัญญาณโทรศัพท์ที่สูงขึ้นในปัจจุบัน ทำให้ทุกค่ายโทรศัพท์ล้วนต้องการรู้ความต้องการของลูกค้า ความสนใจของลูกค้า เพื่อที่จะได้ออกโปรโมชั่น แคมเปญ หรือข้อเสนอต่าง ๆ เพื่อดึงดูดให้ลูกค้าของตนเองเพิ่มขึ้น ซึ่งการที่จะรู้ความต้องการของลูกค้า ว่าลูกค้าต่างมีความต้องการอะไร สนใจหรือมีความชอบด้านไหน คือความท้าทายของแต่ละค่ายที่จะต้องทำการวิเคราะห์ข้อมูลหารูปแบบออกมา เพื่อตอบโจทย์ความต้องการของลูกค้าในปัจจุบัน จากการที่ได้ทำการวิจัยข้อมูลการใช้งานโมบายแอปพลิเคชัน จะพบว่ากลุ่มตัวอย่างที่นำมาวิเคราะห์ มีความสนใจไปในทางด้านแอปพลิเคชันที่ให้ความบันเทิง และโซเชียลต่าง ๆ

5.2 ปัญหาที่พบในระหว่างการดำเนินงาน และแนวทางแก้ปัญหา

การดำเนินงานในการวิจัยครั้งนี้เป็นไปได้อย่างล่าช้า อันเนื่องมาจากเกิดปัญหากับระบบความปลอดภัย (Security) ของทางองค์กร ทำให้ต้องมีการรอการติดตั้งระบบสำหรับเข้าไปใช้งานใหม่ ซึ่งมีการใช้ระยะเวลาค่อนข้างนานประมาณสองเดือน ในการขอสิทธิ์ในการเข้าถึงข้อมูลต่าง ๆ ดังนั้นในระหว่างการรอระบบ และปัญหาที่ตามมาหลังจากเรื่องของระยะเวลาของระบบเสร็จก็คือ การเรียกใช้ข้อมูลหรือการจัดกลุ่มข้อมูลจำเป็นจะต้องมีวิธีที่สะดวกในการซิงโครไนท์ (synchronize) โปรแกรมที่พัฒนากับทางจุฬาฯ ร่วมกับโปรแกรมของเอไอเอส ทางแก้ก็คือใช้เป็นการนำข้อมูลที่ผ่านการตรวจสอบแล้วว่าไม่สามารถที่จะระบุตัวตนได้ ส่งเป็นอีเมลออกไปยังเครื่องข้างนอกระบบละเมื่อประมวลผลเสร็จก็ส่งข้อมูลกลับมาทางอีเมล หรือซีดีเพื่อนำไปวิเคราะห์ผลต่อไป

5.3 ข้อเสนอแนะ

ในการวิเคราะห์ผลจากแอปพลิเคชันนั้น ได้ใช้ข้อมูลแค่เพียงสามวัน ซึ่งอาจจะทำให้ได้ผลลัพธ์ที่ยังไม่แม่นยำมากนัก เนื่องจากความสนใจและข้อมูลในแต่ละวันมีการเปลี่ยนแปลงไปตลอดเวลา ดังนั้นจึงควรคอยทำการดูกลุ่มคลัสเตอร์อย่างสม่ำเสมอ เพื่อให้สามารถนำไปใช้ในการช่วยตัดสินใจสำหรับกระบวนการต่าง ๆ ได้ต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[1] **ขั้นตอนทำเหมืองข้อมูล** [ออนไลน์] เข้าถึงได้จาก :
http://datamining-techniques.blogspot.com/2012/09/data-mining.html
(วันที่ค้นหาข้อมูล 7 พฤศจิกายน 2559)

[2] **การจัดกลุ่ม (Clustering)** [ออนไลน์] เข้าถึงได้จาก :
datamining-techniques.blogspot.com/2012/09/clustering.html
(วันที่ค้นหาข้อมูล 7 พฤศจิกายน 2559)

[3] **การจัดกลุ่มโดยวิธีเคมีน (K-Mean)** [ออนไลน์] เข้าถึงได้จาก :
http://datamining-techniques.blogspot.com/2012/09/k-means-k-means-คลัสเตอร์ing.html
(วันที่ค้นหาข้อมูล 15 พฤศจิกายน 2559)

[4] **เทคนิคการจัดกลุ่มแบบลำดับชั้น** [ออนไลน์] เข้าถึงได้จาก :
www.saruthipong.com/port/document/299-705/299-705-10.pdf
(วันที่ค้นหาข้อมูล 15 พฤศจิกายน 2559)

[5] **ภาพแสดงลิงเกจเมธอดแต่ละวิธี** [ออนไลน์] เข้าถึงได้จาก :
https://www.multid.se/genex/onlinehelp/hs515.htm
(วันที่ค้นหาข้อมูล 20 พฤศจิกายน 2559)

[6] **หลักเกณฑ์ในการรวมกลุ่ม** [ออนไลน์] เข้าถึงได้จาก :
http://slideplayer.in.th/slide/2226485
(วันที่ค้นหาข้อมูล 7 ธันวาคม 2559)

[7] **เทคนิคต่าง ๆ ของ การทำเหมืองข้อมูล** [ออนไลน์] เข้าถึงได้จาก :
llem0nz.blogspot.com/2012/03/data-mining_8699.html
(วันที่ค้นหาข้อมูล 7 ธันวาคม 2559)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง (ต่อ)

[8] ขั้นตอนการทำการจำแนกข้อมูล [ออนไลน์] เข้าถึงได้จาก :
https://wipawanblog.files.wordpress.com/.../chapter-5-classification-decision-tree.pdf
(วันที่ค้นหาข้อมูล 8 ธันวาคม 2559)

[9] **เครือข่ายประสาทเทียมเพื่อวิเคราะห์การเช่าและซื้อบ้านของลูกค้า [ออนไลน์] เข้าถึงได้จาก :**
https://nmcxpress.files.wordpress.com/2012/03/x9.png
(วันที่ค้นหาข้อมูล 8 ธันวาคม 2559)

[10] **ตัวอย่างผลลัพธ์การทำคลัสเตอร์** [ออนไลน์] เข้าถึงได้จาก :
http://2.bp.blogspot.com/LuhBNBJYGCg/T1J95vXyTLI/AAAAAAAAACs/pdGbgcUZjrM/s1600/Untitled.png
(วันที่ค้นหาข้อมูล 12 ธันวาคม 2559)

[11] **ข้อมูลขนาดใหญ่ (Big Data)** [ออนไลน์] เข้าถึงได้จาก :
http://km.buu.ac.th/article/frontend/article_detail/94
(วันที่ค้นหาข้อมูล 12 ธันวาคม 2559)

[12] **องค์ประกอบของข้อมูลขนาดใหญ่** [ออนไลน์] เข้าถึงได้จาก :
https://thanachartdotorg.files.wordpress.com/2013/10/hadoop.jpg
(วันที่ค้นหาข้อมูล 12 ธันวาคม 2559)

[13] **สวิตช์** [ออนไลน์] เข้าถึงได้จาก :
https://krengikung.wordpress.com/
(วันที่ค้นหาข้อมูล 20 ธันวาคม 2559)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก
โปสเตอร์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สหกิจศึกษา ปีการศึกษา 2559

การวิเคราะห์ข้อมูลและสำรวจเชิงลึกเพื่อหาพฤติกรรมการใช้งาน Mobile Application

ชื่อนักศึกษา : นางสาวปิยาภรณ์ เสาให้สกุณีย์ อาจารย์ที่ปรึกษา : ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์
สาขาวิชาวิศวกรรมสารสนเทศ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

องค์กร และบริษัทต่างๆดำเนินธุรกิจภายใต้ภาวะการแข่งขัน และเผชิญกับสถานการณ์ที่ต้องตัดสินใจว่าจะยังคงใช้แผนงานเดิมๆที่มีอยู่ หรือจะหาแผนงานใหม่ๆ เพื่อตอบสนองกับความต้องการของลูกค้าในปัจจุบัน และอนาคต ซึ่งวัตถุประสงค์ที่สำคัญในการวิเคราะห์ด้านความบันเทิงนี้ก็เพื่อเป็นการเพิ่มแนวทางในการดูความสนใจของลูกค้า และนำเสนอบริการได้ตรงตามความสนใจ

วัตถุประสงค์

1.เพื่อจัดกลุ่มลูกค้าตามความสนใจ
2.สามารถนำโครงงานไปประกอบการตัดสินใจต่างๆ ในการออกโปรโมชั่น
3.เพื่ออัพเดทข่าวสารให้กับลูกค้าตามความสนใจ
4.ใช้ R และ Tableau ในการวิเคราะห์และสร้างแบบจำลองจากข้อมูลได้

โปรแกรมและภาษาที่ใช้

โปรแกรมที่ใช้ - Jupyter(from Anaconda)
- Rstudio
- Tableau

ภาษาที่ใช้ - Python
- R
- SQL

วิธีที่ใช้ในการวิเคราะห์ข้อมูล

1. K-mean for Cluster (ใช้ R และ Python)
2. Apriori for Association Rule (ใช้ R)

การวิเคราะห์ข้อมูล

1. การจัดกลุ่มคลัสเตอร์โดย K-mean

การเลือกค่า K (ใช้ภาษา R) : การจะกำหนดกลุ่มที่เหมาะสม จะหาได้จากค่า Wss โดยในที่นี้จะใช้ภาษา R หาค่า K



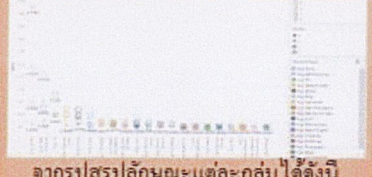
จากรูปทำให้เราเลือกค่า K =4 (ใช้ภาษา Python) และ K = 5(ใช้ภาษา R)

ผลลัพธ์เมื่อ K=5 (R)



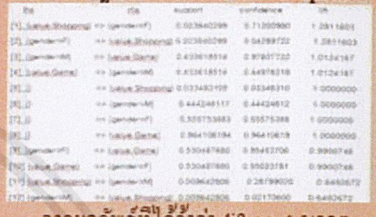
จากรูปสามารถสรุปได้ว่ากลุ่มแรกมีการใช้ Traffic มากแต่จำนวนผู้ใช้งานน้อย และกลุ่มที่จำนวนผู้ใช้งานส่วนใหญ่มาก มี Traffic น้อยซึ่งเป็นไปตามกฎธรรมชาติ และส่วนใหญ่เสียการใช้งานไปกับเรื่องของ Game และ Steaming

ผลลัพธ์เมื่อ K=4 (Python)



จากรูปสรุปลักษณะแต่ละกลุ่มได้ดังนี้
กลุ่ม 1: Social > Web > Media & Video
กลุ่ม 2: IM > Social > Web
กลุ่ม 3: IM > Web > Social
กลุ่ม 4: Social > Web > IM

2. การหากฎความสัมพันธ์โดยวิธี Apriori



จากผลลัพธ์ที่ได้ถ้าค่า lift >> 1 มากๆ จะเห็นว่าถ้าพูดถึงเรื่อง Shopping ก็ไปโฟกัสที่เพศหญิงได้ กรณีค่า lift = 1 คือ อิสระต่อกัน และสำหรับค่า lift << 1 หมายถึงผลลัพธ์ที่ได้มีความสัมพันธ์ที่ตรงข้ามกัน จะเห็นว่าถ้าพูดถึงการ shopping แล้วเพศชายไม่ค่อยมีความสัมพันธ์

ประโยชน์ที่ได้รับ

1. ได้เรียนรู้และฝึกฝนการเขียนโปรแกรมภาษา R และ Python
2. ได้นำความรู้ที่เรียนมาลองใช้งานจริงๆ ในการทำ Data Mining
3. ได้เรียนรู้การทำงานร่วมกับคนอื่นๆ และได้รับความรู้จากนอกตำราเรียน

กิตติกรรมประกาศ

ขอขอบคุณพนักงานฝ่าย BIS/DAE จากทาง AIS ที่ให้ความช่วยเหลือและดูแลเป็นอย่างดี และขอขอบคุณอาจารย์ที่ปรึกษา ผศ.ดร.สุธีรา พันธุ์ธีรานุรักษ์ที่คอยให้คำแนะนำ คำปรึกษาและคอยรับฟังและช่วยเหลือปัญหาต่าง และท้ายที่สุดข้าพเจ้าขอขอบคุณครอบครัวที่คอยให้กำลังใจที่ดีแก่ข้าพเจ้าเสมอมา

ภาพที่ ก.1 โปสเตอร์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้