# ใบรับรองวิทยานิพนธ์

**บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์**

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

**ปริญญา**

| วิศวกรรมคอมพิวเตอร์ | วิศวกรรมคอมพิวเตอร์ |
|---|---|
| **สาขา** | **ภาควิชา** |

**เรื่อง** การตรวจจับความผิดปกติของทราฟฟิกและลักษณะเครือข่ายเพื่อจัดกลุ่มความผิดปกติของทราฟฟิก กรณีศึกษา: บริษัท ทีโอที จำกัด (มหาชน)

Traffic Anomaly Detection and Characterization to Cluster Traffic Anomalies Case Study: TOT Public Company Limited

**นามผู้วิจัย** นางสาวเบญจวรรณ สุขพัฒนศรีกุล

**ได้พิจารณาเห็นชอบโดย**

**อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก** .....

( รองศาสตราจารย์ศิริพร อ่องรุ่งเรือง, M.S. )

**อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม** .....

( ผู้ช่วยศาสตราจารย์จิตร์ทัศน์ ฝักเจริญผล, Ph.D. )

**อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม** .....

( อาจารย์ชัยพร ใจแก้ว, Ph.D. )

**หัวหน้าภาควิชา** .....

( ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D. )

**บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว**

.....

( รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr. )

**คณบดีบัณฑิตวิทยาลัย**

**วันที่** ..... **เดือน** ..... **พ.ศ.** .....

วิทยานิพนธ์

เรื่อง

# การตรวจจับความผิดปกติของทราฟฟิกและลักษณะเครือข่ายเพื่อจัดกลุ่มความผิดปกติของทราฟฟิก กรณีศึกษา: บริษัท ทีโอที จำกัด (มหาชน)

# Traffic Anomaly Detection and Characterization to Cluster Traffic Anomalies Case Study: TOT Public Company Limited


โดย

นางสาวเบญจวรรณ สุขพัฒนศรีกุล


เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์
เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)
พ.ศ. 2552

เบญจวรรณ สุขพัฒเนศรีกุล 2552: การตรวจจับความผิดปกติของทราฟฟิกและลักษณะเครือข่ายเพื่อจัดกลุ่มความผิดปกติของทราฟฟิก กรณีศึกษา: บริษัท ทีโอที จำกัด (มหาชน) ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: รองศาสตราจารย์ศิริพร อ่องรุ่งเรือง, M.S. 85 หน้า

ระบบเครือข่ายขององค์กรมีความเสี่ยงสูงที่จะถูกโจมตีหรือถูกการก่อกวนอย่างไม่ได้คาดคิดมาก่อน แม้ว่าจะมีการใช้งานระบบความปลอดภัยที่เหมาะสมอยู่แล้วก็ตาม วัตถุประสงค์ของโครงงานการค้นคว้าวิจัยนี้จะใช้เทคนิคการทำเหมืองข้อมูลด้วยเทคนิคต่างๆ ซึ่งประกอบด้วยอัลกอริทึม sIB RandomFlatClustering FarthestFirst FilteredClusterer และ K-Means เพื่อนำมาจัดกลุ่มข้อมูลประเภท TCP/IP แพ็กเก็ต และทำการเปรียบเทียบผลการจัดกลุ่มของแต่ละเทคนิคเพื่อหาอัลกอริทึมที่เหมาะสมที่สุดสำหรับใช้ตรวจจับความผิดปกติของทราฟฟิก ผลลัพธ์ของการจัดกลุ่มข้อมูลจะสามารถนำไปสร้างกฎใหม่เพื่อนำใช้ในเครื่องมือที่ใช้ในการตรวจจับความผิดปกติได้

_____ _____ ___ / ___ / ___

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Benjawan Sukpattanasrikul 2009: Traffic Anomaly Detection and Characterization to Cluster Traffic Anomalies Case Study: TOT Public Company Limited. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Siriporn Ongroongrueng, M.S. 85 pages.

Network System of an organization has a high risk to be unexpectedly attacked or disturbed even though an appropriate security system has been provided. The objective of this research project is to apply various clustering data mining techniques, sIB, RandomFlatClustering, FarthestFirst, FilteredClusterer and K-Means to perform TCP/IP packet clustering and compare the anomaly detection efficiency of each technique to find out which algorithm is the most appropriate one for detection of traffic anomaly. The clustering result has been used to create new rules for detection software tool to improve its detection capability.

_____ _____ ___ / ___ / ___

Student's signature Thesis Advisor's signature

# กิตติกรรมประกาศ

กราบขอบพระคุณ รศ.ศิริพรอ่องรุ่งเรือง ประธานกรรมการที่ปรึกษาวิทยานิพนธ์ ผศ.ดร.จิตร์ทัศน์ ฝักเจริญผล กรรมการที่ปรึกษาวิชาเอก และ ดร.ชัยพร ใจแก้ว กรรมการที่ปรึกษาวิชารอง ที่ให้คำปรึกษาในการเรียน การค้นคว้าการวิจัย ตลอดจนการตรวจแก้ไขวิทยานิพนธ์จนกระทั่งเสร็จสมบูรณ์และขอกราบขอบพระคุณผู้แทนบัณฑิตวิทยาลัย ที่ให้ความกรุณาตรวจแก้ไขวิทยานิพนธ์ให้สมบูรณ์ยิ่งขึ้น

วิทยานิพนธ์นี้ได้รับการสนับสนุนจากห้องปฏิบัติการวิจัย Database and Software Technology Research Laboratory โดยงบประมาณของโครงการบัณฑิตวิทยาลัย และคณะวิศวกรรมศาสตร์

เบญจวรรณ สุขพัฒนศรีกุล

พฤษภาคม 2552

# สารบัญ

หน้า


| | |
|---|---|
| สารบัญ | (1) |
| สารบัญตาราง | (2) |
| สารบัญภาพ | (4) |
| คำนำ | 1 |
| วัตถุประสงค์ | 3 |
| การตรวจเอกสาร | 4 |
| อุปกรณ์และวิธีการ | 43 |
| อุปกรณ์ | 43 |
| วิธีการ | 43 |
| ผลและวิจารณ์ | 52 |
| ผล | 52 |
| วิจารณ์ | 79 |
| สรุปและข้อเสนอแนะ | 81 |
| สรุป | 81 |
| ข้อเสนอแนะ | 82 |
| เอกสารและสิ่งอ้างอิง | 83 |
| ประวัติการศึกษาและการทำงาน | 86 |

# สารบัญตาราง


| ตารางที่ | | หน้า |
|---|---|---|
| 1 | ตัวอย่างข้อมูลเพื่อนำมาแบ่งกลุ่มโดย K-means | 25 |
| 2 | ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม | 26 |
| 3 | แสดงค่าเฉลี่ยใหม่ของแต่ละกลุ่ม | 27 |
| 4 | ชุดตัวอย่างสำหรับการเรียนรู้เบย์อย่างง่าย | 33 |
| 5 | ชุดตัวอย่างข้อมูลที่ต้องการจำแนกประเภท | 33 |
| 6 | ตารางสรุปงานวิจัยที่เกี่ยวข้อง | 40 |
| 7 | แสดงตารางการดำเนินงาน | 51 |
| 8 | แสดงผลสรุปการเก็บข้อมูลทราฟฟิกโดยแยกตามโปรโตคอล | 53 |
| 9 | เปรียบเทียบผลการจัดกลุ่มด้วยอัลกอริทึมทั้ง 5 อัลกอริทึม | 67 |
| 10 | การสรุป term context | 67 |
| 11 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision ของข้อมูลกลุ่มที่ 1 | 68 |
| 12 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision ของข้อมูลกลุ่มที่ 2 | 68 |
| 13 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Recall ของข้อมูลกลุ่มที่ 1 | 69 |
| 14 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Recall ของข้อมูลกลุ่มที่ 2 | 69 |
| 15 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision Recall และ F-Measure ของข้อมูลกลุ่มที่ 1 | 70 |
| 16 | ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision Recall และ F-Measure ของข้อมูลกลุ่มที่ 2 | 72 |
| 17 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Destination_port | 74 |
| 18 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Packet_length | 75 |
| 19 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Time | 75 |

# สารบัญตาราง (ต่อ)


| ตารางที่ | | หน้า |
|---|---|---|
| 20 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Destination_port และ Packet_length | 76 |
| 21 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Destination_port และ Time | 76 |
| 22 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Time และ Packet_length | 77 |
| 23 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Source_port Destination_port และ Packet_length | 77 |
| 24 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Time Source_port และ Destination_port | 78 |
| 25 | แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Destination_port Packet_length และ Time | 78 |

# สารบัญภาพ


| ภาพที่ | | หน้า |
|---|---|---|
| 1 | แสดงหน้าจอที่ให้แสดง display filter ในส่วน filter | 15 |
| 2 | แสดงหน้าจอที่ให้แสดง capture filter ในส่วน filter | 16 |
| 3 | แสดงตัวอย่างผลลัพธ์จากการทำ tcpdump | 17 |
| 4 | แสดงขั้นตอนการทำ Data Mining | 21 |
| 5 | แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึมเค-มีน (K-Means) | 25 |
| 6 | แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม FarthestFirst | 28 |
| 7 | แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม RandomFlatClustering | 30 |
| 8 | Pseudocode ของอัลกอริทึม sequential IB Clustering | 31 |
| 9 | แสดง Packet ที่เกิดจากการจับแพกเก็ตโดยใช้โปรแกรม Ethereal | 45 |
| 10 | แสดงขั้นตอนการทำเหมืองข้อมูล | 46 |
| 11 | แสดงโครงสร้างของ Data mining | 46 |
| 12 | แสดงตัวอย่างการเก็บข้อมูลด้วยโปรแกรม Ethereal | 53 |
| 13 | แสดงการเก็บรวบรวมข้อมูลในรูปแบบ CSV file | 54 |
| 14 | ตัวอย่างโปรแกรม RapidMiner | 55 |
| 15 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม K-means ด้วยการ plot ข้อมูลในแบบ 2 มิติ ระหว่าง Attribute Packet_length และ Attribute อื่นๆ ที่เหลือ | 56 |
| 16 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม K-means ด้วยการ plot ข้อมูลในแบบ 2 มิติ ระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ | 57 |
| 17 | แสดงการกระจายตัวของข้อมูลด้วยการ plot แบบวิธี Self-Organizing Map (SOM) | 58 |
| 18 | แสดงการกระจายตัวของข้อมูลด้วยการ plot แบบวิธี SOM แบบ 2 ทิศทาง | 58 |
| 19 | แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port | 59 |
| 20 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม FarthestFirst ด้วยการ plot ข้อมูลในแบบ 2 มิติ ระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ | 60 |

# สารบัญภาพ (ต่อ)


| ภาพที่ | | หน้า |
|---|---|---|
| 21 | แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port | 61 |
| 22 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม FilteredClusterer ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ | 62 |
| 23 | แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port | 63 |
| 24 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม Sequential IB Clustering (sIB) ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ | 64 |
| 25 | แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port | 65 |
| 26 | แสดงผลการจัดกลุ่มด้วยอัลกอริทึม RandomFlatClustering ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ | 66 |
| 27 | แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port | 66 |
| 28 | แสดงกราฟแท่งเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 1 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม | 71 |
| 29 | แสดงกราฟเส้นเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 1 ด้วยค่า Precision Recall และ F-measure | 71 |
| 30 | กราฟแท่งเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 2 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม | 72 |
| 31 | แสดงกราฟเส้นเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 2 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม | 73 |

# การตรวจจับความผิดปกติของทราฟฟิกและลักษณะเครือข่ายเพื่อจัดกลุ่มความผิดปกติของทราฟฟิก กรณีศึกษา: บริษัท ทีโอที จำกัด (มหาชน)

# Traffic Anomaly Detection and Characterization to Cluster Traffic Anomalies Case Study: TOT Public Company Limited

## คำนำ

ปัจจุบันการใช้งานทางด้านเครือข่ายอินเทอร์เน็ตได้รับความนิยมและแพร่หลายอย่างรวดเร็วและครอบคลุมหลากหลายกลุ่มบุคคลมากยิ่งขึ้น เทคโนโลยีทางด้านเครือข่ายอินเทอร์เน็ตที่มีความก้าวหน้าอย่างรวดเร็วนี้อาจทำให้ผู้ใช้ไม่คาดคิดว่ากำลังเสี่ยงต่อการถูกโจมตี การบุกรุกทางเครือข่ายหรือการถูกก่อกวนในการใช้งานทางด้านเครือข่ายได้ หากผู้ใช้งานไม่มีวิธีป้องกันความเสี่ยงเหล่านี้อาจทำให้เกิดความเสียหายกับระบบเครือข่ายของผู้ใช้ได้ ผู้วิจัยจึงได้ศึกษาวิธีที่จะช่วยในการป้องกันการบุกรุก การถูกโจมตี และถูกก่อกวนในการใช้งานระบบเครือข่ายคอมพิวเตอร์ โดยการประยุกต์ใช้งานทางด้านการทำเหมืองข้อมูล (Data mining) เพื่อมาวิเคราะห์และตรวจจับความผิดปกติของทราฟฟิกและลักษณะเครือข่ายเพื่อแยกประเภทความผิดปกติของทราฟฟิก

การทำเหมืองข้อมูลมีการใช้งานกันอย่างแพร่หลายไม่ว่าจะเป็นการใช้งานทางด้านธุรกิจ ด้านเศรษฐศาสตร์ ด้านชีววิทยา ด้านวิทยาศาสตร์คอมพิวเตอร์ เป็นต้น ซึ่งการนำมาใช้งานทางด้านเครือข่ายอินเทอร์เน็ตยังไม่ค่อยได้รับความนิยมมากนัก อาจเนื่องมาจากความยุ่งยากในการเก็บข้อมูลซึ่งมีปริมาณมหาศาล และมีความซับซ้อนมาก แต่ทางผู้วิจัยมองเห็นว่ากระบวนการทางด้านการทำเหมืองข้อมูลเป็นวิธีหนึ่งที่เหมาะสมและสามารถนำมาประยุกต์ใช้งานทางด้านเครือข่ายคอมพิวเตอร์ได้

ผู้วิจัยจึงได้นำการทำเหมืองข้อมูลและการค้นหาความรู้ (Data Mining and Knowledge Discovery) ซึ่งเป็นหลักการที่ใช้ในการค้นหาความสัมพันธ์ระหว่างข้อมูล หรือค้นหาความรู้ที่เป็นประโยชน์ที่อาจซ่อนอยู่ในข้อมูล โดยอาศัยแนวคิดและหลักการด้านฐานข้อมูล ปัญญาประดิษฐ์ และอัลกอริทึมที่มีประสิทธิภาพสำหรับการใช้งานกับฐานข้อมูลขนาดใหญ่ รวมถึงการประยุกต์ใช้วิธีการต่าง ๆ กับสถานการณ์จริงมาเป็นตัวช่วยในการแบ่งแยกประเภทความผิดปกติของข้อมูล

ทราฟฟิกในเครือข่าย และทำให้สามารถแบ่งแยกความผิดปกติได้อย่างอัตโนมัติผ่านการเรียนรู้ โดยแสดงการใช้ลักษณะการกระจายตัว ความผิดปกติที่ตกลงมาโดยธรรมชาติเข้าไปในกลุ่มที่แตกต่างกันและมีความหมาย

ในงานวิจัยนี้ผู้วิจัยได้นำเสนอวิธีการทำเหมืองข้อมูล 5 วิธีด้วยกันประกอบด้วย sIB, RandomFlatClustering, FarthestFirst, FilteredClusterer และ k-means ซึ่ง Erman et al. ได้นำเสนอมาก่อนหน้านี้ในการนำมาใช้กับข้อมูลประเภทเครือข่ายอินเทอร์เน็ต งานวิจัยนี้ได้อธิบายวิธีการวิจัยอย่างเป็นลำดับขั้นตอน เริ่มตั้งแต่การเก็บข้อมูลทราฟฟิกจริง การคัดเลือกและกรองข้อมูล การแปลงข้อมูลเพื่อให้สามารถรองรับหรือใช้งานได้กับอัลกอริทึมแต่ละอัลกอริทึม และได้แสดงผลจากการทดลองอย่างละเอียด สำหรับการวัดประสิทธิภาพผู้วิจัยจะนำผลจากการแบ่งกลุ่มข้อมูลมาวิเคราะห์เพื่อหาอัลกอริทึมที่สามารถแบ่งกลุ่มข้อมูลได้อย่างมีประสิทธิภาพและเหมาะสมที่สุดด้วยการพิจารณาจากค่า Precision Recall และค่า F-Measure ในส่วนสุดท้ายผู้วิจัยได้เสนอแนะถึงข้อบกพร่องและได้ชี้แนวทางในการพัฒนาต่อสำหรับผู้ที่สนใจ

# วัตถุประสงค์

1. เพื่อศึกษาและทดสอบวิธีในการแยกประเภทความผิดปกติของทราฟฟิกที่เหมาะสมได้ โดยใช้ 5 อัลกอริทึม คือ sIB, RandomFlatClustering, FarthestFirst, and FilteredClusterer และ K-Means

2. สามารถนำไปพัฒนาเพื่อเพิ่มความสามารถของเครื่องมือตรวจจับความผิดปกติของทราฟฟิกในอนาคตได้

3. เพื่อแยกประเภทความผิดปกติของข้อมูลทราฟฟิก

# การตรวจเอกสาร

## ความรู้พื้นฐาน

### 1. ระบบเครือข่าย (Network System)

ธีรภัทร์ (2552) ได้อธิบายถึงระบบเครือข่าย ไว้ว่าเป็นการนำคอมพิวเตอร์หลาย ๆ เครื่องมาต่อพ่วงกัน เพื่อใช้ในการสื่อสารถึงกัน ใช้ข้อมูลร่วมกัน รวมทั้งใช้อุปกรณ์ร่วมกัน ทำให้ประหยัดทรัพยากรในการใช้งาน เช่น printer, hard disk เป็นต้น เมื่อมีการเชื่อมต่อกับเครือข่ายอื่นๆ ที่อยู่ห่างไกล เช่น ระบบอินเตอร์เน็ต ซึ่งเป็นเครือข่ายที่เชื่อมต่อคอมพิวเตอร์ทั่วโลก ทำให้สามารถแลกเปลี่ยนข้อมูลได้กับคนทั่วโลก สิ่งที่ต้องคำนึงถึงการใช้งานระบบอินเตอร์เน็ต นั่นก็คือ ความปลอดภัยของข้อมูล เนื่องจากทุกคนสามารถเข้าถึงข้อมูลทุกอย่างที่แลกเปลี่ยนผ่านอินเตอร์เน็ตได้ ดังนั้นอินเตอร์เน็ตจึงเป็นเครือข่ายสาธารณะที่ไม่มีหลักประกันความปลอดภัยของข้อมูลที่ส่งผ่านอินเตอร์เน็ต การรักษาความปลอดภัยในเครือข่ายจึงเป็นสิ่งที่สำคัญและจำเป็นสำหรับองค์กร เนื่องจากข้อมูลและเครื่องมือที่ใช้สำหรับการบุกรุกหรือการโจมตีระบบเครือข่ายนั้น สามารถหาจากอินเตอร์เน็ตได้ง่าย และยังง่ายต่อการใช้งาน ดังนั้นผู้ดูแลระบบจำเป็นที่จะต้องวางแผนและติดตั้งระบบการรักษาความปลอดภัยในเครือข่ายได้ดีด้วย

### 2. การโจมตีเครือข่าย

เครือข่ายเป็นเทคโนโลยีที่น่าอัศจรรย์ แต่ก็ยังคงมีความเสี่ยงอยู่มาก ถ้าไม่มีการควบคุมหรือป้องกันที่ดี การโจมตีหรือการบุกรุกเครือข่ายหมายถึงความพยายามที่จะเข้าใช้ระบบ (Access Attack) การแก้ไขข้อมูลหรือระบบ (Modification Attack) การทำให้ระบบไม่สามารถใช้การได้ (Deny of Service Attack) และการทำให้ข้อมูลเป็นเท็จ (Repudiation Attack) ซึ่งจะกระทำโดยผู้ประสงค์ร้าย ผู้ที่ไม่มีสิทธิ์ หรืออาจเกิดจากความไม่ตั้งใจของผู้ใช้เอง ต่อไปนี้เป็นรูปแบบต่างๆ ที่ผู้ไม่ประสงค์ดีพยายามที่จะบุกรุกเครือข่ายเพื่อลักลอบข้อมูลที่สำคัญหรือเข้าใช้ระบบโดยไม่ได้รับอนุญาต

## 2.1 แพคเก็ตสนิฟเฟอร์

ข้อมูลที่คอมพิวเตอร์ส่งผ่านเครือข่ายนั้นจะถูกแบ่งย่อยเป็นก้อนเล็กๆ ซึ่งจะเรียกว่า “แพกเก็ต (Packet)” แอพพลิเคชันหลายชนิดจะส่งข้อมูลโดยที่ไม่ได้เข้ารหัส (Encryption) หรือในรูปแบบเคลียร์เท็กซ์ (Clear Text) ดังนั้นข้อมูลอาจถูกคัดลอกและโพรเซสโดยแอพพลิเคชันอื่นก็ได้ เน็ตเวิร์คโปรโตคอลเป็นตัวกำหนดหมายเลขของแต่ละแพกเก็ต ซึ่งเป็นสิ่งที่คอมพิวเตอร์ใช้สำหรับบ่งบอกว่าแพกเก็ตนั้นส่งไปไหนหรือมาจากไหน เนื่องจากโปรโตคอลที่ใช้ส่วนใหญ่ เช่น TCP/IP เป็นโปรโตคอลมาตรฐานและเป็นที่รู้จักกันโดยทั่วไป ทำให้บางกลุ่มพัฒนาแอพพลิเคชันตรวจจับแพกเก็ตที่วิ่งบนเครือข่ายได้ ซึ่งเทคนิคนี้เรียกว่า “แพกเก็ตสนิฟเฟอร์ (Packet Sniffer)” ในปัจจุบันนี้โปรแกรมแพกเก็ตสนิฟเฟอร์มีให้ดาวน์โหลดบนอินเตอร์เน็ตมากมาย และผู้ที่ใช้งานไม่จำเป็นต้องมีความรู้เกี่ยวกับคอมพิวเตอร์มากก็สามารถใช้ซอฟต์แวร์เหล่านี้ได้ แพกเก็ตสนิฟเฟอร์เป็นโปรแกรมใช้เน็ตเวิร์คการ์ดในโหมดโพรมิสเชียร์ (Promiscuous mode) ซึ่งในโหมดนี้เน็ตเวิร์คการ์ดจะรับทุกๆแพกเก็ตที่วิ่งบนสายสัญญาณและส่งต่อไปให้ยังแอพพลิเคชันเพื่อโพรเซสต่อไป เนื่องจากแอพพลิเคชันส่วนใหญ่จะส่งแพกเก็ตแบบเคลียร์เท็กซ์ แพกเก็ตสนิฟเฟอร์สามารถตรวจจับข้อมูลที่อาจเป็นประโยชน์ได้ เช่น ชื่อผู้ใช้และรหัสผ่าน เป็นต้น ถ้าหากมีการใช้ฐานข้อมูลผ่านเครือข่าย แพกเก็ตสนิฟเฟอร์อาจโจมตีโดยใช้ชื่อผู้ใช้และรหัสผ่านที่ตรวจจับได้ก็ได้ สิ่งที่กลัวมากกว่าคือ ผู้ใช้มักจะใช้ชื่อผู้ใช้ และรหัสผ่านเดิมกับทุกๆ แอพพลิเคชัน ทำให้ผู้บุกรุกสามารถโจมตีแอพพลิเคชันต่างๆ ได้อย่างง่ายดายและอาจทำให้เครือข่ายเกิดความเสียหายมากกว่าที่คิด

## 2.2 ไอพีสปูฟิง

ไอพีสปูฟิง (IP Spoofing) หมายถึงการที่ผู้บุกรุกอยู่นอกเครือข่ายแล้วแกล้งทำเป็นว่าเป็นคอมพิวเตอร์ที่เชื่อถือได้ (Trusted) โดยอาจจะใช้ไอพีแอดเดรสเหมือนกับที่ใช้ในเครือข่ายหรืออาจจะใช้ไอพีแอดเดรสข้างนอกที่เครือข่ายเชื่อว่าเป็นคอมพิวเตอร์ที่เชื่อถือได้หรืออนุญาตให้เข้าใช้ทรัพยากรในเครือข่ายได้ โดยปกติแล้วการโจมตีแบบไอพีสปูฟิงเป็นการเปลี่ยนแปลง หรือเพิ่มข้อมูลเข้าไปในแพกเก็ตที่รับส่งระหว่างไคลเอนท์และเซิร์ฟเวอร์ หรือคอมพิวเตอร์สื่อสารกันในเครือข่าย การที่จะทำอย่างนี้ได้ผู้บุกรุกจะต้องปรับราท์ติ้งเทเบิ้ลของเราท์เตอร์เพื่อให้ส่งแพกเก็ตไปที่เครื่องของผู้บุกรุก หรืออีกวิธีหนึ่งคือการที่ผู้บุกรุกสามารถแก้ไขให้แอพพลิเคชันส่งข้อมูลที่เป็นประโยชน์ต่อการเข้าถึงแอพพลิเคชันนั้นผ่านทางอีเมลล์ หลังจากนั้นผู้บุกรุกก็สามารถเข้าใช้แอพพลิเคชันได้โดยใช้ข้อมูลดังกล่าว

อย่างไรก็ตาม ถ้าผู้บุกรุกสามารถปรับเปลี่ยนเราท์ติ้งเทเบิ้ลเพื่อให้ส่งข้อมูลไปยังเครื่องปลอมได้ ผู้บุกรุกสามารถรับส่งข้อมูลกับแอพพลิเคชันนั้นเสมือนเป็นหนึ่งในผู้ใช้ทั่วๆ ไปได้ ไอพีสปูฟิงไม่จำเป็นจะต้องเป็นคอมพิวเตอร์ที่อยู่นอกเครือข่ายเท่านั้น แต่อาจจะเป็นผู้ใช้ที่อยู่ข้างในที่ไม่มีสิทธิ์ก็ได้ ซึ่งอย่างที่เป็นที่ทราบกันดีว่า การโจมตีเครือข่ายนั้น 90% จะเป็นการโจมตีที่เกิดจากภายในเครือข่ายเอง

2.3 การโจมตีรหัสผ่าน

การโจมตีรหัสผ่าน (Password Attack) หมายถึง การโจมตีที่ผู้บุกรุกพยายามเดารหัสผ่านของผู้ใช้คนใดคนหนึ่ง ซึ่งวิธีการเดานั้นก็มีหลายวิธี เช่น บรู๊ทฟอร์ซ (Brute-Force), โทรจันฮอร์ส (Trojan Horse), ไอพีสปูฟิง, แพกเก็ตสนิฟเฟอร์ เป็นต้น การเดาแบบบรู๊ทฟอร์ซ หมายถึง การลองผิดลองถูกรหัสผ่านเรื่อยๆ จนกว่าจะถูก บ่อยครั้งที่การโจมตีแบบบรู๊ทฟอร์ซใช้การพยายามล็อกอินเข้าใช้รีซอร์สของเครือข่าย โดยถ้าทำสำเร็จผู้บุกรุกก็จะมีสิทธิ์เหมือนกับเจ้าของแอ็คเคาท์นั้นๆ ถ้าหากแอ็คเคาท์นี้มีสิทธิ์เพียงพอผู้บุกรุกอาจสร้างแอ็คเคาท์ใหม่เพื่อเป็นประตูหลัง (Back Door) และใช้สำหรับการเข้าระบบในอนาคตได้

2.4 การโจมตีแบบ Man-in-the-Middle

การโจมตีแบบ Man-in-the-Middle นั้นผู้โจมตีต้องสามารถเข้าถึงแพกเก็ตที่ส่งระหว่างเครือข่ายได้ เช่น ผู้โจมตีอาจอยู่ที่ ISP ซึ่งสามารถตรวจจับแพกเก็ตที่รับส่งระหว่างเครือข่ายภายในและเครือข่ายอื่นๆ โดยผ่าน ISP การโจมตีนี้จะใช้แพกเก็ตสนิฟเฟอร์เป็นเครื่องมือเพื่อขโมยข้อมูลหรือใช้เซสชันเพื่อแอ็กเซสเครือข่ายภายใน หรือวิเคราะห์การจราจรของเครือข่ายหรือผู้ใช้

2.5 การโจมตีแบบ DOS

การโจมตีแบบดีไนล์ออฟเซอร์วิส หรือ DOS (Denial-of-Service) หมายถึง การโจมตีเซิร์ฟเวอร์โดยการทำให้เซิร์ฟเวอร์นั้นไม่สามารถให้บริการได้ ซึ่งโดยปกติจะทำโดยการใช้รีซอร์สของเซิร์ฟเวอร์จนหมด หรือถึงขีดจำกัดของเซิร์ฟเวอร์ ตัวอย่างเช่น เว็บเซิร์ฟเวอร์ และเอฟทีพีเซิร์ฟเวอร์ การโจมตีจะทำได้โดยการเปิดการเชื่อมต่อ (Connection) กับเซิร์ฟเวอร์จนถึงขีดจำกัดของเซิร์ฟเวอร์ ทำให้ผู้ใช้คนอื่นๆ ไม่สามารถเข้ามาใช้บริการได้ การโจมตีแบบนี้อาจใช้โปรโตคอล

ที่ใช้บนอินเตอร์เน็ตทั่วๆ ไปเช่น TCP (Transmission Control Protocol) หรือ ICMP (Internet Control Message Protocol) การโจมตีแบบดีไนล์ออฟเซอร์วิสเป็นการโจมตีจุดอ่อนของระบบหรือเซิร์ฟเวอร์มากกว่าการโจมตีจุดบกพร่อง (Bug) หรือช่องโหว่ของระบบการรักษาความปลอดภัย อย่างไรก็ตาม การโจมตีอาจทำให้ประสิทธิภาพของเครือข่ายลดลงโดยการส่งแพกเก็ตจำนวนมากเข้าไปในเครือข่าย ซึ่งแพกเก็ตอาจเป็นข้อมูลที่เป็นขยะ

2.6 โทรจันฮอร์ส เวิร์ม และไวรัส

คำว่า “โทรจันฮอร์ส (Trojan Horse)” นี้เป็นคำที่มาจากสงครามโทรจันระหว่างทรอย (Troy) และกรีก (Greek) ซึ่งเปรียบถึงม้าโครงไม้ที่ชาวกรีกสร้างทิ้งไว้แล้วซ่อนทหารไว้ข้างในแล้วถอนทัพกลับ พอชาวโทรจันออกมาดูเห็นม้าโครงไม้ทิ้งไว้ และคิดว่าเป็นของขวัญที่กรีซทิ้งไว้ให้จึงนำกลับเข้าเมืองไปด้วย พอตกดึกทหารกรีกที่ซ่อนอยู่ในม้าโครงไม้นี้ก็ออกมาและเปิดประตูให้กับทหารกรีกเข้าไปทำลายเมืองทรอย สำหรับในความหมายคอมพิวเตอร์แล้ว โทรจันฮอร์สหมายถึงโปรแกรมที่ทำลายคอมพิวเตอร์โดยแฝงมากับโปรแกรมอื่นๆ เช่น เกมส์ สกรีนเซฟเวอร์ เป็นต้น ซึ่งผู้ใช้อาจจะดาวน์โหลดโปรแกรมต่างๆ เหล่านี้มา แต่เมื่อติดตั้งแล้วรันโปรแกรมโทรจันฮอร์สที่แฝงมาด้วยก็จะทำลายระบบคอมพิวเตอร์ เช่น ลบไฟล์ต่างๆ หรืออาจสร้างแบ็คดอร์ให้กับโปรแกรมอื่นเข้ามาทำลายระบบก็ได้

โทรจันฮอร์ส เป็นโปรแกรมที่มุ่งประสงค์ร้ายต่อระบบแต่ได้ปลอมหรือหลอกผู้ใช้ให้เข้าใจว่าเป็นโปรแกรมที่ดีมีประโยชน์ไม่มีอันตรายใดๆ แต่เมื่อโปรแกรมถูกนำเข้าและทำงานก็จะปฏิบัติการอื่นๆ นอกจากหน้าที่เดิม เช่น การขโมยข้อมูลความลับต่างๆ เช่น บัญชีผู้ใช้และรหัสผ่าน เป็นต้น การยึดเป็นฐานที่มั่นเพื่อโจมตีคอมพิวเตอร์เครื่องอื่น หรือเมื่อผู้ใช้เปิดโปรแกรมม้าโทรจันขึ้นมาก็จะเป็นการเปิดช่องทางให้บุคคลอื่นที่มีโปรแกรมควบคุมเข้ามากระทำการมิชอบที่เครื่องของของตนเองได้ ศุภโชค (2548) ได้อธิบายถึงโทรจันฮอร์ส ไว้ว่า โปรแกรมม้าโทรจันถือเป็นโปรแกรมที่ไม่มีคำสั่งหรือการทำงานที่เป็นอันตรายต่อคอมพิวเตอร์โดยตรงและเป็นโปรแกรมที่ไม่สามารถแพร่กระจายได้ ซึ่งจะไม่เหมือนกับไวรัสที่สามารถแพร่กระจายไปยังโปรแกรมหรือไฟล์อื่นๆ บนเครื่องได้ สามารถแบ่งโปรแกรมม้าโทรจันออกเป็นหลายประเภทตามลักษณะการคุกคามและการทำงาน เช่น

1) Client/Server เป็นโปรแกรมม้าโทรจันที่ประกอบด้วยโปรแกรม 2 ส่วน โดยโปรแกรมส่วนที่ถูกส่งไปทำงานบนเครื่องปลายทางที่ต้องการจะเรียกว่าเซิร์ฟเวอร์ (Server) และเมื่อโปรแกรมส่วนนี้ถูกติดตั้งและทำงานแล้วก็จะเปิดช่องทางให้โปรแกรมอีกส่วนที่ทำงานอยู่บนเครื่องของผู้บุกรุก ซึ่งจะเรียกว่า ไคลเอ็นต์ (Client) สามารถควบคุมเครื่องดังกล่าวได้จากระยะไกล โปรแกมม้าโทรจัน ประเภทนี้ ได้แก่ Back Orifice, Subseven และ Netbus เป็นต้น

2) Keylogger โปรแกรมประเภทนี้จะทำหน้าที่บันทึกการกดคีย์ต่างๆ ในขณะที่กำลังใช้คอมพิวเตอร์ ซึ่งรวมถึงการกดคีย์เพื่อใส่ข้อมูลบัญชีผู้ใช้และรหัสผ่านด้วย แล้วบันทึกหรือส่งข้อมูลนี้ไปยังผู้บุกรุก เพื่อให้ผู้บุกรุกสามารถนำข้อมูลเหล่านี้มาใช้เป็นประโยชน์ในการบุกรุกในภายหลัง

โปรแกรมม้าโทรจันส่วนใหญ่จะแฝงตัวมาในรูปแบบของโปรแกรมเกมส์ การ์ดอวยพรรูปภาพสกรีนเซฟเวอร์ (Screen saver) และรูปแบบอื่นๆ ที่ทำให้มีความน่าสนใจที่จะสั่งให้ทำงาน ดังนั้นแนวทางการป้องกันก็คือ จะต้องไม่เปิดหรือสั่งงานโปรแกรมใดก็ตาม ที่ได้รับมาจากแหล่งที่ไม่มีความน่าเชื่อถือหรือไม่รู้จัก และควรจะต้องมีการตรวจสอบความถูกต้องสมบูรณ์ของโปรแกรมก่อนการใช้งาน

เวิร์ม (Worm) หมายถึง โปรแกรมที่เป็นอันตรายต่อระบบคอมพิวเตอร์ โดยจะแพร่กระจายตัวเองไปยังคอมพิวเตอร์เครื่องอื่นๆ ที่ต่ออยู่บนเครือข่าย เวิร์มจะใช้ประโยชน์จากแอพพลิเคชันที่รับส่งไฟล์โดยอัตโนมัติ ลักษณะการแพร่กระจายคล้ายตัวหนอนที่เจาะไชไปยังเครื่องคอมพิวเตอร์ต่างๆ แพร่พันธุ์ด้วยการคัดลอกตัวเองออกเป็นหลาย ๆ โปรแกรมและส่งต่อผ่านเครือข่ายออกไป เวิร์มถือว่าเป็นโปรแกรมคอมพิวเตอร์ชนิดหนึ่งที่แพร่ระบาดได้ด้วยตัวเองจากเครื่องคอมพิวเตอร์หนึ่งสู่เครื่องคอมพิวเตอร์หนึ่งโดยไม่ทำลายไฟล์หรือข้อมูล ต่างจากไวรัสที่อาจจะมีการทำลายไฟล์หรือข้อมูลต่างๆในระบบ และไวรัสจะต้องมีพาหะในแพร่ระบาดไปยังเครื่องอื่นๆ ในปัจจุบันเวิร์มจะมาในรูปแบบของอีเมล์ โดยเมื่อมีผู้ส่งจดหมายอิเล็กทรอนิกส์และแนบโปรแกรมติดมาด้วย ในส่วนของไฟล์ที่แนบมาด้วย (attached file) และสำหรับผู้ใช้บนระบบปฏิบัติการวินโดวส์สามารถคลิกไฟล์เหล่านั้นเพื่อเรียกใช้หรืออ่านไฟล์ได้ ซึ่งเท่ากับเป็นการเรียกโปรแกรมที่ส่งมาให้ทำงาน ถ้าสิ่งที่ส่งมานี้เป็นเวิร์ม ก็จะเป็นการเริ่มต้นการทำงาน โดยจะคัดลอกตัวเองและส่งจดหมายอิเล็กทรอนิกส์ไปให้ผู้อื่นอีก

ไวรัส (Virus) หมายถึง โปรแกรมที่ทำลายระบบคอมพิวเตอร์ โดยจะแพร่กระจายไปยังโปรแกรมอื่นๆ ที่อยู่ในเครื่องเดียวกัน ไวรัสสามารถทำลายเครื่องได้ตั้งแต่ลบไฟล์ทั้งหมดที่อยู่ในฮาร์ดดิสก์ไปจนถึงแค่เป็นโปรแกรมที่สร้างความรำคาญให้กับผู้ใช้ในเครือข่าย เช่น แค่เปิดวินโดวส์แล้วแสดงข้อความบางอย่าง ไวรัสจริงๆ ไม่สามารถที่จะแพร่กระจายไปยังเครื่องอื่นๆ ด้วยตัวเองได้ แต่การแพร่กระจายไปยังเครื่องอื่นต้องอาศัยโปรแกรมอื่นหรือมนุษย์ เช่น การแชร์ไฟล์โดยใช้แผ่นดิสก์เมื่อผู้ใช้รับไฟล์มาใช้ไวรัสก็จะแพร่กระจายไปยังเครื่องของผู้ใช้นั้นและจะเป็นวงจรในลักษณะนี้ต่อไปเรื่อยๆ จุดประสงค์และรายละเอียดการทำงานของไวรัสแต่ละตัวขึ้นอยู่กับผู้เขียนโปรแกรมไวรัสนั้นๆ เช่น อาจสร้างไวรัสสำหรับไปทำลายโปรแกรมหรือข้อมูลอื่นๆ ที่อยู่ในเครื่องคอมพิวเตอร์ หรือแสดงข้อความทางจอภาพ เป็นต้น

จากคำจำกัดความข้างต้น โทรจันฮอร์ส เวิร์ม และไวรัส จะมีความหมายคล้ายๆ กัน ซึ่งบางคนอาจใช้คำทั้งสามนี่แทนกันได้ แต่จริงๆ แล้วทั้งสามคำมีความหมายต่างกัน อย่างไรก็ตามโปรแกรมทำลายคอมพิวเตอร์ในปัจจุบันอาจเป็นได้ทั้ง 3 ชนิดก็ได้

**3. ระบบตรวจจับการบุกรุก (Intrusion Detection System)**

ระบบตรวจจับการบุกรุก หรือ IDS (Intrusion Detection System) (ศรัญญา, 2552) เป็นเครื่องมือสำหรับการรักษาความปลอดภัยอีกประเภทหนึ่งที่ใช้การตรวจจับความพยายามที่จะบุกรุกเครื่องข่าย โดยระบบจะแจ้งเตือนผู้ดูแลระบบเมื่อการบุกรุกหรือพยายามที่จะบุกรุกเครือข่าย IDS นั้นไม่ใช่ระบบที่ใช้ป้องกันการบุกรุกแต่เป็นระบบที่คอยแจ้งเตือนภัยเท่านั้น ถ้าเปรียบได้กับระบบรักษาความปลอดภัยของรถ IDS ก็อาจเปรียบได้กับระบบกันขโมย ซึ่งระบบนี้จะส่งสัญญาณเมื่อมีการตรวจพบความพยายามที่จะขโมยรถ เช่น การงัดประตูหรือกระจก แต่ระบบนี้จะไม่สามารถป้องกันไม่ให้รถถูกขโมยได้ อย่างไรก็ตามโดยธรรมชาติแล้วขโมยก็จะพยายามหลีกเลี่ยงรถที่ติดตั้งระบบนี้ ระบบเครือข่ายก็เช่นกันถ้ามีระบบตรวจจับและแจ้งเตือนการบุกรุก พวกแฮ็กเกอร์ก็จะหลีกเลี่ยงการบุกรุกเครือข่ายนี้

หน้าที่หลักของ IDS คือ แจ้งเตือนการเข้าใช้เครือข่ายที่ผิดปกติ สิ่งที่เป็นประเด็นสำคัญในการออกแบบระบบ IDS ก็คือ เหตุการณ์ใดคือสิ่งที่ถือว่าผิดปกติ ดังนั้นการใช้ IDS นั้นก็ขึ้นอยู่กับว่าอะไรที่จะแจ้งเตือนให้ทราบ คำตอบนั้นไม่ใช่แค่ถูกหรือผิด ขาวหรือดำ แต่จะขึ้นอยู่กับสภาวะของระบบในขณะนั้น

## 3.1 ประเภทของ IDS

IDS แบ่งออกเป็น 2 ประเภทคือ Host-Based IDS และ Network-Based IDS โดยโฮสต์เบสไอดีเอสนั้นคือ ระบบที่ติดตั้งที่โฮสต์และเฝ้าระวังและตรวจจับความพยายามที่จะบุกรุกโฮสต์นั้น ส่วนเน็ตเวิร์คเบสไอดีเอสนั้นคือ ระบบที่ตรวจดูแพ็กเก็ตที่วิ่งอยู่ในเครือข่าย และแจ้งเตือนถ้าพบหลักฐานที่คาดว่าจะเป็นการบุกรุกเครือข่าย

### 3.1.1 Host-Based IDS

โฮสต์เบสไอดีเอสเป็นซอฟต์แวร์ที่รันบนโฮสต์ โดยปกติแล้ว IDS ประเภทนี้จะวิเคราะห์ล็อก (Log) เพื่อค้นหาข้อมูลเกี่ยวกับการบุกรุก ในระบบยูนิกซ์นั้นล็อกที่ IDS และตรวจสอบ เช่น Syslog, Messages, Lastlog และ Wtmp เป็นต้น ส่วนในวินโวส์นั้น IDS ก็จะตรวจสอบอีเวนต์ล็อกต่างๆ เช่น System, Appliction และ Security เป็นต้น โดยปกติ IDS จะอ่านเหตุการณ์ใหม่ที่ขึ้นใน ล็อกและเปรียบเทียบกับกฎที่ตั้งไว้ก่อนหน้า ถ้าตรงก็จะแจ้งเตือนทันที ดังนั้นการที่ IDS จะตรวจจับการบุกรุกได้ระบบจะต้องบันทึกเหตุการณ์ต่างๆที่สำคัญที่เกิดขึ้นกับระบบในล็อกไฟล์ถ้าไม่เช่นนั้น IDS ก็ไม่มีข้อมูลที่จะใช้วิเคราะห์ว่ามีการบุกรุกหรือไม่

นอกจากการตรวจสอบล็อกไฟล์แล้ว IDS บางชนิดอาจสามารถตรวจสอบการเรียกใช้ฟังก์ชันของระบบปฏิบัติการ (System Call) ซึ่งถ้าเหตุการณ์คล้ายหรือตรงกับการบุกรุก IDS ก็จะแจ้งเตือนทันที่ นอกจากนี้ IDS ยังสามารถตรวจสอบแก้ไขไฟล์ในระบบได้ด้วย ซึ่งอาจทำได้โดยการตรวจสอบวันที่เพื่อแก้ไขครั้งสุดท้ายและขนาดของไฟล์ เป็นต้น วิธีที่แน่นอนกว่าก็จะคำนวณเช็คซัม (Checksum) ของไฟล์แล้วเก็บไว้เพื่อเปรียบเทียบเมื่อมีการตรวจสอบความคงสภาพของไฟล์ในระบบโดยเมื่อมีการคำนวณเช็คซัมใหม่แล้วค่าที่ได้ไม่ตรงกับค่าเดิมก็แสดงว่าไฟล์ได้ถูกแก้ไข

ข้อได้เปรียบของโฮสต์เบสไอดีเอส เช่น

1) โฮสต์เบสไอดีเอสสามารถตรวจพบทุกการบุกรุกกับโฮสต์นั้นๆ ได้เสมอถ้าระบบสามารถบันทึกเหตุการณ์ดังกล่าวในล็อกได้ หรือการบุกรุกมีการเรียกใช้ซิสเต็มคอลล์

2) โฮสต์เบสไอดีเอสสามารถบอกได้ว่าการบุกรุกนั้นสำเร็จหรือไม่ โดยการวิเคราะห์ข้อความในล็อกหรือหลักฐานอื่นๆ เช่น มีการแก้ไขไฟล์ที่สำคัญของระบบเป็นต้น

3) โฮสต์เบสไอดีเอสสามารถบ่งชี้ได้ว่ามีการเข้าใช้ระบบอย่างผิดปกติโดยผู้ใช้ของระบบเอง

ข้อเสียเปรียบของโฮสต์เบสไอดีเอสคือ

1) โพรเซสของ IDS อาจถูกโจมตีเองจนอาจไม่สามารถแจ้งเตือนได้

2) โฮสต์เบสไอดีเอสจะแจ้งเตือนก็ต่อเมื่อเหตุการณ์ที่เกิดขึ้นนั้นตรงกับที่กำหนดไว้ก่อนหน้าถ้าแฮ็กเกอร์มีเทคนิคใหม่ๆ IDS อาจไม่แจ้งเตือนการบุกรุกก็ได้

3) การทำงานของโฮสต์เบสไอดีเอสมีผลกระทบต่อประสิทธิภาพของโฮสต์เองเนื่องจากต้องตรวจสอบล็อกไฟล์และซิลเต็มคอลล์

3.1.2 Network-Based IDS

เน็ตเวิร์คเบสไอดีเอสคือ ซอฟต์แวร์พิเศษที่รันบนคอมพิวเตอร์เครื่องหนึ่งต่างหาก IDS ประเภทนี้จะมีเน็ตเวิร์คการ์ดที่ทำงานในโหมดที่เรียกว่า “โพรมิสเซียส (promiscuous Mode)” ซึ่งในโหมดนี้เน็ตเวิร์คการ์ดจะส่งต่อทุกๆ แพ็กเก็ตที่วิ่งอยู่บนเครือข่ายไปให้แอพพลิเคชันโพรเซส ในขณะที่เน็ตเวิร์คการ์ดที่รันในโหมดธรรมดานั้นจะรับเอาเฉพาะแพ็กเก็ตที่มีที่อยู่ปลายทางตรงกับของเครื่องนั้นเท่านั้น เมื่อทุกๆ แพ็กเก็ตส่งผ่านไปให้แอพพลิเคชัน IDS วิเคราะห์ข้อมูลในแพ็กเก็ตเหล่านั้นกับกฎก็จะแจ้งเตือนทันที ในแต่ละ IDS จะมีข้อมูลที่ใช้สำหรับเปรียบเทียบเพื่อบอกว่ามีการพยายามที่จะบุกรุกเครือข่ายหรือไม่ ซึ่งข้อมูลนี้จะเรียกว่า “ซิกเนเจอร์ (Signature)” IDS จะใช้ซิกเนเจอร์ที่เก็บไว้เปรียบเทียบกับข้อมูลที่ได้จากการวิเคราะห์แพ็กเก็ตที่วิ่งบนเครือข่าย ดังนั้นถ้าแฮ็กเกอร์มีเทคนิคใหม่ๆ และ IDS ไม่รู้จักเทคนิคนี้หรือมีซิกเนเจอร์นี้ IDS ก็ไม่สามารถตรวจจับการบุกรุกประเภทนี้ได้

โดยส่วนใหญ่แล้ว IDS ประเภทนี้จะมี 2 เน็ตเวิร์คการ์ด โดยเน็ตเวิร์คการ์ดแรกจะเชื่อมต่อเข้ากับเครือข่ายที่ต้องการเฝ้าระวังหรือตรวจจับการบุกรุก โดยเน็ตเวิร์คการ์ดนี้จะไม่มีหมายเลขไอพี ดังนั้นเครื่องอื่นๆที่อยู่ในเครือข่ายจะมองไม่เห็นเครื่องนี้ ส่วนเน็ตเวิร์คการ์ดอีกอันหนึ่งจะเชื่อมต่อเข้ากับอีกเครือข่ายหนึ่งเพื่อใช้สำหรับส่งการแจ้งเตือนภัยไปยังเซิร์ฟเวอร์ โดยเครือข่ายนี้จะต้องไม่ถูกเชื่อมต่อกับเครือข่ายหลักที่ IDS จะตรวจจับการบุกรุก

ข้อได้เปรียบของเน็ตเวิร์คเบสไอดีเอส เช่น

1) เน็ตเวิร์คเบสไอดีเอสจะแอบซ่อนในเครือข่ายทำให้ผู้บุกรุกไม่รู้ว่ากำลังถูกเฝ้ามองอยู่

2) เน็ตเวิร์คเบสไอดีเอสหนึ่งเครื่องสามารถใช้เฝ้าระวังการบุกรุกได้กับหลายระบบหรือโฮสต์

3) เน็ตเวิร์คเบสไอดีเอสสามารถตรวจจับทุกๆ แพ็กเก็ตที่วิ่งไปยังระบบที่ถูกเฝ้าระวังภัยอยู่

ข้อเสียเปรียบของเน็ตเวิร์คเบสไอดีเอส

1) เน็ตเวิร์คเบสไอดีเอสจะแจ้งเตือนภัยก็ต่อเมื่อตรวจพบแพ็กเก็ตที่ตรงกับซิกเนเจอร์ที่กำหนดไว้ก่อนหน้าเท่านั้น

2) เน็ตเวิร์คเบสไอดีเอสอาจพลาดที่จะตรวจจับแพ็กเก็ตได้ในกรณีที่มีการใช้เครือข่ายมาก จนทำให้ IDS วิเคราะห์แพ็กเก็ตที่วิ่งอยู่บนเครือข่ายไม่ทัน หรือมีเส้นทางข้อมูลอื่นที่ไม่ต้องผ่าน IDS

3) เน็ตเวิร์คเบสไอดีเอสไม่สามารถสรุปได้ว่าการบุกรุกนั้นสำเร็จหรือไม่

4) เน็ตเวิร์คเบสไอดีเอสไม่สามารถตรวจวิเคราะห์แพ็กเก็ตที่เข้ารหัสไว้ได้

5) เครือข่ายที่ใช้สวิตซ์นั้นต้องเซตเพื่อให้พอร์ตที่เชื่อมต่อกับ IDS นั้นสามารถมองเห็นทุกแพ็กเก็ตที่วิ่งผ่านสวิตซ์นั้น

ทั้งโฮสต์เบสไอดีเอสและเน็ตเวิร์คเบสไอดีเอสมีทั้งข้อดีและข้อเสียที่แตกต่างกัน ในขณะที่เน็ตเวิร์คเบสไอดีเอสนั้นสามารถใช้เฝ้าระวังได้ครอบคลุมส่วนของเครือข่ายได้มากกว่า ดังนั้นจึงเป็นทางเลือกที่ประหยัดกว่าแต่โฮสต์เบสไอดีเอสจะเหมาะสมสำหรับการเฝ้าระวังภัยที่อาจจะสร้างโดยผู้ใช้เครือข่ายเองดังนั้นการเลิกประเภทของ IDS ให้เหมะสมนั้นก็ขึ้นอยู่กับภัยคุกคามเครือข่ายขององค์กร

3.2 การแจ้งเตือนภัยของ IDS

IDS จะรายงานเฉพาะสิ่งที่กำหนดให้รายงานเท่านั้น มีอยู่สองสิ่งที่ผู้ดูแลระบบจะต้องคอนฟิกให้กับ IDS สิ่งแรกคือ ซิกเนเจอร์ของการบุกรุก สิ่งที่สองคือเหตุการณ์ที่ผู้ดูแลระบบให้ความสำคัญหรือเหตุการณ์ที่คาดว่าจะเป็นผลไปสู่การบุกรุกในภายหน้า ซึ่งเหตุการณ์ต่างๆ เหล่านี้

อาจเป็นทราฟิกที่ไม่ปกติหรืออาจเป็นบางข้อความในล็อก การคอนฟิกซิกเนเจอร์ให้กับ IDS ของแต่ละองค์กรนั้นอาจจะไม่เหมือนกัน ซึ่งจะขึ้นอยู่กับว่าองค์นั้นจะให้ความสนใจกับการบุกรุกประเภทใด

เมื่อ IDS ได้ถูกคอนฟิกอย่างถูกต้องแล้ว เหตุการณ์ที่ IDS จะรายงานให้ทราบนั้นสามารถแบ่งออกได้เป็น 3 ประเภทคือ

1) การสำรวจเครือข่าย
2) การโจมตี
3) เหตุการณ์ที่น่าสงสัยหรือผิดปกติ

**4. การสำรวจเครือข่าย**

เหตุการณ์ที่เป็นการสำรวจเครือข่ายเป็นการพยายามของผู้บุกรุกที่จะรวบรวมข้อมูลเกี่ยวกับระบบเครือข่ายก่อนที่จะโจมตีจริงๆเช่น

IP Scans: ไอพีสแกนเป็นความพยายามของผู้บุกรุกที่ทราบเกี่ยวกับโฮสต์ต่างๆ ที่อยู่ในเครือข่ายซึ่งระบบที่สแกนนั้นอาจใช้การปิง (Ping) ช่วงของหมายเลขไอพีของเครือข่ายนั้น
Port Scans : หลังจากที่ผู้บุกรุกพอได้ข้อมูลเกี่ยวกับว่าเครือข่ายมีโฮสต์ใดอยู่บ้าง ข้อมูลต่อมาที่ผู้บุกรุกต้องการคือ บริการใดบ้างที่แต่ละโฮสต์บริการอยู่ ซึ่งหมายเลขพอร์ตนั้นจะเป็นสิ่งที่บ่งบิกว่ามีแอพพลิเคชันใดบ้างที่ให้บริการอยู่

Trojan Scans: การสแกนโทรจันนั้นเป็นความพยายามของผู้บุกรุกที่จะตรวจเช็คว่ามีพอร์ตของโทรจันใดบ้างที่เปิดอยู่

Vulnerability Scans: การสแกนหาจุดอ่อนของระบบ เป็นความพยายามที่จะใช้การโจมตีหลายๆ แบบกับระบบหนึ่งเพื่อตรวจสอบเช็คดูระบบนี้ว่ามีจุดอ่อนอย่างไร

File Snooping: ไฟล์สนูปิงเป็นการตรวจเช็คว่ายูสเซอร์มีสิทธิ์อย่างไรกับไฟล์นั้น ซึ่งระบบไฟล์ส่วนใหญ่จะมีการกำหนดสิทธิ์ของผู้ใช้ในการเข้าใช้แต่ละไฟล์
การโจมตี

การโจมตีเครือข่ายหรือระบบควรให้ลำดับความสำคัญสูงสุด เมื่อ IDS รายงานเหตุการณ์นี้ ผู้ดูแลระบบตอบสนองกับเหตุการณ์นี้ทันทีเพื่อป้องกันการสูญเสียมากกว่านี้ บางครั้ง IDS อาจแยกแยะระหว่างการโจมตีจริงๆ กับการสเกนหาจุดอ่อน เนื่องจากสาเหตุทั้งสองนั้น IDS จะตรวจพบซิกเนเจอร์ของการโจมตีเหมือนกัน ผู้ดูแลระบบอาจต้องวิเคราะห์ข้อมูลเพิ่มเติมการสเกนหาจุดอ่อนนั้น IDS จะรายงานการโจมตีหลายๆ รูปแบบในช่วงเวลาสั้นๆ กับระบบใดระบบหนึ่ง ส่วนการโจมตีจริงนั้นอาจมีการรายงานการโจมตีรูปแบบเดียวกับระบบใดระบบหนึ่ง

**5. เหตุการณ์ที่น่าสงสัยหรือผิดปกติ**

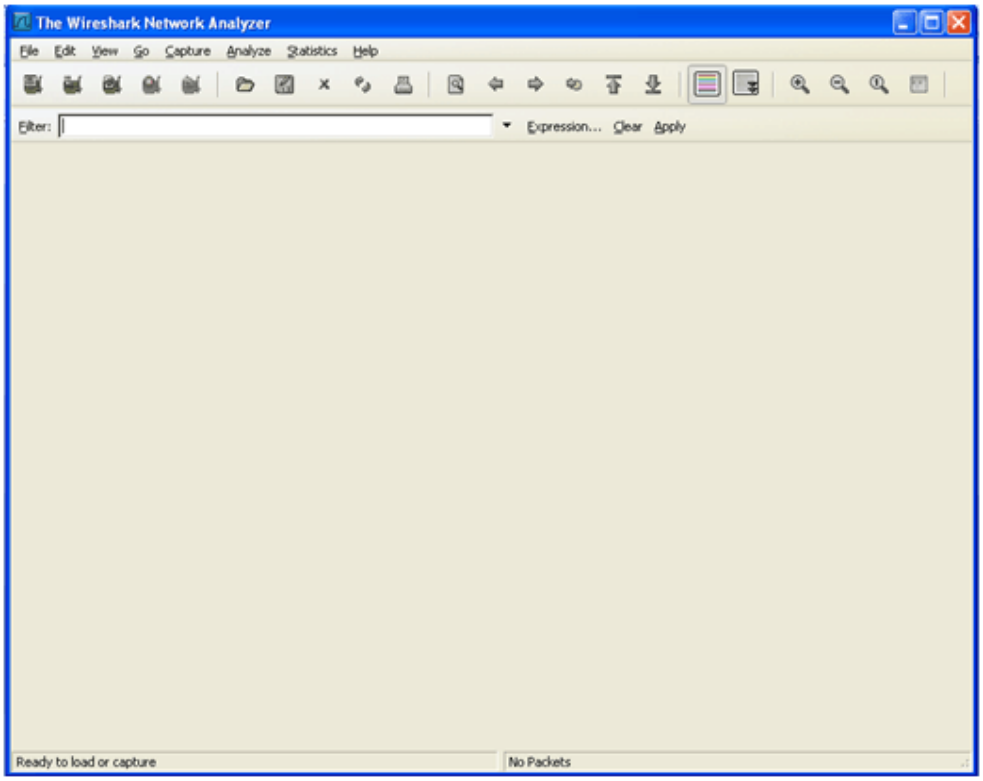
เหตุการณ์อื่นๆ ที่ผิดปกติและไม่ได้จัดอยู่ในประเภทต่างๆ ที่กล่าวมาข้างต้นถือว่าเป็นเหตุการณ์ที่น่าสงสัยว่าอาจมีการโจมตีเครือข่ายเกิดขึ้น ซึ่งผู้ดูแลระบบต้องวิเคราะห์และสืบหาสาเหตุของเหตุการณ์ที่ว่านี้ต่อ ตัวอย่างเช่น บางโฮสต์อาจส่งแพ็กเก็ตที่มีข้อมูลส่วนหัวผิดไปจากที่กำหนดในมาตรฐาน ซึ่งเหตุการณ์นี้อาจเกิดเนื่องจากการโจมตีแบบใหม่ หรือเน็ตการ์ดอาจเสีย หรือข้อมูลอาจเกิดผิดพลาดในระหว่างการส่งผ่านสายสัญญาณ IDS จะไม่มีข้อมูลพอเพียงที่จะบอกได้ว่าเหตุการณ์นี้เกิดขึ้นเพราะอะไร แต่จะแจ้งเตือนให้ผู้ดูแลระบบทราบเพื่อสืบค้นหาสาเหตุที่แท้จริง

**6. โปรแกรม Ethereal**

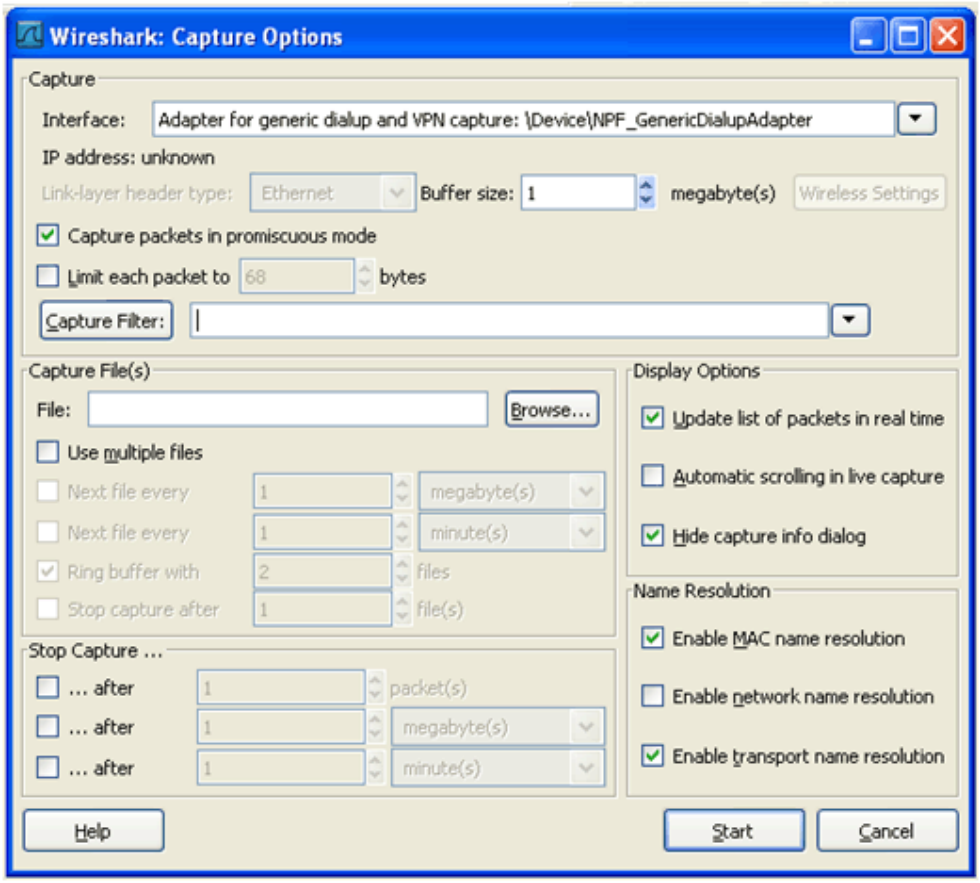
โปรแกรม Ethereal เป็นโปรแกรมวิเคราะห์การจราจรของเครือข่ายโดยจับแพ็กเก็ตจากเครือข่ายและแสดงข้อมูลภายใน โปรแกรม Ethereal จะมี filter 2 แบบให้ใช้ ได้แก่

1) Display filters เป็น filter ที่ใช้สำหรับกรองเอาแต่ Packet ที่ต้องการมาแสดงให้ดู จากจำนวน Packet มากมายที่โปรแกรมจับมาไว้แล้ว เพื่อให้โปรแกรมแสดงผลแต่ละ packet ที่มีคุณสมบัติตามเงื่อนไขที่เราต้องการ แต่เงื่อนไขใน display filter นั้น จะมี syntax ที่แตกต่างจากที่ใช้ใน capture filter

2) Capture filters เป็น filter ที่ใช้สำหรับระบุชนิด หรือประเภทของ Packets ที่จะให้โปรแกรมจับมา เพื่อใช้ลดจำนวน Packets ที่ไม่ต้องการออก เพื่อลดขนาด file หรือหน่วยความจำที่ต้องใช้ในการเก็บ packet ลง ซึ่งเงื่อนไขของ capture filter นั้น จะใช้เงื่อนไขที่มี syntax เดียวกับที่ใช้ในโปรแกรม tcpdump



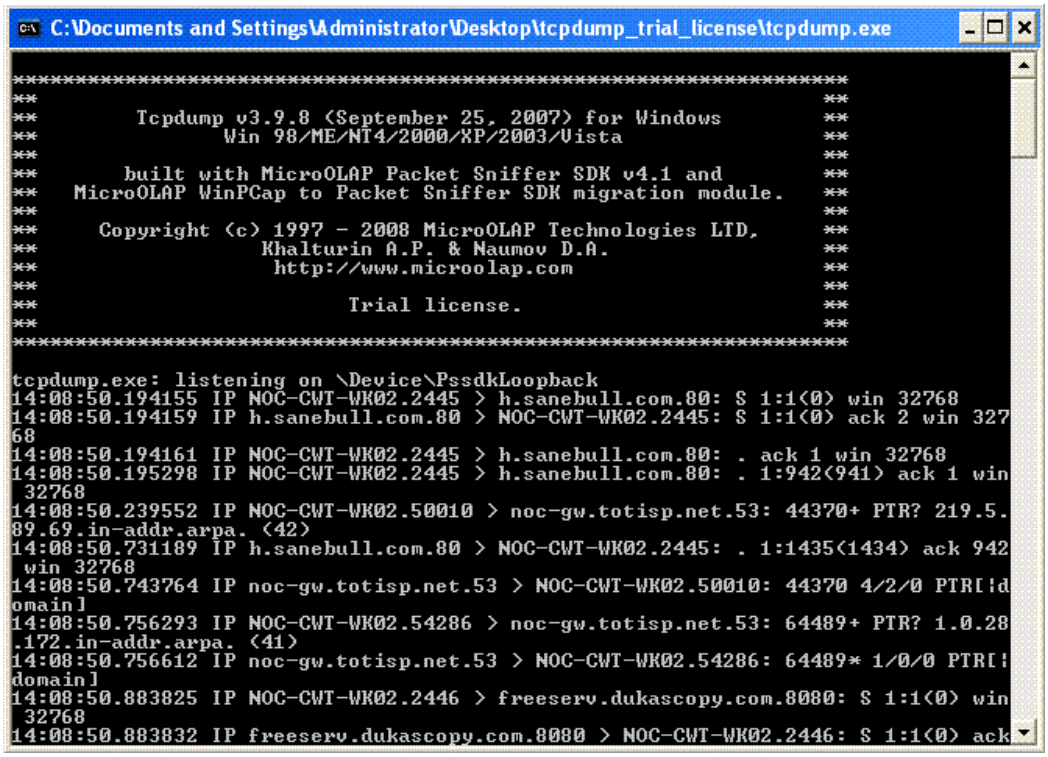

**ภาพที่ 1** แสดงหน้าจอที่ให้แสดง display filter ในส่วน filter




**ภาพที่ 2** แสดงหน้าจอที่ให้แสดง capture filter ในส่วน filter

### 7. โปรแกรม Tcpdump

Tcpdump เป็นโปรแกรม Protocol Analyzers ได้พัฒนามาตั้งแต่ ค.ศ. 1990 ที่ LawrenceBerkeley National Laboratory โดยถูกใช้สำหรับงานทางด้าน network monitor ซึ่งเป็นเครื่องมือที่วิเคราะห์คุณภาพของเครือข่าย (วิเคราะห์ข้อมูลที่อยู่บนเครือข่าย) การทำงานของ tcpdump ตัว tcpdump จะไปเปลี่ยนการทำงานของ Interface (Lan Card) ให้ทำงานในลักษณะ promiscuous mode ซึ่งจะคอยรับข้อมูลจากเครือข่ายและสามารถนำข้อมูลเหล่านั้นมาแสดงผลได้หลายวิธี เช่น แสดงผลบนหน้าจอ (Console) หรือ จะแสดงผลในลักษณะของไฟล์ เพื่อนำกลับมาวิเคราะห์ในภายหลังก็ได้ (คุณาวุฒิ, 2552)




**ภาพที่ 3** แสดงตัวอย่างผลลัพธ์จากการทำ tcpdump

### 8. ความผิดปกติของทราฟฟิก (Traffic Anomaly)

ความผิดปกติของทราฟฟิก คือ ลักษณะข้อมูลการใช้งานเครือข่ายที่มีความผิดปกติไปจาก rule ที่มีอยู่ในฐานข้อมูล อาจเกิดจากแพคเก็ตที่เข้ามานั้นไม่ตรงกับรูปแบบที่กำหนดไว้ในฐานข้อมูลของโปรแกรมสนอท จึงเข้าข่ายว่าเป็นการบุกรุกหรือมีความพยายามในการบุกรุก หรือแพกเก็ตอื่นๆ ที่ผิดปกติและไม่ได้จัดอยู่ในประเภทต่างๆ ที่กล่าวมาข้างต้นถือว่าเป็นแพกเก็ตที่น่าสงสัยว่าอาจมีการโจมตีเครือข่ายเกิดขึ้น ซึ่งไม่มีโอกาสเกิดขึ้นในสภาวะการทำงานปกติได้ แพกเก็ตประเภทนี้เป็นการจงใจเปลี่ยนข้อมูลสำคัญที่ใช้ควบคุมการสื่อสารข้อมูลให้ผิดปกติผิดธรรมชาติของการสื่อสารข้อมูลธรรมดา ตามปกติแต่ละโปรโตคอล เช่น IP TCP UDP ย่อมมีค่าในเฮดเดอร์ที่จะใช้เป็นกลไกการควบคุมการสื่อสารข้อมูล โดยในสภาวะปกติแล้วค่าเฮดเดอร์เหล่านี้จะมีค่าขอบเขตและคาดหมายได้ แพกเก็ตเหล่านี้จะมีการดัดแปลงแพ็กเก็ตด้วยเทคนิคของการควบคุมระดับ Network Layer โดยตรงแบบไม่ผ่านโปรโตคอล ซึ่งเป็นช่องทางให้ผู้โจมตีนำมาใช้โจมตีเป้าหมายให้ทำงานผิดพลาด เพราะต้องจัดการกับแพ็กเก็ตที่ไม่ได้ระบุอยู่ในโปรโตคอลหรืออาจจะ

ใช้เพื่อปกปิดตนเองในการสำรวจเครื่องเป้าหมายที่ต้องการโจมตี เป็นต้นแพ็กเก็ตเหล่านี้เปรียบเสมือนการกำหนดให้ระบบปฏิบัติการให้ทำงานนอกเหนือจากเงื่อนไขที่ได้กำหนดไว้ตามปกติ ดังนั้นผลที่ได้จึงแตกต่างกันโดยขึ้นอยู่กับระบบปฏิบัติการแต่ละชนิด หากระบบปฏิบัติการสามารถจัดการกับแพ็กเก็ตประเภทนี้ได้ก็อาจจะไม่มีผลกระทบใดๆมากนัก แต่หากระบบปฏิบัติการไม่สามารถจัดการกับกรณีของแพ็กเก็ตประหลาดเช่นนี้ได้อย่างเหมาะสม ก็จะส่งผลกระทบอย่างรุนแรงได้ ดังนั้นผู้บุกรุกก็จะใช้คุณสมบัติเหล่านี้ของแพ็กเก็ตเพื่อทำการโจมตีเครื่องเป้าหมายให้หยุดการบริการได้

**9. การทำเหมืองข้อมูล (Data Mining)**

การขุดค้นข้อมูล หรือ การทำเหมืองข้อมูล (data mining) เป็นเทคโนโลยีใหม่ของการประยุกต์ใช้ข้อมูลที่เก็บอยู่ในฐานข้อมูลให้เกิดประโยชน์สูงสุดแก่หน่วยงานที่เป็นเจ้าของข้อมูล การประยุกต์ใช้ข้อมูลที่กล่าวถึงนี้มีได้หลายแนวทาง แต่โดยทั่วไปมักจะเป็นการสรุปภาพรวมของข้อมูลในฐานข้อมูล การวิเคราะห์แนวโน้มการเปลี่ยนแปลงของข้อมูล หรือ การค้นหาความสัมพันธ์ที่ซ่อนอยู่ภายในกลุ่มของข้อมูล

การวิเคราะห์ข้อมูลด้วยโปรแกรมช่วยงาน เช่น SPSS, SAS นักวิเคราะห์ข้อมูลจะต้องเป็นผู้กำหนดว่าจะศึกษาลักษณะใดจากข้อมูล และจะใช้ข้อมูลส่วนใดบ้าง แต่ data mining จะกระทำขั้นตอนต่างๆเหล่านี้ให้โดยอัตโนมัติ โปรแกรม data mining มีความสามารถที่จะค้นหาแนวโน้ม รูปแบบร่วม หรือลักษณะอื่นๆที่น่าสนใจ โดยไม่ต้องพึ่งพาการสั่งงานทุกขั้นตอนจากนักวิเคราะห์ข้อมูล และอาจจะสามารถค้นพบลักษณะที่น่าสนใจจากข้อมูลซึ่งนักวิเคราะห์ข้อมูลไม่ได้คาดหมายมาก่อน

นอกจากนี้ data mining ยังมีความแตกต่างจากระบบผู้เชี่ยวชาญ (expert system) ตรงที่ฐานความรู้ของ data mining ได้จากการสังเคราะห์ขึ้นจากข้อมูลโดยตรง สามารถปรับปรุงฐานความรู้ของตัวเองได้อัตโนมัติตามข้อมูลใหม่ที่ได้รับเพิ่มขึ้น ซึ่งต่างจากระบบผู้เชี่ยวชาญที่ฐานความรู้ถูกป้อนเข้ามาในระบบ และจะคงตัวอยู่เช่นนั้นตลอดการใช้งาน ถึงแม้การประยุกต์ใช้ data mining จะมีได้หลากหลาย แต่สามารถจัดกลุ่มกว้างๆได้เป็นสองกลุ่ม คือ กลุ่มที่ใช้ data mining เพื่อการทำนาย และกลุ่มที่ใช้เพื่อการอธิบาย

การทำ data mining *เพื่อการทำนาย* เป็นการนำความรู้ที่เรียนรู้มาจากข้อมูลที่มีอยู่เพื่อประโยชน์ในการทำนายข้อมูลใหม่ที่จะเกิดขึ้นในอนาคต เช่น จากข้อมูลลูกค้าของแผนกสินเชื่อของธนาคารที่ได้มีการจัดลำดับชั้นของลูกค้าไว้แล้วว่าใครเป็นลูกค้าชั้นดี ใครเป็นลูกค้าในระดับปานกลาง และใครเป็นลูกค้าที่มักจะผิดนัดชำระหนี้ โปรแกรม data mining สามารถเรียนรู้จากข้อมูลเหล่านี้และค้นหาโมเดลที่สามารถใช้อธิบายลักษณะของลูกค้าชั้นดี ลูกค้าระดับปานกลาง และลูกค้าที่ไม่เป็นที่ต้องการ จากโมเดลที่ได้นี้สามารถนำไปใช้ทำนายลูกค้าใหม่ที่มาขอสินเชื่อได้ว่าเขาน่าจะเป็นลูกค้าประเภทใด

การทำ data mining *เพื่อการอธิบาย* เป็นการค้นหารูปแบบที่น่าสนใจจากกลุ่มข้อมูล รูปแบบนี้มักจะเป็นความสัมพันธ์ หรือลักษณะที่เชื่อมโยงกันของข้อมูล การทำ data mining แบบนี้ต่างจากแบบแรกตรงที่ผู้ใช้ไม่ได้กำหนดล่วงหน้าว่าจะให้โปรแกรม data mining ค้นหารูปแบบหรือโมเดลของอะไร แต่ให้ค้นหาทุกรูปแบบที่น่าสนใจจากข้อมูล จะเห็นได้ว่างาน data mining มีได้หลากหลายรูปแบบ ทั้งนี้เนื่องจากความรู้ที่ต้องการได้จากข้อมูลมีได้หลายลักษณะ ตัวอย่างต่อไปนี้แสดงผลสำเร็จของการนำ data mining ไปใช้

1) ด้านการแพทย์: ใช้ data mining ค้นหาผลข้างเคียงของการใช้ยา โดยอาศัยข้อมูลจากแฟ้มประวัติผู้ป่วย, ใช้ในการวิเคราะห์หาความสัมพันธ์ของสารพันธุกรรม

2) ด้านการเงิน: ใช้ data mining ตัดสินว่าควรจะอนุมัติเครดิตให้ลูกค้ารายใดบ้าง, ใช้ในการคาดหมายความน่าจะเป็นว่าธุรกิจนั้นๆมีโอกาสที่จะล้มละลายหรือไม่, ใช้คาดหมายการขึ้น/ลงของหุ้นในตลาดหุ้น

3) ด้านการเกษตร: ใช้จำแนกประเภทของโรคพืชที่เกิดกับถั่วเหลืองและมะเขือเทศ

4) ด้านวิศวกรรม: ใช้วิเคราะห์และวินิจฉัยสาเหตุการทำงานผิดพลาดของเครื่องจักรกล

5) ด้านอาชญวิทยา: ใช้วิเคราะห์หาเจ้าของลายนิ้วมือ

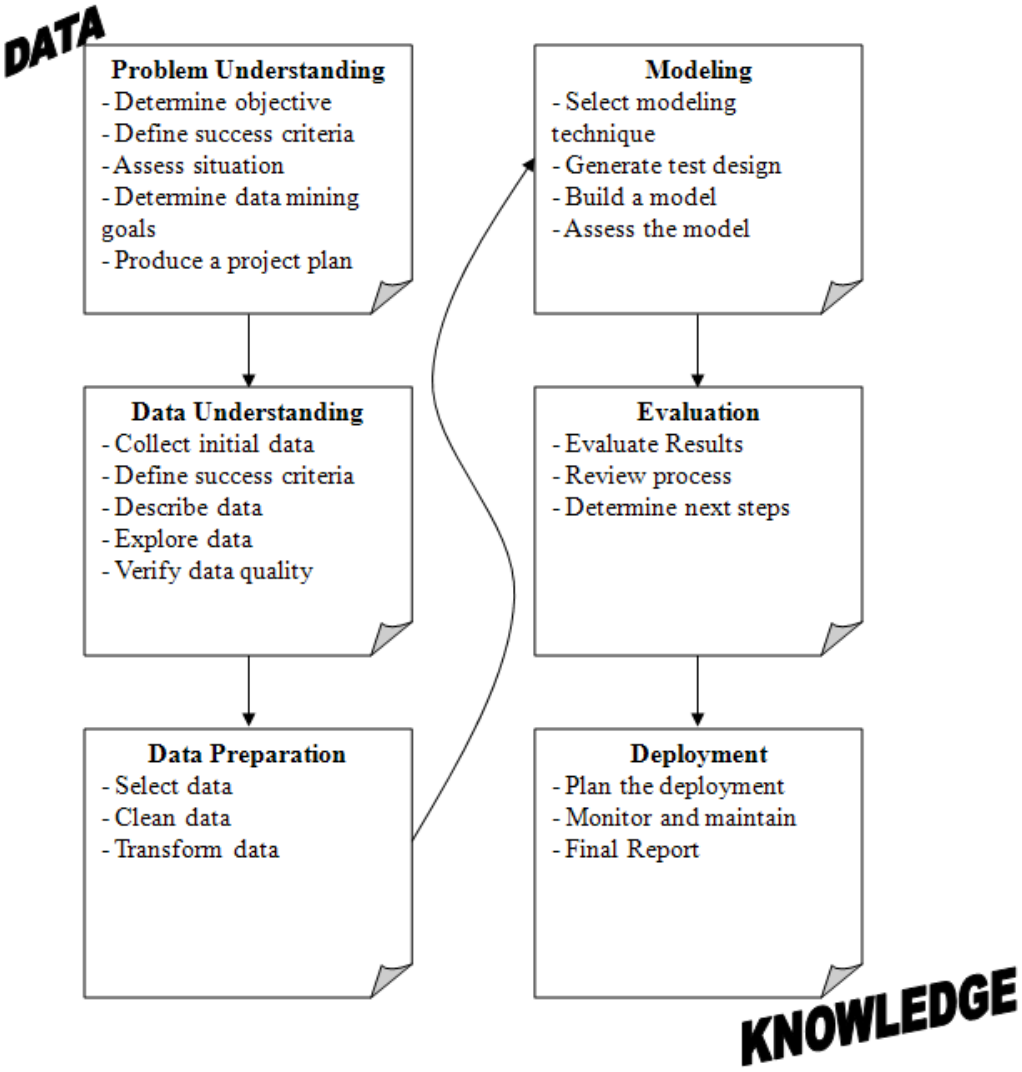
6) ด้านอวกาศ: ใช้วิเคราะห์ข้อมูลที่ส่งมาจากดาวเทียมขององค์การนาซ่า

หัวใจสำคัญของกระบวนการ data mining คือส่วนของโปรแกรมที่ทำหน้าที่สังเคราะห์ความรู้ขึ้นมาจากข้อมูลจำนวนมากในฐานข้อมูล ส่วนสังเคราะห์ความรู้นี้เรียกว่า learning algorithm ซึ่งมีผู้เสนอและพัฒนาอัลกอริทึมส่วนนี้ขึ้นเป็นจำนวนมาก ได้แก่ อัลกอริทึมที่ใช้หลักการของการสร้างต้นไม้ตัดสินใจ (decision-tree induction algorithm) ตัวอย่างเช่น โปรแกรม ID3, C4.5 อัลกอริทึมที่ใช้หลักการทางสถิติและทฤษฎีของเบย์ ตัวอย่างเช่น โปรแกรม naïve

Bayes อัลกอริทึมที่ใช้หลักการของ neural network และอัลกอริทึมอื่นๆอีกมาก ได้มีนักคอมพิวเตอร์ทดสอบเปรียบเทียบความสามารถของอัลกอริทึมแต่ละประเภทเพื่อค้นหาว่าอัลกอริทึมใดมีความสามารถสูงที่สุด ผลการทดสอบส่วนใหญ่ที่ปรากฏจะชี้ว่าไม่มีอัลกอริทึมใดที่ทำงานได้ดีที่สุดในข้อมูลทุกประเภท ทั้งนี้เนื่องจากข้อมูลแต่ละประเภทมีลักษณะเฉพาะตัวที่ต่างกัน เช่น ข้อมูลทางการแพทย์ จะต่างจากข้อมูลด้านกฎหมาย และต่างจากข้อมูลด้านอวกาศ ดังนั้นจึงไม่มีอัลกอริทึมใดที่ดีที่สุดสำหรับข้อมูลทุกประเภท

โครงการวิจัยนี้จึงเสนอขึ้นเพื่อศึกษาและทดสอบวิธีในการแยกประเภทความผิดปกติของข้อมูลทราฟฟิกในระบบเครือข่ายที่เหมาะสมได้ เพื่อเป็นการเพิ่มความสามารถในการตรวจจับทราฟฟิกและลดอัตราการเตือนความผิดปกติ (False alarm rate) ลงได้ และเพื่อนำไปวิเคราะห์ว่าอัลกอริทึมในกลุ่มใดหรือประเภทใดที่มีประสิทธิภาพในการสังเคราะห์หรือการเรียนรู้ที่ดีที่สุด เมื่อคัดเลือกอัลกอริทึมหรือกลุ่มของอัลกอริทึมได้แล้ว ขั้นตอนต่อไปจะเป็นการวิเคราะห์ผลการแยกประเภทความผิดปกติของทราฟฟิก เพื่อให้สามารถเป็นแนวทางในการพัฒนาเครื่องมือที่จะมาช่วยในการตรวจจับความผิดปกติได้ดียิ่งขึ้นในอนาคตคาดคะเนสาเหตุของความผิดปกติของทราฟฟิกได้ใกล้เคียงที่สุด โครงการศึกษาวิจัยนี้ยังจะสามารถใช้เป็นแนวทางสำคัญในการพัฒนาระบบตรวจจับการบุกรุกทางอินเตอร์เน็ตให้มีประสิทธิภาพ และตรงต่อความต้องการของผู้ใช้มากยิ่งขึ้นได้ในอนาคต

9.1 ขั้นตอนการทำเหมืองข้อมูล (Data Mining)




**ภาพที่ 4** แสดงขั้นตอนการทำ Data Mining

จากภาพที่ 4 แสดงขั้นตอนการทำ Data Mining ซึ่งสามารถอธิบายรายละเอียดตามลำดับขั้นตอนต่างๆ ได้ดังนี้

1) ศึกษาปัญหา

การตั้งเป้าหมายการทำเหมืองข้อมูลว่าต้องการแก้ปัญหาใด และจะวัดความสำเร็จได้อย่างไร เช่น เป้าหมายการทำเหมืองข้อมูลเพื่อเพิ่มจำนวนลูกค้า โดยวัดความสำเร็จจากการที่สามารถเพิ่มจำนวนลูกค้าได้ 20% ตามเป้าหมายที่กำหนดไว้ เป็นต้น

2) การรวบรวมข้อมูล

ขั้นตอนนี้จะเป็นการเก็บรวบรวมข้อมูลจากแหล่งข้อมูล กำหนดคุณสมบัติข้อมูลที่รวบรวมมาได้ ตรวจสอบข้อมูลขั้นต้นทั้งความสมบูรณ์และถูกต้อง

3) การเตรียมข้อมูล

หน้าที่ของขั้นตอนนี้คือ การจัดการข้อมูลให้อยู่ในรูปแบบที่สามารถนำเข้าไปวิเคราะห์โดยอัลกอริทึมการทำเหมืองข้อมูลได้โดยขั้นตอนต่างๆ เช่น เลือกข้อมูลที่จะนำมาใช้ หรือแบ่งข้อมูลที่สนใจจากฐานข้อมูลจากนั้นจะเป็นการปรับเปลี่ยนรูปแบบข้อมูลและทำข้อมูลให้สะอาดโดยการแยกข้อมูลที่ไม่เกี่ยวข้อง ซ้ำซ้อน ข้อมูลที่บันทึกไม่ถูกต้อง หรือไม่สอดคล้องกันออกไป

4) การสร้างแบบจำลอง

ขั้นตอนนี้จะเป็นการเลือกอัลกอริทึมที่เหมาะสมในการทำเหมืองข้อมูล โดยในการทำเหมืองข้อมูลนั้นมีเทคนิคและวิธีการหลากหลายรูปแบบ เช่น Database Segmentation, Descriptive Modeling หรือ Link Analysis ซึ่งแต่ละวิธีการก็จะมีอัลกอริทึมต่างๆ ให้เลือกใช้แตกต่างกันไป เช่น การแบ่งกลุ่มข้อมูลอาจใช้ K-Mean Algorithm หรือ Kohonen’s Neural Net แต่ถ้าเป็นการทำ Predictive Modeling อาจใช้ CART (Classification And Regression Tree) หรือ Back Propagation Neural Net หรือถ้าเป็นการทำ Link Analysis ก็จะใช้ Apriori Algorithm เป็นต้น โดยเมื่อเลือกอัลกอริทึมที่ต้องการได้แล้วจะต้องมีการกำหนดรูปแบบการทดสอบและผลลัพธ์ พร้อมทั้งสร้างแบบจำลองตามอัลกอริทึมที่เลือก

5) การประเมินผล

ทำการประเมินแบบจำลองที่สร้างขึ้นด้วยการลองกับสถานการณ์จริง เพื่อดูว่าแบบจำลองนี้ได้ผลเพียงใด

6) เริ่มใช้งานจริง

เริ่มใช้งานจริง เพื่อตรวจผลการทำเหมืองข้อมูลว่าบรรลุตามเป้าหมายที่ตั้งไว้มากน้อยเพียงใด

**10. อัลกอริทึมในการทำเหมืองข้อมูล**

อัลกอริทึมในการทำเหมืองข้อมูล สามารถแบ่งได้เป็น 2 ประเภท ได้แก่

10.1 การสร้างแบบจำลองแบบทำนาย (Predictive Model, Supervised Model)

ในที่นี้ทุกข้อมูลจะมีฉลาก (Label) เป็นคุณสมบัติหนึ่งเพื่อใช้ในการทำนายผลของข้อมูล โดยอัลกอริทึมนี้จะเน้นการจัดกลุ่มข้อมูลตามคุณสมบัติของฉลาก ซึ่งถ้าคุณสมบัติของฉลากไม่ต่อเนื่องก็จะเรียกกระบวนการที่ใช้แบ่งแยกว่า การจำแนก (Classification) แต่ถ้าค่าต่อเนื่องจะเรียกกระบวนการที่ใช้แบ่งแยกว่า การถดถอย (Regression)

10.2 การสร้างแบบจำลองเชิงพรรณนา (Descriptive Model, Unsupervised Model)

ในที่นี้อาจเป็นการหาความสัมพันธ์ (Association) หรือการจัดกลุ่มข้อมูล (Clustering) ซึ่งในการทำงานวิจัยนี้ได้เลือกอัลกอริทึมในการทำเหมืองข้อมูลประเภทนี้มาใช้ โดยอัลกอริทึมที่เลือกใช้ประกอบด้วย

10.2.1 K-Mean Algorithm

เทคนิคการจัดกลุ่มข้อมูลโดยอัลกอริทึมเค-มีน (K-Means Algorithm) นั้นได้มีการพัฒนาและนำเสนอโดย Mac Queen ในปี 1967 ซึ่งแสดงให้เห็นถึงอัลกอริทึมในการจัดกลุ่มที่สมาชิกภายในแต่ละกลุ่มนั้น จะมีระยะใกล้จุดศูนย์กลางหรือตัวแทนของกลุ่ม (Mean) โดยขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม K-Mean นั้นจะประกอบด้วย การกำหนดจำนวนกลุ่มเริ่มต้น กำหนดตัวแทนกลุ่ม การจัดข้อมูลแต่ละตัวเข้ากลุ่ม และสุดท้ายการปรับปรุงตัวแทนกลุ่มในแต่ละกลุ่ม ปัญหาที่พบคือ ปัญหาการกำหนดจำนวนจุดเริ่มต้น ซึ่งส่งผลต่อประสิทธิภาพของการจัดกลุ่ม หรือการที่กลุ่มบางกลุ่มนั้นมีจำนวนสมาชิกน้อยเกินหรืออาจจะไม่มีสมาชิกในกลุ่มเลย จึงได้มีการกำหนดกฎเกณฑ์ว่ากลุ่มที่จะอยู่ในรอบถัดไปนั้นจะต้องมีสมาชิกอยู่ไม่น้อยกว่าค่าคงที่ค่าหนึ่งที่กำหนดขึ้นมา (Bradley, 1998)

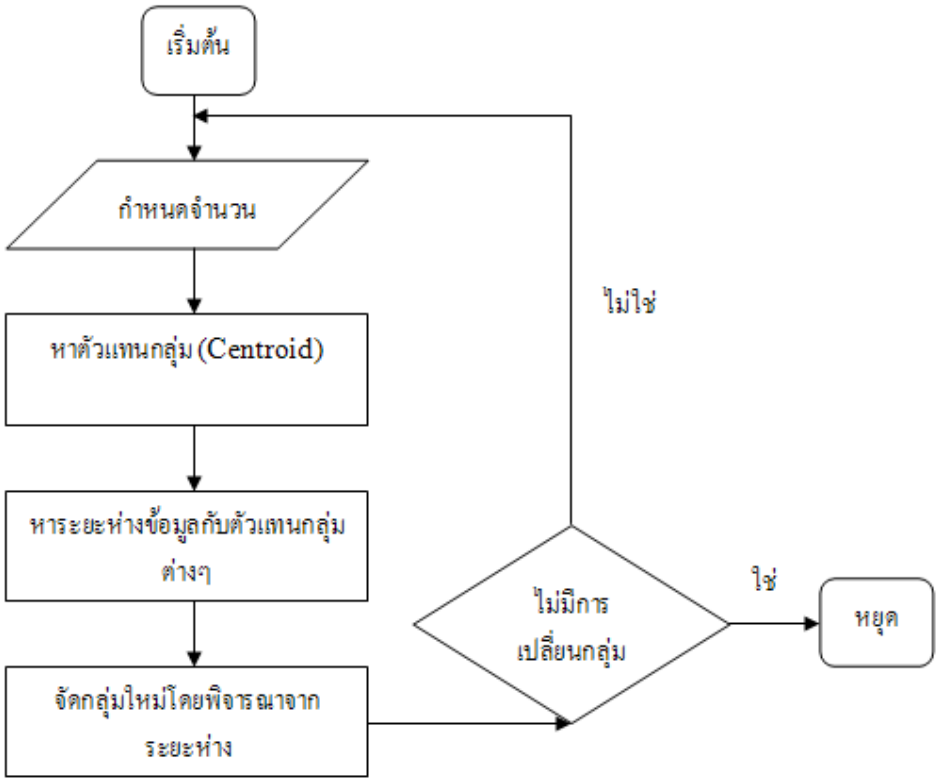
ขั้นตอนการจัดกลุ่ม K-Means

1) กำหนดจำนวนกลุ่มที่ต้องการ โดยกำหนดให้ K เท่ากับจำนวนกลุ่ม โดยที่ในทุกๆ กลุ่มนั้นจะต้องมีสมาชิกภายในกลุ่มเสมอ

2) คำนวณหาตัวแทนกลุ่ม หรือจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม

3) จัดกลุ่มข้อมูลใหม่โดยพิจารณาจากค่าความใกล้ชิดหรือระยะห่างของข้อมูลในกลุ่มกับตัวแทนของกลุ่มต่างๆ ว่าข้อมูลนั้นสมควรที่จะอยู่กลุ่มใด

4) ทำการปรับปรุงการจัดกลุ่มโดยย้อนกลับไปทำข้อ 2 และจะหยุดเมื่อข้อมูลสมาชิกในกลุ่มแต่ละกลุ่มนั้นไม่มีการเปลี่ยนแปลงแล้ว




**ภาพที่ 5** แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึมเค-มีน (K-Means)

ตัวอย่าง K-Means Algorithm

สมมุติมีข้อมูลอยู่ 4 รายการคือ A, B, C และ D โดยแต่ละชุดข้อมูลมี 2 คอลัมน์ คือ X และ Y จากนั้นกำหนดให้มีการจัดกลุ่มข้อมูลเป็น 2 กลุ่ม (K=2) ดังตัวอย่างข้อมูลตารางที่ 1

**ตารางที่ 1** ตัวอย่างข้อมูลเพื่อนำมาแบ่งกลุ่มโดย K-means

| ข้อมูล | X | Y |
|---|---|---|
| **A** | 5 | 3 |
| **B** | -1 | 1 |
| **C** | 1 | -2 |
| **D** | -3 | -2 |

1) ให้สุ่มเลือกกลุ่มโดยสุ่มจับคู่ ดังตัวอย่างจะจับคู่ A กับ B และ C กับ D หลังจากนั้นก็ทำการหาค่าเฉลี่ยจุดกึ่งกลาง จะได้ค่าดังตัวอย่างตารางที่ 2

**ตารางที่ 2** ตัวอย่างการคำนวณค่าจุดกึ่งกลางของกลุ่ม

| กลุ่ม | ค่าเฉลี่ยจุดกึ่งกลาง | กลุ่ม |
|---|---|---|
| | X’ | Y’ |
| **AB** | (5+(-1))/2 = 2 | AB |
| **CD** | (1+(-3))/2 = -1 | CD |

2) หาค่าระยะห่างระหว่างข้อมูลกับกลุ่มโดยใช้สูตร ดังตัวอย่างต่อไปนี้

สูตรการหาระยะห่างแบบ Euclidean Distance

$$A = (x_{1a}, x_{2a}), B = (x_{1b}, x_{2b}) \quad (1)$$

$$d^2(AB) = (x_{1a}, x_{1b})^2 + (x_{2a} - x_{2b})^2 \quad (2)$$

3) นำสูตรมาคำนวณค่าระยะห่างระหว่างจุดข้อมูลกับกลุ่ม จากนั้นจึงจัดข้อมูลให้อยู่ในกลุ่มที่ซึ่งคำนวณค่าได้น้อยกว่า

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$
$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

4) จากนั้นจึงคำนวณสูตรกับข้อมูลที่เหลือให้ครบทั้งหมด

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$
$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

5) เมื่อได้กลุ่มข้อมูลใหม่แล้ว ก็ให้คำนวณหาค่าเฉลี่ยของแต่ละกลุ่มใหม่ดังตาราง 3

**ตารางที่ 3** แสดงค่าเฉลี่ยใหม่ของแต่ละกลุ่ม

| กลุ่ม | ค่าเฉลี่ยจุดกึ่งกลาง | กลุ่ม |
|---|---|---|
| | X’ | Y’ |
| **A** | 5 | A |
| **BCD** | -1 | BCD |

6) ทำซ้ำขั้นตอนที่กล่าวมาข้างต้นจนกว่าข้อมูลจะไม่มีการเปลี่ยนแปลง สุดท้ายแล้วจะได้กลุ่มข้อมูล 2 กลุ่มโดยแบ่งเป็นกลุ่มที่ 1 คือ A และกลุ่มที่ 2 คือ BCD

10.2.2 FarthestFirst Algorithm

เทคนิคการจัดกลุ่มข้อมูลแบบ FarthestFirst Algorithm เป็นอัลกอริทึมที่ทำงานรวดเร็ว ซึ่งจะพิจารณารัศมีสูงสุดและต่ำสุดของกลุ่มข้อมูล ในอัลกอริทึมนี้จุด k จะเป็นจุดแรกที่ถูกเลือกเป็นจุดศูนย์กลางกลุ่ม จุดที่ยังคงอยู่จะถูกเพิ่มเข้าไปยังกลุ่มที่มีจุดศูนย์กลางอยู่ใกล้ที่สุด จุดศูนย์กลางจุดแรกจะถูกเลือกแบบสุ่ม จุดศูนย์กลางจุดที่สองจะถูกเลือกจากจุดที่อยู่ไกลที่สุดจากจุดแรก จุดศูนย์กลางที่ยังคงอยู่แต่ละอันจะถูกตัดสินใจโดยการเลือกจุดที่ไกลที่สุดจากเซตของจุดศูนย์กลางที่ถูกเลือกเรียบร้อยแล้ว ซึ่งให้จุดที่ไกลที่สุดเป็น (X) จากเซต (S) ซึ่งเขียนได้ดังนี้

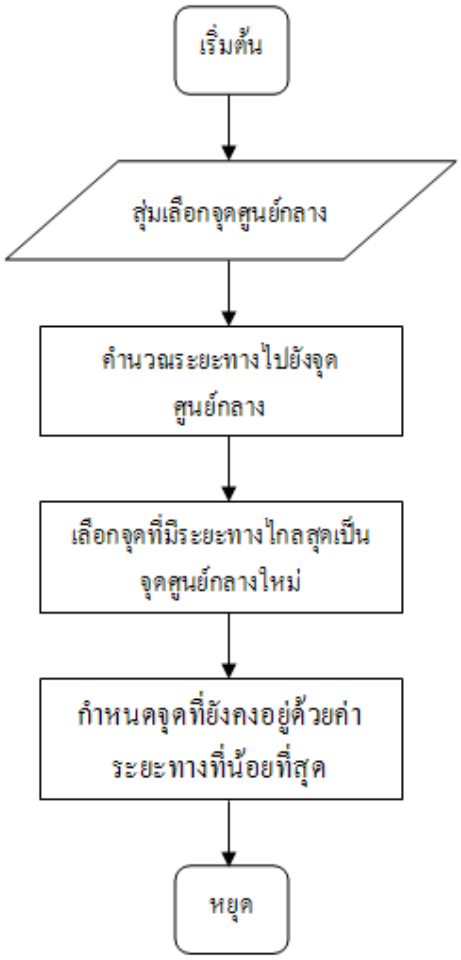
$$max_x\{min\{distance(x, j), j \in S\}\}$$

อัลกอริทึมของ FarthestFirst แสดงได้ดังนี้

1) สุ่มเลือกจุดศุนย์กลางจุดแรก

2) แต่ละจุดให้คำนวณระยะทางไปยังจุดศูนย์กลางของกลุ่มนั้น และเลือกจุดที่มีระยะทางไกลสุด หรือมีค่าระยะทางมากที่สุดเป็นจุดศูนย์กลางใหม่

3) กำหนดจุดที่ยังคงอยู่ด้วยการคำนวณค่าระยะทางไปยังแต่ละจุดศูนย์กลางของแต่ละกลุ่มและใส่ค่าระยะทางที่น้อยที่สุดให้กับกลุ่มนั้นๆ




**ภาพที่ 6** แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม FarthestFirst

10.2.3 RandomFlatClustering

เทคนิค RandomFlatClustering โดยปกติเริ่มจากการสุ่ม (บางส่วน) ของ item เข้าไปยังกลุ่ม และทำการสกัดซ้ำ ซึ่งอัลกอริทึมหลักที่ใช้ คือ K-means แสดงขั้นตอนการทำงานได้ดังนี้

1) อัลกอริทึม Flat มีวิธีคิดคำนวณโดยการแบ่งพื้นที่ของ N รายการเข้าไปในชุดข้อมูลของ K กลุ่ม

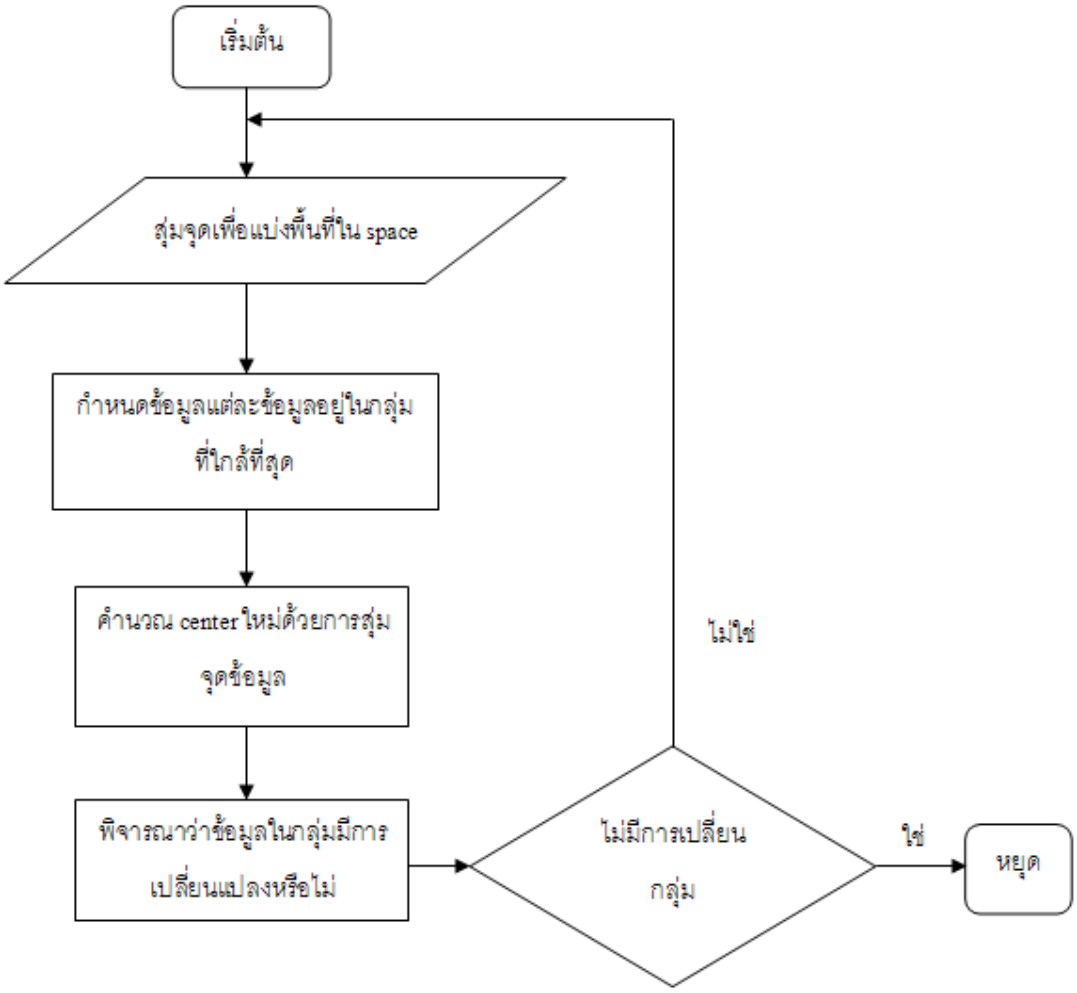
2) กำหนดชุดข้อมูลของรายการและจำนวนกลุ่ม (K) ในการหาให้ทำการแบ่งพื้นที่ใน K กลุ่มที่จะทำให้เกิดความเหมาะสมในการเลือกการแบ่งพื้นที่

3) การพิจารณาผลลัพธ์ที่ดีที่สุดพิจารณาจากความละเอียดในการแบ่งพื้นที่ และเก็บพื้นที่ที่ดีที่สุดนั้นไว้ สำหรับวิธีการวัดประสิทธิภาพไม่มีแนวทางหรือกฎเกณฑ์ที่ชัดเจนจะใช้วิธีการของอัลกอริทึม K-means และ K-medoids มาช่วย

ขั้นตอนการทำ RandomFlatClustering ใช้หลักการเดียวกับ K-means ดังนี้

1) กำหนดกลุ่มเริ่มต้นเป็นศูนย์กลาง
2) ให้กำหนดข้อมูลแต่ละข้อมูลอยู่ในกลุ่มที่ใกล้ที่สุด
3) ทำการคำนวณ center ใหม่ด้วยการหาค่าเฉลี่ยจากจุดข้อมูล
4) ถ้าจุดข้อมูลมีการเปลี่ยนแปลงให้ไปทำตาม Step ที่ 2 ต่อ

ปัญหาของ K-means ที่พบคือ ทำให้การทำงานช้ามาก ดังนั้นเพื่อแก้ไขปัญหานี้จึงได้มีการเริ่มต้นการกำหนดค่าเฉลี่ยของกลุ่มใหม่ด้วยวิธีการสุ่มจุดที่จะวาดพื้นที่ใน space ตามสมควรหลังจากนั้นทำการสุ่มจุดข้อมูลโดยเริ่มต้นจากการสุ่มจุดข้อมูลและนำข้อมูลจุดที่สองให้ไกลออก ไปเท่าที่เป็นไปได้ต่อมานำข้อมูลจุดที่สามออกไปให้ไกลเท่าที่เป็นไปได้อีกครั้ง โดยทั่วๆไปแนวความคิดนี้จะทำให้ K-means มีการทำงานน้อยลง ซึ่งการทำงานหลักก็จะคล้ายกับการทำงานของ k-mediods ด้วยแต่ต่างกันตรงที่เราจะทำการระบุจุด centroid ให้กับจุดข้อมูล




**ภาพที่ 7** แสดงขั้นตอนการจัดกลุ่มข้อมูลโดยใช้อัลกอริทึม RandomFlatClustering

10.2.4 FilteredClusterer

เทคนิค FilteredClusterer คือ การทำ meta- clusterer ที่ใช้ตัว filter โดยตรงก่อนที่ clusterer จะถูกเรียนรู้ โครงสร้างของ filter คือพื้นฐานของข้อมูลที่ฝึกฝน และทดสอบตัวอย่างซึ่งจะถูกประมวลผลโดย filter โดยปราศจากการเปลี่ยนโครงสร้างเดิม

10.2.5 Sequential IB Clustering (sIB)

เทคนิค Sequential IB Clustering เป็นทฤษฎีการแบ่งกลุ่มข้อมูลซึ่งให้การกระจายตัวของข้อมูลเป็นแบบ P(x,y) ของ 2 random variable X และ Y เป็นเทคนิคที่จะทำการกำหนดค่าระยะทางแต่ละค่าไปให้กลุ่มที่มีค่าระยะทางต่ำที่สุด เริ่มต้นจากการสุ่มแบ่งข้อมูลเข้าไปยังกลุ่มแต่ละกลุ่ม และทำการย้ายข้อมูลเมื่ออยู่ใกล้กลุ่มที่มีระยะทางใกล้กับข้อมูลมากที่สุด ทำไปจนกระทั่งข้อมูลในกลุ่มไม่มีการเปลี่ยนกลุ่ม

Pseudocode การทำงานของอัลกอริทึม Sequential IB Clustering เขียนได้ดังนี้

$C \leftarrow$ random partition of $X$ into $K$ clusters
**while** not *done*
  $done \leftarrow TRUE$
  **for every** $x \in X$ :
    Remove $x$ from current cluster $c(x)$
    $c'(x) \leftarrow \text{argmin}_{c \in C} \Delta L(\{x\}, c)$
    **if** $c'(x) \neq c(x)$
      $done \leftarrow FALSE$ .
    Merge $x$ into $c'(x)$
  **end for**
**end while**

**ภาพที่ 8** Pseudocode ของอัลกอริทึม sequential IB Clustering

**11. วิธีการเรียนรู้แบบเบย์ (Bayesian Learning)**

การเรียนรู้แบบเบย์ เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำโมเดลมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ความรู้ก่อนหน้า หมายถึง ความรู้ที่เรามีเกี่ยวกับสมมติฐานแต่ละตัวก่อนที่เราจะเก็บข้อมูล เมื่อใช้งานเราจะนำความน่าจะเป็นของข้อมูลที่เก็บได้มาปรับสมมติฐานซ้ำอีกครั้ง ข้อดีของวิธีการเรียนรู้แบบนี้ก็คือเราสามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior knowledge) เข้ามาช่วยในการเรียนรู้ได้ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดีไม่ด้อยกว่าวิธีการเรียนรู้ประเภทอื่น (บุญเสริม, 2546)

การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกันมีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่นทำให้ผู้วิจัยเลือกวิธีการนี้มาใช้ในงานวิจัย เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและมีอัลกอริทึมในการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่นๆ (Domingos and Pazzani : 1996:105-112) กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม vj สำหรับข้อมูลที่มีคุณสมบัติ n ตัว X={ $a_1$, $a_2$, $\cdots$, $a_n$} หรือ ใช้สัญลักษณ์ว่า P( $a_1$, $a_2$, $\cdots$, $a_n$ | $v_j$) คือ

$$P(a_1, a_2, ..., a_2 \mid v_j) = \prod_{i=1}^{n} P(a_i \mid v_j) \tag{3}$$

โดยที่ $\prod$ หมายถึง ผลคูณของค่า P( $a_i$ | $v_j$) ทั้งหมด และ i = 1,2,3,…,n และ j = 1,2,3,…,n

การนำวิธีการเรียนรู้เบย์อย่างง่ายไปใช้ มีวิธีการดังต่อไปนี้ คือ

1) หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า P( $a_1$, $a_2$, $\cdots$, $a_n$ | $v_j$) จากสมการที่ 3 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ P( $v_j$) ได้เท่ากับ $v_{NB}$

2) นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นเราจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการที่ 4

$$v_{NB} = \arg\max P(v_j) \times \prod_{i=1}^{n} P(a_i \mid v_j); v_j \in V \tag{4}$$

ตัวอย่างการใช้อัลกอริทึมการเรียนรู้เบย์อย่างง่าย โดยใช้ชุดตัวอย่างสอนดังตารางต่อไปนี้

**ตารางที่ 4** ชุดตัวอย่างสำหรับการเรียนรู้เบย์อย่างง่าย

| Name | Hair | Height | Weight | Lotion | Result |
|---|---|---|---|---|---|
| Sarah | blonde | average | light | no | sunburned |
| Dana | blonde | tall | average | yes | none |
| Alex | brown | short | average | yes | none |
| Annie | blonde | short | average | no | sunburned |
| Emily | red | average | heavey | no | sunburned |
| Pete | brown | tall | heavey | no | none |
| John | brown | average | heavey | no | none |
| Katie | blonde | short | light | yes | none |

จากตาราง Name Hair Height Weight Lotion และ Result เป็น Attribute ซึ่ง Result เป็นผลจากการแบ่งลักษณะของ Class และได้บันทึกค่า (Value) ต่างๆ ลงไปแยกตาม Attribute สมมติว่าตัวอย่างที่ต้องการจำแนกประเภท คือ ผลลัพธ์ในตารางที่ 4

**ตารางที่ 5** ชุดตัวอย่างข้อมูลที่ต้องการจำแนกประเภท

| Name | Hair | Height | Weight | Lotion | Result |
|---|---|---|---|---|---|
| Judy | blonde | average | heavey | no | ? |

ค่าความน่าจะเป็นของประเภท Sunburned และ none คำนวณได้จากสมการที่ 4 กรณีความน่าจะเป็นของประเภท Sunburned แทนด้วย $v_{NB} = +$ จะได้ว่า

$$P(+)P(blonde\,|\,+)P(average\,|\,+)P(heavy\,|\,+)P(no\,|\,+) = \frac{3}{8}\times\frac{2}{3}\times\frac{2}{3}\times\frac{1}{3}\times\frac{3}{3} = \frac{1}{18} \quad (5)$$

โดยที่ P(+) หมายถึง ความน่าจะเป็นของประเภท Sunburned ซึ่งผลลัพธ์ที่ได้เป็นประเภท Sunburned จำนวน 3 คน จากทั้งหมด 8 คน

P(blonde|+) หมายถึง ความน่าจะเป็นของคนที่มีผมสี blonde ในประเภท Sunburned

ซึ่งมีจำนวน 2 คน จากทั้งหมด 3 คน

P(average|+) หมายถึง ความน่าจะเป็นของคนที่มีความสูงเฉลี่ย ในประเภท Sunburned ซึ่งมีจำนวน 2 คน จากทั้งหมด 3 คน

P(heavy|+) หมายถึง ความน่าจะเป็นของคนที่มีน้ำหนัก หนัก ในประเภท Sunburned ซึ่งมีจำนวน 1 คน จากทั้งหมด 3 คน

P(no|+) หมายถึง ความน่าจะเป็นของคนที่ไม่มีโลชัน ในประเภท Sunburned ซึ่งมีจำนวน 3 คน จากทั้งหมด 3 คน

กรณีความน่าจะเป็นของประเภท none แทนด้วย $v_{NB}$ = - จะได้ว่า

$$P(-)P(blonde\,|\,-)P(average\,|\,-)P(heavy\,|\,-)P(no\,|\,-) = \frac{5}{8}\times\frac{2}{5}\times\frac{1}{5}\times\frac{2}{5}\times\frac{2}{5} = \frac{1}{125} \quad (6)$$

โดยที่ P(-) หมายถึง ความน่าจะเป็นของประเภท none ซึ่งผลลัพธ์ที่ได้เป็นประเภท none จำนวน 5 คน จากทั้งหมด 8 คน

P(blonde|-) หมายถึง ความน่าจะเป็นของคนที่มีผมสี blonde ในประเภท none ซึ่งมีจำนวน 2 คน จากทั้งหมด 5 คน

P(average|-) หมายถึง ความน่าจะเป็นของคนที่มีความสูงเฉลี่ย ในประเภท none ซึ่งมีจำนวน 1 คน จากทั้งหมด 5 คน

P(heavy|-) หมายถึง ความน่าจะเป็นของคนที่มีน้ำหนัก หนัก ในประเภท none ซึ่งมีจำนวน 2 คน จากทั้งหมด 5 คน

P(no|-) หมายถึง ความน่าจะเป็นของคนที่ไม่มีโลชัน ในประเภท none ซึ่งมีจำนวน 2 คน จากทั้งหมด 5 คน

ดังนั้นคำตอบที่ได้ คือ $v_{NB}$ = + หมายถึง ประเภท Sunburned เนื่องจากประเภท Sunburned มีค่าความน่าจะเป็นมากที่สุด

## งานวิจัยที่เกี่ยวข้อง

งานวิจัยของ Ramah *et al.* (2006) ได้กล่าวถึงความผิดพลาดของทราฟฟิกว่าคือ ลักษณะที่ผิดปกติและเป็นการเปลี่ยนแปลงที่สำคัญของพฤติกรรมของทราฟฟิกในเครือข่าย ซึ่งอาจเกิดจากการมีเจตนามุ่งร้ายหรือไม่มีเจตนามุ่งร้ายก็ได้ ความผิดปกติที่เกิดจากเจตนามุ่งร้ายเป็นสาเหตุให้เกิดการโจมตี เป็นการใช้งานเครือข่ายในทางที่ผิดทำให้เกิดการแพร่ของเวิร์มและไวรัสขึ้น อย่างไรก็ตามการใช้งานเครือข่ายซึ่งไม่ได้มีเจตนามุ่งร้ายก็อาจทำให้เกิดความล้มเหลวได้อาจเกิดจากการคอนฟิกเราเตอร์ผิดพลาด ซึ่งได้นำเสนอระบบการตรวจจับความผิดพลาดที่ได้พัฒนามาจากแบบแผนการตรวจจับความผิดพลาดที่นำเสนอโดย Mei- Ling Shyu และอาศัยการเก็บรวบรวมข้อมูล SNMP เป็นระยะๆ ซึ่งได้ประเมินระบบนี้ว่าสามารถต่อต้านการโจมตีแบบทั่วไปได้ และได้พบการโจมตีจาก Smurf จำนวนหนึ่งที่เข้ามาพร้อมๆกัน ซึ่งตรวจจับได้ดีกว่าวิธีการสแกน ดังนั้นจึงได้ใช้ระบบนี้เพื่อตรวจจับทราฟฟิกที่มีความผิดปกติในเครือข่ายของ Tunisian National University (TNUN) และได้ทำการเก็บทราฟฟิกของเครือข่ายด้วยการค้นหาเส้นทาง (trace) จากระบบการจัดการข้อมูลพื้นฐาน (Management Information Base หรือ MIB) ของไฟร์วอลศูนย์กลางของเครือข่าย TNUN และได้คำนวณการกระจายตัวของเวลาระหว่างที่มีความผิดปกติ และการกระจายตัวของระยะเวลาที่มีความผิดปกติ และได้แสดงให้เห็นว่าความผิดปกติมีการแพร่กระจายในเครือข่าย TNUN และความผิดปกติส่วนมากอยู่ในเครือข่ายน้อยกว่า 5 นาที

ทางผู้วิจัยได้เฝ้าสังเกตุการเพิ่มขึ้นอย่างต่อเนื่องของทราฟฟิกที่มีความผิดปกติในอินเทอร์เน็ตในรูปแบบของการกระจายการปฏิเสธการบริการที่มีการโจมตี เช่น การแพร่ของไวรัสและเวิร์ม การก่อกวนต่างๆ ดังนั้นการป้องกันเครือข่ายเพื่อต่อสู้กับความผิดปกติของทราฟฟิก คือการทำงานวันต่อวันสำหรับโอเปอร์เรเตอร์เครือข่าย ทางผู้วิจัยยังได้ทำการแบ่งแยกความแตกต่างของเทคนิคการตรวจจับการก่อกวนระหว่าง 2 ประเภท คือ Misuse Detection และ Anomaly Detection

สำหรับระบบ Misuse Detection พยายามที่จะตรวจจับการก่อกวนโดยการเปรียบเทียบกิจกรรมปัจจุบันของทรัพยากรการตรวจสอบบัญชีไปยังฐานข้อมูลของการโจมตีที่รู้จัก เทคนิคทั้งสองไม่สามารถตรวจจับการโจมตีที่ไม่รู้จักได้ แต่อย่างไรก็ตามระบบ Anomaly Detection System (ADS) พยายามตรวจจับการก่อกวนโดยการเปรียบเทียบกิจกรรมปัจจุบันของทรัพยากรการตรวจสอบบัญชีเพื่อสร้าง “กิจกรรมปกติ” (normal activity) โดยแสดงในรูปแบบของโครงร่าง

หลักการของเทคนิคเหล่านี้คือต้องการเก็บสถานะต่อหนึ่งการเชื่อมต่อ หรือต่อหนึ่งการไหลเวียนบนลิงค์เดียวหรือโหนดเดียว ซึ่งจำเป็นต้องถูกจัดให้เหมาะสมในทุกๆ โหนดเพื่อให้มีประสิทธิภาพที่สุด และได้พยายามพัฒนาเครื่องมือตรวจจับความผิดปกติที่สามารถจะตรวจจับการโจมตีได้โดยปราศจากการเก็บสถานะต่อหนึ่งการไหลเวียน เครื่องมือนี้ไม่ได้สร้างเพื่อจะระบุความแตกต่างของชนิดของการโจมตีหรือแหล่งกำเนิดของการโจมตี จึงนำมาใช้ให้เป็นประโยชน์ได้สำหรับเส้นทางแรกของเครื่องมือตรวจจับความผิดปกติ ในความเป็นจริงแล้วเครื่องมือนี้สามารถใช้ระบุการก่อกวนได้ก็ต่อเมื่อระบบตรวจจับสิ่งผิดปกติมีสิ่งก่อกวนเป็นจำนวนมาก โดยอาศัยการรวบรวมข้อมูลต่อหนึ่งการไหลเวียน

พวกเขาได้พัฒนาระบบ Anomaly Detection System (ADS) จากแบบแผนการตรวจจับความผิดพลาดซึ่งนำเสนอโดย Mei- Ling Shyu และอาศัยการเก็บรวบรวมข้อมูล SNMP หลังจากนั้นได้ทำการประเมินค่าระบบนี้ถึงการต่อต้านการโจมตีทั่วไป ซึ่งใช้ประโยชน์ของมันมาตรวจจับความผิดปกติของทราฟฟิกในเครือข่าย TNUN และสุดท้ายได้ทำการศึกษารูปแบบชั่วคราวของความผิดปกติของทราฟฟิกในเครือข่าย

ซึ่งได้นำเสนอระบบ Anomaly Detection System (ADS) ซึ่งพัฒนามาจากงานของ Mei-Ling Shyu ซึ่งเสนอข้อคิดเห็นถึงรูปแบบการตรวจจับความผิดปกติที่ไม่มีผู้ดูแลโดยอาศัยหลักการวิเคราะห์จากส่วนประกอบ หรือ Principal Component Analysis (PCA) และตั้งสมมุติฐานว่าจะตรวจจับความผิดปกติได้หมด

PCA คือวิธีการเปลี่ยนแปลงแบบหลากหลาย (multivariate method) เกี่ยวข้องกับการอธิบายโครงสร้าง variance และ covariance ของกลุ่มของตัวแปรผ่านตัวแปรใหม่ไม่กี่ตัวซึ่งรวมกันเป็นเส้นตรงของต้นกำเนิด และได้มีการนำวิธีการตรวจจับความผิดปกติของ Shyu มาใช้ สำหรับวิธีการประเมินผล ADS จะทำโดยการหาเส้นทางความผิดปกติของทราฟฟิกที่ระบุได้ดี โดยได้จัดเครือข่ายให้เหมาะสม ซึ่งประกอบด้วยเครือข่าย LAN สองเครือข่ายที่เชื่อมต่อโดยเราเตอร์ และทำการจำลองทราฟฟิกปกติ และใช้เครื่องกำเนิดทราฟฟิกของเครือข่าย LANTRAFFIC ซึ่งให้มีการเชื่อมต่อแบบ 2 ทิศทางโดยควบคุมให้มีการส่ง 16 TCP และ 16 UDP ระหว่างปลายทางและเครื่องกำเนิดทราฟฟิก ในตอนท้ายได้จัดสถานี ADS ซึ่งประมวลผลการเก็บรวบรวมข้อมูล SNMP จากเราท์เตอร์ศูนย์กลาง ข้อมูลนี้ถูกเก็บรวบรวมโดยโปรแกรมระบบการจัดการ WhatsUP ทุกๆ 20 วินาที

ทางผู้วิจัยได้แสดงระบบการตรวจจับความผิดปกติระดับแรกโดยอาศัยข้อมูล SNMP ซึ่งระบบนี้สามารถตรวจจับความผิดปกติได้โดยอัตโนมัติและเป็นแบบ real time และประเมินระบบนี้ได้ว่าสามารถต่อต้านการโจมตีที่รู้จักจำนวนหนึ่งได้และพบว่ามีประสิทธิภาพในการตรวจจับการโจมตีที่มีการไหลต่อเนื่องซึ่งเป็นตัวทำลายเครือข่าย การโจมตีเหล่านี้ยากต่อการตรวจจับกับระบบการตรวจจับการก่อกวนแบบปกติและป้องกันได้โดยการใช้ firewall เพราะว่าเขาพยายามใช้การเชื่อมต่อแบบปกติ ดังนั้นสามารถพูดได้ว่าการศึกษาของเขาจะเป็นแนวทางเพื่อให้มีการพิสูจน์ต่อสำหรับทราฟฟิกที่มีความผิดปกติในอินเตอร์เน็ต ซึ่งการที่เขาพัฒนาระบบนี้ขึ้นมาก็เพื่อช่วยเหลือผู้ดูแลเครือข่ายใน Tunisian universities เพื่อที่จะให้สามารถตรวจจับการโจมตีได้เร็วขึ้น ในอนาคตเขาวางแผนเพื่อเพิ่มชิ้นส่วนในระบบของเขาสำหรับระบุการโจมตี ดังนั้นผู้ดูและเครือข่ายสามารถนำมาเป็นเครื่องมือเพื่อกรองทราฟฟิกเพื่อจะบรรเทาผลกระทบจากทราฟฟิกที่ผิดปกติเพื่อให้มีทราฟฟิกที่ดี

งานวิจัยของ Lakhina *el at.* (2005) ได้ทำการทดสอบการเพิ่มขึ้นของสเกลการจับการไหลให้มีขนาดใหญ่ขึ้น ทำให้มันเป็นไปได้ที่จะคิดค้นวิธีการวิเคราะห์ traffic ซึ่งตรวจจับและระบุกลุ่มความผิดปกติที่หลากหลายและมีขนาดใหญ่ อย่างไรก็ตามสิ่งที่ท้าทายของการวิเคราะห์อย่างมีประสิทธิภาพก็คือ ที่มาของกลุ่มข้อมูลสำหรับการวินิจฉัยข้อมูลที่ผิดปกติที่ยังหาไม่พบจนบัดนี้ พวกเขาพิสูจน์การกระจายตัวของลักษณะ packet ( IP addresses และพอร์ต) ด้วยการเฝ้าสังเกตเส้นทางที่ปรากฏออกมา ณ ปัจจุบันและโครงสร้างของช่วงกว้างของความผิดปกติ โดยการใช้ entropy เป็นเครื่องมือในการสรุป

ซึ่งได้แสดงการวิเคราะห์ของลักษณะการกระจายตัวซึ่งนำไปสู่สิ่งที่มีนัยสำคัญทั้ง 2 ด้าน คือ (1.) ทำให้การตรวจจับมีความไวสูงต่อช่วงกว้างของความผิดปกติ และขยายการตรวจจับด้วยวิธีพื้นฐานของระดับเสียง (volume- based methods) และ (2.) ทำให้สามารถแบ่งแยกความผิดปกติได้อย่างอัตโนมัติผ่านการเรียนรู้ โดยแสดงการใช้ลักษณะการกระจายตัวความผิดปกติที่ตกลงมาโดยธรรมชาติเข้าไปในกลุ่มที่แตกต่างกันและมีความหมาย

กลุ่มเหล่านี้สามารถใช้แยกประเภทความผิดปกติได้โดยอัตโนมัติ และเป็นการเปิดให้เห็นความผิดปกติชนิดใหม่ พวกเขามีเหตุผลถึงการอ้างข้อมูลจากสองเครือข่ายหลัก (Abilene และ G'eant) และสรุปว่าลักษณะการกระจายตัวแสดงให้เห็นว่าเป็นส่วนประกอบที่สำคัญของการวินิจฉัยความผิดปกติของเครือข่ายทั่วไป

งานวิจัยของเกรียงไกร และคณะ (2006) ได้พัฒนาระบบตรวจจับการบุกรุกทางอินเตอร์เน็ต มีจุดมุ่งหมายเพื่อพัฒนาระบบตรวจจับการบุกรุก สำนักคอมพิวเตอร์ มหาวิทยาลัยราชภัฏจันทรเกษม เพื่อช่วยในการป้องกันการบุกรุก การถูกโจมตี และถูกก่อกวนในการใช้งานระบบเครือข่ายคอมพิวเตอร์และช่วยในการจัดเก็บและการบริหารข้อมูลที่ได้จากการตรวจจับมาวิเคราะห์แก้ไขและปรับปรุงระบบเครือข่ายคอมพิวเตอร์ให้สามารถใช้งานได้อย่างมีประสิทธิภาพ โดยได้เลือกใช้ระบบตรวจจับการบุกรุกแบบ Open Source Softwareโดยระบบปฏิบัติการใช้เป็นโปรแกรมลีนุกส์ (Linux) ระบบฐานข้อมูลใช้เป็นโปรแกรมโพรสเกรดสเอสคิวแอล (PostgreSQL) และใช้โปรแกรมที่เรียกว่า “Snort” ซึ่งเป็นเครื่องมือที่ใช้ในการตรวจจับการบุกรุก เมื่อ สนอท(Snort)จับแพคเก็ต (Package) ที่เข้ามา โดยจะทำการเปรียบเทียบว่าแพคเก็ตนั้นตรงกับรูปแบบของการบุกรุกใดๆ ที่มีอยู่ในฐานข้อมูลของระบบหรือไม่ ซึ่งรูปแบบของการบุกรุกเข้ามาครั้งหนึ่งๆจะเรียกว่าลายเซ็น (Signature) หากแพคเก็ตที่เข้ามานั้นเข้ากับรูปแบบที่กำหนดไว้ในฐานข้อมูลของโปรแกรมสนอท สนอทจะทำการแจ้งให้ทราบโดยการส่งข้อความเตือน จากผลการทดสอบระบบโดยใช้วิธีการประเมินเพื่อหาประสิทธิภาพของระบบ พบว่าระบบสามารถทำงานได้ในระดับดี และครบถ้วนตามวัตถุประสงค์ที่ตั้งไว้ และจากผลการทดสอบระบบตรวจจับการบุกรุก โดยผู้เชี่ยวชาญและผู้ทำงานเกี่ยวกับระบบเครือข่าย สรุปได้ว่าระบบตรวจจับการบุกรุกมีประสิทธิภาพอยู่ในระดับดีในเกือบทุกด้าน และสามารถทำงานได้อย่างมีประสิทธิภาพ และตรงต่อความต้องการของผู้ใช้ระบบตรวจจับการบุกรุกที่ได้พัฒนาขึ้นสามารถใช้งานกับสำนักคอมพิวเตอร์ ได้จริงตามความต้องการ โดยสามารถตรวจจับการบุกรุกเครือข่ายผ่านทางระบบ Internet หลักของมหาวิทยาลัย จึงเป็นไปตามวัตถุประสงค์ของสารนิพนธ์นี้

งานวิจัยของ McGregor *et al.* (2004) ได้นำเสนออัลกอริทึมใหม่ เรียกว่า EM algorithm สำหรับการแยกประเภทของเฮดเดอร์แพ็กเก็ตซึ่งจะเป็นตัวแบ่งแยกทราฟฟิกที่มีชนิดของโปรแกรมเดียวกัน (single transaction, bulk transfer) เขาสร้างสมมุติฐานความสามารถของการใช้การวิเคราะห์กลุ่มที่จะจัดกลุ่มการไหลเวียน (flow) ด้วยการใช้ attribute ในชั้นการขนส่ง (transport layer) แต่เขาไม่ได้ทำการวัดประสิทธิภาพว่า attribute ตัวใดที่มีประสิทธิภาพในการแบ่งแยกทราฟฟิกได้มากที่สุด

งานวิจัยของ Erman *et al.* (2006) ได้นำเสนอการแบ่งแยกทราฟฟิกเครือข่ายด้วยการวิเคราะห์จาก port-based หรือ payload-based โดยการ trace และใช้อัลกอริทึมในการจัดกลุ่ม 2 อัลกอริทึมด้วยกันคือ K-Means และ DBSCAN และได้วัดและประเมินผลเปรียบเทียบกับ

AutoClass algorithm ซึ่งได้มีการใช้งานมาก่อนหน้านี้ จากผลการทดลองทำให้ทราบว่า K-Means และ DBSCAN ทำงานได้ดี แต่ DBSCAN ค่าความถูกต้อง (Accuracy) ที่ต่ำเมื่อเปรียบกับ K-Means และ AutoClass แต่ DBSCAN ให้ผลในการจัดกลุ่มดีกว่า

งานวิจัยของ Silva *et al.* (2006) ได้พัฒนาเพื่อวิเคราะห์ข้อมูลทราฟฟิกของเครือข่าย ซึ่งเลือกข้อมูลในชั้น TCP/IP มาทำการวิเคราะห์ ซึ่งได้มาจากการใช้งานทาง http ซึ่งข้อมูลชนิดนี้มีขนาดใหญ่และพบในชนิดทั่วๆ ไปของข้อมูลทราฟฟิก เขาจึงใช้การจัดกลุ่มข้อมูลมาช่วยในการวิเคราะห์ ซึ่งเลือกใช้อัลกอริทึม SOM ในกระบวนการทำ data reduction ซึ่งข้อดีของ SOM สามารถนำมาใช้ศึกษาพฤติกรรมของทราฟฟิกได้เป็นอย่างดี และใช้ RGCom เป็นเครื่องมือช่วยในการเก็บข้อมูลเพื่อดูพฤติกรรมของข้อมูลทราฟฟิกโดยใช้คุณสมบัติของ content-based มาช่วยและใช้โปรแกรม sniffer ด้วยการทำ tcpdump และโปรแกรม ethereal

**ตารางที่ 6** ตารางสรุปงานวิจัยที่เกี่ยวข้อง

| งานวิจัย | เทคนิค | ข้อมูล | ผล |
|---|---|---|---|
| Khadija RAMAH, Hichem AYARI, Farouk KAMOU, 2006 | ได้พัฒนาระบบ Anomaly Detection System (ADS) จากแบบแผนการตรวจจับความผิดปกติซึ่งนำเสนอโดย Mei- Ling Shyu และอาศัยการเก็บรวบรวมข้อมูล SNMP และได้ทำการเก็บทราฟฟิกของเครือข่ายด้วยการค้นหาเส้นทาง (trace) จากระบบการจัดการข้อมูลพื้นฐาน (MIB) ของไฟร์วอลศูนย์กลาง และตั้งสมมุติฐานว่าจะตรวจจับความผิดปกติได้หมด | ตรวจจับทราฟฟิกที่มีความผิดปกติในเครือข่ายของ Tunisian National University (TNUN) | ไม่สามารถตรวจจับความผิดปกติได้ทั้งหมด ยังมี unknown anomaly ปรากฏอยู่ ยังไม่สามารถระบุได้ว่าเป็นอะไร |
| Anukool Lakhina, Mark Crovella, Christophe Diot, 2005 | ได้ทำการทดสอบการเพิ่มขึ้นของสเกลการจับทราฟฟิกเพื่อที่จะคิดค้นวิธีการวิเคราะห์ traffic และระบุกลุ่มความผิดปกติที่หลากหลายมากยิ่งขึ้นโดยการใช้ entropy เป็นเครื่องมือในการสรุปโครงสร้างของทราฟฟิก | ข้อมูลจากสองเครือข่ายหลัก (Abilene และ G'eant) | ทำให้การตรวจจับมีความไวสูงต่อช่วงกว้างของความผิดปกติ และทำให้สามารถแบ่งแยกความผิดปกติได้อย่างอัตโนมัติผ่านการเรียนรู้ นอกจากนี้ยังได้พบความผิดปกติชนิดใหม่ |

**ตารางที่ 6** (ต่อ)

| งานวิจัย | เทคนิค | ข้อมูล | ผล |
|---|---|---|---|
| นายเกรียงไกร บุตรภักดี, นายจักรินทร์ เรืองไทย และนายนครินทร์ นีละโยธิน, 2006 | ได้พัฒนาระบบตรวจจับการบุกรุกทางอินเตอร์เน็ต เพื่อพัฒนาระบบตรวจจับการบุกรุก เพื่อช่วยในการป้องกันการบุกรุก การถูกโจมตี และถูกก่อกวนในการใช้งานระบบเครือข่ายคอมพิวเตอร์ และช่วยในการจัดเก็บและการบริหารข้อมูลที่ได้จากการตรวจจับมาวิเคราะห์แก้ไขให้สามารถใช้งานได้อย่างมีประสิทธิภาพ โดยใช้โปรแกรม Snort | สำนักคอมพิวเตอร์ มหาวิทยาลัยราชภัฏจันทรเกษม | ระบบสามารถทำงานได้ในระดับดีและครบถ้วนตามวัตถุประสงค์ที่ตั้งไว้ |
| Anthony McGregor, Mark Hall, Perry Lorier, James Brunskill, 2004 | ได้นำเสนออัลกอริทึมใหม่ เรียกว่า EM algorithm สำหรับการแยกประเภทของเฮดเดอร์แพ็กเก็ตซึ่งจะเป็นตัวแบ่งแยก ทราฟฟิกที่มีชนิดของโปรแกรมเดียวกัน (single transaction, bulk transfer) | ห้องวิจัยเครือข่ายคอมพิวเตอร์ | อัลกอริทึม EM algorithm ใช้ได้ดีกับข้อมูลทราฟฟิก แต่ไม่ได้วัดประสิทธิภาพในการนำไปใช้กับแต่ละ attribute ว่า attribute ใดที่เหมาะสมในการนำมาจัดกลุ่มข้อมูลมากที่สุด |

**ตารางที่ 6** (ต่อ)

| งานวิจัย | เทคนิค | ข้อมูล | ผล |
|---|---|---|---|
| Jeffrey Erman, Martin Arlitt, Anirban Mahanti, 2006 | ได้นำเสนอการแบ่งแยกทราฟฟิกเครือข่ายด้วยการวิเคราะห์จาก port-based หรือ payload-based โดยการ trace และใช้อัลกอริทึมในการจัดกลุ่ม 2 อัลกอริทึมด้วยกันคือ K-Means และ DBSCAN | ข้อมูลเก็บด้วยวิธีการ trace จาก Auckland IV และมหาวิทยาลัยของ Calgary | ผลการทดลองทำให้ทราบว่า K-Means และ DBSCAN ทำงานได้ดี แต่ DBSCAN ค่าความถูกต้อง (Accuracy) ที่ต่ำเมื่อเปรียบกับ K-Means และ AutoClass แต่ DBSCAN ให้ผลในการจัดกลุ่มดีกว่า |
| L.S. Silva, T.D. Mancilha, J.D.S. Silva, A.C.F. Santos, e A. Montes5 | ได้เลือกวิเคราะห์ข้อมูลทราฟฟิกของเครือข่าย ในชั้น TCP/IP มาทำการวิเคราะห์ โดยเลือกใช้อัลกอริทึม SOM ในกระบวนการทำ data reduction และใช้ RGCom เป็นเครื่องมือช่วยในการเก็บข้อมูลโดยใช้คุณสมบัติของ content-based มาช่วย และใช้โปรแกรม sniffer ด้วยการทำ tcpdump และโปรแกรม Ethereal | ข้อมูลจากเครือข่ายภายในของ INPE (Brazilian National Institute for Space Research) | เขาค้นพบประโยชน์ของ SOM ซึ่งสามารถช่วยในการวิเคราะห์ของพฤติกรรมระดับกลางของการจราจรที่ monitor ได้และใช้เครื่องมือ RGCom มาช่วยแต่โปรแกรมนี้ไม่ได้ช่วยทำ data reduction |

# อุปกรณ์และวิธีการ

## อุปกรณ์

1. Hardware

   1.1 คอมพิวเตอร์ส่วนบุคคล CPU Intel Core Duo processor 1.66GHz, 1GB DDR2, HD 60 GB

2. Software

| | |
|---|---|
| 2.1 Microsoft Office Excell 2007 | สำหรับเตรียมข้อมูล |
| 2.2 Wireshark Version 0.99.6a (SVN Rev 22276) | สำหรับเก็บข้อมูลทราฟฟิก |
| 2.3 RapidMiner | สำหรับการแบ่งกลุ่มข้อมูล |
| 2.4 Weka | สำหรับการประเมินผล |

3. Database

   3.1 SQL

   3.2 Microsoft ACCESS 2007

4. Operating System

   4.1 ระบบปฏิบัติการ Windows XP

## วิธีการ

**วิธีการดำเนินการวิจัย และสถานที่ทำการทดลอง/เก็บข้อมูล**

1. ศึกษาผลงานวิจัยที่มีมาก่อนหน้านี้ที่เกี่ยวข้องกับงานวิจัยที่จะทำ
2. ศึกษาและค้นหาวิธีหรือโปรแกรมที่จะนำมาใช้เก็บข้อมูลทราฟฟิก
3. ศึกษาการใช้งานโปรแกรม Ethereal เพื่อประสิทธิภาพในการใช้งาน

4. เก็บรวบรวมข้อมูลทราฟฟิกที่มีการใช้งานจริงในบริษัท ทีโอที เป็นเวลา 24 ชั่วโมง

5. ศึกษาขั้นตอนการทำเหมืองข้อมูล

6. ศึกษาวิธีการแบ่งกลุ่มข้อมูลแต่ละวิธี

7. ทำการแยกประเภทของทราฟฟิกที่มีความผิดปกติด้วยวิธีการจัดกลุ่มข้อมูลโดยใช้ 5 อัลกอริทึม คือ sIB, RandomFlatClustering, FarthestFirst, and FilteredClusterer และ K-Means

8. เลือกวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มของทราฟฟิกที่มีความผิดปกติ ด้วยการเปรียบเทียบผลการวัดประสิทธิภาพการแบ่งกลุ่มข้อมูลของแต่ละวิธีโดยใช้อัลกอริทึม NaiveBaye

9. วิเคราะห์ผลการแบ่งกลุ่มทราฟฟิกของแต่ละอัลกอริทึมด้วยการทดสอบกับข้อมูลทราฟฟิกจริงที่ทราบความผิดปกติ (Anomaly Traffic)

10. นิยาม nomaly และ anomaly traffic จากผลการทดลองที่ได้ ดังได้อธิบายไว้ในหน้าที่

11. ระบุ Atrribute ที่เหมาะสมสำหรับการจัดกลุ่มข้อมูลทราฟฟิกดังได้อธิบายไว้ในหน้าที่

12. ส่งผลงานที่นำเสนอแก่งานประชุมทางวิชาการ

## รายละเอียดวิธีการวิจัย

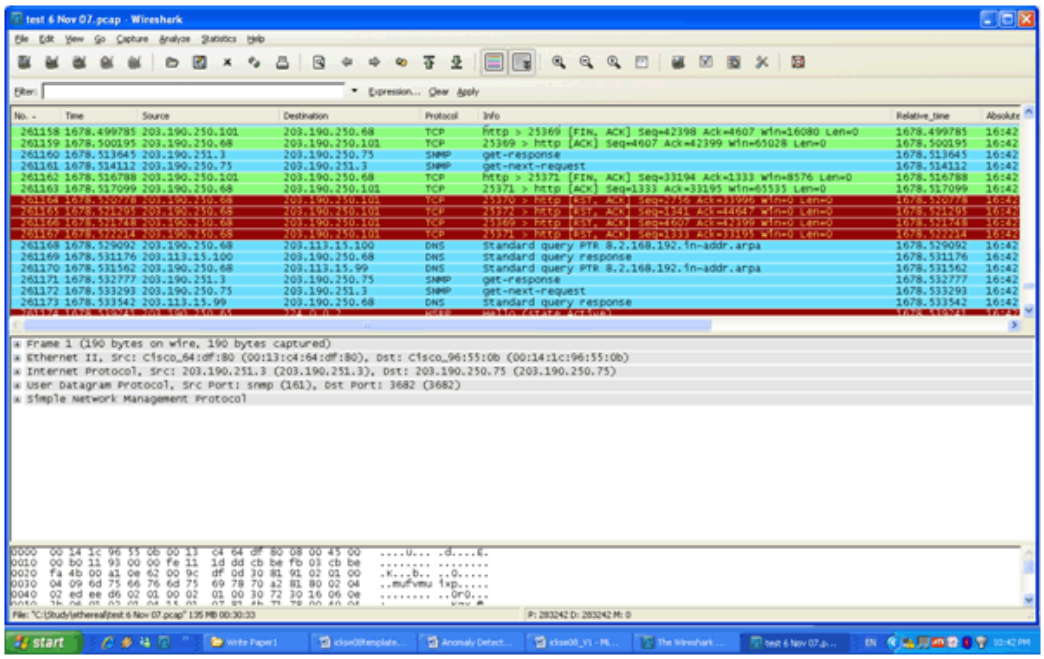
### 1. ศึกษาผลงานวิจัยที่มีมาก่อนหน้านี้ที่เกี่ยวข้องกับงานวิจัยที่จะทำ

ศึกษาหาข้อมูลที่เกี่ยวข้องกับการตรวจจับทราฟฟิกที่มีความผิดปกติให้ได้มากที่สุด เพื่อทำความเข้าใจถึงงานที่ผู้อื่นได้ทำมาแล้ว เพื่อเป็นแนวทางในการศึกษาหาข้อมูลที่จำเป็นต่างๆต่อไป และได้ทราบถึงวิวัฒนาการของงานวิจัยทางด้านนี้ รวมถึงการพัฒนาได้ทันยุคทันสมัยมากที่สุด ทั้งนี้ก็เพื่อนำข้อดีข้อเสียจากผลงานวิจัยอื่นๆมาเป็นประโยชน์ต่อการทำงานวิจัยนี้ได้อีกทางหนึ่ง

### 2. ศึกษาและค้นหาวิธีหรือโปรแกรมที่จะนำมาใช้เก็บข้อมูลทราฟฟิก

โครงงานนี้ได้เลือกใช้โปรแกรมตรวจจับทราฟฟิกที่ชื่อ Ethereal ซึ่งเป็นเครื่องมือที่ใช้ในการวิเคราะห์ข้อมูลเครือข่ายแบบแพกเก็ต ซึ่ง Ethereal จะทำการ capture แพกเก็ตและแสดงรายละเอียดภายในแพกเก็ตออกมาเท่าที่ตัวโปรแกรมจะสามารถทำได้

**3. ศึกษาการใช้งานโปรแกรม Ethereal เพื่อประสิทธิภาพในการใช้งาน**

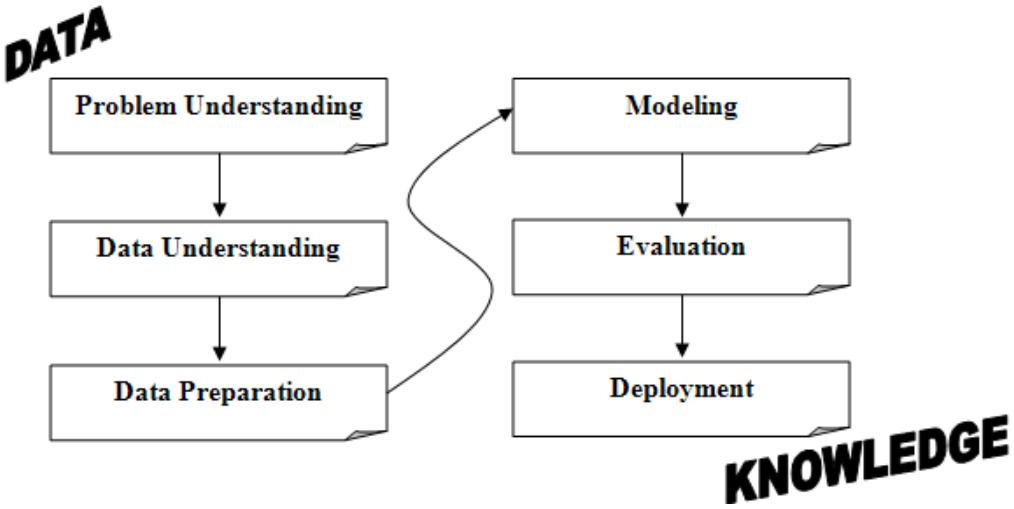


**ภาพที่ 9** แสดง Packet ที่เกิดจากการจับแพกเก็ตโดยใช้โปรแกรม Ethereal

**4. เก็บรวบรวมข้อมูลทราฟฟิกที่มีความผิดปกติในบริษัท ทีโอที เป็นเวลา 24 ชั่วโมง**

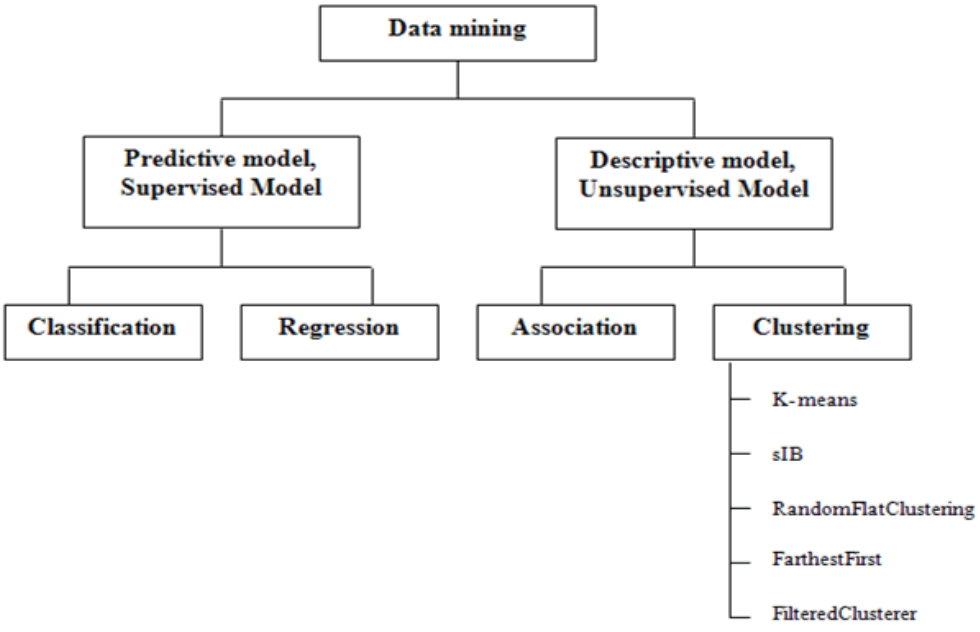
ในขั้นตอนการเก็บข้อมูลทราฟฟิกจะใช้โปรแกรม Ethereal ซึ่งเป็นเครื่องมือที่ใช้ในการตรวจจับแพกเก็ตซึ่งจะทำการเก็บข้อมูลเป็นเวลา 24 ชั่วโมง โดยจะนำข้อมูลแพกเก็ตเหล่านี้มาคัดเลือกและทำการแบ่งกลุ่มข้อมูลต่อไป ซึ่งลักษณะข้อมูลที่เก็บได้จะมีลักษณะดังภาพที่ 6

**5. ศึกษาขั้นตอนการทำเหมืองข้อมูล**




**ภาพที่ 10** แสดงขั้นตอนการทำเหมืองข้อมูล

**6. ศึกษาวิธีการแบ่งแยกประเภทข้อมูลแต่ละวิธี**




**ภาพที่ 11** แสดงโครงสร้างของ Data mining

**7. ทำการแยกประเภทของทราฟฟิกที่มีความผิดปกติด้วยอัลกอริทึม5 อัลกอริทึมด้วยกันคือ K-means, sIB, RandomFlatClustering, FarthestFirst, and FilteredClusterer**

ในขั้นตอนการขุดค้นข้อมูล (Data mining) ผู้วิจัยได้เลือกใช้เทคนิคการจัดกลุ่มข้อมูล (Clustering) เข้ามาใช้ในการทดลองโดยทำการเปรียบเทียบการจัดกลุ่มระหว่างอัลกอริทึมการจัดกลุ่ม (Clustering) 5 ชนิดคือ

7.1 ใช้ K-means Algorithm ในโปรแกรม RapidMiner ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

7.2 ใช้ sIB Algorithm ในโปรแกรม RapidMiner ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

7.3 ใช้ RandomFlatClustering Algorithm ในโปรแกรม RapidMiner ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

7.4 ใช้ FarthestFirst Algorithm ในโปรแกรม RapidMiner ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

7.5 ใช้ FilteredClusterer Algorithm ในโปรแกรม RapidMiner ทำการจัดกลุ่มข้อมูลที่เตรียมไว้

โดยกลุ่มจะถูกแบ่งตามลักษณะการโจมตีและปัจจัยอื่นๆ

**8. เลือกวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มทราฟฟิกที่มีความผิดปกติด้วยการเปรียบเทียบผลการวัดประสิทธิภาพการแบ่งกลุ่มข้อมูลในแต่ละวิธี ใช้อัลกอริทึม NaiveBaye**

เมื่อได้ทำการแยกประเภทของทราฟฟิกที่มีความผิดปกติด้วย 5 วิธีดังกล่าวคือ K-means, sIB, RandomFlatClustering, FarthestFirst, and FilteredClusterer เป็นที่เรียบร้อยแล้ว จากนั้นจะทำการหาวิธีที่เหมาะสมสำหรับการแบ่งกลุ่มของทราฟฟิกที่มีความผิดปกตินั้น ด้วยการวัดประสิทธิภาพการแบ่งกลุ่มข้อมูลของแต่ละวิธีว่าสามารถที่จะแบ่งกลุ่มได้ดีหรือไม่นั้น สามารถทำได้ด้วยการวัดค่าความถูกต้องในการจัดกลุ่มข้อมูล ดังนี้

8.1 การวัดประสิทธิภาพในการจัดกลุ่มข้อมูลด้วยค่า Recall, Precision, F-Score

การวัดประสิทธิภาพในการจัดกลุ่มข้อมูลมีอยู่หลายวิธีแต่วิธีที่นิยมที่สุด คือ การวัดประสิทธิภาพในการจัดกลุ่มข้อมูลด้วยค่า Recall, Precision, F-Score แสดงสมการการคำนวณได้ดังนี้

Precision

ค่า Precision คือ ค่าความแม่นยำในการทดสอบ หลังจากที่ได้ผลจากการจัดกลุ่มข้อมูลเรียบร้อยแล้วจึงนำข้อมูลทุกตัวมาทำการติดป้ายเพื่อนำมาทำการทดสอบความแม่นยำในการจัดกลุ่มข้อมูลที่ได้จากการจัดกลุ่มข้อมูลในแต่อัลกอริทึมโดยกระบวนการทดสอบจะใช้อัลกอริทึมทางด้าน Classification มาช่วยในการทำสอบโดยเรียกใช้อัลกอริทึม NaiveBaye ดังได้อธิบายไว้ในหน้าที่ 29 โดยการพิจารณาจากจำนวนคู่ของข้อมูลที่อัลกอริทึมสามารถทำนายได้ถูกต้องในกลุ่มเดียวกันหารด้วยจำนวนคู่ของข้อมูลทั้งหมดที่ใช้ในการทำนายกลุ่มเดียวกันซึ่งค่า Precision ที่ดีที่สุดจะมีค่าเท่ากับ 1 และค่าที่แย่ที่สุดมีค่าเท่ากับ 0

$$Precision = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster} \quad (7)$$

หรือ

$$Precision = \frac{tp}{tp + fp} \quad (8)$$

Recall

ค่า Recall คือ ค่าความครบถ้วนจากการจัดกลุ่มข้อมูล หลังจากได้ผลจากการจัดกลุ่มข้อมูลเรียบร้อยแล้วจึงนำข้อมูลทุกตัวมาทำการติดป้ายเพื่อนำมาทำการทดสอบหาค่าความครบถ้วนในการจัดกลุ่มข้อมูลที่ได้จากการจัดกลุ่มข้อมูลในแต่อัลกอริทึมโดยกระบวนการทดสอบจะใช้อัลกอริทึมทางด้าน Classification มาช่วยในการทำสอบโดยเรียกใช้อัลกอริทึม NaiveBaye โดยการพิจารณาจากจำนวนคู่ของข้อมูลที่อัลกอริทึมสามารถทำนายได้ถูกต้องในกลุ่มเดียวกันหารด้วย

จำนวนคู่ข้อมูลทั้งหมดตามความเป็นจริงในกลุ่มเดียวกันซึ่งค่า Recall ที่ดีที่สุดจะมีค่าเท่ากับ 1 และค่าที่แย่ที่สุดมีค่าเท่ากับ 0

$$Recall = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsActuallyInSameCluster} \quad (9)$$

หรือ

$$Recall = \frac{tp}{tp + fn} \quad (10)$$

F-measure

ค่า F-measure คือ ค่าที่วัดหาค่าความถูกต้องในการทดสอบ ซึ่งจะพิจารณาจากทั้งค่า Precision และค่า Recall ของการวัดค่าการทดสอบ ซึ่งคำนวณได้จากจำนวนของผลลัพธ์ที่มีความถูกต้องหารด้วยจำนวนผลลัพธ์ที่ได้ และจำนวนของผลลัพธ์ที่มีความถูกต้องหารด้วยจำนวนของผลลัพธ์ที่จะให้ออกมา ค่า F-measure สามารถถูกอธิบายได้ด้วยค่าเฉลี่ยน้ำหนักของค่าความแม่นยำ (Precision) และค่าความครบถ้วน (Recall) ซึ่งค่า F-measure ที่ดีที่สุดจะมีค่าเท่ากับ 1 และค่าที่แย่ที่สุดมีค่าเท่ากับ 0

$$F = \frac{2 \cdot (precision \cdot recall)}{(precision + recall)} \quad (11)$$

$$F_\beta = \frac{(1 + \beta^2) \cdot (precision \cdot recall)}{(\beta^2 \cdot precision + recall)} \quad (12)$$

$$F_\beta = \frac{(1 + \beta^2) \cdot (true\ positive)}{((1 + \beta^2) \cdot true\ positive + \beta^2 \cdot false\ positive + false\ negative)} \quad (13)$$

**9. วิเคราะห์ผลการแบ่งกลุ่มทราฟฟิกของแต่ละอัลกอริทึมด้วยการทดสอบกับข้อมูลทราฟฟิกจริงที่ทราบความผิดปกติ (Anomaly Traffic)**

นำผลการแบ่งกลุ่มทราฟฟิกในแต่ละอัลกอริทึมที่ได้มาทดสอบกับข้อมูลทราฟฟิกจริงที่ทราบว่ามีความผิดปกติ (Anomalous) ด้วยการเก็บข้อมูลจากการทำ tcpdump และนำมาวิเคราะห์ว่าผลการจัดกลุ่มในแต่ละอัลกอริทึมให้ผลลัพธ์ที่มีความถูกต้องตรงกันมากน้อยเพียงใดดังได้แสดงไว้ในตารางที่ 17-25 และยังสามารถระบุได้ว่าผลจากการจัดกลุ่มข้อมูลกลุ่มใดมีโอกาสที่จะเป็นทราฟฟิกที่มีความผิดปกติมากที่สุดโดยการพิจารณาจากค่าร้อยละสูงสุดในแต่ละอัลกอริทึมที่ได้จากการวิเคราะห์จากคู่ของ Attribute ต่างๆ ดังแสดงในหน้าที่ 79-83

**10. นิยาม nomaly และ anomaly traffic จากผลการทดลองที่ได้**

จากผลการทดสอบความถูกต้องตรงกันของข้อมูลที่ได้จากการจัดกลุ่มกับข้อมูลทราฟฟิกจริงที่ทราบว่ามีความผิดปกติ (Anomalous) ดังแสดงไว้ในตารางที่ 17-25 ทำให้ทราบว่าในการแบ่งกลุ่มด้วยอัลกอริทึม FilteredClusterer นั้นข้อมูลในกลุ่มที่ 1 (Cluster0) คือข้อมูลที่มีลักษณะผิดปกติ ซึ่งเป็นข้อมูลที่มีหมายเลขพอร์ตต้นทาง (Source_port) เป็น 8080 และมีขนาดความยาวของแพกเก็ต (Packet_length) เป็น 759 1434 และ 1514 นอกจากนี้ยังสามารถบอกถึงหมายเลขพอร์ตปลายทางที่เป็นของข้อมูลในกลุ่มนี้ได้อีกด้วย ซึ่งหมายเลขพอร์ตปลายทางที่ได้อยู่ในช่วง 24347 ถึง 25505 สาเหตุที่พิจารณาจากการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึม FilteredClusterer นั้นก็เพราะว่าค่าร้อยละที่ได้จากการทดสอบความถูกต้องตรงกันของข้อมูลที่ได้จากการจัดกลุ่มกับข้อมูลทราฟฟิกจริงที่ทราบว่ามีความผิดปกติ (Anomalous) มีค่าสูงที่สุด แต่หากจะพิจารณาถึงอัลกอริทึมอื่นๆ ว่ามีประสิทธิภาพมากน้อยเพียงก็สามารถใช้หลักการวิเคราะห์เช่นเดียวกันกับการวิเคราะห์จากอัลกอริทึม FilteredClusterer ดังได้สรุปไว้ในตารางที่ 17-25

**11. ระบุ Atrribute ที่เหมาะสมสำหรับการจัดกลุ่มข้อมูลทราฟฟิกได้**

สำหรับ Attribute ที่เหมาะสมสำหรับการจัดกลุ่มข้อมูลทราฟฟิกซึ่งให้ผลลัพธ์ค่อนข้างดีถึงดีมากประกอบด้วยคู่ของ Attribute ระหว่าง Source_port และ Time Destination_port และ Packet_length Destination_port และ Time Packet_length และ Time และพบว่า Attribute ระหว่าง Source_port และ Packet_length จากการใช้อัลกอริทึม FilteredCluterer ให้ผลลัพธ์สูงที่สุดคือร้อย

ละ 99.05 ซึ่งเป็นค่าที่ดีที่สุด และเป็นข้อมูลที่อยู่ในกลุ่มที่ 1 (Cluster0) มีจำนวน packet เท่ากับ 6704 packets และเมื่อนำไปทดสอบกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) พบว่ามีข้อมูลที่ตรงกันถึง 6640 packets ซึ่งบ่งบอกได้ว่าข้อมูลกลุ่มนี้มีลักษณะที่บ่งบอกถึงความผิดปกติเมื่อเปรียบเทียบกับข้อมูลที่ได้จากการทำ tcpdump ดังได้แสดงไว้ในหน้าที่ 79-83

**13. อธิบายความสัมพันธ์ระหว่าง Dataset ของเครือข่ายกับการจัดกลุ่ม**

จากผลการจัดกลุ่มข้อมูลทราฟฟิกทำให้สามารถบอกความสัมพันธ์ระหว่าง DataSet ของเครือข่ายกับการจัดกลุ่มข้อมูลได้ว่า ผลจากการแบ่งกลุ่มข้อมูลสามารถระบุได้ว่าข้อมูลกลุ่มใดที่มีลักษณะผิดปกติและมีความผิดปกติอย่างไรจากการเปรียบเทียบกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) โดยสามารถดึงรายละเอียดของ packet ที่ตกอยู่ในกลุ่มที่มีความผิดปกติไม่ว่าจะเป็น Source_port Destination_port หรือ Packet_length ซึ่ง Attribute เหล่านี้มีผลต่อการจัดกลุ่มข้อมูลมากที่สุด ดังได้สรุปไว้ในหน้าที่ 79-83

**12. ส่งผลงานที่นำเสนอแก่งานประชุมทางวิชาการ**

**ตารางที่ 7** แสดงตารางการดำเนินงาน

| ลำดับ | เดือนที่ | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 1 | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | |
| 3 | | | | | | | | | | | | |
| 4 | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | |
| 7 | | | | | | | | | | | | |
| 8 | | | | | | | | | | | | |
| 9 | | | | | | | | | | | | |
| 10 | | | | | | | | | | | | |

# ผลและวิจารณ์

## ผล

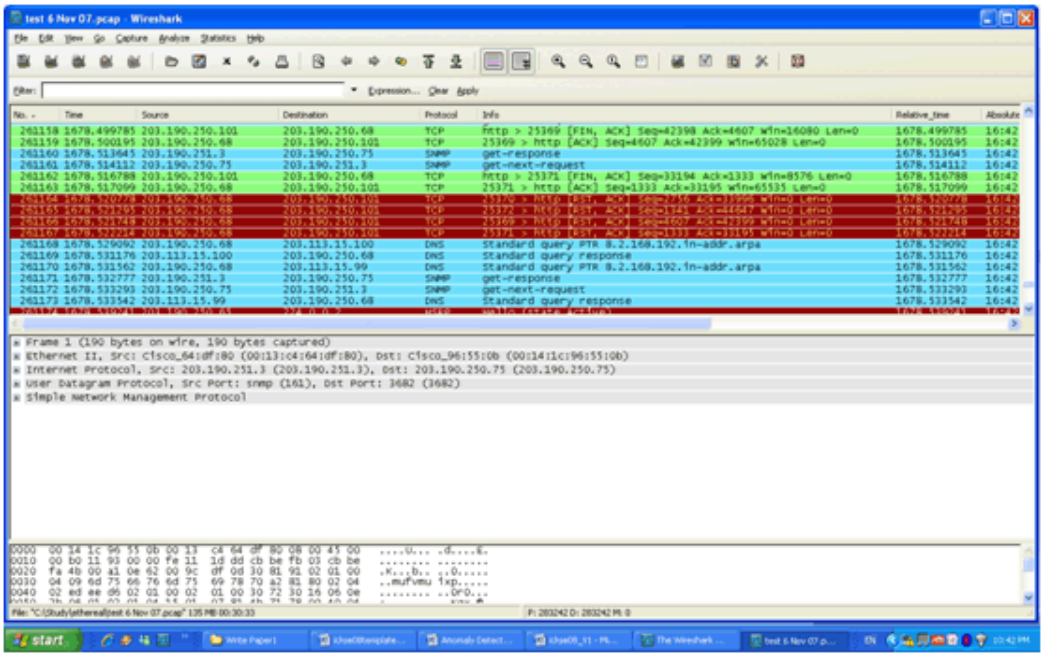
ปัญหาหนึ่งที่มักพบจากการตรวจจับทราฟฟิกแบบอัตโนมัติก็คือทำให้ได้ข้อมูลที่ไม่ถูกต้องและไม่สามารถบ่งบอกได้ว่าปัญหาเหล่านี้สาเหตุจริงๆ แล้วเกิดขึ้นจากอะไร ดังนั้นแนวทางหนึ่งที่จะสามารถช่วยลดปัญหาเหล่านี้ได้จึงต้องใช้การเรียนรู้จากหลายๆ รูปแบบที่แตกต่างกันด้วยกระบวนวิธีทางด้านการทำเหมืองข้อมูลเข้ามาช่วยซึ่งเป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) จากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง การรู้จำแบบ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ โดยงานวิจัยนี้เลือกใช้วิธีการจัดกลุ่มข้อมูลเพื่อแบ่งข้อมูลที่มีลักษณะคล้ายกันออกเป็นกลุ่ม โดยจะใช้อัลกอริทึม 5 อัลกอริทึมมาช่วยในการจัดกลุ่มข้อมูลเพื่อมาเปรียบเทียบหาอัลกอริทึมที่มีประสทธิภาพที่สุด อัลกอริทึมทั้ง 5 ประกอบด้วย K-means sIB FilterdClusterer RandomFlatClustering และ Fasthestfirst

### การเก็บรวบรวมข้อมูล

ผู้วิจัยได้เลือกโปรแกรม Ethereal มาใช้ในการเก็บรวบรวมข้อมูล เนื่องจากเป็นโปรแกรมที่มีการใช้งานหลากหลายและง่ายต่อการใช้งาน นอกจากนี้ยังจะช่วยให้ได้ข้อมูลที่มีความหลากหลายและมีความครอบคลุมของประเภทของข้อมูลได้มากขึ้น ลักษณะข้อมูลที่เก็บได้ด้วยโปรแกรม Ethereal เป็นดังภาพที่ 12 รายละเอียดของ Packet ที่สามารถเก็บได้ประกอบด้วย 8 Attribute ดังนี้ Source address Destination address Protocol Time Packet_length Source Port Destination Port และ Cumulative byte ซึ่งแต่ละ Attribute มีรายละเอียดอธิบายได้ดังนี้

- Source address บอกถึงที่อยู่ต้นทางในการส่งข้อมูล
- Destination address บอกถึงที่อยู่ปลายทางในการส่งข้อมูล
- Protocol บอกถึงรูปแบบที่ใช้ในการสื่อสารกันระหว่างต้นทางกับปลายทาง
- Time บอกถึงเวลา ณ ขณะที่เริ่มมีการส่งข้อมูลระหว่างต้นทางกับปลายทาง
- Packet_length บอกถึงขนาดความยาวของ Packet
- Source Port บอกถึงหมายเลข Port ต้นทาง
- Destination Port บอกถึงหมายเลข Port ปลายทาง

- Cumulative byte บอกถึงขนาดความยาวสะสมของ Packet เป็น Byte



**ภาพที่ 12** แสดงตัวอย่างการเก็บข้อมูลด้วยโปรแกรม Ethereal

ข้อมูลที่เก็บได้ทั้งหมดสามารถสรุปได้ดังตารางที่ 1

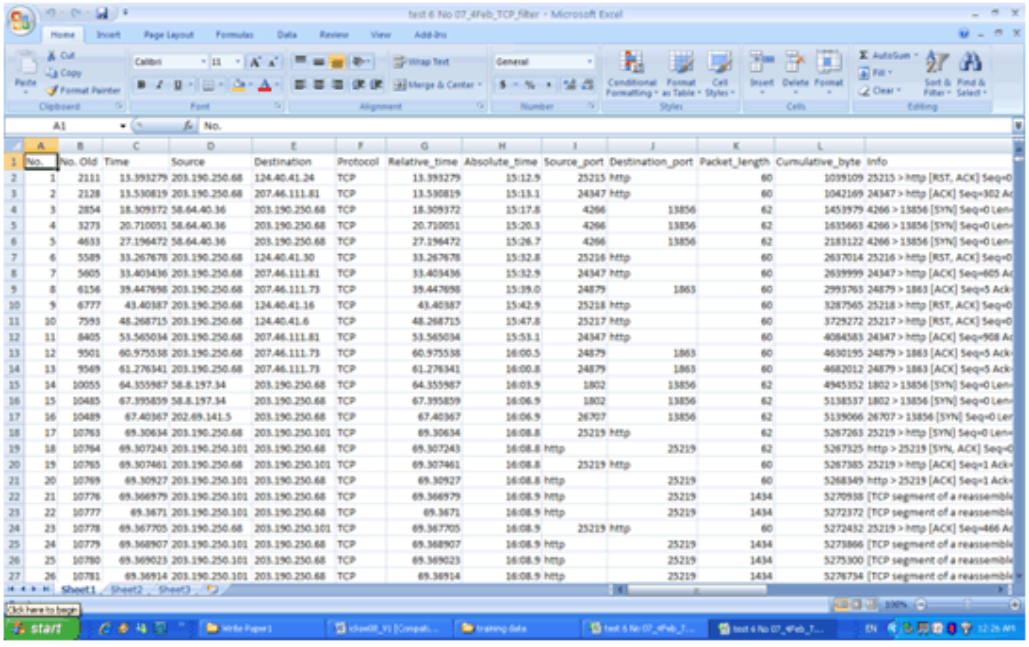
**ตารางที่ 8** แสดงผลสรุปการเก็บข้อมูลทราฟฟิกโดยแยกตามโปรโตคอล

| Protocol | #Packets | #Bytes | %Packets of total | #Bytes of total |
|---|---|---|---|---|
| TCP | 16,899 | 10,359,675 | 1.612 | 2.476 |
| UDP | 263,366 | 391,461,363 | 25.117 | 93.547 |
| ICMP | 20,480 | 2,267,014 | 1.953 | 0.542 |
| DNS | 52,710 | 12,760,619 | 5.027 | 3.049 |
| IP | 20,561 | 1,233,660 | 1.961 | 0.295 |
| Others | 674,559 | 384,724 | 64.331 | 0.092 |
| Total | 1,048,575 | 418,467,055 | 100 | 100 |

ตารางที่ 8 แสดงผลสรุปจากการเก็บข้อมูลเป็นเวลา 24 ชั่วโมง โดยแยกตามลักษณะโปรโตคอลที่ใช้ในการสื่อสารระหว่างกันประกอบด้วยโปรโตคอล TCP UDP ICMP DSN IP และอื่นๆ (SNMP, ARP, HSRP, NTP, SSH) ซึ่งงานวิจัยนี้เลือกที่จะนำ Packet ที่ใช้โปรโตคอล TCP/IP มาทำการทดลอง โดยหลังจากได้ข้อมูลทั้งหมดมาแล้วต่อไปจึงนำเข้ากระบวนการเตรียมข้อมูลก่อนจะนำเข้าอัลกอริทึม

**การเตรียมข้อมูล**

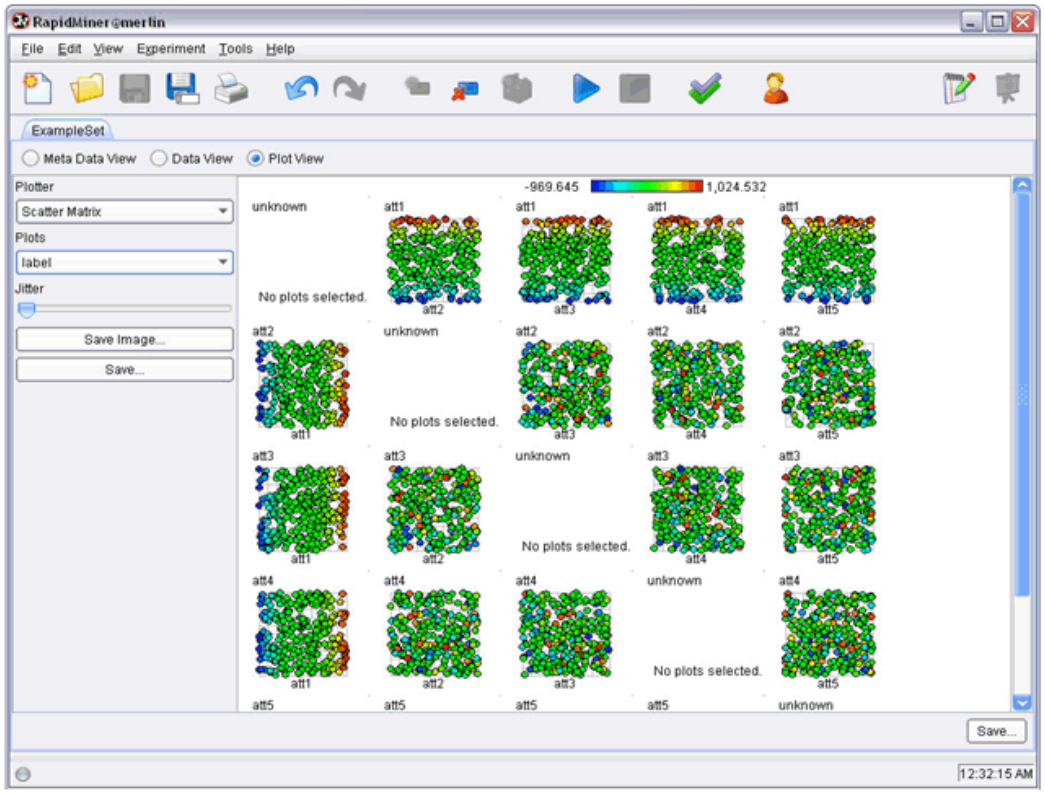
หลังจากได้ข้อมูลที่ต้องการเรียบร้อยแล้วโดยการเลือกทำการทดลอง TCP/IP Packet จากนั้นจึงทำการคัดเลือกข้อมูลที่ไม่สมบูรณ์ออก หรือหากบางข้อมูลมีลักษณะข้อมูลที่แสดงไม่เหมือนกันก็ทำการแปลงข้อมูลนั้นเพื่อให้สามารถนำไปเข้ากระบวนการทางด้าน Data mining ต่อไปได้อย่างมีประสิทธิภาพยิ่งขึ้น ข้อมูลที่ได้จากการเก็บรวบรวมด้วยโปรแกรม Ethereal จะถูกแปลงให้อยู่ในรูปของ CSV ไฟล์ เพื่อให้ง่ายและสะดวกต่อการใช้งาน หรือการแปลงไฟล์ก่อนนำไปใช้กับ Tool อื่นๆ ต่อไป ลักษณะของข้อมูลที่ได้ทำการคัดเลือกและแปลงข้อมูลเรียบร้อยแล้วแสดงได้ดังภาพที่ 13



**ภาพที่ 13** แสดงการเก็บรวบรวมข้อมูลในรูปแบบ CSV file

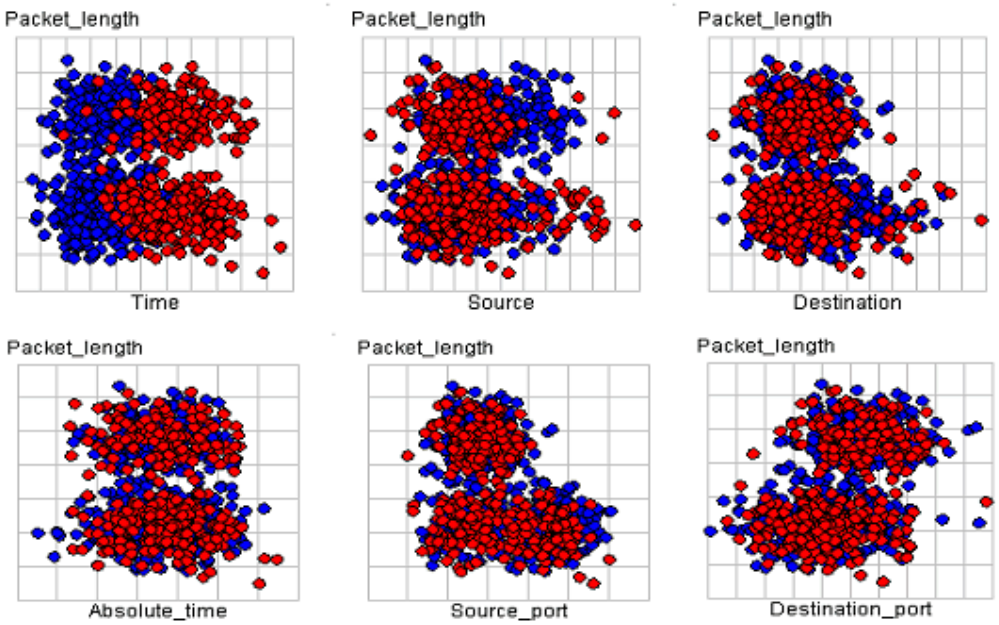
**ซอฟต์แวร์การทำเหมืองข้อมูล**

ในงานวิจัยนี้ได้ค้นพบการใช้วิธีทาง Machine learning ที่เรียกว่า การจัดกลุ่มข้อมูลเพื่อแยกประเภทของทราฟฟิกโดยใช้ข้อมูลสถิติทางด้านการจราจรเครือข่ายในระดับ Transport layer โดยเลือกโปรโตคอล TCP เพื่อศึกษาลักษณะของความผิดปกติของทราฟฟิก เนื่องจากข้อมูลที่ใช้โปรโตคอล TCP มีปริมาณทราฟฟิกจำนวนมากเมื่อเทียบกับโปรโตคอลอื่นๆ และเลือกใช้ซอฟต์แวร์ RapidMiner มาช่วยในการวิเคราะห์ข้อมูล เนื่องจาก RapidMiner มีอัลกอริทึมมากกว่า 400 อัลกอริทึม และรองรับการใช้งานกับไฟล์ชนิดต่างๆ ได้หลากหลายและง่ายต่อการใช้งาน



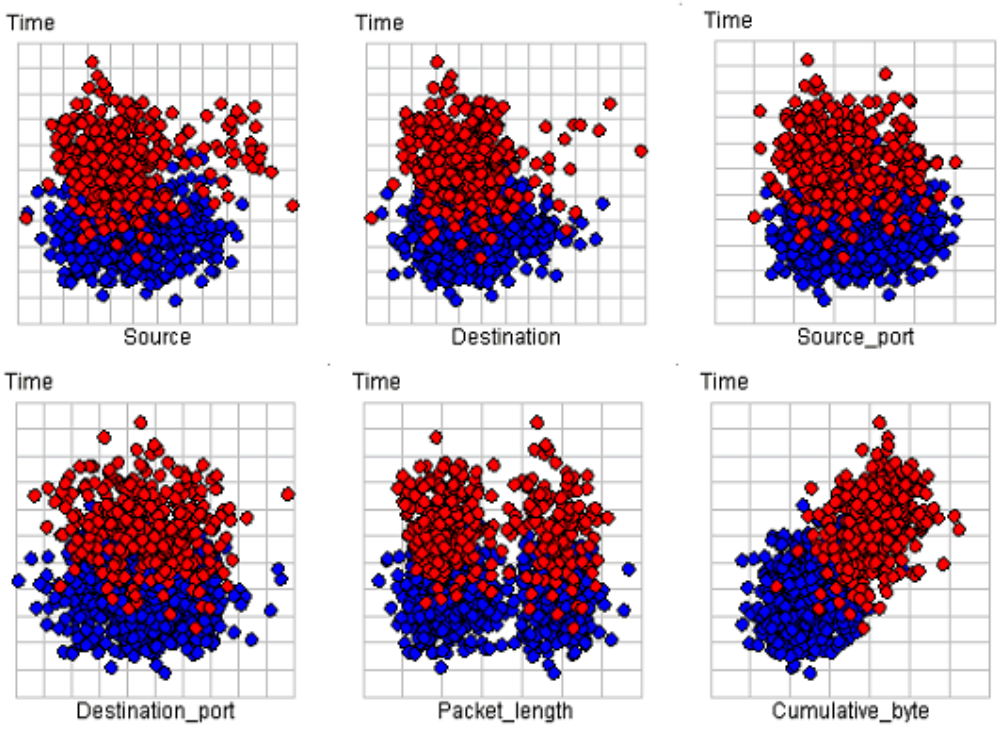

**ภาพที่ 14** ตัวอย่างโปรแกรม RapidMiner

**ผลการแบ่งกลุ่มด้วย K-means**



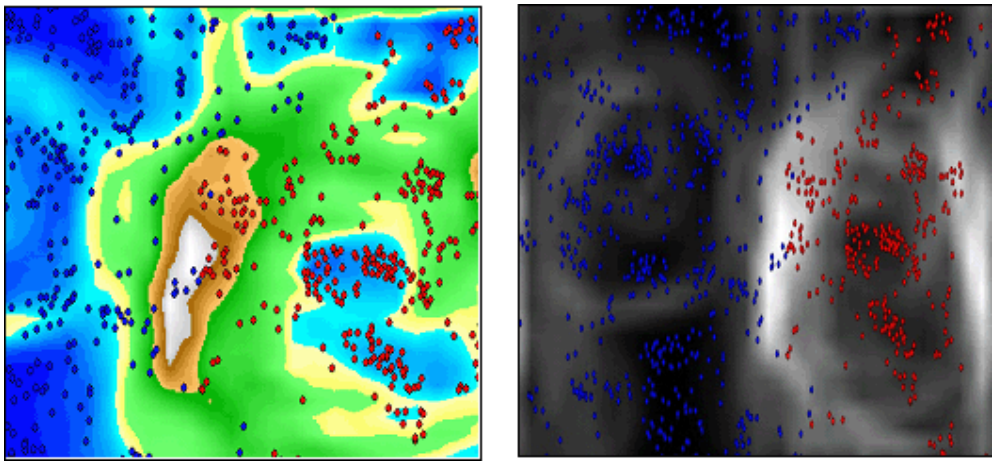

**ภาพที่ 15** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม K-means ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Packet_length และ Attribute อื่นๆ ที่เหลือ

จากภาพที่ 15 แสดงการพล๊อตจุดข้อมูลใน 2 มิติ ระหว่าง attribute ที่ชื่อ Packet_length และทุกๆ attribute ที่เก็บได้จากโปรแกรม Ethereal ข้อมูลแต่ละจุดแสดงถึงข้อมูลทราฟฟิก 1 packet ซึ่งในการทดลองเราจะตั้งสมมุติฐานไว้ว่าข้อมูลที่ใช้ในการทดลองจะมีเพียง 2 กลุ่มเท่านั้นคือ กลุ่มแรกเป็นข้อมูลกลุ่มที่ปกติ และกลุ่มที่สองเป็นกลุ่มที่มีความผิดปกติ ซึ่งได้แสดงด้วยสี 2 สีคือ สีแดง และสีน้ำเงิน ลักษณะการกระจายตัวของข้อมูลจากการใช้อัลกอริทึม K-means จะเห็นว่าจากการ plot ข้อมูลระหว่าง Packet_length และ time เห็นการแบ่งกลุ่มข้อมูลชัดเจนที่สุด แต่ก็ยังมีข้อมูลบางตัวที่มีการกระจายปะปนเข้าไปยังอีกกลุ่มหนึ่ง ซึ่งเบื้องต้นสามารถบอกได้ว่า Packet_length และ time เป็น attribute ที่น่าสนใจที่จะนำมาวิเคราะห์ผลเป็นอย่างมากเนื่องจากผลการจัดกลุ่มมีความชัดเจนมากที่สุดเมื่อเปรียบเทียบกับ attribute อื่นๆ



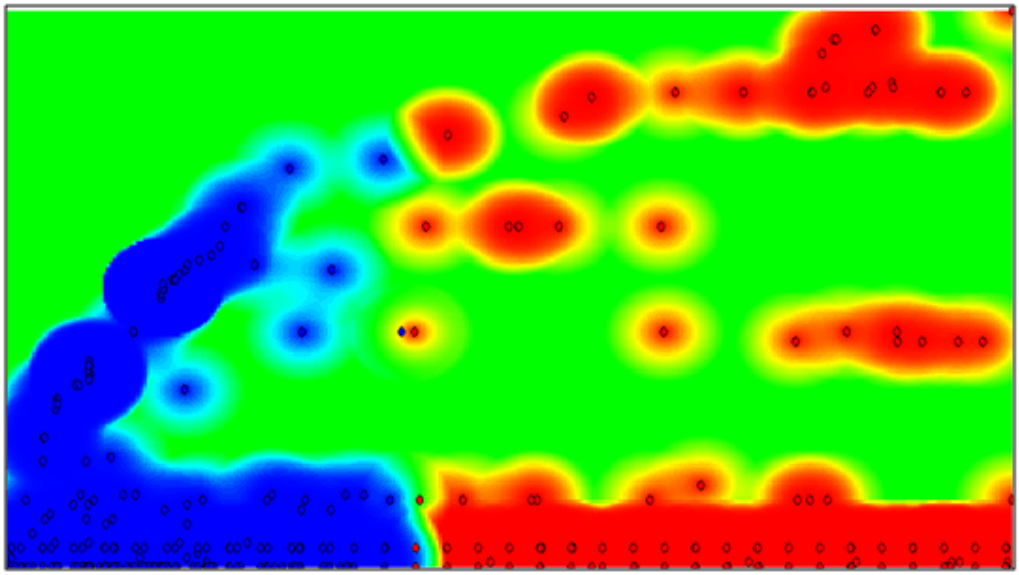

**ภาพที่ 16** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม K-means ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ

ภาพที่ 16 แสดงการกระจายตัวของข้อมูลด้วยการ plot แบบวิธี Self-Organizing Map (SOM) โดยใช้ Kohonen net เพื่อลดมิติของการมอง ซึ่งเป็นการ plot ระหว่าง Attribute Time และ Packet_length ผลจากการ Plot ทำให้เห็นกลุ่มของข้อมูลชัดเจนขึ้น ซึ่งสีแสดงถึงกลุ่มของข้อมูล และทิศทางของจุดข้อมูลแต่จะจุดคือ Attribute ที่ใช้ในการ plot



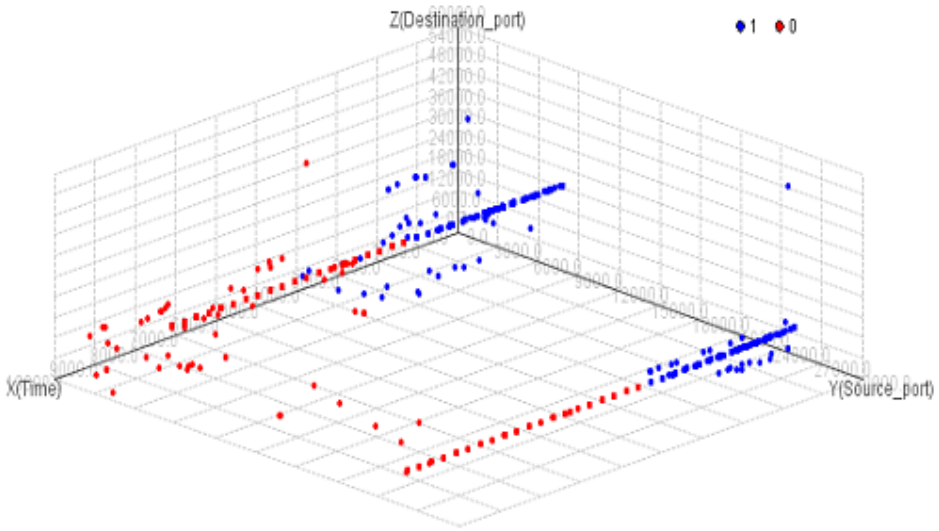
(a) (b)

**ภาพที่ 17** แสดงการกระจายตัวของข้อมูลด้วยการ plot แบบวิธี Self-Organizing Map (SOM)
(a) Gray scale, (b) Landscape



**ภาพที่ 18** แสดงการกระจายตัวของข้อมูลด้วยการ plot แบบวิธี SOM แบบ 2 ทิศทาง

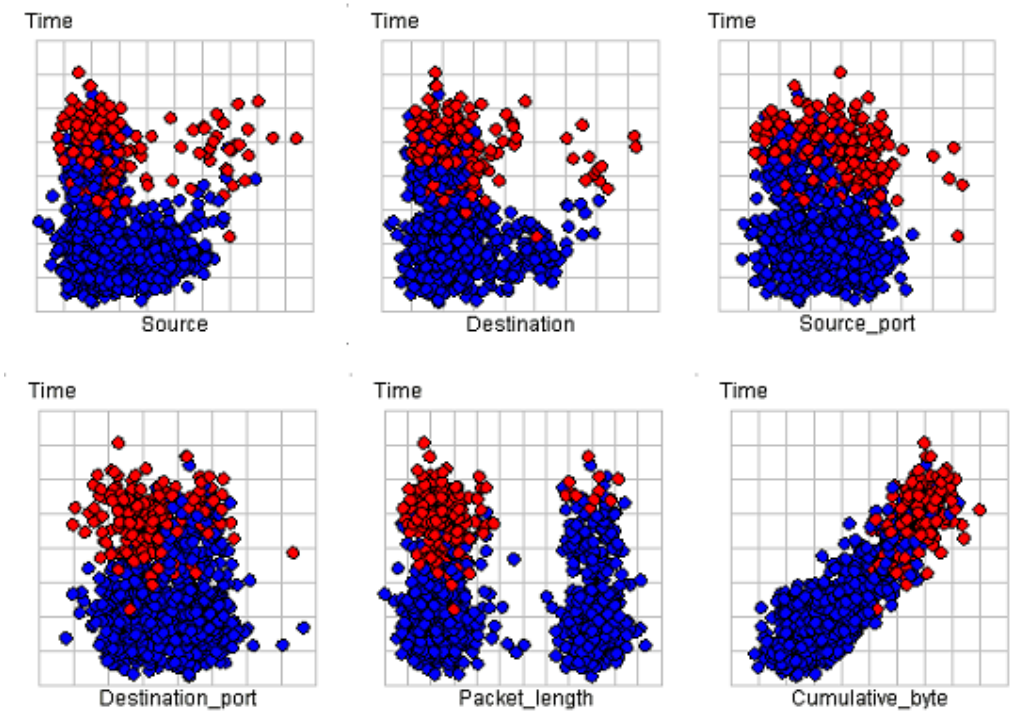
นอกจากนี้ยังสามารถทำการ plot ในรูปแบบ 3 มิติได้ดังแสดงในภาพที่ 19 สีของแต่ละจุดข้อมูลแสดงกลุ่มของข้อมูล ซึ่งเป็นการ plot ระหว่าง 3 attribute คือ แกน X แสดงแกนของ Time แกน Y แสดงแกนของ Source_port และแกน Z แสดงแกนของ Destination_port



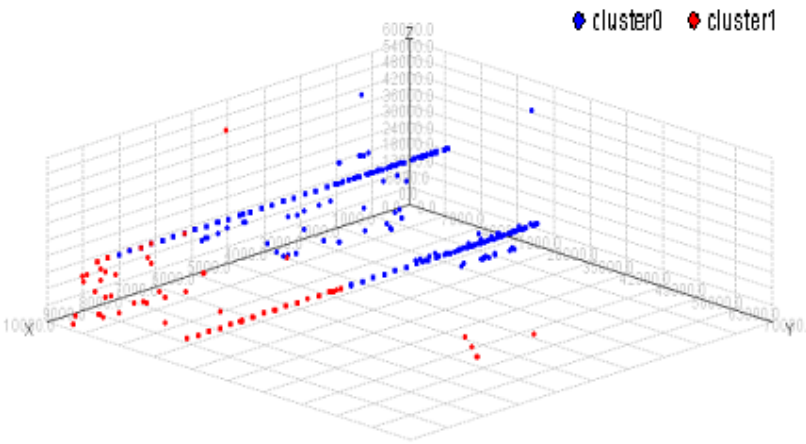

**ภาพที่ 19** แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port

ภาพที่ 19 เป็นการ plot การกระจายของข้อมูลให้อยู่ในรูปแบบ 3 มิติ จากรูปเป็นการ plot ระหว่าง Time Source_port และ Destination_port ในระนาบแกน X Y และ Z ตามลำดับซึ่งกลุ่มที่จัดได้แสดงด้วยจุดข้อมูลสีแดงและสีน้ำเงิน

**ผลการแบ่งกลุ่มด้วย FarthestFirst**



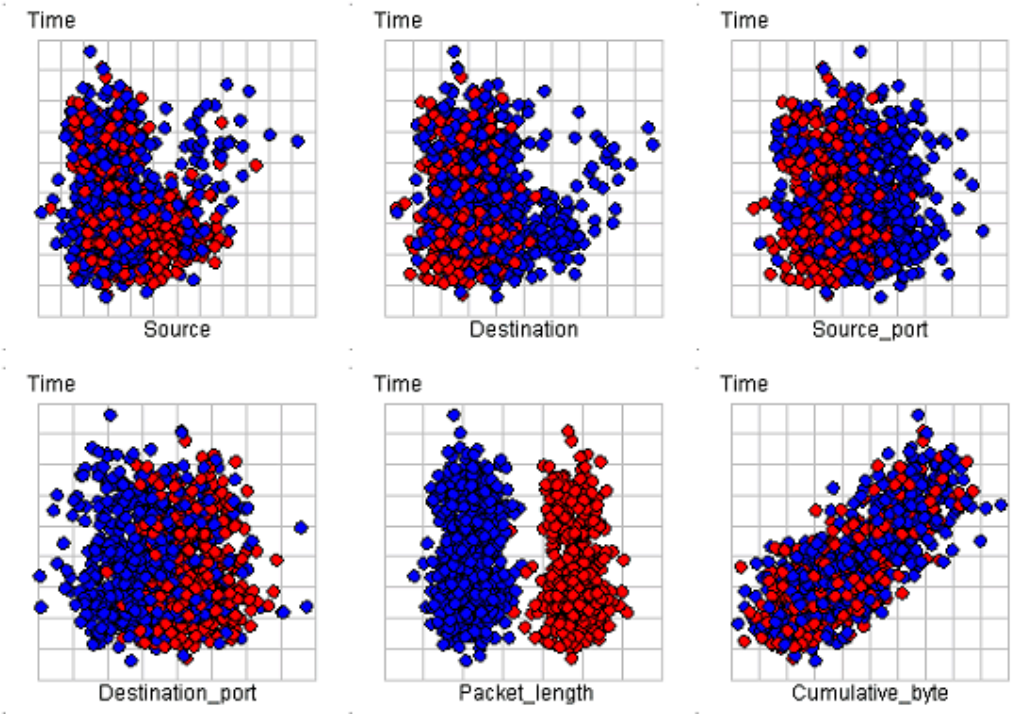

**ภาพที่ 20** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม FarthestFirst ด้วยการ plot ข้อมูลในแบบ 2 มิติ ระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ



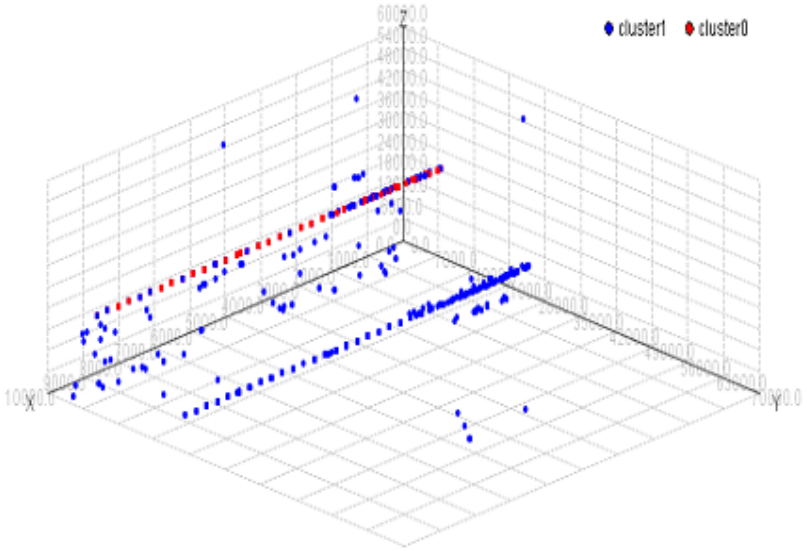

**ภาพที่ 21** แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port

ผลการจัดกลุ่มด้วยอัลกอริทึม FarthestFirst ทำให้ได้ข้อมูลที่อยู่ใน cluster#0 จำนวน 2,746 ข้อมูล ซึ่งคิดเป็น 16.25% และข้อมูลที่อยู่ใน cluster#1 จำนวน 14,154 ข้อมูล ซึ่งคิดเป็น 83.75% ซึ่งจากภาพที่ 21 จะเห็นว่ามีข้อมูลบางข้อมูลยังกระจัดกระจายอยู่

**ผลการแบ่งกลุ่มด้วย FilteredClusterer**



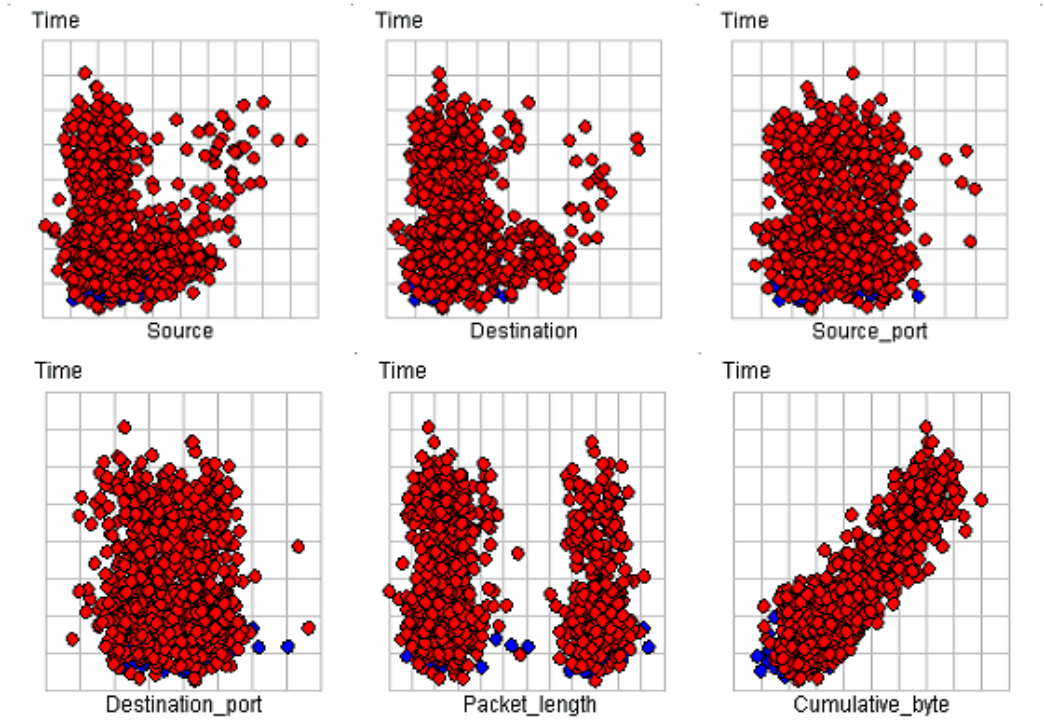

**ภาพที่ 22** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม FilteredClusterer ด้วยการ plot ข้อมูลในแบบ 2 มิติ ระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ



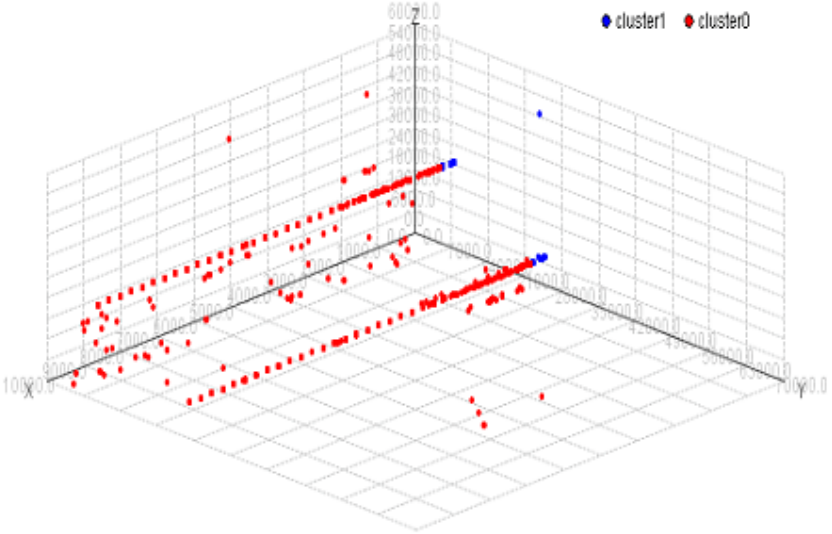

**ภาพที่ 23** แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port

ผลการจัดกลุ่มด้วยอัลกอริทึม FilteredClusterer ทำให้ได้ข้อมูลที่อยู่ใน cluster#0 จำนวน 6,704 ข้อมูล ซึ่งคิดเป็น 39.67% และข้อมูลที่อยู่ใน cluster#1 จำนวน 10,195 ข้อมูล ซึ่งคิดเป็น 60.33% ซึ่งจากภาพที่ 23 จะเห็นว่าในการ plot ระหว่าง attribute Time และ Packet_length นั้นมีการจัดกลุ่มเกือบจะสมบูรณ์ ทำให้มองเห็นการจัดกลุ่มค่อนข้างชัดเจน

**ผลการแบ่งกลุ่มด้วย Sequential IB Clustering (sIB)**



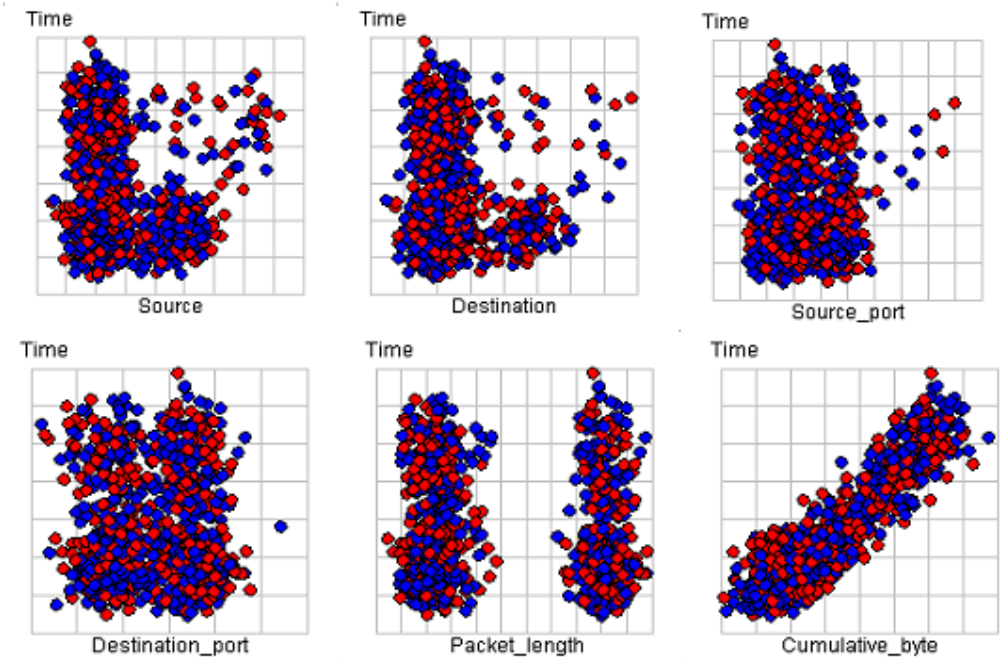

**ภาพที่ 24** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม Sequential IB Clustering (sIB) ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ



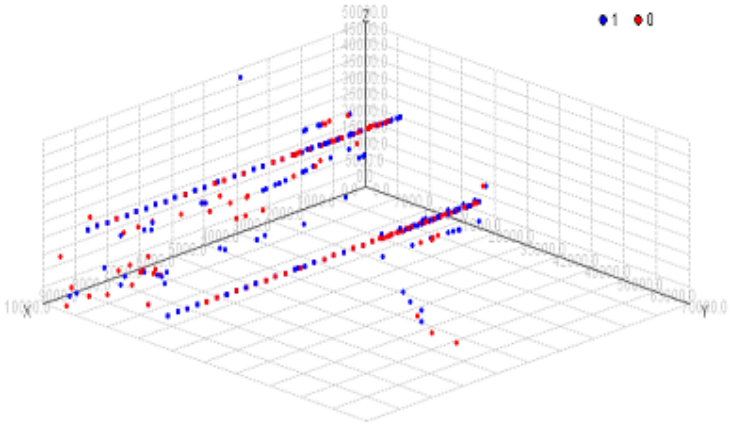

**ภาพที่ 25** แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port

ผลการจัดกลุ่มด้วยอัลกอริทึม Sequential IB Clustering (sIB) ทำให้ได้ข้อมูลที่อยู่ใน cluster#0 จำนวน 986 ข้อมูล ซึ่งคิดเป็น 5.83% และข้อมูลที่อยู่ใน cluster#1 จำนวน 15,913 ข้อมูล ซึ่งคิดเป็น 94.17% ซึ่งจากภาพที่ 25 จะเห็นว่าในการ plot ระหว่าง attribute ต่างๆนั้นแทบไม่เห็นการแบ่งกลุ่มข้อมูลเลย ทำให้ไม่สามารถนำผลจากการจัดกลุ่มด้วยอัลกอริทึมนี้มาใช้งานได้มากนัก

**ผลการแบ่งกลุ่มด้วย RandomFlatClustering**




**ภาพที่ 26** แสดงผลการจัดกลุ่มด้วยอัลกอริทึม RandomFlatClustering ด้วยการ plot ข้อมูลในแบบ 2 มิติระหว่าง Attribute Time และ Attribute อื่นๆ ที่เหลือ




**ภาพที่ 27** แสดงการ plot ในรูปแบบ 3 มิติ ระหว่าง 3 attribute คือ Time Source_port และ Destination_port

ผลการจัดกลุ่มด้วยอัลกอริทึม RandomFlatClustering ทำให้ได้ข้อมูลที่อยู่ใน cluster#0 จำนวน 8,509 ข้อมูล ซึ่งคิดเป็น 50.35% และข้อมูลที่อยู่ใน cluster#1 จำนวน 8,390 ข้อมูล ซึ่งคิดเป็น 49.65% ซึ่งจากภาพที่ 27 จะเห็นว่าในการ plot ระหว่าง attribute ต่างๆนั้นไม่เห็นการแบ่งกลุ่มข้อมูลเลย การกระจายตัวของข้อมูลมีการกระจายตัวแบบไม่เป็นระเบียบ ทำให้ไม่สามารถนำผลจากการจัดกลุ่มด้วยอัลกอริทึมนี้มาใช้งานได้

**ตารางที่ 9** เปรียบเทียบผลการจัดกลุ่มด้วยอัลกอริทึมทั้ง 5 อัลกอริทึม

| Algorithms | #items of Cluster0 | #items of Cluster1 |
|---|---|---|
| **K-means** | 5755 | 11144 |
| **W-sIB** | 986 | 15913 |
| **FilteredClusterer** | 6704 | 10195 |
| **FarthestFirst** | 2745 | 14154 |
| **RadomFlatClustering** | 8509 | 8390 |

**การวัดและประเมิณผล**

**ตารางที่ 10** การสรุป term context

| | **Correct result/ classification** | |
|---|---|---|
| **Obtained result/ classification** | tp (true positive) | fp (false positive) |
| | fn (false negative) | tn (true negative) |

ความถูกต้องของข้อมูลสามารถคำนวณได้จาก

$$Precision = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsPredictedInSameCluster} \quad (14)$$

หรือ

$$Precision = \frac{tp}{tp + fp} \quad (15)$$

**ตารางที่ 11** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision ของข้อมูลกลุ่มที่ 1

| Algorithms | #items of Cluster0 | Tp | tp+fp | Precision |
|---|---|---|---|---|
| **K-means** | 5775 | 5775 | 5775 | 1 |
| **W-sIB** | 986 | 986 | 986 | 1 |
| **FilteredClusterer** | 6704 | 6704 | 6704 | 1 |
| **FarthestFirst** | 2745 | 2739 | 2745 | 0.998 |
| **RadomFlatClustering** | 8509 | 4314 | 8509 | 0.507 |

**ตารางที่ 12** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision ของข้อมูลกลุ่มที่ 2

| Algorithms | #items of Cluster1 | tp | tp+fn | Precision |
|---|---|---|---|---|
| **K-means** | 11144 | 11144 | 11144 | 1 |
| **W-sIB** | 15913 | 15913 | 15913 | 1 |
| **FilteredClusterer** | 10195 | 10195 | 10195 | 1 |
| **FarthestFirst** | 14154 | 14125 | 14154 | 0.998 |
| **RadomFlatClustering** | 8390 | 4253 | 8390 | 0.507 |

$$Recall = \frac{\#PairsCorrectlyPredictedInSameCluster}{\#TotalPairsActuallyInSameCluster} \quad (16)$$

หรือ

$$Recall = \frac{tp}{tp + fn} \quad (17)$$

**ตารางที่ 13** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Recall ของข้อมูลกลุ่มที่ 1

| Algorihtm | # items of Cluster0 | tp | tp + fn | Recall |
|---|---|---|---|---|
| **K-means** | 5755 | 5611 | 5755 | 0.975 |
| **W-sIB** | 986 | 966 | 986 | 0.98 |
| **FilteredCluterer** | 6704 | 6690 | 6704 | 0.998 |
| **FarthestFirst** | 2745 | 2462 | 2745 | 0.897 |
| **RandomFlatClustering** | 8509 | 4884 | 8509 | 0.574 |

**ตารางที่ 14** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Recall ของข้อมูลกลุ่มที่ 2

| Algorithms | # items of Cluster1 | tp | tp + fn | Recall |
|---|---|---|---|---|
| **K-means** | 11144 | 11144 | 11144 | 1 |
| **W-sIB** | 15913 | 15913 | 15913 | 1 |
| **FilteredCluterer** | 10195 | 10195 | 10195 | 1 |
| **FarthestFirst** | 14154 | 14026 | 14154 | 0.991 |
| **RandomFlatClustering** | 8390 | 3632 | 8390 | 0.433 |

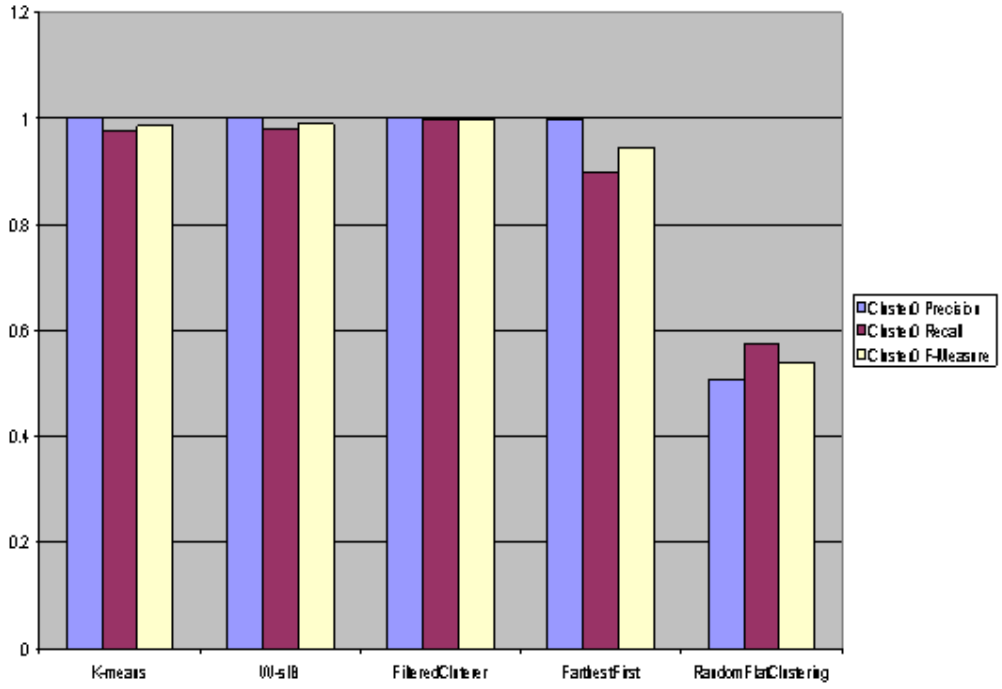
ค่า F-measure จะถูกกำหนดโดยค่า Precision และค่า Recall ซึ่งวัดค่าจากการคืนค้นข้อมูลที่มีการปรับให้เหมาะกับการวัดและประเมินผลการจัดกลุ่ม โดยพิจารณาจากคู่ของจุดทุกๆ คู่ ซึ่งไม่มีกฎเกณฑ์ชัดเจน การตัดสินใจจัดกลุ่มของคู่เหล่านี้ไปยังกลุ่มเดียวกันหรือกลุ่มที่แตกต่างกันถูกพิจารณาจากความถูกต้องถ้ามีความตรงกันภายใต้การทำเครื่องหมาย (Label) ของแต่ละจุด ในทางสถิติค่า F-measure หรือค่า F-score ไว้สำหรับวัดค่าความถูกต้องในการทดสอบ ซึ่งจะพิจารณาจากทั้งค่า Precision และค่า Recall ของการวัดค่าการทดสอบ ซึ่งคำนวณได้จากจำนวนของผลลัพธ์ที่มีความถูกต้องหารด้วยจำนวนผลลัพธ์ที่ได้ และจำนวนของผลลัพธ์ที่มีความถูกต้องหารด้วยจำนวนของผลลัพธ์ที่จะให้ออกมา ค่า F-measure สามารถถูกอธิบายได้ด้วยค่าเฉลี่ยน้ำหนักของค่าความ

แม่นยำ (Precision) และค่าความครบถ้วน (Recall) ซึ่งค่า F-measure ที่ดีที่สุดจะมีค่าเท่ากับ 1 และค่าที่แย่ที่สุดมีค่าเท่ากับ 0 ซึ่งสมการการคำนวณค่า F-measure สามารถคำนวณได้ดังนี้

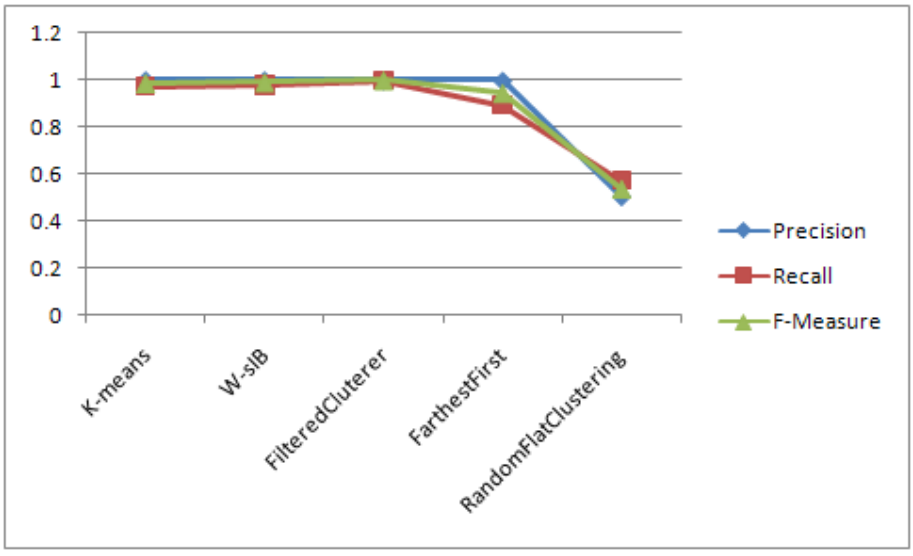
$$F = \frac{2 \cdot (precision \cdot recall)}{(precision + recall)} \quad (18)$$

**ตารางที่ 15** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision Recall และ F-Measure ของข้อมูลกลุ่มที่ 1

| Algorithms | # items of Cluster0 | precision | Recall | F-Measure |
|---|---|---|---|---|
| **K-means** | 5755 | 1 | 0.975 | 0.987 |
| **W-sIB** | 986 | 1 | 0.98 | 0.99 |
| **FilteredCluterer** | 6704 | 1 | 0.998 | 0.999 |
| **FarthestFirst** | 2745 | 0.998 | 0.897 | 0.945 |
| **RandomFlatClustering** | 8509 | 0.507 | 0.574 | 0.539 |



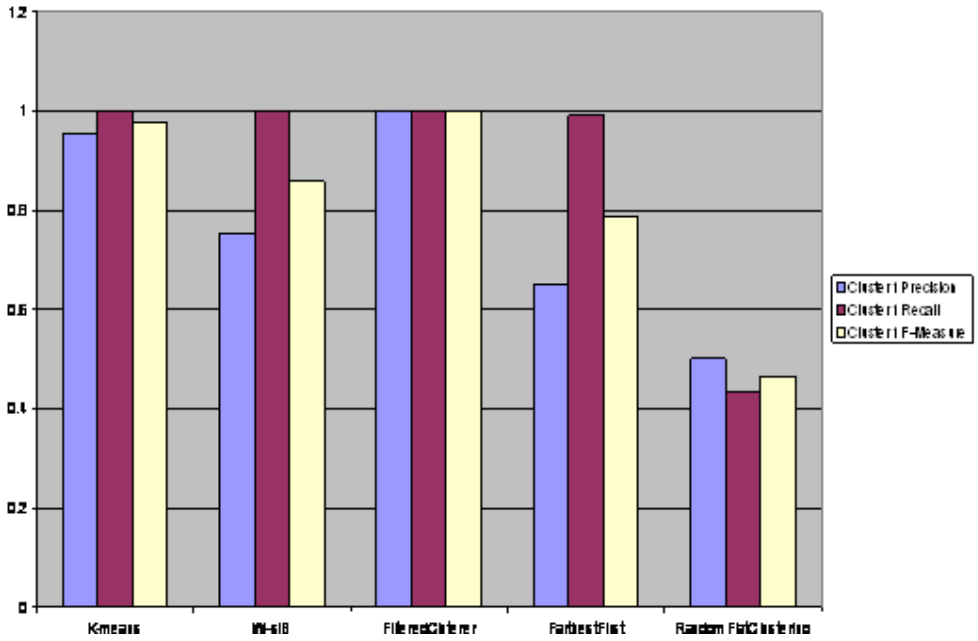

**ภาพที่ 28** แสดงกราฟแท่งเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 1 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม



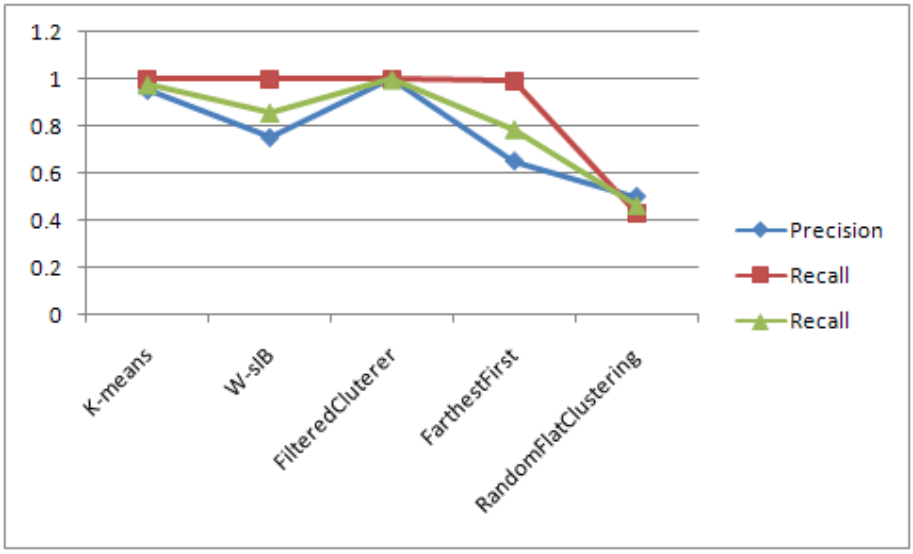

**ภาพที่ 29** แสดงกราฟเส้นเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 1 ด้วยค่า Precision Recall และ F-measure

**ตารางที่ 16** ผลสรุปการวัดค่าความถูกต้องของการจัดกลุ่มด้วยค่า Precision Recall และ F-Measure ของข้อมูลกลุ่มที่ 2

| Algorithms | # items of Cluster0 | precision | Recall | F-Measure |
|---|---|---|---|---|
| **K-means** | 11144 | 0.953 | 1 | 0.976 |
| **W-sIB** | 15913 | 0.752 | 1 | 0.858 |
| **FilteredCluterer** | 10195 | 0.999 | 1 | 0.999 |
| **FarthestFirst** | 14154 | 0.652 | 0.991 | 0.787 |
| **RandomFlatClustering** | 8390 | 0.501 | 0.433 | 0.464 |




**ภาพที่ 30** กราฟแท่งเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 2 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม




**ภาพที่ 31** แสดงกราฟเส้นเพื่อเปรียบเทียบผลการวัดค่าความถูกต้องของกลุ่มข้อมูลที่ 2 ด้วยค่า Precision Recall และ F-measure โดยเปรียบเทียบผลการจัดกลุ่มของแต่ละอัลกอริทึม

สำหรับรูปแบบทั่วไปของสมการ F-measure สำหรับจำนวนจริงที่ไม่เป็นค่าลบ (β) สามารถเขียนได้ดังสมการ

$$F_\beta = \frac{(1+\beta^2)\cdot(precision\cdot recall)}{(\beta^2\cdot precision + recall)} \tag{19}$$

สมการการคำนวณค่า F-measure ในรูป term context สามารถเขียนได้ดังนี้

$$F_\beta = \frac{(1+\beta^2)\cdot(true\ positive)}{((1+\beta^2)\cdot true\ positive+\beta^2\cdot false\ positive+false\ negative)} \tag{20}$$

**การประเมินผลและวัดประสิทธิภาพการจัดกลุ่มของแต่ละอัลกอริทึมด้วยการทดสอบกับข้อมูลทราฟฟิกจริง**

หลังจากที่ได้ผลลัพธ์จากการจัดกลุ่มของแต่ละอัลกอริทึมเรียบร้อยแล้ว ขั้นตอนนี้จะเป็นการนำผลลัพธ์ที่ได้ไปทดสอบกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) จากการเก็บข้อมูลโดยใช้การทำ tcpdump ไปยังเครื่องคอมพิวเตอร์ที่ต้องการจับทราฟฟิกและนำผลลัพธ์

จากการจัดกลุ่มที่ได้มาเปรียบเทียบกับทราฟฟิกที่เก็บได้จากการทำ tcpdump นี้ว่ามีลักษณะของ Packet ใดที่ตรงกันบ้างและตรงกันคิดเป็นกี่เปอร์เซ็นต์ด้วยการพิจารณาความตรงกันในส่วนของ Attribute ต่างๆ ดังนี้ Source_port Destination_port Packet_length และ Time โดยการพิจารณาตั้งแต่ 2 Attribute ขึ้นไปดังได้แสดงรายละเอียดไว้ในตารางที่ 17-25 และทำการวิเคราะห์ด้วยว่าผลจากการเปรียบเทียบที่ได้จะทำให้สามารถบ่งบอกได้ว่าข้อมูลที่อยู่ในกลุ่มใดมีลักษณะที่เข้าข่ายว่าเป็นทราฟฟิกที่มีความผิดปกติ ผลลัพธ์จากการจัดกลุ่มด้วยอัลกอริทึมใดที่มีประสิทธิภาพมากที่สุด และ Attribute ใดมีผลต่อการจัดกลุ่มมากที่สุด

จากตารางที่ 17-25 อธิบายได้ดังนี้ จากตารางทั้งหมดนี้ได้แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่ม (Cluster0 และ Cluster1) และอัลกอริทึม (K-means W-sIB FilteredCluterer FarthestFirst และ RandomFlatClustering) โดยพิจารณาจาก Attribute ตั้งแต่ 2 Attribute ขึ้นไป

โดยการนับจำนวน Packet ที่ตรงกับ Anomaly Traffic ที่ได้จากการทำ tcpdump ทั้งหมด 54934 packets และคำนวณคิดเป็นร้อยละเพื่อให้ง่ายต่อการเปรียบเทียบ อัลกอริทึมใดที่มีจำนวนร้อยละสูงที่สุดแสดงว่าให้ผลลัพธ์ในการจัดกลุ่มถูกต้องตรงกับทราฟฟิกจริงมากที่สุดถือว่าเป็นอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลประเภทเครือข่ายทราฟฟิกมากที่สุด และการจับคู่ระหว่าง Attribute ใดๆ ให้ค่าร้อยละโดยรวมสูงที่สุดแสดงว่า Attribute มีผลต่อการจัดกลุ่มมากที่สุด ดังแสดงได้ตามตารางที่ 17-25 ดังนี้

**ตารางที่ 17** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Destination_port

| **Algorithm** | **Cluster0** | | | **Cluster1** | | |
|---|---|---|---|---|---|---|
| | **#packet** | **#equal** | **%** | **#packet** | **#equal** | **%** |
| K-means | 5755 | 24 | 0.42 | 11144 | 13 | 0.12 |
| W-sIB | 986 | 0 | 0 | 15913 | 37 | 0.23 |
| FilteredCluterer | 6704 | 0 | 0 | 10195 | 37 | 0.36 |
| FarthestFirst | 2745 | 19 | 0.69 | 14154 | 18 | 0.13 |
| RandomFlatClustering | 8509 | 21 | 0.25 | 8390 | 16 | 0.19 |

**ตารางที่ 18** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Packet_length

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 3093 | 53.74 | 11144 | 5644 | 50.65 |
| W-sIB | 986 | 451 | 45.74 | 15913 | 8286 | 52.07 |
| FilteredCluterer | 6704 | 6640 | 99.05 | 10195 | 2097 | 20.57 |
| FarthestFirst | 2745 | 791 | 28.82 | 14154 | 7947 | 56.15 |
| RandomFlatClustering | 8509 | 4376 | 51.43 | 8390 | 4361 | 51.98 |

**ตารางที่ 19** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Source_port และ Time

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 0 | 0.00 | 11144 | 1520 | 13.64 |
| W-sIB | 986 | 287 | 29.11 | 15913 | 1233 | 7.75 |
| FilteredCluterer | 6704 | 1200 | 17.90 | 10195 | 320 | 3.14 |
| FarthestFirst | 2745 | 0 | 0.00 | 14154 | 1520 | 10.74 |
| RandomFlatClustering | 8509 | 781 | 9.18 | 8390 | 739 | 8.81 |

**ตารางที่ 20** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Destination_port และ Packet_length

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 146 | 2.54 | 11144 | 567 | 5.09 |
| W-sIB | 986 | 57 | 5.78 | 15913 | 656 | 4.12 |
| FilteredCluterer | 6704 | 0 | 0.00 | 10195 | 713 | 6.99 |
| FarthestFirst | 2745 | 122 | 4.44 | 14154 | 591 | 4.18 |
| RandomFlatClustering | 8509 | 352 | 4.14 | 8390 | 361 | 4.30 |

**ตารางที่ 21** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Destination_port และ Time

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 0 | 0.00 | 11144 | 1393 | 12.50 |
| W-sIB | 986 | 278 | 28.19 | 15913 | 1115 | 7.01 |
| FilteredCluterer | 6704 | 0 | 0.00 | 10195 | 1393 | 13.66 |
| FarthestFirst | 2745 | 0 | 0.00 | 14154 | 1393 | 9.84 |
| RandomFlatClustering | 8509 | 719 | 8.45 | 8390 | 674 | 8.03 |

**ตารางที่ 22** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา จาก Time และ Packet_length

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 0 | 0.00 | 11144 | 1889 | 16.95 |
| W-sIB | 986 | 349 | 35.40 | 15913 | 1540 | 9.68 |
| FilteredCluterer | 6704 | 14 | 0.21 | 10195 | 1875 | 18.39 |
| FarthestFirst | 2745 | 0 | 0.00 | 14154 | 1889 | 13.35 |
| RandomFlatClustering | 8509 | 959 | 11.27 | 8390 | 930 | 11.08 |

**ตารางที่ 23** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Source_port Destination_port และ Packet_length

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 13 | 0.23 | 11144 | 6 | 0.05 |
| W-sIB | 986 | 6 | 0.61 | 15913 | 19 | 0.12 |
| FilteredCluterer | 6704 | 0 | 0.00 | 10195 | 19 | 0.19 |
| FarthestFirst | 2745 | 9 | 0.33 | 14154 | 10 | 0.07 |
| RandomFlatClustering | 8509 | 13 | 0.15 | 8390 | 6 | 0.07 |

**ตารางที่ 24** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Time Source_port และ Destination_port

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 0 | 0.00 | 11144 | 1520 | 13.64 |
| W-sIB | 986 | 287 | 29.11 | 15913 | 1233 | 7.75 |
| FilteredCluterer | 6704 | 1200 | 17.90 | 10195 | 320 | 3.14 |
| FarthestFirst | 2745 | 1520 | 55.37 | 14154 | 0 | 0.00 |
| RandomFlatClustering | 8509 | 781 | 9.18 | 8390 | 739 | 8.81 |

**ตารางที่ 25** แสดงจำนวน Packet ที่ตรงกับ Anomaly Traffic แบ่งตามกลุ่มและอัลกอริทึม โดยพิจารณา Destination_port Packet_length และ Time

| Algorithm | Cluster0 | | | Cluster1 | | |
|---|---|---|---|---|---|---|
| | # packet | #equal | % | #packet | #equal | % |
| K-means | 5755 | 0 | 0.00 | 11144 | 25 | 0.22 |
| W-sIB | 986 | 10 | 1.01 | 15913 | 15 | 0.09 |
| FilteredCluterer | 6704 | 0 | 0.00 | 10195 | 25 | 0.25 |
| FarthestFirst | 2745 | 0 | 0.00 | 14154 | 25 | 0.18 |
| RandomFlatClustering | 8509 | 11 | 0.13 | 8390 | 14 | 0.17 |

## วิจารณ์

จากผลการจัดกลุ่มของกลุ่มที่ 1 จะเห็นว่าอัลกอริทึมที่ให้ค่า precision สูงที่สุดคือ อัลกอริทึม K-means W-sIB และ FilteredCluterer ซึ่งมีค่าเท่า 1 นั่นแสดงว่าอัลกอริทึม K-means มีประสิทธิในการจัดกลุ่มมากที่สุดสำหรับกลุ่มที่ 1 และอัลกอริทึมที่ให้ค่า precision ต่ำสุดก็คือ อัลกอริทึม RandomFlatClustering ซึ่งมีค่า precision เท่ากับ 0.507 สำหรับอัลกอริทึม Filterclusterer เป็นอัลกอริทึมที่ให้ค่า recall และค่า F-measure สูงที่สุด คือ 0.998 และ 0.999 ตามลำดับ

อัลกอริทึม RandomFlatClustering ให้ค่า precision recall และ F-measure ต่ำที่สุด ภาพที่ 10 เป็นการเปรียบเทียบค่า precision recall และ F-measure ของกลุ่มที่ 1 ซึ่งจากภาพสามารถบอกได้ว่าอัลกอริทึม Filterclusterer มีประสิทธิภาพมากที่สุดซึ่งได้ผลสรุปเช่นเดียวกันกับข้อมูลในกลุ่มที่ 2 โดยค่า precision recall และ F-measure ที่ได้มีค่าเท่ากับ 0.999 1 และ 0.999 ตามลำดับ และจากผลการทดลองทางผู้วิจัยมีความเห็นว่าอัลกอริทึม RandomFlatClustering ไม่เหมาะสมที่จะนำมาใช้ในการจัดกลุ่มข้อมูลประเภททราฟฟิก แต่อาจจะเหมาะกับข้อมูลชนิดอื่นๆก็เป็นได้

จากผลการทดลองในการประเมินผลและวัดประสิทธิภาพการจัดกลุ่มของแต่ละอัลกอริทึมด้วยการทดสอบกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) นั้นพบว่า Attribute แต่ละ Attribute มีผลต่อการจัดกลุ่มข้อมูลทราฟฟิกมากน้อยต่างกัน และอัลกอริทึมแต่ละอัลกอริทึมก็ให้ผลลัพธ์ที่แตกต่างกันอย่างเห็นได้ชัด ดังในตารางที่ 17-25 จะเห็นว่าบาง Attribute แทบไม่มีผลต่อการจัดกลุ่มเลยเมื่อเรานำไปวัดประสิทธิภาพกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) อย่างเช่น คู่ของ Attribute ระหว่าง Source_port และ Destination_port การจับกลุ่มด้วยสาม Attribute อย่างเช่น Source_port Destination_port และ Packet_length และสุดท้าย Destination_port Packet_length และ Time ซึ่งให้ผลลัพธ์ในแต่ละอัลกอริทึมไม่ถึง 10% ซึ่งถือว่าเป็นค่าที่น้อยมากและในการทดลองนี้ไม่มีผลต่อการนำมาพิจารณาลักษณะข้อมูลที่มีความผิดปกติ สำหรับ Attribute ที่ให้ผลลัพธ์ค่อนข้างดีถึงดีมากประกอบด้วยคู่ของ Attribute ระหว่าง Source_port และ Time Destination_port และ Packet_length Destination_port และ Time Packet_length และ Time และพบว่า Attribute ระหว่าง Source_port และ Packet_length ด้วยการใช้อัลกอริทึม FilteredCluterer ให้ผลลัพธ์สูงที่สุดคือร้อยละ 99.05 ซึ่งเป็นค่าที่ดีที่สุดในการทดลอง และเป็นข้อมูลที่อยู่ในกลุ่มที่ 1 (Cluster0) มีจำนวน packet เท่ากับ 6704 packets และเมื่อนำไปทดสอบกับข้อมูลทราฟฟิกที่ทราบความผิดปกติอยู่แล้ว (Anomalous) พบว่ามีข้อมูลที่ตรงกันถึง

6640 packets ซึ่งบ่งบอกได้ว่าข้อมูลกลุ่มนี้มีลักษณะที่บ่งบอกถึงความผิดปกติเมื่อเปรียบเทียบกับข้อมูลที่ได้จากการทำ tcpdump และจากผลการจัดกลุ่มข้อมูลทราฟฟิกยังสามารถดึงรายละเอียดของ packet ที่ตกอยู่ในกลุ่มที่มีความผิดปกติได้อีกด้วยไม่ว่าจะเป็น Source_port Destination_port หรือ Packet_length เมื่อทราบรายละเอียดของ Attribute ต่างๆ อย่างละเอียดก็จะสามารถนำข้อมูลเหล่านี้ไปพัฒนาสร้างกฎที่ใช้ในเครื่องมือตรวจจับความผิดปกติได้

# สรุปและข้อเสนอแนะ

## สรุป

จากผลการทดลองซึ่งได้นำเทคนิคการจัดกลุ่มข้อมูลมาช่วยในการแบ่งประเภทของทราฟฟิก ทั้งนี้ก็เพื่อวิเคราะห์หาวิธีที่เหมาะสมที่สุดที่จะนำมาใช้ในการแบ่งกลุ่มข้อมูลทราฟฟิกซึ่งจะเป็นประโยชน์ต่อการนำไปสร้างกฎในการพัฒนาเครื่องมือที่ช่วยในการตรวจจับความผิดปกติของทราฟฟิกในอนาคตได้ ในการทดลองนี้ได้นำอัลกอริทึมการจัดกลุ่มข้อมูลมาใช้ 5 อัลกอริทึมด้วยกัน คือ sIB RandomFlatClustering FarthestFirst FilteredClusterer และ K-means ซึ่งเป็นอัลกอริทึมที่ไม่เคยมีการนำมาใช้กับข้อมูลประเภททราฟฟิกของเครือข่ายมาก่อน (ยกเว้น K-means) และจากผลการทดลองพบว่าอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลประเภททราฟฟิกของเครือข่ายมากที่สุดคือ อัลกอริทึม FilteredClusterer เนื่องจากให้ค่า Precision Recall และ F-measure สูงที่สุด ซึ่งการวัดประสิทธิภาพในการจัดกลุ่มด้วยค่า Precision Recall และ F-measure สามารถบ่งบอกได้ว่าอัลกอริทึมใดมีความถูกต้องแม่นยำ และมีความครบถ้วนสมบูรณ์ของข้อมูลมากที่สุด ซึ่งเป็นสิ่งจำเป็นก่อนที่จะมีการนำผลจากการจัดกลุ่มไปประยุกต์ใช้งานจริงในอนาคต นอกจากนี้ในการวัดประสิทธิภาพของแต่ละอัลกอทึมนั้นได้นำอัลกอริทึม K-means มาเป็นตัวเปรียบเทียบด้วย ซึ่งอัลกอริทึม K-means เป็นอัลกอริทึมที่มีการนำมาพัฒนาและประยุกต์ใช้งานกับข้อมูลทราฟฟิกเครือข่ายมาก่อน และสำหรับในการทดลองนี้อัลกอริทึม K-means เป็นอัลกอริทึมที่ให้ผลลัพธ์ในการจัดกลุ่มค่อนข้างดีแต่น้อยกว่าอัลกอริทึม FilteredClusterer ข้อดีของอัลกอริทึม K-means ก็คือ มีการทำงานที่รวดเร็วแม้ว่าข้อมูลทราฟฟิกจะมีปริมาณมหาศาลก็ตาม แต่ทั้งนี้ทั้งนั้นก็อาจขึ้นอยู่กับลักษณะของเครื่องคอมพิวเตอร์ที่ใช้ในการทดลองด้วย

จากการทดลองพบว่าเมื่อนำผลลัพธ์ที่ได้จากการแบ่งกลุ่มข้อมูลมาทดสอบกับข้อมูลทราฟฟิกที่ทราบว่ามีความผิดปกติ (known anomaly) ซึ่งได้เก็บข้อมูลทราฟฟิกจากการทำ tcpdump ระหว่างเครื่องคอมพิวเตอร์ที่ต้องการเก็บข้อมูลทราฟฟิกทำให้ทราบว่า Attribute ที่มีผลต่อการจัดกลุ่มข้อมูลประเภททราฟฟิกของเครือข่ายมากที่สุดก็คือ Attribute Source_port และ Packet_length เนื่องจากให้ค่าร้อยละในการทดสอบกับข้อมูลทราฟฟิกที่ทราบว่ามีความผิดปกติ (known anomaly) สูงที่สุดในทุกๆ อัลกอริทึม และจากการพิจารณาจากตารางที่ 17 ยังพบอีกว่าอัลกอริทึมที่มีประสิทธิภาพในการจัดกลุ่มมากที่สุดนั่นก็คือ อัลกอริทึม FilteredClusterer ซึ่งข้อมูลใน Cluster0 มีข้อมูลทราฟฟิกที่ตรงกับข้อมูลทราฟฟิกที่ทราบว่ามีความผิดปกติ (known anomaly) ถึงร้อยละ

99.05 และข้อมูลใน Cluster1 มีข้อมูลทราฟฟิกที่ตรงกับข้อมูลทราฟฟิกที่ทราบว่ามีความผิดปกติ (known anomaly) ร้อยละ 20.57 ซึ่งค่าร้อยละนี้บ่งบอกถึงข้อมูลใน Cluser0 มีข้อมูลทราฟฟิกที่มีลักษณะผิดปกติอยู่สูงมาก ผู้วิจัยจึงสรุปว่าในการแบ่งกลุ่มด้วยอัลกอริทึม FilteredClusterer นั้น ข้อมูลในกลุ่มที่ 1 (Cluster0) คือข้อมูลที่มีลักษณะผิดปกติ ซึ่งเป็นข้อมูลที่มีหมายเลขพอร์ตต้นทาง (Source_port) เป็น 8080 และมีขนาดความยาวของแพกเก็ต (Packet_length) เป็น 759 1434 และ 1514 นอกจากนี้ยังสามารถบอกถึงหมายเลขพอร์ตปลายทางที่เป็นของข้อมูลในกลุ่มนี้ได้อีกด้วย ซึ่งหมายเลขพอร์ตปลายทางที่ได้อยู่ในช่วง 24347 ถึง 25505

สำหรับงานวิทยานิพนธ์นี้ได้ตอบสนองต่อวัตถุประสงค์ที่ตั้งไว้เป็นอย่างดี ไม่ว่าจะเป็นผลลัพธ์ของการจัดกลุ่มข้อมูลที่ได้จากอัลกอริทึม FilteredClusterer ที่ได้ผลลัพธ์ที่ถือว่ามีประสิทธิภาพสูงมาก ซึ่งจะสามารถนำผลจากการจัดกลุ่มนี้ไปประยุกต์ใช้กับเครื่องมือช่วยตรวจจับความผิดปกติในอนาคตต่อไปได้เป็นอย่างดี นอกจากนี้ยังสามารถวิเคราะห์หาความผิดปกติของทราฟฟิกที่เกิดขึ้นได้อีกด้วยโดยการวิเคราะห์ถึงรายละเอียดของข้อมูล Packet ที่กระจายตัวอยู่ในกลุ่มที่คาดว่าจะมีความผิดปกติ

## ข้อเสนอแนะ

สำหรับการวิจัยต่อไปในอนาคตนั้น ควรทดลองสร้างกฎที่ได้จากการจัดกลุ่มข้อมูลและนำไปใช้กับเครื่องมือที่ช่วยในการตรวจจับทราฟฟิกที่มีความผิดปกติเพื่อวัดประสิทธิภาพในการนำไปใช้งานจริง สำหรับผู้ที่สนใจสามารถศึกษา Attribute อื่นๆ ที่มีผลต่อการจัดกลุ่มเพิ่มเติม อย่างเช่น ส่วนหัวของแพกเก็ต (packet header) มาทำการทดลองต่อไป นอกจากนี้ยังสามารถศึกษาหรือคิดค้นอัลกอริทึมใหม่ๆ ที่มีทั้งประสิทธิภาพและความรวดเร็วในการทำงานได้ดียิ่งขึ้นรวมถึงการหาเครื่องมือที่ดีกว่านี้มาช่วยในการทำการทดลองด้วย

# เอกสารและสิ่งอ้างอิง

Andreopoulos, William. 2006. Clustering Algorithms for Categorical Data.

Basu, S. 2003. Semi-supervised Clustering: Learning with Limited User Feedback. *In* **Technical Report UT-AI-TR-03-307**

Berry, M.W., U. Dayal, C. Kamath and D. Skillicorn. 2004. pp. 338. *In* **Proceedings of the Fourth SIAM International Conference on Data Mining**.

Davis, J. And M. Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. *In* **ICML'06**, USA.

Erman, J., M. Arlitt and A. Mahanti. 2006. Traffic Classification Using Clustering Algorithms. *In* **SIGCOMM'06 MineNet Workshop,** Canada.

Ester, M., H. Kriegel, J. Sander and X. Xu. 1996. pp. 226-231. A density-based Algorithm for discovering Clusters in Large Spatial Databases with Noise. *In* **Knowledge Discovery and Data Mining (KDD 96)**

F1 score. Available Source: http://en.wikipedia.org/wiki/F-score, March 22, 2009.

Kerdprasop, N. 2006. **A proper algorithm and technique for mining the medical diagnosis data sets.** Available Source: http://www.sut.ac.th/Engineering/Computer/faculty/nittaya/myweb/files%20to%20upload/ResearchConcept_4.doc, December 18, 2008.

Lakhina, A., M. Crovella and C. Diot. 2005. Mining Anomalies Using Traffic Feature Distributions. *In* **Technical Report BUCS-TR-2005-002**

MacQueen, J. 1976. Some methods for classification and analysis of multivariate observations. pp. 281-297. *In* **Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability.**

McGregor, A., M. Hall, P. Lorier and J. Brunskill. 2004. Flow Clustering Using Machine Learning Techniques. *In* **PAM 2004**

Münz, G., S. Li, G. Carle. 2007. Traffic Anomaly Detection Using K-Means Clustering. *In* **GI/ITG Workshop MMBnet**

Niu, Z.Y., D.H. Ji and C.L. Tan. 2006. Using cluster validation criterion to identify optimal feature subset and cluster number for document clusterin. *In* **Information Processing and Management' 06**

O'Rourke, S., G. Chechik, R. Friedman and E. Eskin. 2006. **Discrete profile comparison using information bottleneck.** BMC Bioinformatics. Available Source: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1810319, January 24, 2009.

Pirooznia, M., J.Y. Yang, M.Q. Yang and Y. Deng. 2007. A comparative study of different machine learning methods on microarray gene expression data. *In* **BIOCOMP'07**

Precision and recall. Available Source: http://en.wikipedia.org/wiki/Precision_and_recall, 2008.

Ramah, K., H. Ayari and F. Kamoun. 2006. Traffic Anomaly Detection and Characterization in the Tunisian National University Network. pp. 136-147. *In* **Networking 2006.**

RapidMiner. Available Source: http://downloads.sourceforge.net/yale/rapidminer -4.2 - guimanual.pdf, 2008.

RapidMiner. Available Source: http://rapid-i.com/content/ blogcategory/38/69/,

Sharpe, R., E. Warnicke and U. Lamping. 2004. **Ethereal User's Guide V2.00 for Ethereal 0.10.5.** Available Source: http://www.rootsecure.net/content/downloads/pdf/ethereal_guide.pdf, 2008.

Shyu, M., S. Chen and K. Sarinnapakorn. 2003. pp. 172-179. A Novel Anomaly Detection Scheme Based on Principal Component Classifier. *In* **Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop.**

Silva, L.S., T.D. Mancilha, J.D.S. Silva, A.C.F. Santos and eA. Montes. 2006. A Framework for Analysis of Anomalies in the Network Traffic. *In* **INPE'06.**

Tan, P.N., M. Steinbach and V. Kuman. 2006. Introduction to Data Mining. *In* **Addison Wesley**

Winston, H., H. Chang and S. Chang. 2005. pp. 82-91. Visual Cue Cluster Construction via Information Bottleneck Principle and Kernel Density Estimation. *In* **CIVR 2005**, USA.

Zander, S., T. Nguyen and G. Armitage. 2005. Automatic Traffic Classification and Application Identification using Machine Learning. *In* **LCN'05**

## ประวัติการศึกษา และการทำงาน

| | |
|---|---|
| ชื่อ –นามสกุล | นางสาวเบญจวรรณ สุขพัฒนศรีกุล |
| วัน เดือน ปี ที่เกิด | วันที่ 4 กุมภาพันธ์ 2525 |
| สถานที่เกิด | จังหวัดพังงา |
| ประวัติการศึกษา | ปริญญาตรี มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตหาดใหญ่ ปี 2548 |
| ตำแหน่งหน้าที่การงานปัจจุบัน | วิศกร 4 |
| สถานที่ทำงานปัจจุบัน | บริษัท ทีโอที จำกัด (มหาชน) |
| ผลงานดีเด่นและรางวัลทางวิชาการ | World Academy of Science, Engineering and Technology , Volume 37, January 2009, the WCSET conference proceedings (ISSN 2070-3740) |
| ทุนการศึกษาที่ได้รับ | ทุนวิจัยคณะวิศวกรรมศาสตร์ |