การคัดแยกเอกสารสำคัญออกจากเอกสารทั่วไปด้วยวิธีเอสวีเอ็ม

อรทิพย์ เลื่อยงาม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม คณะวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2555

Confidential Document Classification from General Document Using SVM

ORATHIP LUEYNGAM

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Computer and Telecommunication

Faculty of Engineering, Dhurakij Pundit University

2012

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ประสบผลสำเร็จลงได้ ด้วยความช่วยเหลืออย่างดียิ่งของอาจารย์ในภาควิชาวิศวกรรมคอมพิวเตอร์ และโทรคมนาคม ท่านอาจารย์ ดร.ชัยพร เขมะภาตะพันธ์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ให้คำแนะนำ ชี้แนะข้อคิดต่างๆ ในการศึกษาวิจัย ตลอดจนแนวทางในการแก้ไขปัญหาต่างๆ ทบทวนวิธีการวิจัย เพื่อความสมบูรณ์ และถูกต้อง และเอาใจใส่ข้าพเจ้ามาโดยตลอด จึงขอกราบขอบพระคุณเป็นอย่างสูงมา ณ โอกาสนี้

ขอขอบพระคุณ ดร.ประศาสน์ จันทราทิพย์ ประธานคณะกรรมการสอบวิทยานิพนธ์ รวมทั้ง ดร.ณรงค์เดช กีรติพรานนท์ และ ดร.เจนจบ วีระพานิชเจริญ ที่สละเวลามาเป็นกรรมการสอบวิทยานิพนธ์ และกรุณาให้คำแนะนำที่เป็นประโยชน์ต่องานวิจัย ขอขอบคุณเจ้าหน้าที่ทุกท่านที่ช่วยดำเนินเรื่องต่างๆ ให้เป็นอย่างดี

ขอขอบพระคุณอาจารย์ และเจ้าหน้าที่ในภาควิชาวิศวกรรมคอมพิวเตอร์ และโทรคมนาคมทุกท่านที่ได้ประสิทธิประสาทวิชาความรู้ และให้ความช่วยเหลือมาโดยตลอด ขอขอบคุณเพื่อนทุกท่าน ที่ให้คำปรึกษา ให้ความช่วยเหลือ ตลอดระยะเวลาการศึกษา รวมทั้งเจ้าหน้าที่ของสายงานเทคโนโลยีสารสนเทศ การประปานครหลวง ที่ได้ให้กำลังใจ และให้ความช่วยเหลือข้าพเจ้ามาโดยตลอด

ความดีอันเกิดจากการศึกษาค้นคว้านี้ ผู้เขียนขอมอบแด่บิดา มารดา ครู อาจารย์ และผู้มีพระคุณทุกท่าน ผู้เขียนมีความซาบซึ้งในความกรุณาอันดียิ่งจากทุกท่านที่ได้กล่าวมา และขอกราบขอบพระคุณมา ณ โอกาสนี้

อรทิพย์ เลื่อยงาม

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูป

รูปที่ **หน้า**

สารบัญรูป (ต่อ)

รูปที่ **หน้า**

หัวข้อวิทยานิพนธ์	การคัดแยกเอกสารสำคัญออกจากเอกสารทั่วไปด้วยวิธีเอสวีเอ็ม
ชื่อผู้เขียน	อรทิพย์ เลื่อยงาม
อาจารย์ที่ปรึกษา	ดร.ชัยพร เขมะภาตะพันธ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์และโทรคมนาคม
ปีการศึกษา	2554

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอและออกแบบกระบวนการคัดแยกเอกสารสำคัญออกจากเอกสารทั่วไปด้วยวิธีเอสวีเอ็ม เพื่อใช้เป็นต้นแบบสำหรับการกรองเอกสารในระบบการป้องกันข้อมูลรั่วไหลด้วย Data Loss Prevention (DLP) โดยเป็นผลิตภัณฑ์ของ Symantec เพื่อใช้ในการบริหารจัดการความมั่นคงของข้อมูล ซึ่งการประปานครหลวงได้นำมาพัฒนาใช้งานเพื่อป้องกันเอกสารที่เป็นความลับรั่วไหล อย่างไรก็ตามระบบดังกล่าวยังขาดประเภทเอกสารที่เป็นต้นแบบของการกรองข้อมูลสำคัญ และเพื่อลดเวลาการคัดแยกเอกสารซึ่งมีปริมาณมากและยากต่อการกำหนดระดับความสำคัญ จึงทำการพัฒนาวิธีการคัดแยกประเภทเอกสาร โดยใช้เทคนิคการแยกประเภทแบบมีผู้สอนด้วยวิธี SVM โดยประมวลเอกสารเป็นเวกเตอร์ค่าน้ำหนัก TF-IDF นำค่าน้ำหนักที่ได้เข้าสู่กระบวนการแยกประเภทเอกสารด้วย LS-SVM โดยมีข้อจำกัดว่าเอกสารต้นแบบดังกล่าวต้องไม่มีการแก้ไขระดับความสำคัญอีกในภายหลัง ผลการทดสอบประสิทธิภาพการคัดแยกประเภทเอกสาร โดยแบ่งออกเป็น 2 กรณีได้แก่ กรณีที่ 1 คือการคัดแยกประเภทเอกสารกับชุดทดสอบนอกกลุ่มเป้าหมาย ได้ค่าความถูกต้องตามคุณลักษณะที่เลือกใช้คำสำคัญจำนวน 10, 20, 30 และ 40 คำ เป็น 83.33%, 81.67%, 80.00% และ 83.33% ตามลำดับ ขณะที่กรณีที่ 2 คือการคัดแยกประเภทเอกสารกับชุดทดสอบกลุ่มเป้าหมาย ให้ค่าความถูกต้องตามคำสำคัญที่เลือกใช้คำสำคัญจำนวน 10, 20, 30 และ 40 คำ เป็น 83.33%, 88.33%, 86.67% และ 86.67% ตามลำดับ จากผลการทดสอบแสดงให้เห็นว่ากระบวนการที่นำเสนอมีค่าความถูกต้องอยู่ในเกณฑ์สูงและยอมรับได้ อย่างไรก็ตามค่าความถูกต้องของการคัดแยกประเภทเอกสารในกลุ่มเป้าหมายสูงกว่าเมื่อเทียบกับการทดสอบกับข้อมูลนอกกลุ่มเป้าหมาย ดังนั้นถ้าสามารถจำกัดขอบเขตของกลุ่มเป้าหมายและคัดเลือกคุณลักษณะเป็นตัวแทนที่ดีของชุดเอกสาร รวมทั้งจำนวนคุณลักษณะที่เลือกใช้อย่างเหมาะสม ก็จะมีประสิทธิภาพในการคัดแยกประเภทเอกสารดีขึ้น

Thesis Title	Confidential Document Classification from General Documents Using SVM
Author	Orathip Lueyngam
Thesis Advisor	Dr. Chiyaporn Khemapatapan
Department	Computer and Telecommunication Engineering
Academic Year	2011

ABSTRACT

This thesis proposes and designs the procedure for classifying confidential documents from general documents by using SVM method. The classified document will be used as a reference for filtering document in Data Loss Prevention (DLP) system, the Symantec's product, used in security management. Metropolitan Waterworks Authority (MWA) adopt this Symantec's system to use in the organization to protect the leakage of confidential documents. However, the system needs many reference documents to be used as filtering references. Thus, the system takes much time before use if references are manually selected. So, document classifying procedure based on training using SVM method has been proposed. The procedure will process the documents for finding TF-IDF weight and classify documents by using LS-SVM. The limitation of this procedure is that reference documents cannot be revised later. The results from 2 testing cases found that 1) testing documents out of target documents will provide accuracy about 83.33%, 81.67%, 80.00% and 83.33% for the feature set of 10, 20, 30 and 40 words, respectively, 2) testing documents in target documents will provide accuracy about 83.33%, 88.33%, 86.67% and 86.67% for the feature set of 10, 20, 30 and 40 words, respectively. It can be noted that the proposed procedure has high accuracy and is acceptable. However, testing documents in target group can provide higher accuracy than testing documents out of target group. Thus, if document can be targeted and grouped before processing, the procedure will be more efficiency.

บทที่ 1
บทนำ

1.1 ที่มาของงานวิจัย

ในสังคมยุคดิจิตอลทุกวันนี้ องค์กรต่างๆ มีการปรับตัว โดยนำระบบคอมพิวเตอร์มาใช้ในการจัดการในเรื่องของข้อมูลอย่างแพร่หลาย ไม่ว่าจะเป็นสื่อสารผ่านอินเตอร์เน็ต การสื่อสารผ่านสื่ออิเล็กทรอนิกส์ นั่นแสดงว่าข้อมูล และสารสนเทศต่างๆ มีการจัดทำขึ้นมากมาย เพื่อใช้ในการดำเนินธุรกิจ และถือว่าเป็นสินทรัพย์ของธุรกิจ สินทรัพย์เหล่านี้มีความสำคัญที่องค์กรต้องป้องกัน จากสภาพแวดล้อมในปัจจุบันที่ต้องติดต่อสื่อสารกันมากขึ้น ทำให้เกิดปัญหา และจุดอ่อนหลากหลายรูปแบบที่ทำให้สารสนเทศนั้นเสียหาย หากมองอีกมุมมองหนึ่งก็เป็นภัยร้ายแรงต่อองค์กรได้เช่นกัน หากข้อมูลต่างๆ เหล่านั้นตกอยู่ในมือของผู้ที่ไม่ประสงค์ดี หรือผู้ที่ไม่สมควร

สิ่งเหล่านี้นำไปสู่การรั่วไหลของข้อมูลได้ทั้งนั้น และพบว่าสาเหตุของการรั่วไหลส่วนมากเกิดจากบุคคลภายในองค์กร สิ่งสำคัญ คือความเข้าใจในเรื่องของช่องทางการรั่วไหลของข้อมูลว่ามีช่องทางใดบ้าง จะได้กำหนด และจัดทำมาตรการควบคุมที่เหมาะสม การจัดหมวดหมู่ของข้อมูล การแยกประเภทข้อมูลตามคุณค่าที่มีต่อองค์กร และผลกระทบถ้าหากข้อมูลเหล่านั้นถูกนำไปเผยแพร่ หรือแก้ไข

จากปัญหาข้างต้น จึงได้นำเสนอระบบการจัดประเภทเอกสารเพื่อนำเข้าสู่กระบวนการทำงานร่วมกับโซลูชั่นด้านการรักษาความมั่นคงปลอดภัยของเอกสาร ในส่วนของการป้องกันข้อมูลรั่วไหล DLP (Data Leakage Prevention) เพื่อควบคุมการไหลของเอกสารสำคัญ และบังคับใช้นโยบายการป้องกันข้อมูลสำคัญ โดยศึกษาเทคนิคที่ใช้กันอยู่ในปัจจุบันคือเทคนิคการจำแนกประเภทเอกสาร โดยมีการกำหนดประเภทเอกสารไว้ล่วงหน้า เทคนิคนี้จะทำการสอนให้ระบบรู้จำรูปแบบเอกสาร ในแต่ละประเภทเสียก่อน หลังจากนั้นจึงนำเอกสารที่ต้องการจำแนกประเภทเข้าไปในระบบ ให้ระบบทำการแยกประเภทเอกสารเพื่อสร้างโครงรูป (Template) ตามประเภทเอกสารที่กำหนด (เอกสารสำคัญ และเอกสารทั่วไป) โดยใช้ทฤษฎีที่เกี่ยวข้องคือ TF-IDF (Term Frequency and Inverse Document Frequency) เป็นวิธีการคำนวณค่าน้ำหนักจากความถี่ของคำที่ปรากฏในเอกสาร และ SVM (Support Vector Machine) เป็นขั้นตอนการใช้งานตัวจำแนก นำมาสนับสนุนการแยกประเภทเอกสาร ก่อนนำเข้าสู่โซลูชั่นของการป้องกันข้อมูลรั่วไหลเพื่อลดเวลา

ในการคัดแยกประเภทเอกสารเพื่อสร้างโครงรูปให้ระบบ และทดสอบประสิทธิภาพของการคัดแยกประเภทเอกสารโดยหาค่าความถูกต้อง (Accuracy Rate)

จากการแก้ปัญหาข้างต้น ยังส่งผลเพื่อสร้างความมั่นใจว่าข้อมูลความลับ และสารสนเทศ ยังถูกเก็บรักษาอยู่ครบถ้วนปลอดภัย และเพื่อป้องกันการดำเนินธุรกิจให้เป็นไปอย่างต่อเนื่อง ซึ่งสอดคล้องกับการนำระบบต่างๆ มาใช้เพื่อให้มีประสิทธิภาพ และเกิดความน่าเชื่อถือในการที่จะควบคุมข้อมูลต่างๆ ให้เป็นไปอย่างสากล เช่น COBIT, ITIL, ISO/IEC27001:2005, ISO17799:2005 แต่ระบบที่เป็นที่นิยมใช้กันอย่างแพร่หลาย ครอบคลุมข้อมูลทุกประเภท และมีความน่าเชื่อถือในระดับสากลคือ ISO/IEC27001:2005 และ ISO17799:2005

1.2 วัตถุประสงค์ของงานวิจัย

1 ศึกษาการบริหารจัดการเกี่ยวกับการป้องกันข้อมูลรั่วไหล เพื่อให้เกิดความชัดเจนในเรื่องของกระบวนการทำงานของระบบ

2 ศึกษาเทคนิคที่นำมาปรับใช้สำหรับการแยกประเภทเอกสาร เพื่อนำไปใช้งานเป็นโครงรูปในการคัดกรองของระบบการป้องกันเอกสารรั่วไหล

3. พัฒนาระบบการแยกประเภทเอกสารสำคัญออกจากเอกสารทั่วไปก่อนนำเข้าระบบการป้องกันเอกสารรั่วไหลของการประปานครหลวงผ่านทางอุปกรณ์ Symantec DLP

4. ทดสอบประสิทธิภาพของการแยกประเภทเอกสาร โดยใช้ค่าความถูกต้อง (Accuracy Rate)

1.3 ขอบเขตของการวิจัย

1. วิเคราะห์และศึกษาหลักการทำงานของระบบการป้องกันการรั่วไหลของข้อมูล เทคนิคการให้น้ำหนักคำ และเทคนิคการแยกประเภทเอกสาร เพื่อนำผลลัพธ์จากการแยกประเภทเอกสารไปเป็นโครงรูปในการคัดกรองของระบบการป้องกันเอกสารรั่วไหล

2. พัฒนาระบบการแยกประเภทเอกสารลับออกจากเอกสารทั่วไปก่อนนำเข้าระบบการป้องกันเอกสารรั่วไหลของการประปานครหลวงผ่านทางอุปกรณ์ Symantec DLP

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ความชัดเจนเกี่ยวกับหลักการทำงานของระบบการป้องกันการรั่วไหลของข้อมูล และสามารถแยกประเภทเอกสารลับออกจากเอกสารทั่วไป เพื่อกำหนดประเภทของข้อมูลจากเอกสารสำคัญที่ต้องการป้องกัน โดยสามารถลดระยะเวลาที่ใช้ในการคัดแยกประเภทเอกสาร และนำ

ผลลัพธ์ที่ได้จากการคัดแยกไปเป็นโครงรูปในการคัดกรองของระบบการป้องกันข้อมูลรั่วไหล Symantec DLP เพื่อเพิ่มประสิทธิภาพการทำงานของระบบการป้องกันข้อมูลรั่วไหลต่อไป

1.5 แผนการดำเนินงาน

เดือน	2553		2554						2555	
	ตค.	ธค.	กพ.	เมย	มิย.	สค.	ตค.	ธค.	กพ.	เมย
1. ศึกษาระบบการป้องกันข้อมูลรั่วไหล	←	→								
1.1 เทคโนโลยี DLP	←	→								
1.2 นโยบายการป้องกันข้อมูลรั่วไหล	←	→								
1.3 การทำงานของ Symantec DLP	←	→								
2. ศึกษาการแยกประเภทเอกสาร	←	→								
2.1 เทคนิคการตัดคำภาษาไทย	←	→								
2.2 เทคนิคการให้น้ำหนักคำ TF-IDF	←	→								
2.3 ทฤษฎี SVM	←	→								
3. วิเคราะห์และศึกษาแนวทางการพัฒนา	←	→								
3.1 พัฒนาระบบการแยกประเภทเอกสาร	←	→								
3.2 กระบวนการนำเข้าเอกสาร	←	→								
3.3 กระบวนการแยกประเภทเอกสาร	←	→								
4. วิเคราะห์ผลการทดสอบ	←	→								
5. ศึกษาวิธีการตรวจสอบความน่าเชื่อถือของการทดสอบ Accuracy Rate	←	→								
6. รวบรวมข้อมูลจัดทำวิทยานิพนธ์	←	→								

บทที่ 2
ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

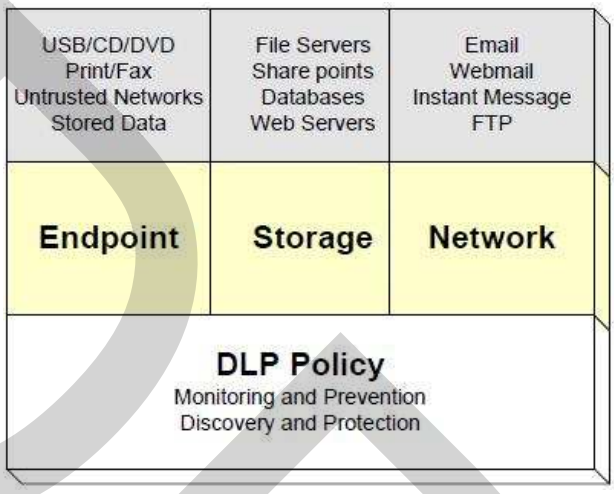
2.1 เทคโนโลยีการป้องกันข้อมูลรั่วไหล (Data Leakage Prevention : DLP)

กระบวนการทำงานและวิธีการป้องกันการรั่วไหลของข้อมูลสำคัญ ในการถ่ายโอนโดยไม่ได้รับอนุญาต หรือเปิดเผยข้อมูลสำคัญ เพื่อลดความเสี่ยงจากการถูกขโมย และสูญเสียโอกาส ความสามารถทางธุรกิจ โดยการตรวจสอบและควบคุมการใช้งานข้อมูล รวมทั้งกำหนดนโยบายให้สอดคล้องกับกลยุทธ์และกระบวนการทางธุรกิจ

2.1.1 เทคโนโลยีพื้นฐานของ DLP (Data Leakage Prevention) เทคโนโลยีพื้นฐานของระบบ DLP จะมีโครงสร้างการทำงานดังแสดงในรูปที่ 2.1 และในการสร้างรูปแบบข้อมูลสำคัญนั้น ต้องมีการระบุข้อมูลก่อนการถ่ายโอนโดยใช้อัลกอริทึมในการระบุข้อมูลที่มีรูปแบบ หรือประเภทเดียวกับข้อมูลสำคัญที่ได้ระบุไว้ สิ่งเหล่านี้อาจทำให้การประมวลผลช้ากว่าปรกติ อัลกอริทึมที่นำมาใช้ปรับใช้ควรมีลักษณะการทำงานที่เป็นระบบ และสามารถระบุข้อมูลสำคัญ โดยลดข้อมูลเกี่ยวกับเอกสารให้เล็กลง เพื่อง่ายต่อการนำไปใช้ในระบบ ส่วนประกอบที่ได้รับการพิจารณาเพื่อป้องกันการรั่วไหลของข้อมูลภายในองค์กรประกอบด้วยข้อมูลเคลื่อนไหว (Data in motion) คือ ข้อมูลใด ๆ ที่มีการเคลื่อนไหวผ่านทางเครือข่ายไปยังภายนอกผ่านทางอินเทอร์เน็ต ข้อมูลที่มีอยู่ (Data at rest) คือ ข้อมูลที่อยู่ในระบบแฟ้มฐานข้อมูล หรือการจัดเก็บข้อมูลด้วยวิธีการอื่น และข้อมูล ณ จุดปลายทาง (Data in Use) คือ ข้อมูล ณ จุดสิ้นสุดของเครือข่ายสำหรับเครื่องคอมพิวเตอร์ และอุปกรณ์ต่อพ่วง รวมถึงอุปกรณ์จัดเก็บข้อมูลแบบพกพา โดยพื้นฐานแล้วจะต้องทราบเกี่ยวกับความหมายของการรั่วไหล ช่องทางการรั่วไหล และวิธีการที่ทำให้ข้อมูลรั่วไหล รวมทั้งผลกระทบที่อาจเกิดขึ้นได้ ดังต่อไปนี้

2.1.1.1 การรั่วไหลของข้อมูล เกิดจากช่องโหว่ของการจัดเก็บไฟล์ข้อมูล ซึ่งเป็นข้อมูลที่มีค่าต่อองค์กร และเกิดผลกกระทบถ้าหากข้อมูลถูกนำไปเผยแพร่หรือนำไปใช้ประโยชน์เพื่อแสวงหาผลประโยชน์ของผู้ประสงค์ร้ายหรือนำไปใช้อย่างไม่เหมาะสม โดยแบ่งระดับความสำคัญของข้อมูลออกเป็น 4 ระดับคือ ข้อมูลสาธารณะ (เปิดเผยสู่สาธารณะชนได้โดยไม่ส่งผลกระทบใดๆ ต่อองค์กร) ข้อมูลภายใน (เอกสารที่ใช้ติดต่อภายในองค์กร บันทึกข้อความ หรือข้อมูลที่เกิดจากการดำเนินธุรกิจ) ข้อมูลลับ (ข้อมูลทางธุรกิจ การเงิน และข้อมูลอื่นที่ไม่ใช่ข้อมูลหวงห้าม หาก

ข้อมูลเหล่านี้ถูกเผยแพร่อาจส่งผลเสียต่อองค์กรได้) ข้อมูลหวงห้าม (ข้อมูลที่ประกอบด้วยข้อมูลที่มีการควบคุมการเข้าถึง การจัดเก็บ หรือประมวลผลตามกฎหมาย หรือข้อบังคับที่องค์กรกำหนด)



รูปที่ 2.1 รูปแบบของ DLP (Data Loss Prevention System Model)

2.1.1.2 การรั่วไหลแบ่งออกได้เป็น การรั่วไหลโดยตั้งใจ เช่นการส่งข้อมูลออกภายนอกองค์กร ซึ่งอาจจะเป็นการขโมย ลบ หรือแก้ไขข้อมูลสำคัญ หรือการรั่วไหลโดยไม่ตั้งใจ เช่นการแนบไฟล์เอกสารไปกับจดหมายอิเล็กทรอนิกส์ การนำไปแสดงไว้บนเว็บบอร์ด หรือเว็บไซต์ที่ให้บริการอัพโหลด ดาวน์โหลด โดยรู้เท่าไม่ถึงการณ์ สิ่งเหล่านี้อาจเกิดจากบุคคลในองค์กรที่ขาดความรู้ความเข้าใจในเรื่องของการใช้งานเทคโนโลยี และความรู้เท่าไม่ถึงการณ์ หรือบุคคลภายนอกองค์กร ซึ่งถือว่าเป็นภัยคุกคามที่มากับการใช้งานเครือข่ายโดยขาดความรู้ความเข้าใจและตระหนักในเรื่องของความปลอดภัย ทำให้นำพาภัยคุกคามเข้ามาติดตั้งที่เครื่องลูกข่าย กลายเป็นช่องโหว่ของระบบ และเป็นอีกจุดหนึ่งที่ทำให้ข้อมูลสำคัญหลุดรั่วออกไป

2.1.1.3 วิธีการรั่วไหล แบ่งออกได้เป็น 2 ช่องทางใหญ่ๆ คือการรั่วไหลผ่านทางเครือข่ายและอุปกรณ์ปลายทาง รวมทั้งอุปกรณ์ต่อพ่วงด้วย การรั่วไหลผ่านทางเครือข่าย (การแนบไฟล์ไปกับจดหมายอิเล็กทรอนิกส์ หรือนำไปวางไว้บนเครือข่ายอินเตอร์เน็ตที่ให้บริการอัพโหลด หรือดาวน์โหลด) การรั่วไหลผ่านทางอุปกรณ์ปลายทาง (คอมพิวเตอร์ โทรศัพท์มือถือ สื่อจัดเก็บข้อมูล เช่น CD/DVD, USB, MP3)

2.1.1.4 ผลกระทบ เมื่อบุคคลผู้ไม่มีสิทธิ์ได้รับข้อมูลสำคัญเหล่านี้ ไม่เพียงแต่จะทำให้เกิดความเสียหายต่อความสามารถในการแข่งขันทางธุรกิจเท่านั้น แต่ในปัจจุบันด้วยกฎระเบียบใน

เรื่องการป้องกันความลับของข้อมูล โดยเฉพาะงานสำคัญ เช่น ธุรกิจการบริการซึ่งถือครองข้อมูลส่วนตัวของลูกค้าไว้ ธุรกิจการเงิน และการแพทย์ ถึงแม้จะมีการรั่วไหลเพียงเรคคอร์ดเดียวก็อาจจะทำให้องค์กรต้องถูกฟ้องร้องจากเจ้าของข้อมูล และหน่วยงานภาครัฐที่ควบคุมกฎระเบียบในเรื่องของการป้องกันความลับของข้อมูลด้วย

2.1.2 พื้นฐานของนโยบายการรักษาความปลอดภัย ในการรักษาของข้อมูลสำคัญนั้น พื้นฐานของนโยบายการรักษาความปลอดภัยมีความสำคัญด้วยเหตุหลายประการ ในเรื่องของบทบาทความรับผิดชอบของเจ้าหน้าที่ขององค์กร และผู้ที่เข้ามาปฏิบัติงานจากภายนอก ซึ่งช่องโหว่จากภัยคุกคามที่เกิดขึ้นกับข้อมูลสำคัญ เป็นปัญหาที่เกี่ยวข้องกับระบบสารสนเทศที่พบในปัจจุบัน ขั้นตอนในการป้องกัน

2.1.2.1 การประเมินความเสี่ยงเพื่อกำหนดกลยุทธ์ และวิธีการในการป้องกันข้อมูลที่เป็นความลับ ในขั้นตอนการดำเนินการสำรวจและประเมินความเสี่ยง

2.1.2.2 ระบุและแยกประเภทข้อมูลที่เป็นความลับ เพื่อกำหนดข้อมูลที่เป็นความลับหรือไม่เป็นความลับ และกำหนดระดับการป้องกันที่เหมาะสมโดยใช้การจำแนกประเภทข้อมูล

2.1.2.3 พัฒนานโยบายและวิธีการในการป้องกันข้อมูล กำหนดความรับผิดชอบในการป้องกันข้อมูล ทำการเปรียบเทียบข้อมูลที่มีอยู่กับนโยบาย และสร้างนโยบายสำหรับการใช้งานระบบ โดยใช้ซอฟต์แวร์มาเป็นเครื่องมือในการปฏิบัติตามนโยบาย ค้นหาว่าข้อมูลที่เป็นความลับเก็บไว้ที่ไหน (Data Discover) เฝ้าระวังการใช้ข้อมูลที่เป็นความลับ (Data Monitor) ปกป้องและป้องกันข้อมูลที่เป็นความลับ (Data Protection) และบริหารจัดการและบังคับใช้นโยบายในการรักษาความปลอดภัยของข้อมูล

2.1.2.4 ปรับใช้เทคโนโลยีเข้ามาช่วยในการบังคับใช้และปฏิบัติตามนโยบาย โดยการนำเทคโนโลยีมาปรับใช้เพื่อช่วยในการดำเนินการตามนโยบาย และบังคับใช้แบบอัตโนมัติ

2.1.2.5 สื่อสารและศึกษาผู้มีส่วนได้ส่วนเสียในการสร้างวัฒนธรรมในการใช้งาน โดยแจ้งให้พนักงานและผู้มีส่วนได้ส่วนเสียในการป้องกันข้อมูลสำคัญ และกระตุ้นพฤติกรรมในการป้องกัน โดยแจ้งให้ทราบ ฝึกอบรม และมีการสื่อสารกันอย่างต่อเนื่อง

2.1.2.6 รวบรวมวิธีการป้องกันข้อมูลในกระบวนการธุรกิจ เพื่อป้องกันข้อมูลที่เป็นความลับขององค์กร โดยรวบรวมกระบวนการทางธุรกิจ และข้อมูลที่เป็นความเสี่ยง นำเข้าสู่ระบบการรักษาข้อมูล

2.1.2.7 ตรวจสอบการใช้งานของผู้เกี่ยวข้อง ตรวจสอบค่าการติดตั้ง และดำเนินการตรวจสอบวิธีการในการประเมินขั้นตอนการป้องกันข้อมูล โดยใช้เทคโนโลยีในการเฝ้าระวัง และเครื่องมือในการตรวจสอบ

2.1.3 วิธีการป้องกันการรั่วไหลของข้อมูล แบ่งออกได้เป็น 3 ขั้นตอนคือ

2.1.3.1 การค้นหาข้อมูล (Data Discover) เป็นกระบวนการค้นหาข้อมูลความลับ ไฟล์เอกสาร และข้อมูลสำคัญต่างๆ ที่กระจัดกระจายอยู่ตามเครื่องลูกข่าย และไฟล์เซฟเวอร์ต่างๆ เช่น ข้อมูลลูกค้า ข้อมูลทรัพย์สินทางปัญญา รวมทั้งข้อมูลสำคัญอื่นๆ วัตถุประสงค์หลักคือทำให้องค์กรทราบว่าข้อมูลสำคัญขององค์กรถูกจัดเก็บอยู่ที่ใดบ้าง ใครเป็นผู้ใช้ข้อมูล และแต่ละข้อมูลมีระดับความเสี่ยงที่จะก่อให้เกิดความสูญเสียในระดับใด การค้นหาบนระบบเครือข่าย (ข้อมูลในขณะมีการโอนถ่าย) การค้นหาข้อมูลที่อยู่บนระบบ(ข้อมูลที่จัดเก็บอยู่บน Disk Drive) และการวิเคราะห์ข้อมูล

2.1.3.2 การแยกประเภทข้อมูล (Data Classification) เป็นกระบวนการในการจำแนกข้อมูลตามมูลค่ารวมถึงผลกระทบที่มีต่อองค์กรเมื่อมีการรั่วไหลของข้อมูลจำเป็นต้องมีการจัดลำดับความสำคัญ ในระดับของการควบคุมที่เหมาะสมโดยมีการจำแนกตามการพัฒนามาตรฐาน หรือนโยบายสำหรับการจำแนกข้อมูล ระบุชนิดของข้อมูลโดยแบ่งตามหน่วยงาน ผู้ดูแลระบบ ผู้ใช้งานแต่ละชนิดข้อมูล การบำรุงรักษาระบบ การประมวลผล หรือการจัดเก็บข้อมูลแต่ละประเภท การจำแนกประเภทของข้อมูลจะเพิ่มตัวควบคุม เพื่อจำกัดการเข้าถึง และการโอนถ่ายข้อมูลสำคัญ ซึ่งจะช่วยลดปริมาณการรั่วไหลของข้อมูลภายในองค์กร

2.1.3.3 การเฝ้าระวังข้อมูล (Data Monitor) เป็นกระบวนการในการตรวจสอบช่องทางการสื่อสารต่างๆ ขององค์กร การเข้าเว็บไซต์ การใช้งานอีเมลล์ การใช้งานโปรแกรมสนทนา และอื่นๆ ตามนโยบายที่กำหนด เพื่อช่วยตรวจสอบข้อมูลในลักษณะของข้อมูลที่มีการถ่ายโอน ซึ่งจะบอกถึงต้นทาง ปลายทาง และช่องทางที่ทำให้ข้อมูลรั่วไหลออกไปได้ รวมทั้งเพิ่มความสามารถในการจัดการ และควบคุมข้อมูลรั่วไหลในระดับการใช้งานที่เครื่องคอมพิวเตอร์ส่วนบุคคล เป็นการบริหารจัดการเครื่องลูกข่าย และอุปกรณ์ต่อพ่วงต่างๆ ที่ข้อมูลถูกจัดเก็บไว้บนตัวเครื่องและที่อยู่บนระบบเครือข่าย รวมไปถึงการกำหนดสิทธิ์ของแต่ละบุคคลในการเข้าถึงข้อมูล และการใช้งานอุปกรณ์ต่อพ่วง

2.1.3.4 การป้องกันข้อมูล (Data Protection) เป็นกระบวนการที่ดำเนินการ ตามนโยบายที่ได้กำหนดไว้ในเชิงป้องกัน ที่ระบุต้นทาง ปลายทาง และวิธีการที่อาจทำให้ข้อมูลรั่วไหลออกไปได้ ซึ่งจะเชื่อมโยงนโยบายการป้องกันข้อมูลกับกระบวนการทางธุรกิจ ให้สามารถรู้ได้ว่าใคร ทำอะไร ที่ไหน อย่างไร ทำให้สามารถป้องกันข้อมูลสำคัญ และข้อมูลความลับในลักษณะของข้อมูลที่ใช้งาน และข้อมูลที่มีการถ่ายโอนได้โดยกำหนดให้มีการตอบสนองต่อการเข้าถึงข้อมูล เช่น การจำกัดการใช้งานกักข้อมูล เพื่อพิจารณาก่อนอนุญาตให้กระทำการใดๆกับข้อมูลหรือการเข้ารหัสข้อมูลก่อนการถ่ายโอน รวมทั้งแจ้งเตือนไปยังผู้เกี่ยวข้อง หรือเจ้าของข้อมูลด้วย

จากการศึกษาพบว่าผลิตภัณฑ์ด้านการรักษาความมั่นคงปลอดภัย ในส่วนของการป้องกันข้อมูลรั่วไหลที่มีใช้งานอยู่ในปัจจุบัน มักจะใช้เครื่องมือที่แตกต่างกัน เช่น รูปแบบของข้อมูลสำคัญ ขั้นตอนวิธีการ และกลไกการออกรายงาน สำหรับระบบข้อมูลเพื่อสร้างโครงสร้างพื้นฐานให้เห็นภาพรวมของระบบการป้องกันข้อมูลรั่วไหล จากการค้นคว้าคุณลักษณะของผลิตภัณฑ์ต่างๆที่มีอยู่ในปัจจุบันพบว่าบางบริษัททั้ง Code Green Network, GTB Technologies, Vericept, Websense จะนำเสนอผลิตภัณฑ์ DLP ที่ทำในส่วนของการป้องกันข้อมูลรั่วไหลโดยตรง และในส่วนของผู้ผลิตรายใหญ่ เช่น IBM, McAfee, Symantec และ TrendMicro มีการเพิ่มฟังก์ชั่นของ DLP เพื่อขยายขอบเขตของความปลอดภัย โดยมีการควบรวม เช่น Cisco system ผ่านทาง IronPort ซึ่งเป็นบริษัทย่อยของทาง Cisco เอง โดยใส่คุณสมบัติของ DLP เข้าไปในผลิตภัณฑ์ที่เกี่ยวข้องกับการป้องกันเครือข่าย ดังจะนำเสนอคุณลักษณะความสามารถของผลิตภัณฑ์ต่างๆตารางที่ 2.1 จะเห็นได้ว่าผลิตภัณฑ์ที่มีอยู่ในปัจจุบันมีฟังก์ชั่นการใช้งาน ตามคุณสมบัติหลักของระบบการป้องกันข้อมูลรั่วไหลคือ การค้นหาข้อมูล การเฝ้าระวัง และการป้องกัน ซึ่งในแต่ละคุณสมบัติจะมีการนำเทคโนโลยีเข้ามาช่วย ในการแยกประเภทข้อมูลเพื่อระบุ หรือจำแนกข้อมูลที่จะกำหนดให้เป็นข้อมูลระดับไหนข้อมูลที่เป็นความลับ หรือข้อมูลทั่วไป

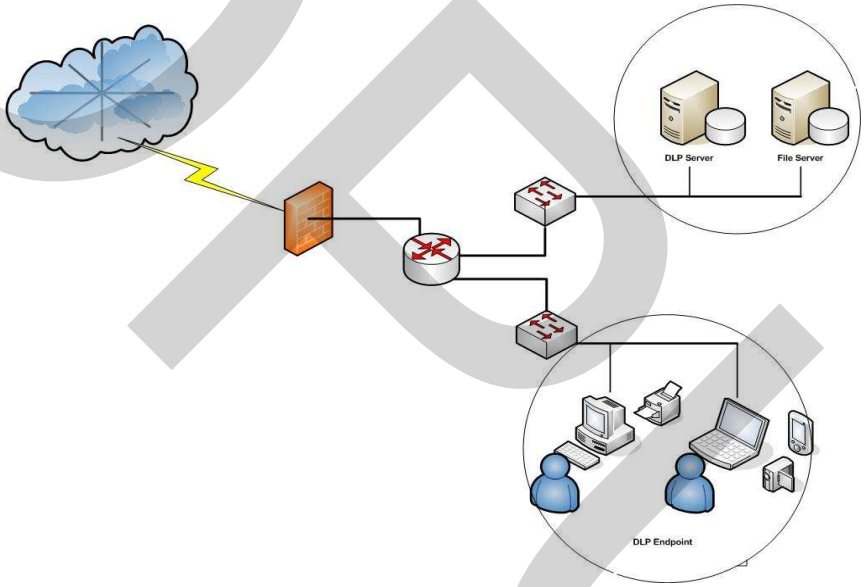
ตารางที่ 2.1 คุณสมบัติการทำงานของผลิตภัณฑ์ DLP ที่มีจำหน่ายอยู่ในปัจจุบัน

Solution	Classificate	Discover	Monitor											Data Protection				
			Data In Motion					Data In Use										
			TCP	SMTP	FTP	HTTP	HTTPS	USB	CD/DVD	Copy/Past	PrintScreen	Print	Fax	Allow	Block	Notification	Quanrantine	Encryp
BitArmor	✓	✓	✓	✓	✓			✓	✓						✓			✓
Bluecoat	✓	✓	✓	✓											✓			
CA	✓	✓	✓	✓	✓			✓	✓		✓	✓			✓	✓	✓	
Checkpoint								✓	✓									✓
CiscoIronport			✓	✓	✓	✓	✓							✓	✓	✓		
CodeGreen		✓	✓	✓	✓	✓	✓	✓	✓			✓		✓	✓	✓	✓	✓
EndpointProtection			✓			✓	✓	✓	✓			✓			✓	✓		✓
GTB		✓						✓	✓	✓				✓	✓			✓
NextTier	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓					✓
NextLabs	✓	✓		✓	✓	✓	✓	✓										✓
PalisadeSystem	✓	✓	✓	✓	✓	✓	✓								✓		✓	✓
RSA	✓	✓	✓	✓						✓		✓			✓	✓		✓
Sophos	✓		✓					✓	✓					✓	✓			✓
Symantec		✓			✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓
Trustwave	✓	✓	✓	✓	✓	✓	✓	✓	✓						✓	✓	✓	✓
Vericept	✓	✓						✓	✓					✓		✓		✓
Websense	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓	✓

ระบบการป้องกันข้อมูลรั่วไหล มีการออกแบบเพื่อให้ผู้ใช้งานตรวจสอบการเข้าถึงข้อมูลโดยโปรแกรมบนเครือข่าย จำกัดการใช้งาน เช่นการตรวจสอบช่องทาง และโปรแกรมการกรองเนื้อหาสำคัญ ก่อนที่จะสามารถส่งออกผ่านทางอุปกรณ์ต่อพ่วง โดยสามารถออกรายงานเหตุการณ์การเข้าสู่ระบบ หรือการจัดการความปลอดภัยของช่องทางการติดต่อ เพื่อแก้ไขปัญหาการ

สูญเสียข้อมูลสำคัญ การควบคุมการเคลื่อนไหวของข้อมูล โหมดในการติดต่อสื่อสารทางอิเล็กทรอนิกส์ผ่านตัวกรองที่กำหนดไว้ล่วงหน้า และถูกใช้กับข้อมูลที่มีการจำแนกประเภทข้อมูลไว้แล้ว เพื่อเพิ่มประสิทธิภาพ และความสามารถในการคัดกรองได้อย่างถูกต้อง

จากการศึกษาพบว่าในปัจจุบันยังไม่มีองค์กรใดออกมาทำงานเกี่ยวกับระบบงานนี้ และไม่มีมาตรฐานของระบบการป้องกันข้อมูลรั่วไหลอย่างชัดเจน ซึ่งถือว่าเป็นเรื่องยากสำหรับเครื่องมือในการป้องกันข้อมูลรั่วไหล เพื่อจะอธิบายรูปแบบของข้อมูลสำคัญ จึงได้นำเสนอการจำแนกประเภทเอกสาร (Data Classification) เพื่อจัดรูปแบบของข้อมูลก่อนที่จะนำเข้าสู่ระบบการป้องกันข้อมูลรั่วไหลต่อไป

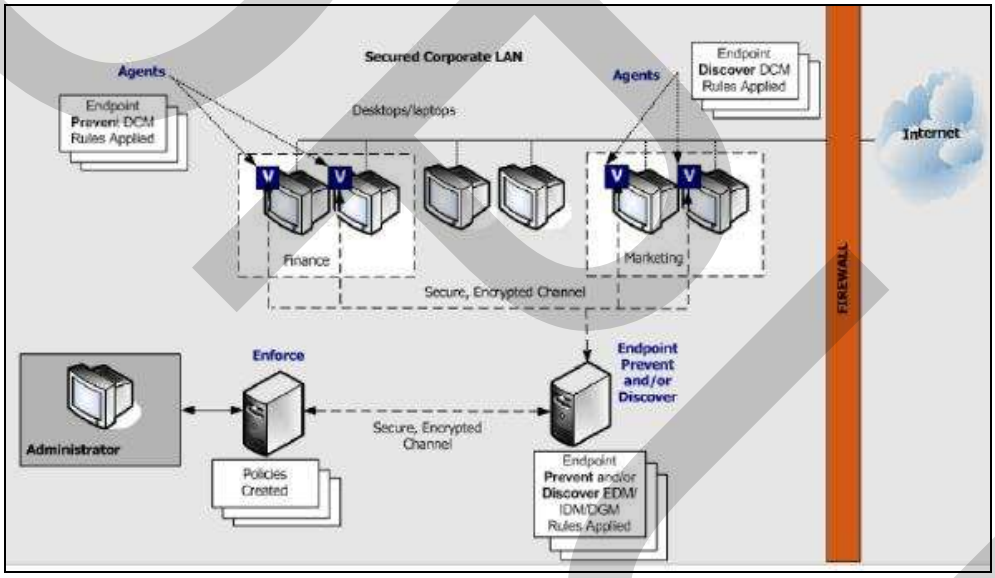


รูปที่ 2.2 โครงสร้างเครือข่ายของการใช้งานระบบ DLP

2.2 Symantec DLP Solution

Symantec Data Loss Prevention คือระบบการรักษาความปลอดภัยให้กับข้อมูลสำคัญขององค์การ ซึ่งมีขั้นตอนการทำงานหลักๆ คือ การค้นหาข้อมูล และการเฝ้าระวังข้อมูลสำคัญ โดยในขั้นตอนของการค้นหาข้อมูลสำคัญ จะทำการตรวจหาข้อมูลสำคัญผ่านทางโปรแกรมย่อย (Agent) ที่ติดตั้งไว้ที่เครื่องลูกข่าย และจะทำการเฝ้าระวังอยู่อย่างสม่ำเสมอ โดยการกำหนดความถี่ของการตรวจสอบผลที่ได้คือทราบตำแหน่งที่อยู่ของเอกสารสำคัญ และจะมีการรวบรวมตำแหน่งการจัดเก็บของเอกสารสำคัญไว้ และจะมีการบังคับใช้นโยบายการป้องกันข้อมูล โดยมีการวางโครงสร้างเครือข่ายดังรูปที่ 2.2

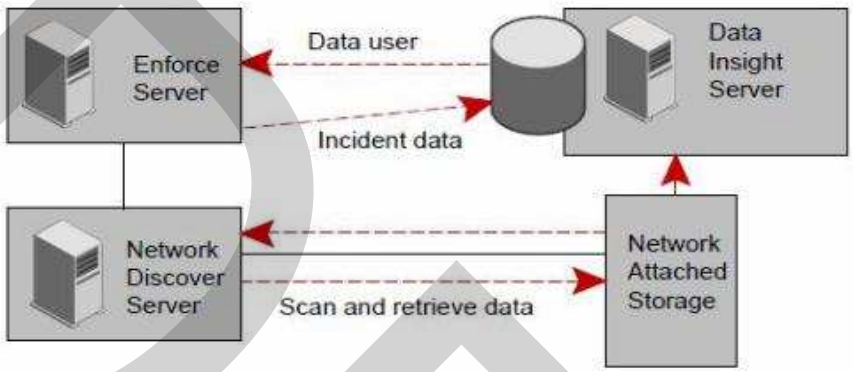
แสดงรูปแบบการติดตั้ง Symantec DLP Server (DLP Server + File Server) บนเครือข่าย จะบอกถึงตำแหน่งของเครื่องปลายทางที่ติดตั้งโปรแกรมย่อย (Agent) ไว้เพื่อเฝ้าระวังข้อมูลสำคัญ DLP Endpoint และอุปกรณ์ทางเครือข่ายที่เข้ามาเกี่ยวข้องกับระบบการป้องกันข้อมูลรั่วไหล โดยในลักษณะการวางจะอยู่ในระบบเครือข่ายภายในหลังอุปกรณ์ Firewall ในส่วนของขั้นตอนการเฝ้าระวังนั้นจะทำการตรวจสอบการใช้งานข้อมูล ไม่ว่าจะเป็นการใช้งานอีเมลล์ เว็บไซต์ โปรแกรมสนทนา หรือการโอนถ่ายข้อมูล โดยจะทำการตรวจสอบเนื้อหาของข้อมูลสำคัญ และแจ้งเตือนเมื่อพบเหตุการณ์ต้องสงสัย ผลที่ได้คือทราบถึงกิจกรรมการใช้งานที่เกิดขึ้นในเครือข่าย ทำให้การทำงานร่วมกับหน่วยงานภายนอกองค์กร เช่น ลูกค้าหรือคู่ค้า มีความปลอดภัยมากขึ้น



รูปที่ 2.3 แผนผังการทำงานของ Symantec Data Loss Prevention

ระบบโครงสร้างของ Symantec DLP ดังแสดงในรูปที่ 2.3 ประกอบไปด้วยส่วนหลักๆ 2 ส่วนคือ DLP Server ซึ่งทำหน้าที่ในการบังคับใช้นโยบาย(Enforce) และในส่วนของโปรแกรมย่อย (Agent) ที่ติดตั้งอยู่ที่เครื่องลูกข่ายเพื่อรองรับนโยบายที่ได้รับจากส่วนกลาง รวมทั้งทำหน้าที่ในการรับคำสั่งในการตรวจสอบการใช้งานข้อมูลสำคัญดังที่ได้กล่าวไปข้างต้นด้วย การทำงานร่วมกันของ DLP Server และโปรแกรมย่อยที่ติดตั้งไว้ที่เครื่องลูกข่าย หรือคอมพิวเตอร์ส่วนบุคคล ดังแสดงในรูปที่ 2.4 แสดงการไหลของข้อมูลระหว่าง DLP Server ซึ่งแบ่งการทำงานออกเป็น 3 ส่วน คือในส่วนงานการบังคับใช้นโยบาย (Enforce Server) การค้นหาข้อมูลสำคัญ (Network

Discover Server) จะทำการสแกนไฟล์ และโฟล์เดอร์ที่ถูกจัดเก็บในเครือข่าย และเฝ้าระวังการใช้งานข้อมูลสำคัญ (Data Insight Server) ที่จัดเก็บข้อมูลที่กำหนดไว้ให้เป็นข้อมูลสำคัญ และจัดเก็บข้อมูลในส่วนผู้ใช้งานด้วย



รูปที่ 2.4 การทำงานระหว่าง Data Insight และ Symantec Data Loss Prevention Server

รูปแบบของนโยบายเกี่ยวกับเอกสารสำคัญใน Symantec DLP จะอยู่ในรูปแบบของ IDM Rule หรือการทำดัชนีเอกสารสำคัญ กฎนี้จะมองเนื้อหาจากเอกสารเฉพาะที่มีการลงทะเบียนไว้ (เอกสารสำคัญ) และจะมีการตรวจสอบความเหมือนต้องมีค่าเป็น 80% หรือมากกว่าเมื่อเทียบกับเอกสารต้นฉบับ โดยจะประกอบไปด้วย 2 เงื่อนไขซึ่งทั้งสองเงื่อนไขนี้จะต้องตรงกับกฎที่ตั้งไว้ถึงจะเรียกว่าเหตุการณ์ต้องสงสัย เงื่อนไขแรกเกิดจากการรวมกันของคำหลักจากคำหลักที่เป็นความลับ (Confidential Keywords) ตามเอกสารต้นฉบับที่ได้กำหนดไว้แล้วว่าเป็นความลับ ตามการเปรียบเทียบกับคำในพจนานุกรมที่กำหนดไว้ หรือตามชนิดของไฟล์ที่ได้กำหนด และระบบรองรับ เช่น Excel_macro, xls, works_spred, sylk, Quattro pro, mod, csv, applix_spread, 123, doc, pdf, ppt จะมีในส่วนของการกำหนดการบังคับใช้เมื่อเกิดเหตุการณ์ต้องสงสัยดังรูปที่ 2.5 ซึ่งแบ่งออกเป็น 4 การตอบสนองต่อการตรวจพบเหตุการณ์ต้องสงสัย การเก็บรักษาข้อมูลเหตุการณ์ที่เกิดขึ้น (Limit incident data retention) ช่วยในการระบุส่วนของเนื้อหาที่จะจัดเก็บในการบันทึกเหตุการณ์ที่เกิดขึ้น การป้องกันที่จุดปลายทาง การปิดกั้น (Block) โดยจะบล็อกหรือปิดกั้นการถ่ายโอนข้อมูลที่ละเมิดนโยบายที่ได้กำหนดไว้ การป้องกันที่จุดปลายทาง การแจ้งเตือน (Notify) ที่หน้าจอจะแสดงผลการแจ้งเตือนไปยังผู้ใช้งานปลายทางเมื่อมีข้อมูลที่เป็นความลับกำลังจะถูกถ่ายโอนผ่านโปรโตคอลไปยังปลายทางการป้องกันที่จุดปลายทาง การยกเลิกโดยผู้ใช้งาน (User

Cancel) ช่วยให้ผู้ใช้งานสามารถยกเลิกนโยบายเพื่อที่จะทำการถ่ายโอนข้อมูลที่ซึ่งละเมิดนโยบายออกไปได้ ซึ่งถือว่าเป็นตัวเลือกที่ต้องใช้งานอย่างระมัดระวังมาก

Response Rule	
All: Limit Incident Data Retention	Lets you specify the parts of the content to retain in the incident record. See "All: Limit Incident Data Retention response rule action" on page 326.
Endpoint Prevent: Block	Blocks the transfer of data that violates the policy. For example, a copy of confidential data from an endpoint computer to removable media (a USB flash drive, for example). Optionally, it displays an on-screen notification to the endpoint user. See "Endpoint Prevent: Block response rule action" on page 330.
Endpoint Prevent: Notify	Displays an on-screen notification to the endpoint user when confidential data is transferred from the endpoint protocol or destination (except local drive). See "Endpoint Prevent: Notify response rule action" on page 332.
Endpoint Prevent: User Cancel	Lets users override a policy to allow transfer of data that violates the policy. The override option is time sensitive. See "Endpoint Prevent: User Cancel" on page 335.

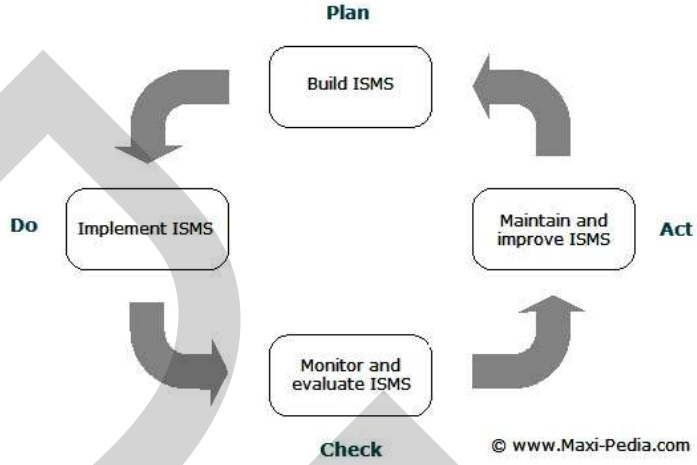
รูปที่ 2.5 การตอบสนองแบบอัตโนมัติต่อเหตุการณ์ต้องสงสัย

2.3 มาตรฐานด้านความมั่นคงปลอดภัยทางด้านสารสนเทศ ISO/IEC27001:2005

การนำระบบคอมพิวเตอร์มาใช้ในการจัดการ เรื่องของข้อมูลอย่างแพร่หลาย ไม่ว่าจะเป็นอินเตอร์เน็ต การสื่อสารอิเล็กทรอนิกส์ ก่อให้เกิดความเสี่ยงสำคัญรูปแบบหนึ่งคือการโจรกรรมข้อมูล หรือการสูญหายของข้อมูลที่มีความสำคัญ ISO/IEC 27001:2005 เป็นมาตรฐานที่ดีที่สุด ในการจัดการด้านระบบรักษาความปลอดภัยข้อมูลองค์กร ด้วยมาตรฐานนี้ สามารถสร้างกลยุทธ์ และกำหนดทิศทางสำหรับการประเมิน การวัดค่า และการป้องกันการคุกคามจากภายนอก โดยผ่านกระบวนการจัดการความเสี่ยงของมาตรฐานได้

ISO/IEC 27001:2005 หรือ Information Security Management System (ISMS) ซึ่งข้อกำหนดต่างๆกำหนดขึ้นโดยองค์กรที่มีชื่อเสียง และมีความน่าเชื่อถือระหว่างประเทศ คือ ISO (The International Organization for Standardization) และ IEC(The International Electrotechnical Commission) เป็นมาตรฐานสากลที่กล่าถึงมาตรฐานของระบบบริหารจัดการเพื่อความมั่นคงปลอดภัยของข้อมูล มีพื้นฐานมาจากแนวทางการจัดการความเสี่ยง (Risk Approach) เพื่อรักษาไว้ซึ่งความลับ (Confidentiality) ความถูกต้องครบถ้วน (Integrity) และความพร้อมใช้งาน (Availability)

ของข้อมูลสารสนเทศ รวมทั้งทรัพย์สินอื่นๆ ที่มีความสำคัญขององค์กรตามหลัก Plan-Do-Check-Action (PDCA Model) แสดงในรูปที่ 2.6



รูปที่ 2.6 Plan-Do-Check-Act Cycle

ที่มา: http://www.maxi-pedia.com/PDCA

เพื่อให้เกิดวิธีปฏิบัติงานที่เป็นระบบ และมีการพัฒนาขึ้นอย่างต่อเนื่อง (Continuous Improvement)โดยเริ่มตั้งแต่ การสร้าง การดำเนินงาน การนำระบบมาใช้ การตรวจสอบ การวัดผล การทบทวน การบำรุงรักษา และการปรับปรุงระบบบริหารความมั่นคงปลอดภัย โดยจุดประสงค์ของมาตรฐานนี้เพื่อจะทำให้องค์กร สามารถบริหารจัดการทางด้านความปลอดภัยได้อย่างมีระบบและเพียงพอเหมาะสมต่อการดำเนินธุรกิจขององค์กร โดยเริ่มแรกองค์กรต้องทำการวิเคราะห์ความเสี่ยงของระบบจากภัยคุกคาม และจุดอ่อนต่างๆในระบบจากนั้นจึงวิเคราะห์และเลือกแนวทางการควบคุม และป้องกันสารสนเทศต่างๆอย่างเหมาะสม

2.3.1 ISO/IEC 27001:2005 หรือ Information Security Management System (ISMS) เป็นระบบการจัดการความปลอดภัยข้อมูลเพื่อให้ระบบข้อมูลสารสนเทศขององค์กรมีคุณสมบัติในด้านต่างๆดังต่อไปนี้

2.3.1.1 Confidentiality ข้อมูลต่างๆ สามารถเข้าถึงได้เฉพาะผู้ที่มีสิทธิที่จะเข้าเท่านั้น

2.3.1.2 Integrity ข้อมูลมีความถูกต้อง ครบถ้วนสมบูรณ์ โดยไม่ได้ถูกเปลี่ยนแปลงหรือแก้ไข จากผู้ที่ไม่ได้รับอนุญาต

2.3.1.3 Avalilability ข้อมูลพร้อมที่จะใช้งานโดยผู้ที่มีสิทธิในการเข้าถึงข้อมูลสามารถเข้าถึงข้อมูลได้ทุกเมื่อหากต้องการ

2.3.2 สำหรับระบบ ISO/IEC 17799:2005 เป็นกรอบด้านการควบคุมระบบ ความปลอดภัยข้อมูล ซึ่งแบ่งออกเป็น 11 การควบคุมหลัก ดังต่อไปนี้

2.3.2.1 Security Policy นโยบายความมั่นคงปลอดภัยขององค์กร

2.3.2.2 Organization Information Security โครงสร้างความมั่นคงปลอดภัยขององค์กร

2.3.2.3 Asset Management การจัดหมวดหมู่ และการควบคุมทรัพย์สินขององค์กร

2.3.2.4 Human Resource Security มาตรฐานของบุคลากรเพื่อสร้างความมั่นคงปลอดภัยให้กับองค์กร

2.3.2.5 Physical and Environment Security ความมั่นคงปลอดภัยทางกายภาพ และสิ่งแวดล้อมขององค์กร

2.3.2.6 Communications and operations management เป็นการบริหารจัดการด้านการสื่อสาร และการดำเนินงานของเครือข่ายสารสนเทศขององค์กร

2.3.2.7 Access Control การควบคุมการเข้าถึงระบบสารสนเทศขององค์กร

2.3.2.8 Information System acquisition, development and maintenance การพัฒนาและดูแลระบบสารสนเทศ

2.3.2.9 Information security incident management การบริหารจัดการเหตุการณ์ละเมิดความมั่นคงปลอดภัย

2.3.2.10 Business continuity management การบริหารความต่อเนื่องในการดำเนินงานขององค์กร

2.3.2.11 Compliance การปฏิบัติตามข้อกำหนดทางด้านกฎหมาย และบทลงโทษของการละเมิดนโยบาย

โดยระบบ ISO/IEC 27001:2005 จะต้องมีการแนะนำให้ประยุกต์ตามข้อกำหนดของ ISO/IEC17799:2005 มาใช้ในการควบคุมและจัดการเกี่ยวกับความเสี่ยงที่เกิดขึ้น (ตามข้อกำหนด 4.2g ของ ISO/IEC27001:2005) ระบบการจัดการความปลอดภัยข้อมูล ISO/IEC27001:2005 หรือ ISMS เป็นระบบ dynamic system ที่มีการประยุกต์หลักการ PDCA Cycle ที่สามารถประยุกต์ใช้ได้กับทุกธุรกิจ เพื่อให้ระบบข้อมูลขององค์การ มี Confidentiality ให้แน่ใจว่าข้อมูลต่างๆ สามารถเข้าถึงได้เฉพาะผู้ที่มีสิทธิ์ที่จะเข้าเท่านั้น มี Integrity ป้องกันให้ข้อมูล มีความถูกต้อง และความสมบูรณ์ และ Availability ผู้ที่มีสิทธิ ในการเข้าถึงข้อมูล สามารถเข้าถึงได้เมื่อมีความต้องการ โดยไม่ใช่ให้ระบบไม่มีความเสี่ยงเลย หรือไม่เกิดปัญหาเลย ทำให้เกิดประสิทธิภาพในการใช้ทรัพยากรในการลงทุนสำหรับการจัดการความปลอดภัยของข้อมูลอย่างมีประสิทธิภาพ โดยส่วนใหญ่จะมีการใช้งานร่วมกับ ระบบ ISO/IEC17799:2005 เพื่อให้เกิดประสิทธิภาพในการดำเนินงาน

2.4 กฎหมายกับความมั่นคงปลอดภัยสารสนเทศ

พระราชกฤษฎีกากำหนดหลักเกณฑ์ และวิธีการในการทำธุรกรรมทางอิเล็กทรอนิกส์ภาครัฐ พ.ศ.2549 ประกาศบังคับใช้เมื่อ 10 มกราคม พ.ศ.2550 การประกาศบังคับใช้กฎหมายฉบับนี้เป็นกฎหมายลำดับรองของ พรบ.ว่าด้วยธุรกรรมอิเล็กทรอนิกส์ โดยมีจุดมุ่งหมายเพื่อให้การทำธุรกรรมอิเล็กทรอนิกส์ภาครัฐมีผลทางกฎหมาย เพื่อพัฒนาการทำธุรกรรมทางออนไลน์ของภาครัฐให้อยู่ภายใต้มาตรฐาน หรือทิศทางเดียวกัน การจัดทำแนวนโยบาย/แนวทางปฏิบัติในการรักษาความมั่นคงปลอดภัยสารสนเทศ ทั้งในด้าน Physical และ Electronically โดยต้องมีการควบคุมการเข้าถึง และใช้สารสนเทศ การทำการสำรองข้อมูลในสภาพพร้อมใช้งาน และมีแผนฉุกเฉิน รวมไปถึงการตรวจสอบ และประเมินความเสี่ยงอย่างต่อเนื่อง (มาตรา 5) จัดทำแนวนโยบาย/แนวทางปฏิบัติในการคุ้มครองข้อมูลส่วนบุคคล ในกรณีที่มีการรวบรวมจัดเก็บ ใช้หรือเผยแพร่ข้อมูลที่ระบุตัวตนได้ (มาตรา 6)

2.4.1 บทบาทหน้าที่ของผู้ใช้บริการ เนื้อความของ พรบ.จะครอบคลุมไปยังบุคคลที่ถือได้ว่าเป็นผู้ใช้บริการทุกคน มีประเด็นความผิดสำคัญ ดังนี้ การเข้าถึงระบบคอมพิวเตอร์ผู้อื่นโดยมิชอบ (มาตรา 5) การล่วงรู้ และเปิดเผยข้อมูลมาตรการป้องกันการเข้าถึงระบบคอมพิวเตอร์ที่ผู้อื่นจัดทำขึ้นเป็นการเฉพาะ (มาตรา 6) การเข้าถึงข้อมูลคอมพิวเตอร์โดยไม่ชอบ (มาตรา 7) การลักลอบดักข้อมูลคอมพิวเตอร์ของผู้อื่น (มาตรา 8) เช่นการทำ Packet Sniffing การทำให้เสียหาย ทำลาย แก้ไขเปลี่ยนแปลง เพิ่มเติมข้อมูลคอมพิวเตอร์ของผู้อื่น โดยไม่ชอบ (มาตรา 9) การกระทำเพื่อให้การทำงานของระบบคอมพิวเตอร์ของผู้อื่นไม่สามารถทำงานได้ตามปกติ (มาตรา 10) เช่นการทำ Denial of Services เป็นต้น การจำหน่าย หรือเผยแพร่โปรแกรมที่ใช้ในการกระทำความผิด (มาตรา 13) การเผยแพร่ โดยการนำข้อมูลปลอม เท็จ ลามก เข้าสู่ระบบคอมพิวเตอร์ ซึ่งก่อให้เกิดความเสียหายต่อผู้อื่น ต่อความมั่นคงของประเทศ ต่อความสงบสุขของประชาชน (มาตรา 14) การเผยแพร่ภาพจากการตัดต่อ ดัดแปลง ตกแต่งข้อมูลคอมพิวเตอร์ที่เป็นภาพของบุคคล (มาตรา 16) อันทำให้ผู้อื่นเสียชื่อเสียง ถูกดูหมิ่น เกลียดชัง หรือได้รับความอับอาย ฐานความผิดของ พรบ. ฉบับนี้ถือเป็นความผิดทางอาญา

2.5 เทคโนโลยีการแยกประเภทเอกสาร (Document Classification)

เทคนิคการแยกประเภทเอกสาร นำมาปรับใช้ในการคัดแยกเอกสารภาษาไทย โดยพิจารณาจากรูปแบบของความน่าจะเป็นที่เกี่ยวข้องกับการตัดสินใจของมนุษย์ ซึ่งประกอบด้วยขั้นตอนการคัดแยกตั้งแต่ขั้นตอนการตัดคำในเอกสารภาษาไทย การให้น้ำหนักคำของแต่ละเอกสาร และการนำค่าน้ำหนักเหล่านั้นมาเป็นอินพุตของระบบการคัดแยกประเภทเอกสาร โดยกำหนด

ลักษณะสำคัญในการแยกประเภทเอกสารเป็นอีกอินพุตหนึ่งของระบบการคัดแยกเพื่อระบุประเภทของเอกสาร

2.5.1 การตัดคำในเอกสารภาษาไทย (Thai Word Segmentation) คือกระบวนการในการแยกแต่ละคำในเอกสารภาษาไทย ซึ่งประกอบไปด้วยตัวหนังสือภาษาไทย ตัวหนังสือภาษาอังกฤษ ตัวเลข และสัญลักษณ์พิเศษต่างๆออกมาเป็นแต่ละคำเพื่อนำไปใช้ประโยชน์ในการหาความถี่ของคำ ลักษณะของประโยคในภาษาไทยประกอบไปด้วยคำ ซึ่งในแต่ละคำจะประกอบไปด้วยส่วนต่างๆ แบ่งออกได้เป็นสามส่วน สี่ส่วน และห้าส่วน ตามอักขระวิธี ได้แก่พยัญชนะ สระ และวรรณยุกต์ เช่นคำว่า กา ก่า ก้า ก๊า ก๋า เป็นต้น โดยคำว่า กา ไม่มีรูปวรรณยุกต์ แต่มีเสียงวรรณยุกต์สามัญ ส่วนแบบสี่ส่วน จะเพิ่มเติมตัวสะกดมา เช่น กาย บิน รวม และสุดท้ายแบบห้าส่วน จะเพิ่มในส่วนของการันต์ได้แก่คำว่า การณ์ จลน์ เป็นต้น การประสมคำกันระหว่างตัวอักษรจะมีหลักการเพิ่มเติมอีก เช่นการใช้ อักษรสูง อักษรกลาง และอักษรต่ำ หรือการวางตำแหน่งของสระในแต่ละคำ

การตัดคำภาษาไทยโดยการใช้อักขระวิธีในการตัดคำนั้นสามารถทำได้ในระดับหนึ่งเท่านั้นเนื่องจากคำบางคำเป็นคำที่เลียนเสียงจากภาษาต่างประเทศ ดังนั้นอาจจะเกิดจากการประสมคำนอกเหนือจากอักขระวิธี โดยเทคนิคการตัดคำโดยทั่วไปจะแบ่งออกเป็น 3 แบบใหญ่ๆ คือ วิธีการตัดคำที่ยาวที่สุด (Longest Matching) เป็นการตัดคำด้วยการค้นหาคำเริ่มจากตัวอักษรที่อยู่ซ้ายสุดของข้อความนั้นไปยังตัวอักษรถัดไปจนกว่าจะพบคำที่ยาวที่สุดที่มีอยู่ในพจนานุกรม วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) เป็นวิธีการตัดคำโดยคำที่ตัดได้เหล่านั้นอาจจะเป็นไปได้ทั้งหมด แล้วเลือกข้อความที่ตัดได้จำนวนคำน้อยที่สุดมาใช้งาน วิธีการตัดคำแบบคำนวณเชิงสถิติเพื่อหาความเป็นไปได้ (Probabilistic Model) วิธีการนี้นำเอาค่าสถิติของการเกิดของคำนั้นๆและลำดับหน้าที่ของคำ เข้ามาช่วยในการคำนวณหาความน่าจะเป็นเพื่อที่จะเลือกแบบที่มีโอกาสการเกิดมากที่สุด ซึ่งสามารถตัดคำได้ดีกว่า 2 แบบแรก แต่ข้อจำกัดของวิธีการนี้คือ จะต้องมีฐานข้อมูลที่มีการตัดคำที่ถูกต้องและมีการกำหนดหน้าที่ของคำ เพื่อนำไปใช้ในการสร้างสถิติ และวิธีการตัดคำแบบใช้คุณลักษณะ (Feature-Based Approach) วิธีการตัดคำแบบนี้จะพิจารณาจากบริบท (Context Words) และการเกิดร่วมกันของคำ หรือหน้าที่ของคำ (Collocation) เข้ามาช่วยในการตัดคำ ซึ่งโดยรวมแล้ววิธีการหลักสำหรับตัดคำในเอกสารภาษาไทยจะถูกพิจารณาตามขั้นตอนดังนี้

2.5.1.1 การใช้กฎ ลักษณะของการใช้กฎเพื่อตัดคำในภาษาไทย จะใช้ไวยากรณ์ทางภาษา โดยภาษาไทยจะแบ่งตัวอักษรเป็นหมวดหมู่ตามลักษณะการใช้งาน ได้แก่ กลุ่มพยัญชนะ

กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลข และกลุ่มตัวอักษรพิเศษ ขั้นตอนการตัดพยางค์จะทำจากซ้ายไปขวาเป็นส่วนใหญ่

2.5.1.2 การใช้พจนานุกรม ซึ่งถูกนำมาปรับใช้ในการวิจัยนี้ผลลัพธ์ที่ได้จะอยู่ในระดับคำ โดยมีหลักการว่าจะทำการตรวจสอบสายอักขระ (String) ซึ่งเป็นชุดของตัวอักษรที่ได้จากเอกสารโดยการค้นหาคำเริ่มจากอักษรตัวซ้ายสุดของข้อความนั้นไปยังอักษรตัวถัดไป และนำไปเทียบกับคำในพจนานุกรม หากพบคำในพจนานุกรมที่สามารถเป็นคำในสายอักขระนั้นได้มากกว่าหนึ่งคำ จะทำการเลือกคำที่ยาวที่สุด (Longest matching) หากอักษรตัวต่อมาไม่พบว่าเป็นคำที่ตรงในพจนานุกรมที่มีอยู่จะทำการย้อนกลับไปเลือกคำที่สั้นกว่าแทน เรียกวิธีการนี้ว่าการย้อนรอย (Back tracking) การค้นหาคำศัพท์จากพจนานุกรมจะนำคำที่ได้จากเอกสาร และนำไปแบ่งครั้งแรกด้วยช่องว่างระหว่างประโยค จากนั้นนำไปแยกในระดับพยางค์โดยการใช้กฎ ผลที่ได้จะถูกเก็บเป็นสองส่วน คือส่วนที่ต้องนำไปค้นหาต่อไปในพจนานุกรม และไม่ต้องนำไปค้นหาต่อ คำที่ไม่ต้องนำไปค้นหาต่อ ได้แก่คำดังนี้ คำที่ปรากฏสัญลักษณ์ - / , . ซึ่งจะถูกตัดออกในขั้นตอนการตัดคำหยุด (Stop Word) และในส่วนของคำที่จะนำไปค้นหาในพจนานุกรม การค้นหาคำหากพบมากกว่าหนึ่งคำปรากฏอยู่จะใช้เทคนิค Longest Matching เพื่อเลือกคำศัพท์ที่ยาวที่สุด

2.5.2 การให้น้ำหนักคำ (Word Weighting) วัตถุประสงค์เพื่อสร้างเนื้อหาของเอกสาร ให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเรียนรู้ได้ รูปแบบของค่าน้ำหนักจะถูกใช้ในการจำลอง Vector Space สำหรับการกำหนดความคล้ายคลึงกันของเอกสาร สามารถหาได้จากรูปแบบของความน่าจะเป็นที่เกี่ยวข้องกับการตัดสินใจของมนุษย์ เพื่อใช้ในกระบวนการเรียนรู้ลักษณะของตัวแทนเอกสาร โดยจะอยู่ในรูปแบบของเวกเตอร์น้ำหนักคำ TF-IDF (Term Frequency Inverse Document Frequency)

$$w_{ij} = tf_{ij} \times \log_2 \frac{N}{n_j} \qquad (2\text{-}1)$$

โดยที่ w_{ij} = น้ำหนักของคำ T_j ในเอกสาร D_i

tf_{ij} = ความถี่ของคำ T_j ในเอกสาร D_i

N = จำนวนเอกสารทั้งหมดในระบบ

n_j = จำนวนเอกสสารที่มีคำ T_j ปรากฏอยู่อย่างน้อยหนึ่งครั้ง

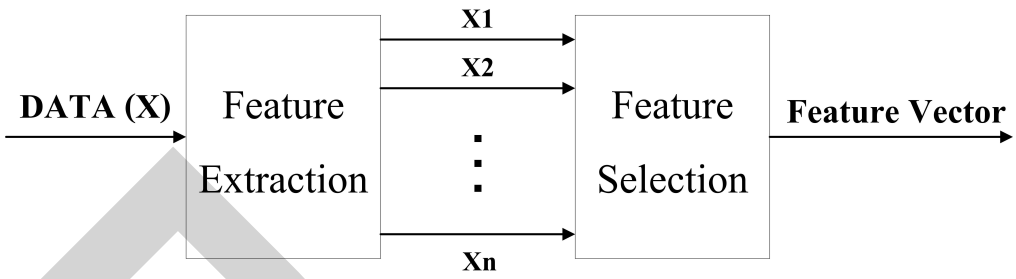
เป็นวิธีที่คำนวณน้ำหนักจากความถี่ ของการปรากฏของคำ T_j ในเอกสาร D_i และพิจารณาความถี่ของคำ T_j ที่ปรากฏในเอกสารอื่นร่วมด้วย โดยมีแนวคิดที่ว่าคำที่ปรากฏใน

เอกสารน้อยฉบับจะมีค่าน้ำหนักสูง ส่วนคำที่ปรากฏในเอกสารหลายฉบับจะมีค่าน้ำหนักน้อย เนื่องจากเป็นค่าที่ไม่แสดงลักษณะเฉพาะของเอกสารนั้น ดังสมการ 2-1

พารามิเตอร์ n_j ในสมการจะหาค่าได้เมื่อระบบได้รู้ทุกเอกสารที่มีในระบบ ตัวอย่างเช่น ถ้าทำการพิจารณาเอกสารที่ประกอบไปด้วยคำ 100 คำที่ปรากฏคำที่กำหนด 3 ครั้ง ทำการหาค่าตามสูตร เทอมของความถี่ tf_{ij} สำหรับคำที่เกิดขึ้น (3/100)=0.03 ถ้ามีเอกสารจำนวน 1,000 เอกสารและคำๆนั้นปรากฏใน 100 เอกสารสามารถหาค่าความถี่ของเอกสารได้จาก $\log(1{,}000/100) = 10$ ดังนั้นค่าน้ำหนัก tf-idf หรือ $w_{ij} = 0.03 \times 10 = 0.3$

2.5.3 ขั้นตอนการตัดคำหยุด (stop word) ออกเพื่อกรองคำที่ไม่สื่อถึงความสำคัญของเอกสาร และลบสัญลักษณ์พิเศษ เช่น $,#,@,?,! , ตัวเลข หรือคำเชื่อมที่เป็นคำฟุ่มเฟือย เช่น เป็น อยู่ คือ และ ก็ ฯลฯ ดังนั้นคำที่เหลือจะถูกเลือกเป็นคำสำคัญ (significant word) และจะถูกเก็บไว้ และนำความถี่ของคำที่ได้ ไปเป็นตัวแทนของเอกสาร ในขั้นตอนนี้จะทำการกำหนดคำหยุดไว้ในอาเรย์ของโปรแกรม เพื่อกรองคำหยุดเหล่านี้ทิ้งก่อนนำคำที่เหลือไปประมวลผลในขั้นตอนต่อไป

2.5.4 การจำแนกประเภทเอกสาร มีวัตถุประสงค์เพื่อปรับปรุงความถูกต้อง และความรวดเร็วในการจำแนกเอกสาร อันเนื่องมาจากจำนวนเอกสารที่มีมากมาย ประกอบกับจำนวนกลุ่มของเอกสารที่มีหลายกลุ่ม เมื่อแปลงเอกสารให้อยู่ในรูปของเมตริกซ์ความถี่ของคำสำคัญ ปัญหาที่เกิดขึ้นคือจำนวนของคำสำคัญที่มีมาก ทำให้ใช้เวลานานในการจำแนกเอกสาร การแก้ปัญหาจำนวนคำสำคัญที่มีจำนวนมาก ประกอบไปด้วย การเลือกลักษณะสำคัญ (Feature Selection) คือ การเลือกคำสำคัญบางคำจากคำสำคัญทั้งหมด โดยพิจารณาจากค่าน้ำหนักของคำสำคัญนั้นๆ โดยค่าลักษณะสำคัญจะอยู่ในรูปของเวกเตอร์ ซึ่งจะเรียกเวกเตอร์นี้ว่าเวกเตอร์ลักษณะสำคัญ (Feature Vector) เป็นจำนวนที่นำมาใช้ในการจำแนกวัตถุหรือเหตุการณ์ จำนวนของลักษณะสำคัญขึ้นอยู่กับผู้ใช้ต้องการความละเอียดของการแบ่งแยกระหว่างกลุ่มข้อมูลอยู่ในระดับใด อีกวิธีคือ วิธีการสกัดลักษณะสำคัญ (Feature Extraction) เป็นวิธีการแปลงลักษณะของเวกเตอร์ค่าน้ำหนักของคำสำคัญ ให้อยู่ในรูปแบบใหม่ที่มีมิติที่น้อยลงผลจากการสกัดลักษณะสำคัญมีสองประการ คือ สามารถลดจำนวนข้อมูลนำเข้า โดยมีน้อยกว่าคำสำคัญทั้งหมด การแทนเอกสารให้อยู่ในรูปแบบที่มีขนาดเล็กลง ด้วยการคัดเลือก และสกัดลักษณะสำคัญ สามารถแสดงได้ดังรูปที่ 2.7



รูปที่ 2.7 กระบวนการของการสกัดลักษณะสำคัญและแบ่งกลุ่มข้อมูล

ลักษณะสำคัญที่ดีนั้นควรมีลักษณะที่เหมือนกันสำหรับแบบอย่างทั้งหมดในกลุ่มเดียวกัน และแตกต่างกันสำหรับแบบอย่างที่อยู่ต่างกลุ่มกัน ในการแบ่งประเภทเอกสารที่ใช้ในวิทยานิพนธ์ฉบับนี้ จะทำการแบ่งประเภทตามลักษณะสำคัญของเอกสาร โดยจะแบ่งออกเป็น 2 ประเภท คือ

2.5.4.1 เอกสารสำคัญที่ถือเป็นเอกสารสำคัญ (Secret)

2.5.4.2 เอกสารทั่วไปที่ถือว่าไม่เป็นเอกสารสำคัญ (No Secret)

เพื่อให้สามารถแยกประเภทของเอกสารออกจากกันได้ ดังนั้นจะได้เวกเตอร์ลักษณะสำคัญ X ดังนี้

$$X = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ . \\ . \\ . \\ x_n \end{bmatrix} \qquad (2\text{-}2)$$

เมื่อ $\{x_1, x_2, ..., x_n\}$ คือลักษณะสำคัญ (ค่าน้ำหนัก) เป็นคำสำคัญที่ถูกเลือกมาเป็นตัวแทนของเอกสารเพื่อการจำแนกหมวดหมู่

2.5.5 การเลือกคุณลักษณะ (Feature Selection) การเลือกคำที่มีความสำคัญน้อยออก เพื่อดูประสิทธิภาพในการทำนายหลังจากที่ได้ตัดคำบางตัวออกซึ่งส่วนใหญ่จะให้ค่าความถูกต้องสูงขึ้น เพราะคำสำคัญที่เหลือจะเป็นคำที่มีเป็นตัวแทนที่ดีของเอกสาร การลดขนาดเอกสารคือการนำคำที่

ไม่มีนัยสำคัญออก ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้สร้างตัวจำแนกหมวดหมู่ โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี การลดขนาดจึงเป็นขั้นตอนหนึ่งที่จะต้องทำก่อน

2.5.6 การสกัดคุณลักษณะ (Feature Extraction) เป็นการดึงคุณลักษณะ (Feature) ของเอกสารออกมา กับการลดขนาดเอกสารลง ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่าจะใช้อะไรเป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศ และต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของเอกสาร และใช้พื้นฐานค่าความถี่ หรือค่าน้ำหนักคำ เป็นค่าของคุณลักษณะ ตัวแทนคุณลักษณะของเอกสารจะเก็บอยู่ในรูปแบบเวกเตอร์ โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความถี่คำ (Word Frequency) หรือแทนด้วยค่าน้ำหนักคำ (Word Weighting) งานวิจัยนี้ใช้การเลือกคุณลักษณะแบบคำเดียว (Single word) ซึ่งได้จากขั้นตอนการตัดคำโดยใช้พจนานุกรมเรียบร้อยแล้ว ผลลัพธ์ที่ได้จากการตัดคำจะได้เป็นคำเดียวจำนวนมาก เพื่อคัดเลือกมาใช้เป็นตัวแทนเอกสารในการเรียนรู้ โดยใช้วิธีเลือกคำที่มีความถี่มากที่สุด หรือคำที่ปรากฏในเอกสารมากฉบับ

2.5.7 การสร้างดัชนี (Indexing) คอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์ใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือ การคำนวณค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร อาจจะเรียกได้ว่าการหาค่าน้ำหนัก การสร้างดัชนี จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม ซึ่งงานวิจัยนี้ใช้วิธีหาความถี่ของคำที่ปรากฏในเอกสาร ถ้าคำมีค่าความถี่มากจะส่งผลให้ได้ค่าน้ำหนักที่มีค่าสูงตามไปด้วยเมื่อถึงขั้นตอนนี้จะได้รูปแบบที่มีลักษณะของการแสดงความสัมพันธ์ระหว่างคำ และเอกสารทั้งหมดด้วยเวกเตอร์ 2 มิติ ดังตัวอย่างในรูปที่ 2.8

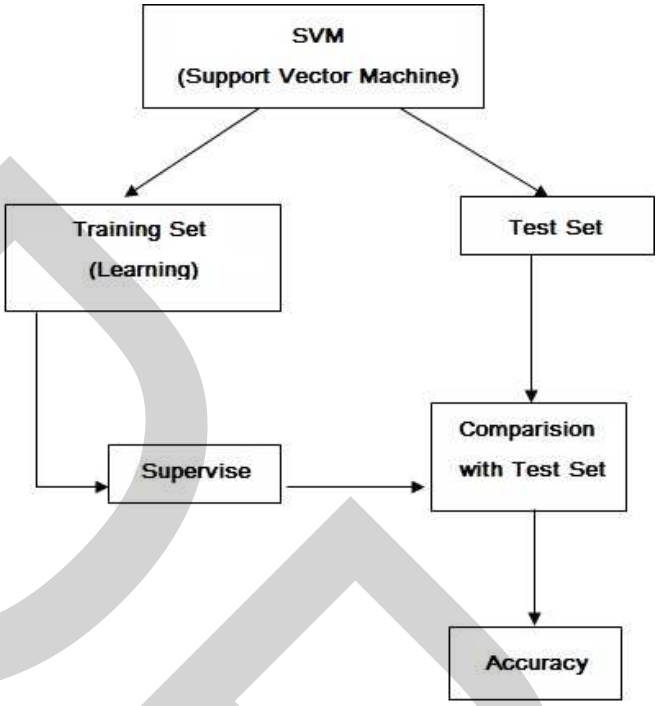
	W_1	W_2	...	W_k	...	W_V
D_1	W_{11}	W_{12}	...	W_{1k}	...	W_{1V}
D_2	W_{21}	W_{22}	...	W_{2k}	...	W_{2v}
D_3	W_{31}	W_{32}	...	W_{3k}	...	W_{3V}
...	...	...	...	...	...	...
D_N	W_{N1}	W_{N2}	...	W_{Nk}	...	W_{NV}

รูปที่ 2.8 รูปแบบแสดงความสัมพันธ์ระหว่างคำ และเอกสารทั้งหมดด้วยเวกเตอร์ 2 มิติ

2.6 ทฤษฎี SVM (Support Vector Machines)

เป็นเวกเตอร์ที่ระบุลักษณะเฉพาะ (Identifies) เช่น ใช้สำหรับการหาดัชนี เพื่อเป็นข้อมูลที่ใช้ในการกรองการสืบค้นข้อมูลในการทำดัชนี และจัดอันดับความเกี่ยวข้องโดยใช้การกำหนดขอบเขตไว้ล่วงหน้า เป็นการแยกประเภทแบบมีผู้สอน โดยแนวคิดหลักของวิธีการนี้ ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลอออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม๊ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Spaceและสร้างฟังก์ชั่นวัดความคล้าย เรียกว่า เคอร์เนลฟังก์ชั่น (Kernel Function) บน Feature Space เหมาะใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง วิทยานิพนธ์นี้เลือกใช้วิธี SVM เพราะเป็นวิธีการที่ได้รับความนิยม ได้ผลลัพธ์ที่น่าพึงพอใจ และง่ายต่อการพัฒนา ดังแสดงแผนผังการทำงานของ SVM ดังรูปที่ 2.9 โดยทั่วไปแล้วจะนำ SVM มาแบ่งกลุ่มข้อมูลออกเป็น 2 กลุ่ม สมมติให้มีชุดข้อมูลมากลุ่มหนึ่ง และจะสร้างเส้นแบ่งข้อมูลออกเป็น 2 กลุ่ม โดยมีคำตอบที่เป็นไปได้คือ $y = \{+1, -1\}$ โดยค่า y นี้จะเป็นผลลัพธ์ที่ต้องการให้ SVM เรียนรู้ เพื่อใช้สำหรับแยกกลุ่มข้อมูลทั้งสองกลุ่มออกจากกัน สำหรับรากฐานเดิมของ SVM ถูกนำมาใช้กับข้อมูลที่เป็นลักษณะเชิงเส้น แต่ในความเป็นจริงบางครั้งข้อมูลที่นำมาใช้อาจจะมีลักษณะแบบไม่เป็นเชิงเส้น ซึ่งสามารถแก้ไขปัญหาดังกล่าวด้วยการนำ เคอร์แนลฟังก์ชัน (Kernel Function) มาใช้โดยจะกล่าวต่อไปในเรื่องเคอร์แนลฟังก์ชัน แนวคิดพื้นฐานของทฤษฎีสามารถอธิบายดังนี้

ให้กลุ่มข้อมูลทดลอง $D = \{x_i, y_i ; i = 1,2,...,n\}$ ในขณะที่ $x_i = (x_{i1}, x_{i2},..., x_{in}) \in R^n$ นั้นเป็นข้อมูลนำเข้า และ $y_i \in \{+1, -1\}$ เมื่อ (+1 ="เอกสารสำคัญ", -1="เอกสารทั่วไป") ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ SVM โดยที่ SVM นั้นมุ่งเป้าเพื่อหาฟังก์ชั่นการตัดสินใจที่สามารถแยกแยะค่าที่ไม่ทราบได้



รูปที่ 2.9 แผนผังการทำงานของ Support Vector Machines

โดยใช้สมการที่

$$y = sign\left\{\sum_{i=1}^{n} w_i \phi_i(x) + b\right\} \quad (2\text{-}3)$$

$$\phi(x) = [\phi_1(x), \phi_2(x), ..., \phi_n(x)]^T \quad (2\text{-}4)$$

เมื่อ $sign$ หมายถึง ถ้าค่าที่ได้มากกว่า 0 จะเป็น +1 แต่ถ้าน้อยกว่า 0 จะเป็น -1

w_i เป็นเวกเตอร์น้ำหนัก (Weighting) ที่เชื่อมโยงจาก feature space ไปสู่ output space

x เป็นกลุ่มข้อมูลซึ่งไม่สามารถแยกได้ด้วยสมการเส้นตรงให้อยู่ในรูปแบบที่สามารถใช้สมการแบ่งแยกได้

$\phi(x)$ แทนฟังก์ชั่นสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นเป็นข้อมูลที่อยู่ในรูปแบบที่สมการเชิงเส้นสามารถแยกแยะได้

b เป็นค่าไบอัส (bias) เป็นค่าโน้มเอียงหรือ threshold ที่ตั้งไว้ทำให้สมการที่ (2-5) สำหรับจำแนกข้อมูล

$$f(x)=\sum_{i=1}^{n} w_i \phi_i(x)+b \qquad (2\text{-}5)$$

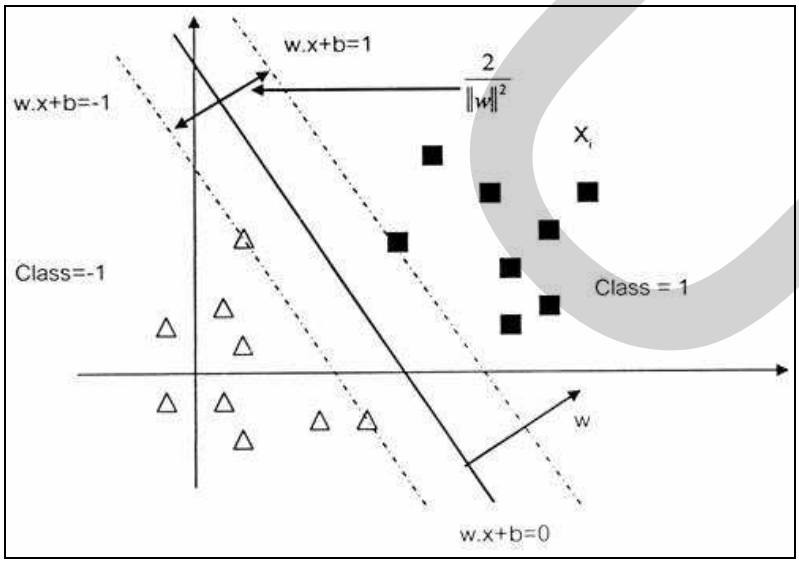
ฟังก์ชั่นที่สร้างการแบ่งประเภทของข้อมูลเชิงเส้น และได้ฟังก์ชั่นการแบ่ง $f(x)$ ถ้าข้อมูลเซต D นั้นสามารถแยกโดยใช้สมการแบบเส้นตรง ก็แสดงว่ามีการแบ่งแยกประเภทที่สมบูรณ์ กรณีการจัดแบ่งประเภทสำหรับค่า y ถูกกำหนดให้มีค่าเป็น +1 และ -1 ดังนั้นในสมการที่ (2-6) และ (2-7) สามารถเขียนรวมกันได้ดังสมการ (2-8)

$$w^T \cdot x + b \geq +1 \; ; \; y_i = +1 \qquad (2\text{-}6)$$

$$w^T \cdot x + b \leq -1 \; ; \; y_i = -1 \qquad (2\text{-}7)$$

$$y\left(w^T \cdot x + b\right) - 1 \geq 0 \; ; \; \forall i \qquad (2\text{-}8)$$

ทำให้สามารถแบ่งแยกข้อมูลทั้ง 2 กลุ่มได้อย่างชัดเจน รวมทั้งเป็นการแก้ปัญหาซึ่งเน้นเรื่อง การกำหนดระยะแยกแยะ (Margin) ให้มีขนาดกว้างที่สุด โดยการเลือกไฮเปอร์เพลน (Hyperplane) ที่เหมาะสม โดยมีข้อแม้ว่าในช่วงระยะแยกแยะ ไม่มีข้อมูลอยู่ ดังรูปที่ 2.10 ดังสมการที่เป็นเชิงเส้น และสามารถแยกแยะข้อมูลได้เกิดจากการเพิ่มขอบเขตเส้นแบ่งที่มีความกว้างเท่ากับ $2/\|w\|^2$



รูปที่ 2.10 ตัวอย่างการแยกแยะข้อมูลด้วย SVM

หากแต่บางครั้งก็ไม่สามารถที่จะแยกแยะข้อมูลได้ถูกต้องทั้งหมด ทำให้ต้องมีการกำหนดตัวแปรเพื่อยอมรับค่าความผิดพลาด โดยทำการเพิ่มตัวแปล ξ (Slack Variable) โดยเขียนแสดงรายละเอียดได้ดังสมการ (2-9) และ (2-10)

$$w^T \cdot x + b \geq +1 - \xi \; ; \; y_i = +1 \qquad (2\text{-}9)$$

$$w^T \cdot x + b \geq -1 + \xi \; ; \; y_i = -1 \qquad (2\text{-}10)$$

โดย $\xi > 0$

ทำให้ได้โครงสร้างของ SVM ซึ่งประกอบด้วย 2 ส่วนหลักคือ การเพิ่มระยะแยกแยะมากที่สุด และการแก้ปัญหาด้วยการลดข้อผิดพลาดให้ต่ำที่สุด ดังแสดงในสมการที่ (2-11)

$$\underset{w,b,\xi}{Minimize} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \qquad (2\text{-}11)$$

โดยที่ : $y_i\left(w^T\phi(x_i)+b\right)+\xi_i - 1 \geq 0$

เมื่อ $\xi_i \geq 0, i = 1,2,...,N$

จากสมการ (2-11) พบว่าสมการมีอยู่ 2 ส่วนด้วยกัน ได้แก่ สมการที่ใช้แทนการปรับค่าระยะแยกแยะข้อมูล (Margin) เพื่อใช้หาช่องว่างสำหรับระยะแยกแยะข้อมูลสูงที่สุด ส่วนที่สองแทนการปรับค่าผิดพลาดให้ต่ำสุด โดยมีค่า C ซึ่งเป็นค่าตัวแปรที่ผู้ใช้สามารถกำหนดค่าได้เองเพื่อปรับความสมดุลระหว่างการให้ความสำคัญของระยะแยกแยะสูงสุด หรือให้ความสำคัญกับค่าความผิดพลาดที่ต้องการให้ต่ำที่สุด โดยปกติค่า C จะกำหนดให้มีค่ามากกว่า 1 จากนั้นทำการแก้ปัญหาด้วยฟังก์ชั่นลากรองจ์ (Lagrangian) ด้วยการกำหนดค่าตัวแปรแบบเซตคู่ (Dual Sets) เพิ่มเติม แล้วทำการแก้ปัญหาจากการกำหนดข้อจำกัดที่ดีที่สุด (Constrained Optimization) ทำให้ได้ผลดังสมการที่ (2-12)

$$\underset{\alpha}{Minimize} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j K\left(x_i, x_j\right) - \sum_{i=1}^{N}\alpha_i \qquad (2\text{-}12)$$

โดยที่ : $\sum_{i=1}^{N} y_i \alpha_i = 0$

เมื่อ $0 \leq \alpha_i \leq C \; , i = 1,2,...,N$

โดยที่ α_i (แอลฟา) เป็น Lagrange multipliers เพื่อที่จะใช้สำหรับการแก้ปัญหาที่ดีที่สุด α_i^* เพื่อให้สำเร็จในการจำแนกข้อมูลที่ไม่ได้เป็นฟังก์ชั่นการจำแนกข้อมูลแบบเชิงเส้นดังสมการ (2-13)

$$f(x) = \sum_{j=1}^{N} w_j \alpha_i^* K(x, x_j + b) \qquad (2\text{-}13)$$

ในขณะที่ $K(x, x_j) = \phi^T(x)\phi(x_j)$ นั้นเป็น kernel function และมี kernel function ที่พบได้บ่อย 3 แบบด้วยกันดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 ชนิดของเคอร์เนลฟังก์ชัน

ชนิดของเคอร์เนลฟังก์ชัน	เคอร์เนลฟังก์ชัน
Radial Basis Function (RBF)	$K(x, x_i) = \exp\left(-\|x - x_i\|^2 / 2\sigma^2\right)$
Polynomial	$K(x, x_i) = \left(x_i^T x + r\right)^d$
Linear	$K(x, x_i) = x_i^T x$

สำหรับในวิทยานิพนธ์ฉบับนี้ได้นำเคอร์เนลแบบ Radial Basis Function (RBF) เป็น Kernel Function มาใช้เพื่อทำการเปรียบเทียบกับการแบ่ง SVM แบบ Linear เนื่องจากได้มีผู้วิจัยหลายท่านได้ทดลองแล้วพบว่า Kernel function นี้ให้ผลลัพธ์ที่ดี จากนั้นจะทำการทดสอบตามขั้นตอนที่ผู้วิจัยกำหนดขึ้นซึ่งขั้นตอนการดำเนินงานในการทดสอบจะได้กล่าวถึงในบทต่อไป

2.7 งานวิจัยที่เกี่ยวข้อง

ได้ทำการศึกษาเกี่ยวกับลักษณะการทำงานของระบบงานการป้องกันข้อมูลรั่วไหล และให้ความสนใจในประเด็นในส่วนของการแยกประเภทเอกสารก่อนนำเข้าระบบ งานวิจัยที่เกี่ยวกับการออกแบบระบบป้องกันข้อมูลรั่วไหล ผลกระทบของการรั่วไหลของข้อมูลสำคัญ และช่องทางการรั่วไหลของข้อมูล ออกมาในช่วงหลังนี้อย่างต่อเนื่อง รวมทั้งงานวิจัยเกี่ยวกับการแบ่งประเภทข้อมูลออกมาในช่วงหลายปีที่ผ่านมา ซึ่งสามารถนำมาใช้ในการวิเคราะห์และศึกษาได้เป็นอย่างดี

Gilberto และคณะ (2010) ได้นำเสนอ Security Framework สำหรับการป้องกันภัยคุกคามที่มากับ Social Network โดยกล่าวถึงเทคโนโลยี Data Loss Prevention เป็นวิธีป้องกันการรั่วไหลของข้อมูลสำคัญ วิธีการที่ใช้โดยทั่วไปประกอบด้วยโฮสต์ องค์ประกอบเครือข่าย หรือทั้ง

สองส่วน กล่าวถึงพื้นฐานของนโยบายการรักษาความปลอดภัย ความปลอดภัยของเอกสารสำคัญ กล่าวถึงความสำคัญของนโยบายที่เกินกว่าคำนิยามของบทบาทความรับผิดชอบสำหรับเจ้าหน้าที่และบุคคลผู้เกี่ยวข้อง หรือคู่ค้า เอกสารสำคัญเป็นสาเหตุทำให้เกิดช่องโหว่จากภัยคุกคาม ที่อาจก่อให้เกิดปัญหาและส่งผลกระทบต่อสารสนเทศที่พัฒนาขึ้นอย่างรวดเร็วในปัจจุบัน พื้นฐานของการป้องกันการสูญหายของข้อมูล ครอบคลุมทุกอย่างตั้งแต่ข้อมูลที่เป็นความลับเกี่ยวข้องกับลูกค้าหนึ่งไปยังไฟล์ ยุทธศาสตร์ของผลิตภัณฑ์ของบริษัท ที่ถูกส่งไปยังคู่แข่ง การสูญหายของข้อมูลพนักงานที่มีโอกาสเกิดขึ้นได้ตลอดเวลา กับบุคคลที่สาม หรือบุคคลภายในที่มีการลักลอบนำข้อมูลลูกค้า การเงินหรือทรัพย์สินทางปัญญา และข้อมูลที่เป็นความลับอื่นๆ ซึ่งถือเป็นการละเมิดนโยบายขององค์กร โดยมีอยู่ 3 ส่วนที่นำมาพิจารณาเพื่อลดการสูญเสีย คือ ข้อมูลเคลื่อนไหว เป็นข้อมูลใดๆที่มีการเคลื่อนไหวผ่านทางเครือข่ายไปยังภายนอก ผ่านทางเครือข่ายอินเตอร์เน็ต ข้อมูลที่มีอยู่ในระบบแฟ้มฐานข้อมูลหรือวิธีการจัดเก็บอื่นๆ ข้อมูล ณ จุดปลายทางซึ่งเป็นจุดสิ้นสุดของเครือข่าย สำหรับอุปกรณ์ USB และแล็ปท็อป แสดงให้เห็นรูปแบบ DLP ระบบข้อมูลการป้องกันและแก้ไขปัญหาการสูญเสียและใช้การควบคุมข้อมูลเคลื่อนไหวในโหมดของการติดต่อสื่อสารทางอิเล็กทรอนิกส์ผ่านตัวกรอง ที่กำหนดไว้ล่วงหน้า หรือป้ายชื่อในการจำแนกข้อมูลลงในไฟล์

Daeseon และคณะ (2006) ได้นำเสนอ method สำหรับการป้องกันข้อมูลส่วนบุคคลที่มีการรั่วไหลผ่านทางระบบอินเตอร์เน็ต กล่าวว่าทุกๆ packet ที่ถูกส่งผ่านทางเครือข่ายจากคอมพิวเตอร์ส่วนบุคคลออกไปยังอินเตอร์เน็ตเซิฟเวอร์ จะถูกเช็คถ้าหาก packet ดังกล่าวประกอบด้วยข้อมูลส่วนบุคคลจะถูกตรวจจับและมีการตัดสินใจเกี่ยวกับการส่งข้อมูลนั้นๆ โดยการตัดสินใจดังกล่าว ขึ้นอยู่บนพื้นฐานของนโยบายที่ได้ทำการควบคุม เป็นการนำเสนอรูปแบบการควบคุมข้อมูลส่วนบุคคลในการ transfer และ description ของระบบโครงสร้าง กล่าวถึงข้อมูลส่วนบุคคล เช่น ID และ password เป็นเหตุให้มีการสวมสิทธิ์ เนื้อหาข้อมูล เช่น อีเมลล์แอดเดรส และหมายเลขโทรศัพท์เป็นเหตุให้มีอีเมลล์ขยะ หรือโทรศัพท์ก่อกวน หมายเลขบัตรเครดิต หรือเลขบัญชีธนาคาร ที่ถูกใช้สำหรับอาชญากรรมบนโลกอินเตอร์เน็ต มีหลายช่องทางที่ข้อมูลเหล่านี้จะหลุดรอดออกไป อาจจะโดยการหลอกลวงผ่านทาง อินเตอร์เน็ตหรือทางเว็บไซต์ปลอมที่ทำให้เหมือนเป็นหน้าเว็บเพจจริง และหลอกให้มีการกรอกข้อมูลสำคัญลงไป ซอร์ฟแวร์หลอกลวงที่ถูกติดตั้งลงในคอมพิวเตอร์ส่วนบุคคลและคอยดักจับข้อมูลผ่านการกดแป้นพิมพ์ โดยซอร์ฟแวร์เหล่านี้จะส่งข้อมูลไปยังผู้ประสงค์ร้าย ที่ได้ทำการติดตั้งซอร์ฟแวร์หลอกลวงนี้ไว้ โดยผู้วิจัยได้นำเสนอวิธีสำหรับการป้องกันการรั่วไหล โดยใช้พื้นฐานที่ว่าจะไม่ส่งข้อมูลส่วนบุคคลออกไป “Do not send the personal information to the hazardous recipient” มีการนำเสนอ รูปแบบการควบคุมการถ่ายโอนข้อมูลส่วนบุคคลซึ่งประกอบด้วยองค์ประกอบดังนี้ ข้อมูลส่วนบุคคล ผู้ใช้งาน

โปรแกรมที่ใช้ส่ง หรือรูปแบบการส่ง ปลายทาง ข้อมูล log ของการถ่ายโอนข้อมูลส่วนบุคคล นโยบาย และขั้นตอนในการควบคุมด้วย ในส่วนของ ระบบควบคุมการถ่ายโอนข้อมูลส่วนบุคคล ประกอบด้วยองค์ประกอบดังต่อไปนี้ เครื่องมือในการตัดสินใจ นโยบายในการบริหารจัดการ ความน่าเชื่อถือของการบริหารจัดการ ระบบเฝ้าระวังของ packet และ http traffic และกลไกของการทำงาน ประกอบด้วย เครื่องมือในการติดป้ายกำกับ การจับคู่กับต้นแบบ การจับคู่กับมูลค่า และการวิเคราะห์ของการรักษาความมั่นคงปลอดภัย

Yuguo Wang (2008) ได้นำเสนอการปรับใช้พื้นฐานของการจัดหมวดหมู่แบบ SVM มาใช้สำหรับการจัดหมวดหมู่สำหรับเอกสารห้องสมุดอิเล็กทรอนิกส์ ได้นำเสนอทฤษฎีของ SVM ที่ใช้ในการฝึกสอนและทดสอบในการแยกประเภทเอกสาร เนื่องจาก SVM เป็นเครื่องมือที่ได้รับความนิยม และมีความสามารถในการเรียนรู้จากชุดข้อมูลตัวอย่างขนาดเล็กในขั้นตอนการฝึกสอน และมีประสิทธิภาพสูงในการคำนวณ โดยทดสอบกับเอกสารจำนวน 10,000 ฉบับจาก 10 หัวข้อของห้องสมุดอิเล็กทรอนิกส์ ในเอกสาร 1,000 ฉบับของแต่ละหัวข้อ โดยแต่ละเอกสารจะถูกแทนด้วยค่าเวกเตอร์ TFIDF และนำเข้ากระบวนการ SVM เพื่อจำแนกประเภท ทำการทดลองกับ 4 ลักษณะ ทำการแยกประเภทที่แตกต่างกัน คือ 1vs1, DSAGSVM, 1vs.rest และ Tree SVM ประสิทธิภาพที่ได้จากการทดลองพบว่า 1vs1 มีประสิทธิภาพในการแยกประเภทมากที่สุด ในแง่ของความเร็วประมวลผล ส่วน Tree SVM ใช้เวลาน้อยกว่าแบบ 1vs1 ถึง 13 เท่าแต่มีประสิทธิภาพในการแยกประเภทน้อยกว่าถึง 20% เมื่อเทียบกับการแยกประเภทโดยวิธีอื่น

Zhijie และคณะ (2010) นำเสนอขั้นตอนวิธีการจัดหมวดหมู่ตามตัวอักษร SVM เปรียบเทียบกับการจัดหมวดหมู่ด้วยวิธีการอื่น โดยในขั้นตอนของกระบวนการแยกประเภทข้อความ จะทำการเตรียมชุมข้อมูลและสกัดให้อยู่ในรูปแบบของคุณลักษณะ ตามรูปแบบที่ระบบต้องการ ขนาดของชุดข้อมูลจะปรับช่วงของข้อมูลเป็น [-1,1] เพื่อหลีกเลี่ยงคุณลักษณะที่มีขนาดใหญ่ และลดปัญหาในการคำนวณเชิงตัวเลขขณะการฝึกสอนชุดข้อมูล รูปแบบที่ได้จากการดำเนินการทดสอบและพยากรณ์ สุดท้ายจะได้ผลลัพธ์จากการจัดหมวดหมู่ กล่าวถึงกระบวนการทั่วไปของการจัดหมวดหมู่ข้อความ ด้วย 4 โมดูล การเตรียมข้อความ การสกัดคุณลักษณะ การฝึกอบรม และรูปแบบการฝึกอบรม โดยอธิบายพื้นฐานของ SVM ในขั้นตอนการจำแนกข้อความและนำเสนอรูปแบบการจัดหมวดหมู่ การวิเคราะห์ข้อมูล และการจัดหมวดหมู่ข้อความผ่านการจัดหมวดหมู่อัตโนมัติ ทำให้สามารถแยกประเภทข้อความได้ ช่วยให้ผู้เกี่ยวข้องสามารถค้นหาและวิเคราะห์ความสำคัญของเอกสารได้ มีการใช้ SVM ซึ่งอยู่บนพื้นฐานของทฤษฎีการเรียนรู้ ข้อดีของ SVM คือมีโครงสร้างการทำงานที่ง่าย และมีประสิทธิภาพ กล่าวถึงเทคนิคการจัดหมวดหมู่ที่ประกอบด้วยวิธีการแบบ Descision Tree, KNN,Naïve Bayesian และ SVM กล่าวถึงลักษณะของ

วิธีการแต่ละประเภท มีการนำเสนอการจัดหมวดหมู่ข้อความบนพื้นฐานของ SVM กระบวนการประมวลผลข้อความ การดึงลักษณะ การฝึกสอน และการทดสอบ ผลที่ได้จากการทดลองการแยกประเภทแบบ KNN, Naïve Bayesian และ SVM พบว่า อัตราความถูกต้อง (Accuracy Rate) และอัตราการเรียกคืน (Recall rate) ของคุณภาพการจัดหมวดหมู่ SVM ให้ค่าความถูกต้องมากกว่าการจัดหมวดหมู่แบบ KNN และ Naïve Baysian

พรพล ธรรมรงค์รัตน์ (2552) นำเสนองานวิจัย การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและซัพพอร์ตเวกเตอร์แมชชีน ได้นำเสนอขั้นตอนการจำแนกประเภทเว็บเพจ โดยแบ่งขั้นตอนการทำงานเป็น 4 ขั้นตอนคือการเตรียมข้อมูลเว็บ การสร้างลักษณะเฉพาะ การลดขนาดลักษณะเฉพาะ การจำแนกประเภท และการให้คะแนนเสียง โดยการวิจัยได้ใช้เทคนิคการให้น้ำหนักคำ TFIDF การลดขนาดลักษณะเฉพาะ และใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน SVM ร่วมกับอัลกอริทึม การให้คะแนนเสียงมาสร้างแบบจำลอง เพื่อให้ได้ผลการจำแนกประเภทที่ถูกต้องมากขึ้น

ปโยธร และกานดา (2548) ได้นำเสนองานวิจัยเกี่ยวกับการปรับปรุงกฎสำหรับการตัดคำในเอกสารภาษาไทย กล่าวถึงลักษณะของคำในภาษาไทย ที่นอกจากจะมีคำไทยแท้แล้ว ยังมีบางคำที่มาจากภาษาต่างประเทศ ที่ถูกสะกดอยู่ในรูปของคำอ่านภาษาไทย และบางคำจะมีการผสมอักษรที่แตกต่าง ลักษณะของประโยคในภาษาไทยมีการเขียนติดกันทำให้การแยกแยะคำมีปัญหา จึงได้นำหลักเกณฑ์ ที่เรียกว่าอักขระวิธีมาใช้ในการแยกแยะระดับย่อยของคำ วิธีการหลักสำหรับการตัดคำในเอกสารภาษาไทย คือ การใช้กฎเพื่อตัดคำภาษาไทยโดยการใช้ไวยากรณ์ทางภาษาแบ่งตัวอักษรเป็นหมวดหมู่ตามลักษณะการใช้งาน ได้แก่ กลุ่มพยัญชนะ กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลข และอักษรพิเศษ ขั้นตอนการตัดพยางค์จากซ้ายไปขวาเป็นส่วนใหญ่ การใช้พจนานุกรม ผลลัพธ์ที่ได้จะอยู่ในระดับคำ โดยทำการตรวจสอบสายอักขระ ซึ่งเป็นชุดของตัวอักษรที่ได้จากเอกสาร จากนั้นนำไปค้นหาคำในพจนานุกรม หากพบคำในพจนานุกรมที่สามารถเป็นคำในสายอักขระนั้น ได้มากกว่าหนึ่งคำ จะทำการเลือกคำที่ยาวที่สุด (Longest Matching) หากอักษรตัวต่อมาไม่พบคำที่ตรงในพจนานุกรมที่มีอยู่ จะทำการย้อนกลับไปเลือกคำที่สั้นกว่าแทน และกฎที่งานวิจัยได้นำเสนอเพิ่มเติมจากกฎที่มีอยู่ คือการตัดคำที่มีความเกี่ยวข้องกับภาษาต่างประเทศ และคำที่อยู่นอกเหนือจากการประสมตามอักขระวิธี คำที่ได้จากเอกสารจะถูกนำไปแบ่งครั้งแรกด้วยช่องว่างระหว่างประโยค และนำไปแยกในระดับพยางค์โดยใช้กฎ ผลที่ได้จากการแบ่งด้วยกฎจะถูกจัดเก็บเป็นสองส่วน คือส่วนที่ต้องนำไปค้นหาต่อในพจนานุกรม และไม่ต้องนำไปค้นหาต่อ ได้แก่สัญลักษณ์ - / , . คำที่มีตัวเลขและคำที่ตัวอักษรต่างประเทศอยู่ระหว่างประโยคที่ถูกแบ่งด้วยช่องว่างซึ่งถือเป็นคำที่สมบูรณ์แล้ว

อภิชาต ขานทอง และคณะ (2546) ได้นำเสนองานวิจัย การสรุปใจความสำคัญของเอกสาร โดยเลือกประโยคที่แสดงถึงเนื้อหาหลักของเอกสาร และนำประโยคที่ได้มาสร้างเป็นใจความสำคัญของเอกสาร เป็นการนำทฤษฎีการให้น้ำหนักคำโดยใช้สูตรการหาค่า TFIDF และการประเมินใจความสำคัญที่ได้ เพื่อหาแนวทางในการสรุปใจความสำคัญของเอกสารภาษาไทย เนื่องจากโครงสร้างประโยคภาษาไทยมีความซับซ้อน การหาขอบเขตของประโยคจึงทำได้ยาก ซึ่งแนวทางของการสรุปใจความสำคัญจะแบ่งตามลักษณะได้เป็น การยึดตามค่าความถี่เป็นหลักสำคัญ (Frequency-Base) โดยวิธีนี้จะดูที่ความถี่และตำแหน่งของคำในเอกสารเป็นหลัก และยึดหลักตามฐานความรู้ (Knowledge -Base) โครงสร้างหลักของเอกสารที่ได้ จะขึ้นอยู่กับการตีความจากฐานความรู้ที่มีอยู่โดยใช้เทคนิคของการแยกแยะ (Classification) มีการสร้างโครงรูป (Template) ตามประเภทเอกสาร และยึดตามความสอดคล้องและความเข้ากันได้ (Discourse-Base) โดยยึดตามโครงสร้างของประโยคภาษานั้นๆ โดยขั้นตอนในการสรุปใจความสำคัญ แบบดึงจากต้นฉบับ มีขั้นตอนหลักๆอยู่ 3 ขั้นตอน คือการแยกคำสำคัญ การแยกประโยค และการเรียงประโยคเป็นใจความสำคัญ ในขั้นตอนการแยกคำสำคัญ นั้นจะขึ้นอยู่กับการพิจารณาความสำคัญของเอกสาร ความสำคัญของคำ จะขึ้นอยู่กับจำนวนครั้งของคำที่ปรากฏอยู่ในประโยค โดยปกติแล้วคำที่อยู่ในประโยคจะถูกพิจารณาว่าเป็นคำสำคัญถ้าคำนั้นปรากฏหลายครั้งในประโยค แต่ปรากฏน้อยครั้งในเอกสารทั้งหมด โดยใช้สูตรการหาค่า TFIDF ในการแยกประโยค และนำมาวิเคราะห์ถึงความสำคัญของประโยค เป็นผลต่อเนื่องจากความถี่ของคำสำคัญ และตำแหน่งของคำสำคัญ ที่พบในประโยคนั้น โดยในส่วนของคำที่มีความถี่สูงเกินไป และความถี่น้อยเกินไป จะถูกตัดทิ้ง ก็จะได้ขอบเขตของคำที่มีประโยชน์ในการพิจารณาความสำคัญของประโยค

นนท์ บุญนิธิประเสริฐ (2553) นำเสนองานวิจัยเกี่ยวกับการกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่ เป็นการพัฒนาวิธีการกรองข้อความสั้น ที่สามารถใช้งานได้กับข้อความภาษาไทย ภาษาอังกฤษ และภาษาไทยปนภาษาอังกฤษ ให้สามารถกรองข้อความสแปมออกจากบริการส่งข้อความสั้นในเครือข่ายโทรศัพท์เคลื่อนที่ โดยการทำงานจะอธิบายการกรองข้อความ เมื่อตัวกรองได้รับข้อความจากชุมสาย (BTS/MSC) และถอดข้อความจาก SMPP Message ให้อยู่ในรูปของ text เข้าสู่กระบวนการ TN เพื่อลบ Character ที่ไม่สามารถตัดเป็นคำออกไป เช่น @ # , ! . รวมถึงตัวเลข เมื่อข้อความพร้อมจะทำการตัดคำภาษาอังกฤษด้วยการตรวจสอบการเว้นวรรคจากนั้นจะลบคำที่จัดอยู่ในประเภท Stop words ออก และทำการเปรียบเทียบคำเข้ากับ TFIDF ตามอัลกอริทึมการกรองเพื่อสรุปผลของข้อความว่าจัดเป็นข้อความปรกติหรือข้อความสแปม ใช้วิธีการค้นหาลักษณะเด่นของเอกสาร ให้อยู่ในรูปของกลุ่มข้อมูล (Feature Vector) โดยอ้างอิงจากชุดตัวอักษรหรือคำในเอกสาร

และจำนวนเอกสารทั้งหมดที่ถูกกำหนดให้เป็นข้อมูลฝึกสอน ใช้วิธี TFIDF ในการแปลงเอกสารที่ต้องการนำไปคำนวณให้เป็นชุดข้อมูล Vector เพื่อนำไประบุการจำแนก นำข้อมูลเข้าสู่กระบวนการตัดคำ และนำเข้าสู่กระบวนการจำแนกโดยใช้เทคนิค SVM เนื่องจากง่ายต่อการพัฒนาและให้ผลลัพธ์เป็นที่น่าพึงพอใจ ก่อนจะนำตัวจำแนกมาใช้ต้องมีการฝึกฝนโดยใช้ลักษณะที่จะใช้บ่งบอกแต่ละพฤติกรรม ลักษณะดังกล่าวจะถูกปรับค่าเพื่อให้อยู่ในช่วงที่เหมาะสม โดยกำหนดช่วงที่เหมาะสมเป็น [-1,1] เพื่อหลีกเลี่ยงปัญหาความซับซ้อนจากการคำนวณ โดยการปรับค่าจะถูกใช้สำหรับลักษณะที่ใช้ในการฝึกฝนและลักษณะที่จะทดสอบ

จันทิมา พลพินิจ และคณะ (2549) ได้นำเสนองานวิจัย เกี่ยวกับการสร้างตัวกรองเว็บอนาจาร แบบอัตโนมัติ และวิธีการกรองเว็บโดยใช้ฐานข้อมูล มีการประยุกต์ในเรื่องของการจัดกลุ่มเอกสารโดยใช้ ความเป็นไปได้ในการแยกประเภท ได้กล่าวถึงระบบการกรองเว็บซึ่งแบ่งออกเป็น 2 ลักษณะ คือ พื้นฐานของMetadata ซึ่งเป็นระบบการกรองเว็บโดยอาศัยข้อมูลหรือสารสนเทศจากภายนอกเว็บ และการใช้ฐานข้อมูล เป็นระบบการกรองเว็บที่ใช้สารสนเทศบนหน้าเว็บ เช่นข้อความ รูปภาพ เป็นส่วนสำคัญในการสร้างโมเดลหรือตัวกรอง โดยนำเสนอกระบวนการวิจัยแบ่งขั้นตอนเป็นการตัดคำแบบพจนานุกรม โดยวิธีการเทียบคำที่ยาวที่สุด (Longest matching) การหาความถี่และน้ำหนักคำโดยสมการหาความถี่และน้ำหนักคำ TFIDF การวัดประสิทธิภาพโดยการหาค่า F-Measure และจำแนกประเภทโดยใช้ตัวแยกประเภท นาอีฟเบล์ (Naïve Bayes) และผลลัพธ์ที่ได้จะเกิดจากการประมวลผลค่าในเชิงสถิติ ถูกเรียกว่า “Probabilistic Classifier” โดยโครงสร้างโมเดลกรองจาก Web Site จำนวน 200 Web Site (Training Set) และทดสอบจาก 120 Web Site (Test Set) ขนาดเวกเตอร์ของคำ 6,189 คำ จากผลการทดลองพบว่าจำนวนคำที่พบในเอกสารเป็นปัจจัยสำคัญในการสร้างโมเดลสำหรับการกรองเว็บโดยอาศัยข้อความ ดังนั้นหากเพิ่มขนาดของ bag of words ให้มีขนาดใหญ่ขึ้นอาจจะหมายถึงความถูกต้องในการทดลองที่เพิ่มขึ้นด้วย

นิเวศ จิระวิชิตชัย (2554) การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ เป็นงานวิจัยที่นำเสนอแบบจำลองการจัดหมวดหมู่เอกสารภาษาไทย โดยการลดคุณลักษณะของเอกสารก่อนนำไปประมวลผลด้วยเครื่องจักการเรียนรู้ เพื่อลดมิติของข้อมูล และลดระยะเวลาประมวลผล ประหยัดทรัพยากรของระบบและเพิ่มประสิทธิภาพในการจัดหมวดหมู่ โดยวัดประสิทธิภาพการจัดหมวดหมู่เอกสารที่ระดับคุณลักษณะที่ดีที่สุดพบว่าอัลกอริทึม SVM ให้ประสิทธิภาพสูงสุดคือ 94.3% รองลงมาเป็นอัลกอริทึม Naive Baye 86.2%และอัลกอริทึม RBF 86.1% อัลกอริทึม J48 79.7% อัลกอริทึม Ripper 78.9% และอัลกอริทึม KNN 70.3% ตามลำดับ

ภรัณยา อำมฤครัตน์ (2552) การเปรียบเทียบประสิทธิภาพการคัดเลือก และจำแนกข้อมูลด้วยวิธีการทางเครือข่ายประสาทเทียม เป็นการเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลโดยใช้เครื่องมือทางเครือข่ายประสาทเทียม ผลที่ได้อยู่ในระดับที่ยอมรับได้ที่จะนำวิธีการดังกล่าวไปใช้เพื่อพยากรณ์สถานะ (Class) ว่าปกติหรือผิดปกติ จากผลการทดลองที่ได้ จะเห็นว่าค่าประสิทธิภาพของแต่ละวิธีจากากรใช้ข้อมูลชุดเดียวกันในการทดสอบนั้นๆ พบว่าค่าประสิทธิภาพการจัดกลุ่มของชุดข้อมูลที่ใช้ในการทดสอบแบบ SVM ที่ใช้ Kernel แบบ RBF จะให้ผลการทดสอบที่สูงกว่า MLP ได้ค่าความถูกต้อง 96.55 - 100 เปอร์เซ็นต์ การนำเทคนิคการคัดเลือกข้อมูลที่เหมาะสม เพื่อเพิ่มประสิทธิภาพการทำนาย เพื่อสังเคราะห์โมเดลได้อย่างรวดเร็ว และเพื่อลดความซับซ้อนของรูปแบบโมเดลนั้น จะพบว่าผลของประสิทธิภาพการจำแนกไม่ต่างกับการเลือกใช้แอทริบิวต์ทั้งหมดมากนัก โดยเฉพาะในการทดสอบแบบ SVM

บทที่ 3
ระเบียบวิธีวิจัย

3.1 แนวทางการวิจัยและพัฒนา

เนื่องด้วยมีการพัฒนาระบบงาน การป้องกันข้อมูลรั่วไหล เข้ามาใช้ในองค์กรด้วยอุปกรณ์ Symantec DLP เพื่อบริหารจัดการความปลอดภัยของข้อมูล สนันสนุนด้วยมาตรฐานด้านความมั่นคงปลอดภัยด้านเทคโนโลยีสารสนเทศ ISO/IEC27001:2005 เพื่อรักษาไว้ซึ่งความลับ (Confidentiality) ความถูกต้องครบถ้วน(Integrity) และความพร้อมใช้ (Availability) ของข้อมูลโดยข้อมูลจะถูกเข้าถึงโดยผู้ที่มีสิทธิ์เท่านั้น และจะไม่ถูกแก้ไขดัดแปลง หรือเคลื่อนย้าย ถ้าไม่ได้รับอนุญาต โดยตรงกับหลักควบคุมหลักของระบบ ISO/IEC27001:2005ในเรื่องของ นโยบายความมั่นคงปลอดภัย การจัดหมวดหมู่ของสินทรัพย์ ซึ่งข้อมูลสำคัญก็ถือเป็นสินทรัพย์ขององค์กรรวมทั้งการบริหารจัดการด้านการบริหารจัดการด้านการสื่อสาร และการเข้าถึงระบบข้อมูลสารสนเทศด้วย ยังสนับสนุน กฎหมายความมั่นคงปลอดภัยสารสนเทศ ที่ออกเป็นพรบ. ว่าด้วยธุรกรรมอิเล็กทรอนิกส์ ที่ประกาศบังคับใช้เมื่อ พ.ศ.2550 ที่ว่าด้วยการควบคุมการเข้าถึง และใช้สารสนเทศรวมไปถึงการตรวจสอบ และประเมินความเสี่ยง (มาตรา 5) การคุ้มครองข้อมูลส่วนบุคคลในกรณีที่มีการรวบรวมจัดเก็บข้อมูลไว้ มิให้ทำการเผยแพร่ข้อมูลที่ระบุตัวตน เมื่อไม่ได้รับอนุญาต (มาตรา 6) การลักลอบดักข้อมูลคอมพิวเตอร์ของผู้อื่น ทั้งด้วยทาง Physical เช่น การก๊อปปี้ไฟล์ข้อมูลผ่านทางไดร์พกพา ส่งผ่านแนบไฟล์ทางสื่ออิเล็กทรอนิกส์ หรือทาง Logical เช่นการทำ Packet Sniffing ข้อมูลคอมพิวเตอร์ของผู้อื่น การทำให้เสียหาย ทำลาย แก้ไข เปลี่ยนแปลง หรือเพิ่มเติมข้อมูลคอมพิวเตอร์ของผู้อื่น โดยไม่ชอบ (มาตรา 9)

ด้วยความสามารถของอุปกรณ์ Symantec DLP จะทำงานในส่วนของการเฝ้าระวังข้อมูลสำคัญ และบังคับใช้นโยบายที่ได้กำหนดขึ้นเท่านั้น ยังขาดขั้นตอนการกำหนดข้อมูลสำคัญและเป็นความลับขององค์กรก่อนนำเข้าสู่ระบบเพื่อเป็นข้อมูลต้นแบบในการกรองของระบบ ซึ่งในปัจจุบันมีการจัดเก็บในรูปแบบอิเล็กทรอนิกส์เป็นปริมาณมาก ยากต่อการระบุ และแยกประเภทของเอกสาร จึงมีวัตถุประสงค์ในการออกแบบวิธีการแยกประเภทเอกสารในหน่วยงานย่อยต่างๆ โดยการสร้างระบบการคัดแยกประเภทเอกสาร เพื่อคัดแยกเอกสารสำคัญ ออกจากเอกสารทั่วไป ดังกระบวนการของขั้นตอนการดำเนินงานวิจัยในตารางที่ 3.1

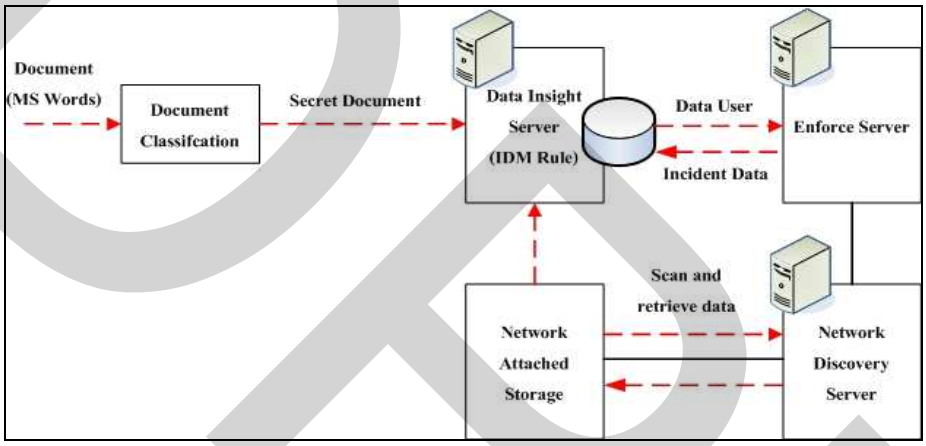
ตารางที่ 3.1 รายละเอียดกระบวนการของขั้นตอนในการดำเนินงานวิจัย

ขั้นตอนการทำงาน	วิธีการทำงาน
1	นำเอกสาร Microsoft Word (.doc) เข้าระบบ
2	แปลงไฟล์เอกสารจาก .doc ไปเป็นไฟล์ .txt โดยใช้โปรแกรม Antiword
3	นำเข้าไฟล์ .txt เข้าสู่กระบวนการตัดคำภาษาไทยโดยใช้โปรแกรม SWATH
4	นำคำที่ได้จากการตัดคำมาเปรียบเทียบกับข้อมูลคำหยุด และลบคำหยุดออก
5	คำนวณหาค่าน้ำหนักคำของแต่ละเอกสาร โดยใช้สมการ TF-IDF
6	กำหนดคุณลักษณะสำคัญ (Feature Selection) และสกัดคุณลักษณะสำคัญ (Feature Extraction) สำหรับเป็นข้อมูลนำเข้าในการแยกประเภทเอกสาร
7	นำเข้าค่าคุณลักษณะสำคัญ (ค่าน้ำหนักคำ) เพื่อคำนวณการแยกประเภทเอกสาร โดยใช้ LS-SVM
8	คำนวณอัตราความถูกต้องของการคัดแยกประเภทเอกสาร
9	นำเอกสารสำคัญที่ได้จากการแยกประเภทแล้วเข้าสู่ระบบป้องกันการรั่วไหลข้อมูล Symantec DLP

นำเอกสารที่ถูกกำหนดว่าเป็นเอกสารสำคัญ โดยหน่วยงานเจ้าของเอกสารนั้นๆ และเอกสารทั่วไป นำมาเป็นข้อมูลชุดฝึกสอน และชุดทดสอบ เรานำต้นแบบเอกสาร145 ฉบับ มาทำการฝึกสอนเพื่อทำการแยกประเภทเอกสาร และจะนำชุดฝึกสอนนั้นมาเป็นต้นแบบในการทดสอบกับชุดข้อมูลอื่นๆต่อไป ในวิทยานิพนธ์นี้ใช้เทคนิคการให้น้ำหนักคำ TF-IDF ร่วมกับเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) มาสร้างแบบจำลอง และพัฒนาโปรแกรมสำหรับแยกประเภทเอกสาร เพื่อความถูกต้องของผลการแยกประเภทเอกสาร และลดเวลาการคัดแยกเอกสารสำคัญออกจากเอกสารทั่วไป โดยมีเงื่อนไขว่าต้องเป็นเอกสารที่ถือว่าสิ้นสุดไม่มีการเปลี่ยนแปลงเนื้อหา และระดับความสำคัญของเอกสาร ถ้าหากมีการเปลี่ยนแปลงจะต้องทำการแยกประเภทเอกสารใหม่อีกครั้ง เมื่อได้ผลลัพธ์จากการแยกประเภทเอกสารแล้ว จะนำเอกสารสำคัญมาเป็นโครงรูป ในการกรองเอกสารของระบบ การป้องกันข้อมูลรั่วไหล Symantec DLP ในส่วนของ IDM (Indexed Document Matching) Rule หรือการทำดัชนีเอกสารสำคัญบนเครื่อง Data Insight Server ซึ่งจะมี Database Relation ที่ทำงานร่วมกับ Enforce Server หรือเครื่องเซิฟเวอร์ที่ทำหน้าที่กำหนด และ

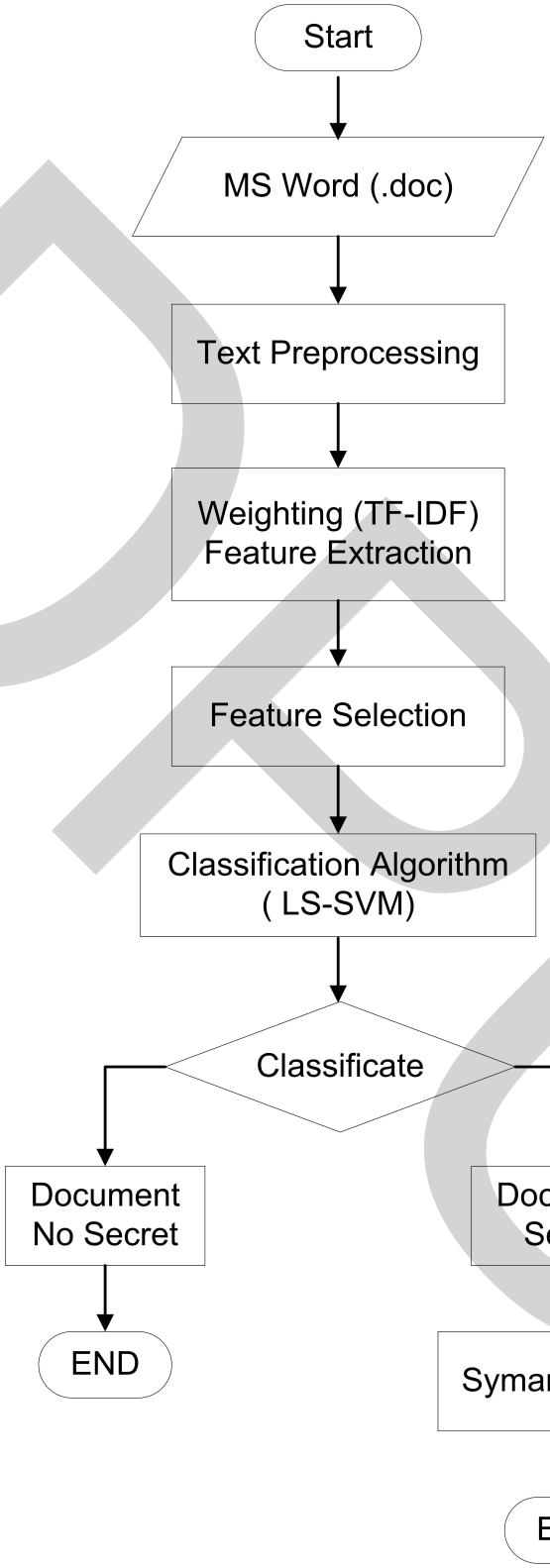
บังคับใช้นโยบายการป้องกันข้อมูล ส่วนของ Network Discovery Server ช่วยตรวจสอบการไหลของข้อมูลผ่านทาง DLP Agent ที่ติดตั้งอยู่บนเครื่องลูกข่าย

กฎของ IDM Rule จะทำการพิจารณาเนื้อหาจากเอกสารเฉพาะที่มีการลงทะเบียนไว้ว่าสำคัญ (เอกสารสำคัญที่เป็นเอกสารต้นฉบับ) และมีการตรวจสอบความเหมือนของเนื้อหาในเอกสารที่มีค่าความเหมือน 80% หรือมากกว่าจากเอกสารต้นฉบับที่ได้ระบุไว้ใน IDM Rule ซึ่งเป็นเอกสารสำคัญที่ได้จากการแยกประเภทไว้ก่อนนำเข้าสู่ระบบ ดังแสดงการทำงานในรูปที่ 3.1



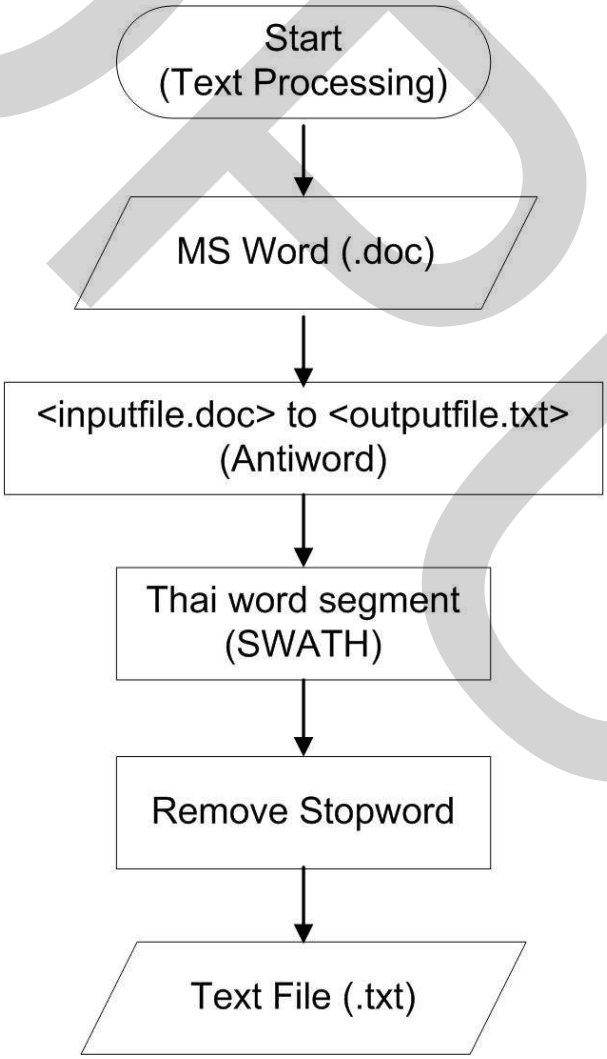
รูปที่ 3.1 แผนผังการทำงานของการนำเอกสารที่ถูกแยกประเภทแล้วเข้าสู่ Symantec DLP

มีข้อจำกัดว่าเอกสารที่นำมาเข้าสู่กระบวนการจะต้อง ไม่มีการเปลี่ยนแปลง และแก้ไขระดับความสำคัญ เช่นการระบุว่าเป็นเอกสารสำคัญในภายหลัง หลังจากกำหนดเป็นเอกสารทั่วไปแล้ว และไม่ครอบคลุมถึงเอกสารที่มีการเปลี่ยนแปลงในภายหลังจากการแยกประเภทแล้วจะทำการทดสอบการแยกประเภทเอกสาร ตามขั้นตอนในการดำเนินงานวิจัยดังรูปที่ 3.2 จากรายละเอียดกระบวนการของขั้นตอนในการดำเนินงานวิจัย ในตารางที่ 3.1 จะอธิบาย กระบวนของขั้นตอนการทำงานของระบบการแยกประเภทเอกสาร แสดงผังกระบวนการทำงาน ในรูปที่ 3.2 ขั้นตอนแรกเป็นขั้นตอนนำเข้าเอกสารไมโครซอฟท์เวิร์ด ผ่านกระบวนการ Text Processing ซึ่งจะกล่าวต่อไปตามรูปที่ 3.3 ผลลัพธ์ที่ได้จะได้คำภาษาไทยที่ถูกประมวลผลในขั้นตอนการตัดคำภาษาไทย และจำกัดคำหยุด ออกแล้ว คำที่เหลือจากกระบวนการ Text Processing จะนำเข้าสู่กระบวนการหาค่าน้ำหนักคำ TF-IDF ในขั้นตอนนี้จะนำเข้าไฟล์เอกสาร .txt เข้าสู่โปรแกรม PHP เรียกใช้ฟังก์ชั่น TF-IDF เพื่อหาค่าน้ำหนักคำ เมื่อได้ค่าน้ำหนักคำ ซึ่งคือว่าเป็นขั้นตอนในการทำ Feature Extraction ด้วย นั่นก็คือการหาค่าคุณลักษณะนั่นเอง



รูปที่ 3.2 กระบวนการของขั้นตอนในการดำเนินงานวิจัย

มีการคัดเลือกคุณลักษณะ (Feature Selection) ตามความถี่ของคำที่ปรากฏในเอกสารมากฉบับ ของเอกสารแต่ละประเภท และเมื่อได้ค่าคุณลักษณะมาเป็นตัวแทนของเอกสารแล้ว จะนำเข้าสู่กระบวนการแยกประเภทเอกสาร โดยนำเข้าโปรแกรม MATLAB ซึ่งเป็นภาษาที่มีประสิทธิภาพสูงทางด้านเทคนิคในการคำนวณ มีรูปแบบโปรแกรมที่ง่ายต่อการใช้งาน และมีเครื่องมือที่หลากหลาย ในที่นี้จะใช้เครื่องมือซัพพอร์ตเวกเตอร์แมชชีนแบบค่ากำลังสองน้อยที่สุด (Least Squares Support Vector Machine: LS-SVM) เป็นการพัฒนาต่อจาก SVM เพื่อสนับสนุนการทำงานกับข้อมูลปริมาณมาก มาใช้ในการเรียนรู้เพื่อแบ่งเขตข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้ โดยใช้ฟังก์ชั่นเคอร์เนลชนิด ฟังก์ชั่นเรเดียลเบซิส (RBF Kernel Function) มาประยุกต์ใช้สำหรับการพยากรณ์



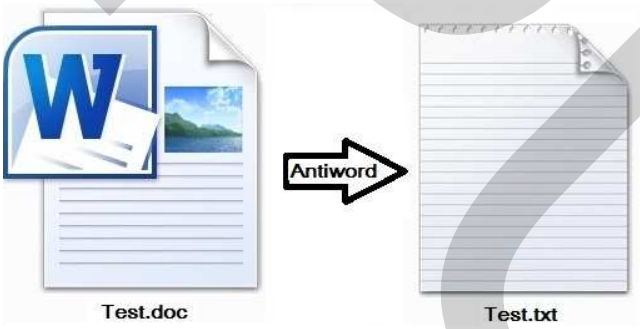
รูปที่ 3.3 ขั้นตอนการแปลงไฟล์เอกสาร (Text Processing)

และในขั้นตอนการทดสอบจะนำข้อมูลชุดฝึกสอนทดสอบกับข้อมูลชุดทดสอบเพื่อแยกประเภทเอกสารสำคัญ ออกจากเอกสารทั่วไป พร้อมหาอัตราความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสารจากชุดข้อมูลฝึกสอน และนำเอกสารสำคัญที่แยกประเภทออกมาได้ไปเป็นต้นแบบในการกรองเอกสารสำคัญของระบบ Symantec DLP ต่อไป

3.2 กระบวนการนำเข้าเอกสาร

เอกสารที่จะนำเข้าระบบจะอยู่ในรูปแบบของเอกสาร Microsoft Word (.doc) ต้องทำการแปลงไฟล์เอกสารให้อยู่ในรูปแบบของไฟล์ข้อความธรรมดา (.txt) ก่อนดังแสดงขั้นตอนกระบวนการแปลงไฟล์เอกสาร (Text Processing) ในรูป 3.3 แสดงการทำงานในการแปลงไฟล์เอกสาร Microsoft words ให้เป็นไฟล์ข้อความ (.txt) เพื่อเป็นอินพุตการหาค่าน้ำหนักคำ TF-IDF การสกัดลักษณะสำคัญ และการคัดเลือกคุณลักษณะสำคัญต่อไป

3.2.1 อ่านไฟล์เอกสาร Microsoft Word ให้เป็นไฟล์ข้อความ Text ขั้นตอนแรกแสดงขั้นตอนการอ่านไฟล์เอกสารไมโครซอฟท์เวิร์ด (.doc) ให้อยู่ในรูปไฟล์ข้อความ (.txt) โดยผ่านโปรแกรมย่อย Antiword ดังแสดงในรูปที่ 3.4

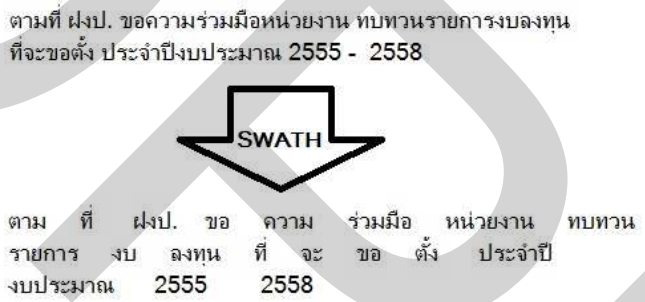


รูปที่ 3.4 แปลงไฟล์เอกสาร Microsoft word .doc เป็นไฟล์ข้อความ .txt

จากกระบวนการนำเข้าเอกสาร ซึ่งเป็นเอกสาร Microsoft Word (.doc) และทำการแปลงไฟล์จากเอกสาร .doc เป็นไฟล์ข้อความ .txt โดยใช้โปรแกรม Antiword เพื่อเป็นข้อมูลนำเข้า นำไฟล์เอกสารไมโครซอฟท์เวิร์ดเข้าสู่โปรแกรมย่อม Antiword Version 3.7 ใน PHP Program ซึ่งเป็นฟรีโปรแกรมที่ถูกนำมาใช้เพื่อแสดงความหมายของเอกสารไมโครซอฟท์เวิร์ดให้แสดงผลอยู่ในรูปแบบของไฟล์ข้อความ (plain text) โปรแกรม Antiword จะถูกเรียกผ่านทาง command line

ตามรูปแบบ Antiword <source_path>.doc <destination_path>.txt เมื่อ source_path คือ ไดร์ที่เก็บไฟล์ข้อมูล .doc ส่วน destination_path คือ ไดร์ปลายทางที่จะเก็บไฟล์ข้อมูล .txt

3.2.2 การตัดคำ (Word Segmentation) นำไฟล์ข้อความที่ได้นำเข้าสู่กระบวนการตัดคำภาษาไทยเพื่อเพิ่มประสิทธิภาพของการจำแนกหมวดหมู่เอกสารภาษาไทย การตัดคำภาษาไทย ซึ่งมีลักษณะการเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่แบ่งสายอักขระเหล่านี้ออกเป็นคำๆ ในที่นี้จึงเรียกใช้ฟังก์ชั่น SWATH ซึ่งเป็นฟรีโปรแกรมจากเนคเทค ที่จะทำการตัดคำแบบยาวที่สุด (Longest Matching) ที่เข้ามาจากซ้ายไปขวา และเทียบกับคำที่ปรากฏในพจนานุกรม แล้วทำต่อไปเรื่อยๆ จนจบสายอักขระ ดังแสดงตัวอย่างการตัดคำภาษาไทย ในรูปที่ 3.5



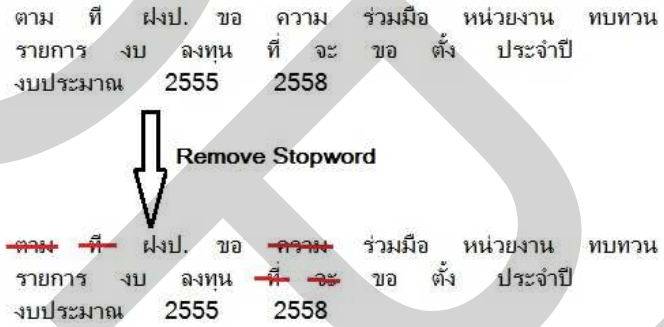
รูปที่ 3.5 การตัดคำภาษาไทย โดยใช้โปรแกรม SWATH

โปรแกรมการตัดคำภาษาไทย SWATH โปรแกรมตัดคำภาษาไทย ถูกพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ หรือ NECTEC ในวิทยานิพนธ์นี้ใช้ SWATH Version 2.0 ลักษณะการทำงานจะทำการเปรียบเทียบคำที่ปรากฏในเอกสาร กับข้อมูลที่มีอยู่ในฐานข้อมูลพจนานุกรม โดยข้อมูลที่นำมาใช้กับโปรแกรมนี้ต้องเป็นไฟล์ข้อความเท่านั้นซึ่งได้มาจากผลลัพธ์จากกระบวนการแปลงไฟล์จาก MS Word เป็นไฟล์ txt โดยจะเลือกตัดคำแบบเลือกคำที่ยาวที่สุด (Longest Matching) โดยมีรูปแบบการเรียกใช้โดย swath.exe -b "|" -d data < text.txt > out.txt เมื่อ text.txt เป็นไฟล์นำเข้าที่ได้รับจากกระบวนการอ่านไฟล์ข้อความจากเอกสารไมโครซอฟท์เวิร์ด และ out.txt คือไฟล์เอาท์พุตที่ได้จากการตัดคำภาษาไทย

3.2.3 การกำจัดคำหยุด (Remove Stop Word) เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไป เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจคามของเอกสารเปลี่ยนแปลงตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำ หรือกลุ่มคำให้สัมพันธ์กัน คำสันธานที่ทำหน้าที่เชื่อมคำกับคำสรรพนามที่ใช้แทนคำนาม

ที่กล่าวถึงมาแล้วในประโยคก่อนหน้า เป็นต้น ดังนั้นการจำกัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการหาค่าน้ำหนัก เพื่อจำกัดคุณลักษะที่ไม่เป็นประโยชน์ออก ตัวอย่างคำหยุดเช่น ที่ ใน ว่า และ จะ มี ได้ ตาม ให้ ความ เป็นต้น ตัวอย่างการจำกัดคำหยุด ดังแสดงในรูปที่ 3.6

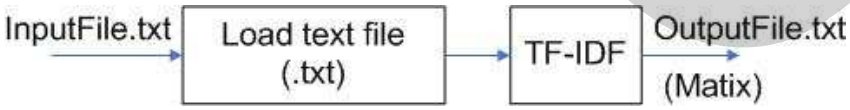
การกำจัดคำหยุด (Remove Stop word) หรือคำฟุ่มเฟือยที่ไม่สื่อความหมายโดยทำการระบุคำหยุดเหล่านี้ไว้ในอาร์เรย์ ในส่วนของโปรแกรม PHP ในขั้นตอนนี้จะตัดคำที่ไม่มีความหมายออก และคงคำที่สื่อความหมายไว้เพื่อเพิ่มประสิทธิภาพในการหาค่าน้ำหนักคำต่อไปนำคำที่เหลือทั้งหมดมาหาค่าน้ำหนักคำ เมื่อได้ผลลัพธ์ออกมาอยู่ในรูปแบบของเมตริกซ์เวกเตอร์ เป็นข้อมูลนำเข้าระบบการแยกประเภทเอกสาร LS-SVMต่อไป



รูปที่ 3.6 การกำจัดคำหยุด

3.3 กระบวนการหาค่าน้ำหนักคำ และเลือกคุณลักษณะ

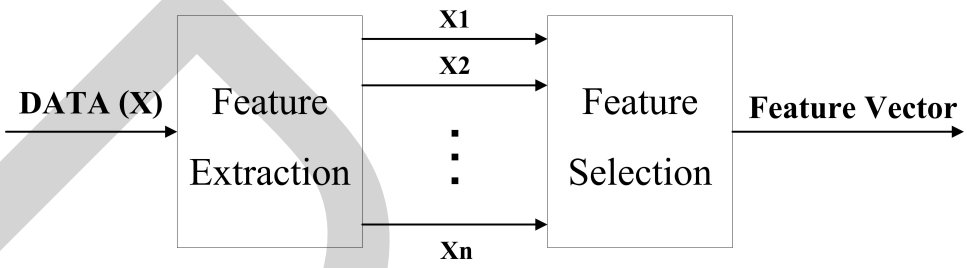
นำความถี่ของคำที่เหลือจากการกำจัดคำหยุด เข้าสู่กระบวนการหาค่าน้ำหนักคำ ในเอกสารนำเข้า TFIDF สมการ (3-1) ดังแสดงกระบวนการในรูปที่ 3.7 นำค่าน้ำหนักที่อยู่ในรูปของเมตริกซ์เวกเตอร์ เป็นอินพุตเข้ากระบวนการคัดแยกประเภทเอกสารต่อไป



รูปที่ 3.7 การหาค่าน้ำหนักคำของแต่ละเอกสาร

วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะเอกสารคือการดึงคุณลักษณะ (Feature) ของเอกสารออกมา กับการลดขนาดเอกสารลง การดึงคุณลักษณะในงานวิจัยนี้จะใช้คำเดี่ยวเป็น

ตัวแทนคุณลักษณะของเอกสาร และให้ค่าน้ำหนักคำแทนค่าคุณลักษณะ (Feature vector) ของเอกสารนั้นๆ ดังแสดงกระบวนการคัดเลือก และสกัดคุณลักษณะในรูปที่ 3.8



รูปที่ 3.8 การเลือกคุณลักษณะ และสกัดค่าคุณลักษณะ

การคัดเลือกคุณลักษณะ คือการนำคำที่ไม่มีนัยสำคัญออก จำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่ โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี

3.3.1 การคัดเลือกคุณลักษณะ เป็นขั้นตอนแรกที่จะต้องทำก่อนในขั้นตอนการแยกประเภท งานวิจัยนี้จึงเลือกคำที่มีความถี่ของคำที่ปรากฏในเอกสารมากฉบับ เลือกมาเป็นตัวแทนของเอกสารทั้งในส่วนของเอกสารสำคัญ และเอกสารทั่วไป ดังแสดงตัวอย่างของคุณลักษณะ (Feature) ดังต่อไปนี้

3.3.1.1 เอกสารสำคัญ แต่ละเอกสารจะปรากฏด้วยคำสำคัญที่ชี้ได้ว่าเป็นเอกสารสำคัญ ดังต่อไปนี้ [หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง งบประมาณ เบิกเอกสาร ฝบบ เห็นชอบ โปรแกรม สาขา SAP สารสนเทศ ศทส]

3.3.1.2 เอกสารทั่วไป แต่ละเอกสารปรากฏคำสำคัญที่ชี้ว่าเป็นเอกสารทั่วไป ดังต่อไปนี้ [ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุน เทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข]

ตารางที่ 3.2 แสดงตัวอย่างค่าเวกเตอร์ลักษณะสำคัญ (Feature Vector)

เอกสาร	คุณลักษณะที่คัดเลือกจากเอกสารสำคัญ						คุณลักษณะที่คัดเลือกจากเอกสารทั่วไป					
	งบประมาณ	หน่วยงาน	ดำเนินการ	...	SAP	เบิก	ส่งเสริม	เงื่อนไข	ลงทุน	...	ระดับ	ปัญหา
Doc 1	0.307	0.271	0.271	...	0.000	0.000	0.000	0.000	0.236	...	0.000	0.000
Doc 2	0.000	0.236	0.236	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
Doc 3	0.000	0.379	0.236	...	0.236	0.000	0.000	0.000	0.000	...	0.000	0.000
Doc 4	0.000	0.236	0.214	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
Doc 5	0.236	0.271	0.236	...	0.286	0.000	0.000	0.000	0.000	...	0.000	0.307
Doc 61	0.000	0.000	0.000	...	0.000	0.000	0.000	-0.162	0.000	...	-0.162	0.000
Doc 62	0.000	0.000	0.000	...	0.000	0.000	-0.162	0.000	0.000	...	-0.162	0.000
Doc 63	0.000	0.000	0.000	...	0.000	0.000	-0.162	0.000	-0.162	...	0.000	0.000
Doc 64	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.038
Doc 65	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.000	-0.085

นำคำทั้งหมดมาจัดเรียงตามลำดับจากความถี่ที่ปรากฏจากน้อยไปมาก ทำการเลือกคำสำคัญที่มีความถี่มากที่สุดมาเป็นคุณลักษณะสำคัญ (Feature) จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งกลุ่ม ถือว่าเป็นการสร้าง เวกเตอร์ตัวแทนของเอกสาร ด้วยค่าน้ำหนักคำตามคุณลักษณะที่ได้ทำการคัดเลือกไว้ ดังแสดงตัวอย่างเวกเตอร์คุณลักษณะ ดังตารางที่ 3.2 งานวิจัยนี้ใช้วิธี TF-IDF ซึ่งคำนวณจากค่าความถี่ของคำที่ปรากฏในเอกสารหนึ่งๆ เทียบกับคำที่ปรากฏในเอกสารทุกฉบับในชุดเอกสาร คำที่ปรากฏในเอกสารมาก มีความถี่มาก ก็จะส่งผลให้ได้ค่าน้ำหนักมีค่าสูงมากตาม เมื่อถึงขั้นตอนนี้จะได้รูปแบบที่มีลักษณะของการแสดงความสัมพันธ์ระหว่างคำ (Words :Feature) และเอกสารทั้งหมด (Document: DOC) ทำการเลือกคุณลักษณะที่เลือกมาจากชุดเอกสารสำคัญ {หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ...} และคุณลักษณะที่ถูกเลือกจากชุดเอกสารทั่วไป {ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ...} มาเป็นตัวแทนของเอกสารแต่ละฉบับ ซึ่งจะให้ค่าน้ำหนักคำของแต่ละคำ ในแต่ละเอกสาร เพื่อเป็นตัวแทนของเอกสาร

ค่าน้ำหนักคำเป็นค่าคุณลักษณะนำมาเป็นตัวแทนของเอกสาร การลดขนาดของค่าคุณลักษณะ (Dimention Reduction) เพื่อเตรียมข้อมูล สำหรับเป็นตัวแทนเอกสารในการคัดแยกประเภทเอกสารสาเหตุที่ต้องลดขนาดของค่าคุณลักษณะเนื่องจากเมื่อข้อมูลมีค่าคุณลักษณะสูง และไม่ได้ไปในทิศทางเดียวกันทำให้ข้อมูลเกิดการกระจาย จะทำให้บางจุดอาจจะไม่มีข้อมูลอยู่เลย ดังนั้นการที่มีข้อมูล และขนาดของข้อมูลจำนวนมาก นอกจากจะทำให้ใช้เวลาในการคำนวณมาก

แล้ว ยังทำให้เกิดปัญหาในเรื่องของการตีความกลุ่มข้อมูล ซึ่งถ้ามีตัวแปรของข้อมูลต่ำ และค่าของข้อมูลที่ไม่ห่างกันมาก ก็จะทำให้ข้อมูลมีลักษณะของการเกาะกลุ่มกัน ในทางตรงกันข้ามถ้าข้อมูลมีค่าสูง และแตกต่างกันมาก ก็อาจจะทำให้ถูกตีความว่าเป็นคนละกลุ่มกันได้ งานวิจัยนี้จึงได้ทำการลดค่าคุณลักษณะของข้อมูลเพื่อลดตัวแปรนำเข้า

การลดค่าคุณลักษณะของข้อมูลนำเข้า เพื่อให้การสร้างโมเดลทำได้ง่าย และช่วยให้เพิ่มความแม่นยำของการวิเคราะห์ข้อมูล โดยมีจุดมุ่งหมายุในการลดค่าคุณลักษณะเพื่อเจาะจงกลุ่มของคุณลักษณะ และทำให้ได้ผลลัพธ์ ของการพยากรณ์สามารถนำไปใช้ประโยชน์ได้เต็มที่ ก่อนจะนำค่าคุณลักษณะไปเป็นตัวแทนของเอกสาร และนำเข้าสู่กระบวนการฝึกสอนการคัดแยกประเภทเอกสารใน LS-SVM

3.3.2 การเลือกคุณลักษณะสำคัญ การเลือกคำที่มีความสำคัญน้อยออก เพื่อดูประสิทธิภาพในการทำนายหลังจากที่ได้ตัดคำบางตัวออกซึ่งส่วนใหญ่จะให้ค่าความถูกต้องสูงขึ้น จำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร ตัวแทนคุณลักษณะของเอกสารและใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศ และต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของเอกสาร และใช้พื้นฐานค่าความถี่หรือค่าน้ำหนักคำ เป็นค่าของคุณลักษณะ ตัวแทนคุณลักษณะของเอกสารจะเก็บอยู่ในรูปแบบเวกเตอร์ งานวิจัยนี้ใช้การเลือกคุณลักษณะแบบคำเดียว (Single word) ซึ่งได้จากขั้นตอนการตัดคำโดยใช้พจนานุกรมเรียบร้อยแล้ว ผลลัพธ์ที่ได้จากการตัดคำจะได้เป็นคำเดียวจำนวนมาก เพื่อคัดเลือกมาใช้เป็นตัวแทนเอกสารในการเรียนรู้ โดยใช้วิธีเลือกคำที่มีความถี่มากที่สุด หรือคำที่ปรากฏในเอกสารมากฉบับ ดังแสดงตามสมการ 3-1, 3-2 และตารางที่ 3.3 แสดงตัวอย่างของตัวแปรที่ใช้ในสมการ เมื่อค่าคุณลักษณะถูกแทนด้วยค่าน้ำหนักคำ w_{ij}

ตารางที่ 3.3 ค่าน้ำหนักคำของเอกสาร D_i และคำ W_j

Document/*Word*	W_1	W_2	W_3	…	W_n
D_1	w_{11}	w_{12}	w_{13}	…	w_{1n}
D_2	w_{21}	w_{22}	w_{23}	…	w_{2n}
D_3	w_{31}	w_{32}	w_{33}	…	w_{3n}
…	…	…	…	…	…
D_m	w_{m1}	w_{m2}	w_{m3}	…	w_{mn}

จากตาราง 3.3 แสดงจำนวนเอกสาร $D_1, D_2, ..., D_m$ ที่ปรากฏคำ $W_1, W_2, ..., W_n$ และค่าน้ำหนักคำ w_{mn} ซึ่งแสดค่าน้ำหนักคำของเอกสาร D_m และคำ W_n และแสดงวิธีการหาความถี่ C_j ของการปรากฏคำ W_n ในเอกสาร $D_1, D_2, ..., D_m$ ดังสมการ (3-1) โดยถ้าหากคำ W_n มีค่ามากกว่าค่า 0 จะถูกนับเป็น 1 $(x_{ji} = 1)$ แต่ถ้าหากคำ W_n ไม่ค่าเท่ากับ 0 $(x_{ji} = 0)$จะไม่ถูกนับ ดังสมการ (3-2)

$$C_j = \sum_{i=1}^{m} x_{ij} \tag{3-1}$$

$$x_{ij} = \begin{cases} 1 & if \quad w_{mn} > 0 \\ 0 & if \quad w_{mn} = 0 \end{cases} \tag{3-2}$$

กำหนดให้ w_{ij} เป็นค่าน้ำหนักคำที่ j เอกสารที่ i

โดยที่ $j = 1,2,...,n$

$i = 1,2,...,m$

จากสมการ 3-1 แสดงวิธีการคัดเลือกคุณลักษณะ เมื่อ C_j คือคุณลักษณะที่คำนวณจากความถี่ของคำ x_{ij} จากเอกสาร D_m ที่ $i = 1,...,m$ และคำ W_n ที่ $j = 1,...,n$ เมื่อ x_{ij} จะถูกนับเป็น 1 เมื่อค่าน้ำหนักคำ w_{mn} มีค่ามากกว่า 0 ตามสมการที่ 3-2 คัดเลือกคุณลักษณะโดยเรียงความถี่คุณลักษณะที่ปรากฏ C_j เลือกคุณลักษณะที่มีความถี่ที่ปรากฏคำในเอกสารมากฉบับ ที่เรียงจากมากไปยังน้อย ตามจำนวนคุณลักษณะที่ต้องการใช้ เมื่อค่าคุณลักษณะถูกแทนด้วยค่าน้ำหนักคำ w_{mn} ของคำที่ n และเอกสารที่ m

Assume sorted data $M(C_1), ..., M(C_j)$

Binsearch (x)

put left, right fingers on 0 and C_j

while x not equal M(left finger)

set midpoint halfway between fingers

if M(midpoint) < x

then move left finger to midpoint

else move right finger to midpoint

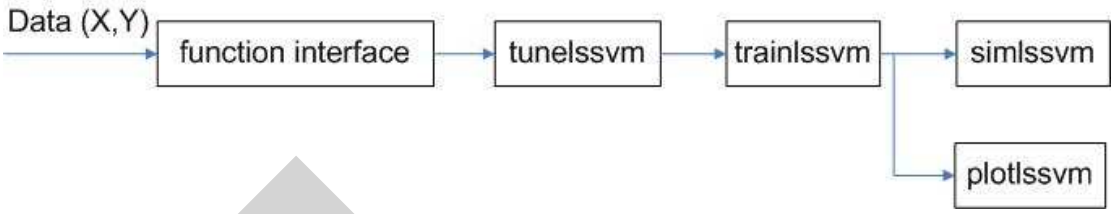
end while

แสดงขั้นตอนจำลองการคัดเลือกคุณลักษณะโดยการเรียงลำดับตามความถี่ของคำที่ปรากฏในเอกสารมากฉบับจากมากไปหาน้อย โดยทำการเรียงลำดับข้อมูล C_1 ถึง C_j ตามรหัสเทียม (Pseudo Code) ที่แสดงถึงขั้นตอนการเรียงลำดับด้วยภาษาที่เข้าใจง่าย เรียงลำดับแบบไบนารี่เสิร์ช นำข้อมูลมาแบ่งเป็น 2 ส่วน แล้วนำข้อมูลเปรียบเทียบกับข้อมูลตำแหน่งกลาง M(midpoint) โดยให้ค่า $M(C_j)$ เริ่มต้นเป็นค่า x ถ้ามีค่าน้อยกว่าให้นำมาเรียงในส่วนหน้า แต่ถ้ามากกว่าให้จัดเรียงไว้ในส่วนหลัง ซึ่งเป็นวิธีการเรียงลำดับที่มีประสิทธิภาพสูงเพราะไม่จำเป็นต้องคิดตั้งแต่ตัวแรกทำให้สามารถลดระยะเวลาในการจัดลำดับได้เป็นอย่างดี

3.4 กระบวนการแยกประเภทเอกสาร

วิธีการที่ใช้ในการแยกประเภทเอกสารในวิทยานิพนธ์เล่มนี้คือ LS-SVM เนื่องจากมีความแม่นยำในการแยกประเภท กับข้อมูลปริมาณมาก เมื่อได้ค่าน้ำหนัก TF-IDF ของค่าในเอกสารทั้งหมดแล้วแต่ละฉบับ นำค่าที่ได้นี้เป็นอินพุตเข้ากระบวนการแยกประเภท โดยใช้วิธี LS-SVM วิธีนี้จะต้องให้โปรแกรมทำการเรียนรู้ข้อมูลก่อนที่จะทำการเปรียบเทียบเพื่อแบ่งประเภท ซึ่งผู้วิจัยได้ทำการคัดเลือกคุณลักษณะสำคัญ (Feature Selector) และนำค่าน้ำหนักคำ (Feature Extraction) เพื่อใช้แทนคุณลักษณะของเอกสาร โดยกำหนดประเภทของเอกสารไว้ ให้ประเภทที่ 1 คือประเภทเอกสารสำคัญ (Secret) ประเภทที่ 2 คือประเภทเอกสารทั่วไป (No Secret) แทนด้วยค่าตัวเลข 1 และ -1 ตามลำดับ จากนั้นนำไฟล์น้ำหนักคำตามคุณลักษณะที่ได้คัดเลือกไว้แล้วเป็นข้อมูลนำเข้า ตามทฤษฎีของ LS-SVM

จากทฤษฎี SVM แนวคิดหลักของวิธีการนี้ ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลอออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม๊ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space โดยมีคำตอบที่เป็นไปได้คือ $y = \{+1, -1\}$ โดยค่า y นี้จะเป็นผลลัพธ์ที่ต้องการให้ SVM เรียนรู้ เพื่อใช้สำหรับแยกกลุ่มข้อมูลทั้งสองกลุ่มออกจากกัน เมื่อทำการกำหนดกลุ่มของข้อมูลที่ต้องการจะแบ่งเรียบร้อยแล้ว จะมีการสร้างเส้นแบ่งข้อมูลที่เหมาะสม (Optimal Separate Hyperplane) เพื่อแยกประเภท การสร้างเส้นแบ่งเขตนั้นเพื่อทำการตรวจสอบการแบ่งข้อมูล แสดงขั้นตอนการทำงาน ระบุรายการของฟังก์ชั่นที่ถูกใช้ใน LS-SVM Model ที่นำมาใช้เป็นเครื่องมือ (tool) ในโปรแกรม MATLAB ดังรูปที่ 3.9

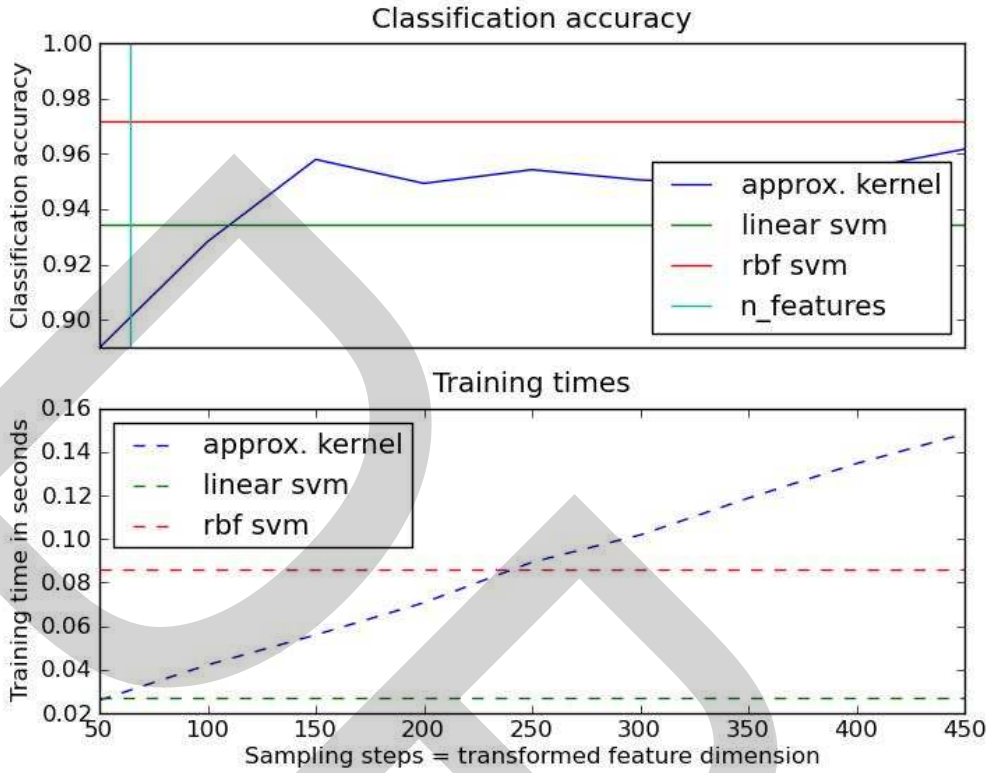


รูปที่ 3.9 ขั้นตอนการทำงานของ LS-SVM Model

อัลกอริทึมการจัดหมวดหมู่ (Classifier Algorithm) อัลกอริทึมในการจัดหมวดหมู่การเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจัดหมวดหมู่เอกสารแบ่งออกเป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มเอกสารต้นแบบ และแยกหมวดหมู่ของเอกสารที่สนใจ โดยตรวจสอบหาความคล้ายกับกลุ่มเอกสารต้นแบบ โดยข้อมูลนำเข้าอัลกอริทึมเครื่องจักรการเรียนรู้ คือคุณลักษณะของคำเดียวที่อยู่ในรูปแบบเวกเตอร์ ที่ผ่านการลดคุณลักษณะของค่าน้ำหนักมาเรียบร้อยแล้ว

ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) วิธีการนี้ใช้หาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูลสองกลุ่มออกจากกัน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม๊ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space อย่างไรก็ตาม SVM มี เคอร์เนลฟังก์ชั่น (Kernel Function) ที่สามารถประยุกต์ใช้ในการแก้ปัญหาได้หลายวิธี Kernel Function ที่พบได้บ่อย 3 แบบด้วยกัน เช่น Radial Basis Function (RBF), Polynomial และ Sigmoid สำหรับงานวิจัยนี้ได้เลือกใช้ RBF Kernel (Radial Basis Function) เป็นฟังก์ชั่นที่ใช้ในการทดลอง เนื่องจากมีผู้วิจัยหลายท่านได้ทำการทดสอบประสิทธิภาพการใช้งาน Kernel Function พบว่า RBF เป็น Kernel Function ที่ให้ผลการทดสอบมีประสิทธิภาพมากที่สุด ดูได้จากกราฟเปรียบเทียบประสิทธิภาพการทำงานของ Kernel ดังรูปที่ 3.10

โดยทั่วไปแล้วการแยกประเภทเอกสาร จะมีการเลือกใช้เคอร์แนลอยู่ 2 อย่างคือ Linear Kernel และ RBF Kernel จากรูปที่ 3.10 เป็นการเปรียบเทียบประสิทธิภาพการทำงานของ Kernel Function ในส่วนของความถูกต้องในการคัดแยกประเภท (Classify Accuracy) และเวลาที่ใช้ในกระบวนการฝึกสอน (Training Time) ระหว่าง Linear Kernel และ RBF Kernel จากกราฟแสดงเวลา และค่าความถูกต้องสำหรับการแยกประเภทกับ Kernel การประมาณการ โดยทำการแยกประเภทกับชุดข้อมูลตัวอย่าง พบว่าเรเดียลเบสิสฟังก์ชั่น (RBF Kernel) ให้ค่าความถูกต้อง (Classify Accuracy) มากกว่า และใช้เวลาในการประมวลผลน้อยกว่าเคอร์เนลเชิงเส้น (Linear Kernel)



รูปที่ 3.10 กราฟเปรียบเทียบประสิทธิภาพของ Kernel Function

ที่มา: Explicit feature map approximation for RBF kernels http://scikit-learn.sourceforge.net

3.5 อัตราความถูกต้อง (Accuracy Rate)

วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพ คือความสามารถในการเข้าใกล้ค่าจริง (True Value) ของการวัดผล โดยหาได้จาก สมการ 3-3

$$\frac{Total\ Document\ read - Total\ Error}{Total\ Document\ read} \times 100 = Accuracy\ Rate\ (\%) \qquad (3\text{-}4)$$

บทที่ 4
ผลการศึกษา

ในส่วนนี้จะกล่าวถึงการทดสอบ ผลการทดสอบการแยกประเภทเอกสาร และรวมถึงการวิเคราะห์ความแม่นยำของข้อมูลชุดฝึกสอน โดยข้อมูลที่ใช้ในการทดสอบจะประกอบไปด้วยข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบ โดยแต่ละชุดข้อมูลจะประกอบด้วยข้อมูลที่ถูกกำหนดประเภทไว้ล่วงหน้าทั้งข้อมูลที่เป็นเอกสารสำคัญ (Secret) และข้อมูลที่เป็นเอกสารทั่วไป (No Secret) ข้อมูลนำเข้าจะเป็นไฟล์เอกสาร Microsoft word (.doc) เท่านั้น เป็นเอกสารที่ถือว่าสิ้นสุดไม่มีการเปลี่ยนแปลงเนื้อหาในเอกสาร และระดับความสำคัญ จะแสดงวิธีการทดสอบตามกระบวนการที่ได้ออกแบบไว้ในแต่ละขั้นตอน แสดงผลการทดสอบการแยกประเภทเอกสาร ทั้งการนำไฟล์เอกสาร Microsoft word (.doc) เข้าไปยังส่วนของโปรแกรม PHP ที่ประกอบไปด้วยโปรแกรมย่อย Antiword สำหรับดูไฟล์เอกสาร Microsoft word ในรูปแบบของไฟล์ข้อความธรรมดา Text mode ทำการแปลงไฟล์ (.doc) ให้อยู่ในรูปแบบไฟล์ข้อความ (.txt) นำไฟล์ข้อความเข้าสู่ขั้นตอนการตัดคำในเอกสารภาษาไทย โดยใช้โปรแกรม SWATH การทำงานของโปรแกรมจะทำการเปรียบเทียบคำในเอกสารกับข้อมูลที่มีอยู่ในพจนานุกรมในลักษณะเรียงจากซ้ายไปขวา จะทำการเลือกคำที่ยาวที่สุดที่ตรงกัน และตัดคำนั้นออกมา คำที่ได้ทั้งหมดจะนำเข้าสู่กระบวนการตัดคำหยุด หรือคำที่ไม่สื่อความหมายออก โดยมีการระบุคำเหล่านั้นไว้ในอาร์เรย์ส่วนของการตัดคำในโปรแกรม PHP นำคำที่เหลือจากการตัดคำหยุดมาหาค่าน้ำหนัก TF-IDF (Term Frequency and Inverse Document Frequency) เป็นกระบวนการคำนวณค่าน้ำหนักจากความถี่ของคำที่ปรากฏในเอกสารนั้นๆ เทียบกับเอกสารอื่นที่เป็นชุดข้อมูลเดียวกัน ทำการคัดเลือกคุณลักษณะสำคัญ (Feature Selector) และนำค่าน้ำหนักคำ (Feature Vector) เพื่อใช้แทนคุณลักษณะของเอกสารด้วยทำการลดขนาดค่าคุณลักษณะเพื่อเพิ่มประสิทธิภาพในการคัดแยกประเภทเอกสาร ค่าคุณลักษณะจะมีรูปแบบเป็นไฟล์ข้อมูลเมตริกซ์เพื่อทำการแยกประเภทเอกสาร โดยใช้ LS-SVM (Least Squares Support Vector Machine) เป็นกระบวนการแยกประเภทแบบมีผู้สอน (Supervised Learning) ผ่านทางโปรแกรม MATLAB ภาษาคอมพิวเตอร์ที่มีประสิทธิภาพสูงทางด้านเทคนิคในการคำนวณ มีรูปแบบโปรแกรมที่ง่ายต่อการใช้งาน ผลลัพธ์ที่ได้จะสามารถแยกประเภทของเอกสารออกเป็นสองส่วนตามที่ได้กำหนดไว้ข้างต้น คือ เอกสารสำคัญ Secret (1) และเอกสาร

ทั่วไป No Secret(-1) ทำการทดสอบอัตราความถูกต้อง (Accuracy Rate) ของการกรองเอกสาร โดยนำเอกสารชุดทดสอบมาแยกประเภทผ่านข้อมูลชุดฝึกสอน

4.1 กระบวนการทดสอบ

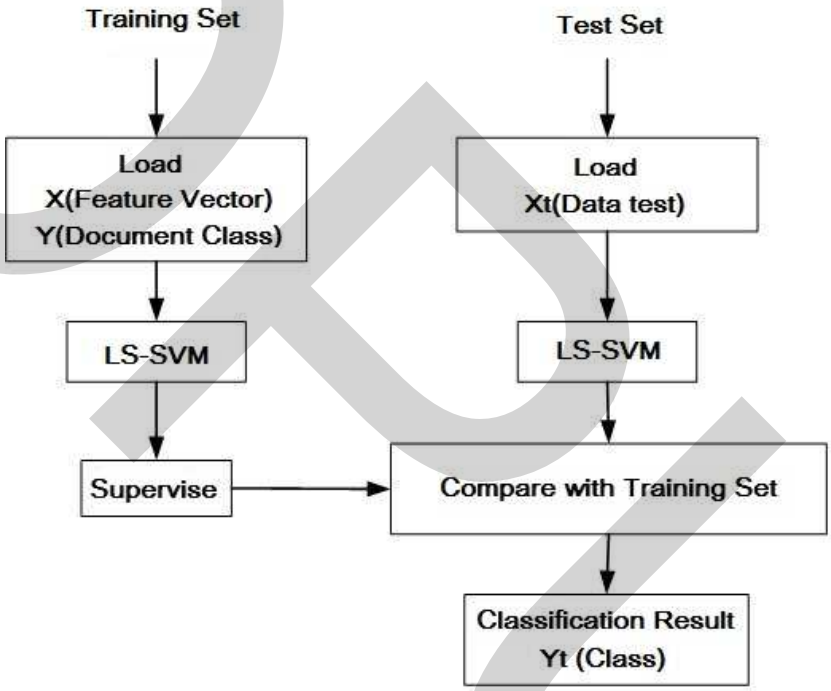
วิธีการแยกประเภทคือ การนำข้อมูลทั้งหมดมาทำการฝึกสอนโดยแยกประเภทให้ข้อมูลชุดฝึกสอนก่อนว่าจัดอยู่ในประเภทเป็นเอกสารความสำคัญ Secret (1) หรือเอกสารทั่วไป No Secret (-1) ทำการทดสอบโดยการนำข้อมูลส่วนหนึ่งของข้อมูลชุดฝึกสอนมาทำการทดสอบตามประเภทที่ได้กำหนดไว้ เริ่มจากนำข้อมูลเอกสารทั้งหมด 145 ฉบับซึ่งถือว่าเป็นข้อมูลชุดฝึกสอน โดยในข้อมูลชุดฝึกสอน จะประกอบด้วยข้อมูลที่เป็นเอกสารสำคัญ และเอกสารทั่วไปประกอบอยู่ ซึ่งแบ่งเป็นประเภทเอกสารดังแสดงในตารางที่ 4.1

ตารางที่ 4.1 รายละเอียดของเอกสารที่ใช้ในการฝึกสอนและทดสอบการแยกประเภทเอกสาร

ชุดข้อมูลสำหรับ (Data Set)	**ชุดข้อมูลตั้งต้น**
ข้อมูลชุดฝึกสอน (Training Set)	145 ข้อมูล
ข้อมูลชุดทดสอบ (Test Set)	60 ข้อมูล/เอกสารสำคัญ 85 ข้อมูล/เอกสารทั่วไป

นำเอกสาร Microsoft Word (.doc) เข้าสู่กระบวนการแปลงให้อยู่ในรูปแบบไฟล์ข้อความธรรมดา (.txt) โดยผ่านโปรแกรม PHP ด้วยโปรแกรมย่อย Antiword นำไฟล์ข้อความที่ได้เข้าสู่กระบวนการตัดคำภาษาไทยโดยการเปรียบเทียบกับพจนานุกรมที่มีโดยใช้โปรแกรมย่อย SWATH นำคำที่ได้มาเปรียบเทียบกับชุดคำหยุด (Stop word) ที่ได้กำหนดไว้ นำคำที่เหลือจากการตัดคำหยุดมาหาค่าน้ำหนักคำโดยใช้ความถี่ของคำในเอกสาร มีแนวคิดที่ว่าคำที่ปรากฏในเอกสารน้อยฉบับจะมีค่าน้ำหนักสูงส่วนคำที่ปรากฏในเอกสารหลายฉบับ จะมีค่าน้ำหนักต่ำ เนื่องจากเป็นคำ ที่ไม่แสดงถึงลักษณะเฉพาะของเอกสารนั้นโดยใช้สมการ TF-IDF เมื่อได้ค่าน้ำหนักคำในเอกสาร ตัวอย่างค่าน้ำหนักดังตารางที่ 3.2 จะแสดงลำดับเอกสาร โดยแต่ละฉบับจะแสดงค่าน้ำหนักของแต่ละคำที่ค้นพบในแต่ละเอกสาร คำนวณจากความถี่ของคำที่ปรากฏในเอกสารนั้นๆเทียบกับเอกสารทั้งหมดในชุดข้อมูล และทำการคัดเลือกคุณลักษณะสำคัญ จากความถี่ของคำที่ปรากฏมากครั้งในชุดข้อมูลนั้นๆ ทั้งในส่วนของเอกสารสำคัญ และเอกสารทั่วไป และนำค่าน้ำหนักมาเป็นค่าคุณลักษณะ แต่ก่อนจะนำเข้าไปเข้ากระบวนการฝึกสอนเพื่อคัดแยก

ประเภทเอกสารนั้น จำเป็นต้องทำการลดค่าคุณลักษณะ มีจุดมุ่งหมายเพื่อเจาะจงกลุ่มของคุณลักษณะ และทำให้ได้ผลลัพธ์ที่มีความแม่นยำ และนำค่าคุณลักษณะนี้ไปเป็นเมตริกซ์อินพุตเข้าสู่กระบวนการแยกประเภทเอกสารด้วย LS-SVM ผ่านโปรแกรม MATLAB เมื่อทำการฝึกสอนแล้ว จะทำการนำข้อมูลชุดทดสอบเข้าสู่กระบวนการแยกประเภทอีกครั้ง และดูผลการทดสอบว่าผลการแยกประเภทจะตรงกับประเภทเอกสารที่ระบบได้เรียนรู้ไว้หรือไม่ ขั้นตอนที่กล่าวมาแสดงขั้นตอนได้ดังรูปที่ 4.1



รูปที่ 4.1 ขั้นตอนการฝึกสอนและทดสอบการแยกประเภทเอกสาร โดยใช้ LS-SVM

โดยที่ $X = N \times d$ เมตริกซ์ค่าลักษณะสำคัญ (Feature Vector) สำหรับข้อมูลฝึกสอน

$Y = N \times 1$ เป็นเมตริกซ์ของประเภทสำของข้อมูลฝึกสอน

$Xt = M \times d$ เป็นเมตริกซ์ของค่านำเข้าสำหรับข้อมูลชุดทดสอบ

$Yt = M \times 1$ เป็นเมตริกซ์ของผลลัพธ์ที่ได้จากการทดสอบ

เมื่อ $N =$ เอกสารนำเข้า

$d =$ คุณลักษณะสำคัญ

$M =$ เอกสารนำเข้าของข้อมูลชุดทดสอบ

$Y = \{+1,-1\}$, $Yt = \{+1,-1\}$

2.2.1 ขั้นตอนการทดสอบ การคัดเลือกคุณลักษณะสำคัญ (Feature Selection) โดยเลือกจากสองส่วน จำนวน 40 คุณลักษณะ โดยแบ่งเป็นคุณลักษณะจากเอกสารสำคัญ จำนวน 20 คำ มีคุณลักษณะ (Feature) เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับ เรียงจากมากไปน้อย [หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง งบประมาณ เบิก เอกสาร ฝบบ เห็นชอบ โปรแกรม สาขา SAP สารสนเทศ ศทส] โดยมีค่าน้ำหนักคำในแต่ละเอกสาร และคุณลักษณะจากเอกสารทั่วไป จำนวน 20 คำ มีคุณลักษณะ (Feature) เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุน เทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข] เมื่อเลือกคุณลักษณะเพื่อเป็นตัวแทนของเอกสารได้แล้ว ขั้นตอนต่อมาคือนำค่าน้ำหนักของคำเหล่านี้มาเป็นค่าคุณลักษณะ (Feature Vector) ดังแสดงในตารางที่ 4.2 น้ำหนักคำของแต่ละเอกสารโดยผ่านกระบวนการหาค่าน้ำหนักคำ TF-IDF

ตารางที่ 4.2 ค่าน้ำหนักคำของแต่ละเอกสารที่แบ่งตามค่าคุณลักษณะ (Feature) ที่ได้เลือกไว้

ประเภทเอกสาร		คุณลักษณะที่คัดเลือกจากเอกสารสำคัญ						คุณลักษณะที่คัดเลือกจากเอกสารทั่วไป					
		งบประมาณ	หน่วยงาน	ดำเนินการ	...	SAP	เบิก	ส่งเสริม	เงื่อนไข	พาณิชย์	...	ระดับ	ปัญหา
เอกสารสำคัญ	Doc 1	0.307	0.271	0.271	...	0.000	0.000	0.000	0.000	0.236	...	0.000	0.000
	Doc 2	0.000	0.236	0.236	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
	Doc 3	0.000	0.379	0.236	...	0.236	0.000	0.000	0.000	0.000	...	0.000	0.000
	Doc 4	0.000	0.236	0.214	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.000
	Doc 5	0.236	0.271	0.236	...	0.286	0.000	0.000	0.000	0.000	...	0.000	0.307
เอกสารทั่วไป	Doc 61	0.000	0.000	0.000	...	0.000	0.000	0.000	0.038	0.000	...	0.038	0.000
	Doc 62	0.000	0.000	0.000	...	0.000	0.000	0.038	0.000	0.000	...	0.038	0.000
	Doc 63	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.077
	Doc 64	0.000	0.000	0.000	...	0.000	0.000	0.038	0.000	0.038	...	0.000	0.000
	Doc 65	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.115

จากค่าน้ำหนักคำของแต่ละคำที่เลือกมาเป็นคุณลักษณะ โดยผ่านกระบวนการหาค่าน้ำหนักคำ TF-IDF จำนวน 145 เอกสาร โดยแบ่งเป็นเอกสารสำคัญ 60 เอกสาร และเอกสารทั่วไป จำนวน 85 เอกสาร โดยคัดเลือกคุณลักษณะของแต่ละประเภทเอกสารออกมา เมื่อได้ค่าคุณลักษณะออกมาแล้ว สังเกตได้ว่าคุณลักษณะจะเป็นตัวบ่งบอกประเภทของเอกสารได้อย่างชัดเจน คือค่าน้ำหนักของคำคุณลักษณะจะมีค่ามากในประเภทของเอกสารนั้นๆ ยกตัวอย่างเช่นในส่วนของคุณลักษณะที่ถูกเลือกจากเอกสารสำคัญ ซึ่งก็คือเอกสารที่ 1 – 60 ค่าน้ำหนักคำของคำที่ถูกเลือกมา

เป็นคุณลักษณะจะมีค่าน้ำหนักคำปรากฏอยู่ แต่ในส่วนของค่าน้ำหนักคำของคำที่ถูกเลือกเป็นคุณลักษณะของเอกสารทั่วไป จะมีค่าน้ำหนักคำปรากฏเป็น 0 หรือปรากฏค่าน้อยมากในเอกสารที่ 1-60 ในกรณีตรงกันข้าม ในส่วนของเอกสารทั่วไป 61-145 จะเห็นได้ว่า ค่าน้ำหนักจะไปปรากฏในส่วนของคำคุณลักษณะที่ถูกเลือกคุณลักษณะตัวแทนของเอกสารทั่วไป แต่ในส่วนของคำคุณลักษณะตัวแทนของเอกสารสำคัญ จะปรากฏค่าเป็น 0

ขั้นตอนต่อมาคือการลดขนาดของค่าคุณลักษณะ(Normalize) เนื่องจากข้อมูลดิบที่ได้มาอาจจะทำให้เกิดปัญหาในเรื่องของการตีความกลุ่มข้อมูล ซึ่งถ้ามีตัวแปรของข้อมูลต่ำ และค่าของข้อมูลที่ไม่ห่างกันมาก ก็จะทำให้ข้อมูลมีลักษณะของการเกาะกลุ่มกัน ในทางตรงกันข้ามถ้าข้อมูลมีค่าสูง และแตกต่างกันมาก ก็อาจจะจะทำให้ถูกตีความว่าเป็นคนละกลุ่มกันได้ งานวิจัยนี้จึงได้ทำการลดค่าคุณลักษณะของข้อมูลเพื่อลดตัวแปรนำเข้า เพื่อเพิ่มประสิทธิภาพของการคัดแยกประเภท

จากค่าคุณลักษณะที่ได้ หลังจากผ่านการลดค่าคุณลักษณะแล้ว ค่าที่ได้จะมีการกระจายตัวดีมากยิ่งขึ้น และค่าของข้อมูลไม่ห่างกันมากอยู่ระหว่างค่า 0 และ 1 นำข้อมูลค่าน้ำหนักเข้ากระบวนการคัดแยกประเภทเอกสารด้วยวิธี LS-SVM ร่วมกับค่าผลลัพธ์ของประเภทเอกสารแต่ละฉบับ ที่อยู่ในรูปเมตริกซ์ ถูกกำหนดประเภทไว้แทนด้วยตัวเลข (-1) คือเอกสารทั่วไป และ (+1) คือเอกสารสำคัญ เป็นข้อมูลนำเข้าอีกตัวหนึ่งในชุดการฝึกสอน

ผลการทดสอบดูจากค่า Yt (ผลที่ได้จากการแยกประเภท) และเปรียบเทียบกับค่า Y (ประเภทของข้อมูลชุดฝึกสอน) ที่กำหนดไว้ ทำการทดสอบอัตราความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสาร โดยนำเอกสารชุดทดสอบมาคัดแยกประเภทเอกสารผ่านข้อมูลชุดฝึกสอน และหาค่าความถูกต้องจากผลลัพธ์ของการแยกประเภทที่ได้

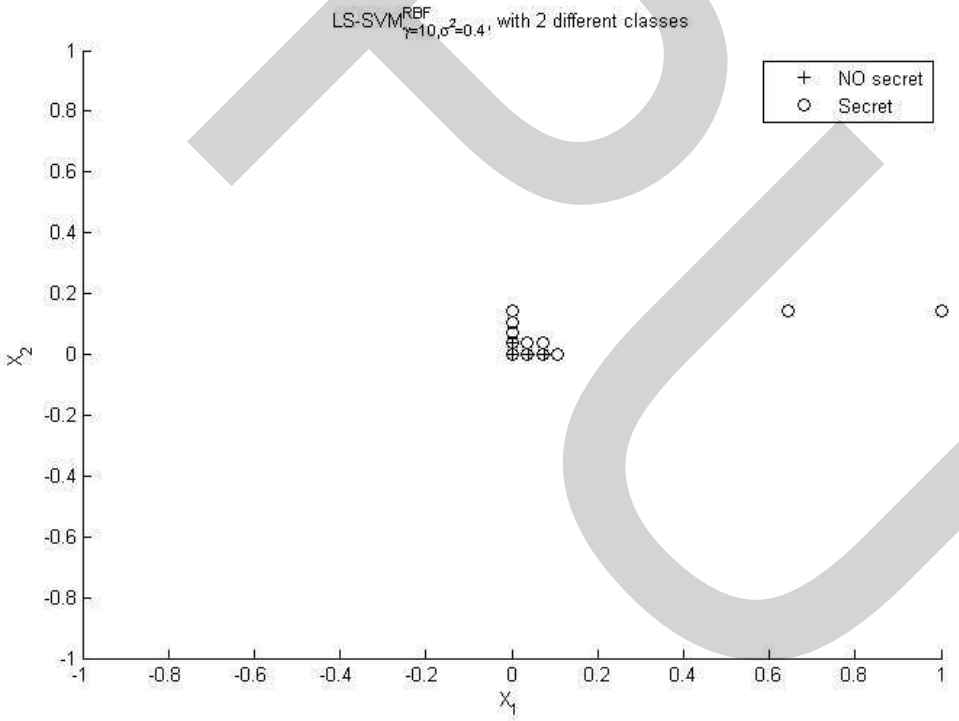
4.2 การวัดประสิทธิภาพ

การวัดประสิทธิภาพการคัดแยกประเภทเอกสารสำคัญ ออกจากเอกสารทั่วไป โดยคำนวณอัตราความถูกต้อง (Accuracy Rate) เป็นวิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพ ของการเลือกใช้จำนวนคุณลักษณะ ในแต่ละการทดลองซึ่งการหาอัตราความถูกต้อง คือความสามารถในการเข้าใกล้ค่าจริง (True Value) ของการวัดผล โดยหาได้จาก สมการ 4-1

$$\frac{Total\ Document\ read - Total\ Error}{Total\ Document\ read} \times 100 = Accuracy\ Rate\ (\%) \qquad (4\text{-}1)$$

4.3 ผลการทดสอบ

การทดสอบการแยกประเภทเอกสารของข้อมูลชุดฝึกสอน ซึ่งประกอบไปด้วย เอกสารสำคัญ 60 เอกสาร และเอกสารทั่วไปจำนวน 85 เอกสาร จำนวนคุณลักษณะสำคัญ (Feature) 40 คุณลักษณะ โดยเลือกคุณลักษณะจากชุดเอกสารสำคัญ 20 คุณลักษณะ และเอกสารทั่วไป 20 คุณลักษณะ หาประสิทธิภาพการคัดแยกประเภทเอกสารสำคัญ ออกจากเอกสารทั่วไป โดยวิธีการหาค่าความถูกต้อง (Accuracy Rate) โดยมีตัวแปรเป็นจำนวนคุณลักษณะที่เลือกใช้ในการทดสอบแต่ละครั้ง ในทุกๆเอกสารจะถูกแทนด้วยค่าน้ำหนักของแต่ละคำในชุดเอกสาร และถูกเลือกคำสำคัญมาเป็นคุณลักษณะ (Feature) เพื่อเป็นตัวแทนของเอกสาร หาค่าน้ำหนักคำ (Feature Vector) โดยใช้ทฤษฎี TF-IDF ประเภทเอกสารทั่วไปถูกแทนด้วยตัวเลข (-1) ส่วนเอกสารสำคัญถูกแทนด้วยตัวเลข (1) นำข้อมูลดังกล่าวที่อยู่ในรูปแบบของเมตริกซ์เข้าสู่ระบบการแยกประเภทเอกสาร LS-SVM ดังกราฟแสดงผลการแยกประเภทเอกสารของชุดข้อมูลฝึกสอนตามรูปที่ 4.2



รูปที่ 4.2 กราฟแสดงผลการแยกประเภทเอกสารชุดฝึกสอนด้วยวิธี LS-SVM จำนวน 145 เอกสาร โดยใช้จำนวนคุณลักษณะ 40 คุณลักษณะ

การคัดแยกประเภทของข้อมูลชุดฝึกสอน แสดงออกมาในรูปที่ 4.2 กราฟ LS-SVM ตามฟังก์ชั่นการพล๊อตกราฟของ LS-SVM Model ในโปรแกรม MATLAB ที่แสดงคลาส ที่

แตกต่างกัน 2 คลาส เครื่องหมาย + แสดงคลาสของเอกสารทั่วไป และ ○ แสดงคลาสของเอกสารสำคัญ บน Feature Space X_1 และ X_2 ที่เกิดจากฟังก์ชั่นแม๊ป SVM ที่ย้ายข้อมูลจาก Input Space (ข้อมูลต้นฉบับ) ไปยัง Feature Space (ข้อมูลที่ถูกแยกประเภทแล้ว) จากกราฟจะเห็นได้ว่าจุดข้อมูลเอกสารทั่วไปในส่วนของ Feature Space มีลักษณะเกาะกลุ่มกัน เนื่องจากค่าน้ำหนักคำมีค่าใกล้เคียงกัน และมีค่าน้ำหนักคำเป็น 0 จำนวนมาก รวมทั้งในส่วนของข้อมูลสำคัญ ที่มีการเกาะกลุ่มของข้อมูลเพราะค่าน้ำหนักใน Input Space มีค่าไม่ต่างกันมาก และมีค่าน้ำหนักเป็น 0 จำนวนมากเช่นกัน

ค่าสัมประสิทธิ์ที่ใช้ในการคำนวณของ LS-SVM ประกอบไปด้วยค่าแอลฟา (alpha) หรือ Support Value สนับสนุนการเพิ่มประสิทธิภาพในการคัดแยกประเภท , แกมมา (gam) เป็นค่า trade-off เพื่อลดความผิดพลาดในการคัดแยกประเภท ,ซิกมา (sig2) Kernel parameter หรือ ในกรณีใช้เคอร์เนลฟังชั่นเป็น เรเดียลเบสิสฟังก์ชั่น RBF และค่าไบอัส b หรือค่าความโน้มเอียงใช้สำหรับการจำแนกข้อมูล เป็นสัมประสิทธิ์ที่ได้จากการประมวลผลของ LS-SVM และค่าน้ำหนักเอกสารชุดฝึกสอนจำนวน 145 ฉบับ ซึ่งแบ่งเป็นเอกสารสำคัญ 60 เอกสาร เอกสารทั่วไป 85 เอกสาร คุณลักษณะสำคัญ 40 คำ และให้ผลลัพธ์ของประเภทเอกสารสำคัญแทนด้วย (1) และเอกสารทั่วไปแทนด้วย (-1) No Secret เพื่อแยกประเภทเอกสาร แสดงค่าดังตารางที่ 4.3 แสดงจำนวนชุดข้อมูลชุดฝึกสอน (Feature Vector) มีเมตริกซ์เป็น $X = 145 \times 40$ มีเอกสาร 145 ฉบับและมีคำสำคัญ (Feature) 40 คำในทุกๆ เอกสาร $Y = 145 \times 1$ แสดงประเภทของเอกสารชุดฝึกสอนของแต่ละเอกสาร

ตารางที่ 4.3 ค่าตัวแปรที่ได้จากผลการแยกประเภทเอกสารชุดฝึกสอน

ตัวแปร	ค่า	ต่ำสุด	สูงสุด
ค่าคุณลักษณะเอกสารชุดฝึกสอน	145×40	0	1
ประเภทของเอกสารชุดฝึกสอน	145×1	-1	1
Support Value (α)	145×1	-10.5819	15.1833
ค่าความโน้มเอียง (b)	0.1667	0.1667	0.1667
ค่า trade-off (gam)	10	10	10
Kernel Parameter (sig2)	0.4000	0.4000	0.4000
Type	*Classification*		

จากตารางที่ 4.3 แสดงผลลัพธ์ที่ได้จากการนำเวกเตอร์คุณลักษณะสำคัญของข้อมูลชุดฝึกสอน 145×40 กำหนดค่าคุณลักษณะจำนวน 40 คำ โดยมีค่าน้ำหนักต่ำสุดอยู่ที่ 0 และค่าน้ำหนักสูงสุดที่ 1 เวกเตอร์ประเภทของเอกสารชุดฝึกสอนของแต่ละเอกสาร $Y(145 \times 1)$ โดยกำหนดให้เป็น $[-1,1]$ กำหนดค่าสัมประสิทธิ์ trade off ($gam = 10$) และ Kernel Parameter ($sig2 = 0.4000$) เพื่อหาสัมประสิทธิ์ ค่าแอลฟา หรอ Support Value และค่าไบอัส b โดยค่าสัมประสิทธิ์ที่ได้จากการคำนวณ $\alpha = -10.5819$ 15.1833 และค่าไบอัส (bias) หรือค่าความโน้มเอียงสำหรับให้ระบบรู้จำ $b = 0.1667$ เพื่อให้เส้นแบ่งสามารถทำการแบ่งเอกสารได้เป็น 2 ประเภท คือ เอกสารสำคัญ (1) และเอกสารทั่วไป (-1) ตามที่ได้กำหนดไว้

นำข้อมูลชุดฝึกสอนที่ได้ นำมาทดสอบกับข้อมูลชุดทดสอบ ข้อมูลชุดทดสอบแบ่งออกเป็น 2 ประเภทคือ เอกสารชุดทดสอบที่อยู่ในกลุ่มเป้าหมาย และเอกสารชุดทดสอบที่อยู่นอกกลุ่มเป้าหมาย โดยจะแบ่งการทดสอบเป็น 4 กรณี คือ การทดสอบโดยใช้ค่าคุณลักษณะ 10, 20, 30 และ 40 คำ โดยเป็นคุณลักษณะที่ถูกเลือกจากคำที่ปรากฏความถี่ที่ปรากฏในเอกสารมากฉบับ โดยเรียงจากมากไปน้อย ซึ่งถูกคัดเลือกโดยใช้ข้อมูลชุดเดียวกัน เอกสารชุดทดสอบประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไปที่ยังไม่ได้ทำการคัดแยกประเภทดังตารางที่ 4.4

ตารางที่ 4.4 ค่าน้ำหนักคำตามคุณลักษณะของเอกสารชุดทดสอบจำนวน 60 เอกสาร

ประเภทเอกสาร		คุณลักษณะที่คัดเลือกจากเอกสารสำคัญ						คุณลักษณะที่คัดเลือกจากเอกสารทั่วไป					
		งบประมาณ	หน่วยงาน	ดำเนินการ	...	SAP	เบิก	ส่งเสริม	เงื่อนไข	พาณิชย์	...	ระดับ	ปัญหา
เอกสารสำคัญ	Doc 1	0.000	0.179	0.036	...	0.036	0.000	0.000	0.000	0.000	...	0.000	0.000
	Doc 2	0.036	0.072	0.036	...	0.286	0.000	0.000	0.000	0.000	...	0.000	0.107
	Doc 3	0.036	0.464	0.072	...	0.072	0.286	0.036	0.036	0.000	...	0.036	0.000
	Doc 4	0.000	0.036	0.036	...	0.036	0.000	0.000	0.000	0.000	...	0.000	0.000
	Doc 5	0.000	0.072	0.072	...	0.072	0.000	0.000	0.000	0.000	...	0.000	0.000
เอกสารทั่วไป	Doc 61	0.000	0.000	0.000	...	0.000	0.000	0.000	0.100	0.000	...	0.100	0.000
	Doc 62	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.100	0.000
	Doc 63	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.000	...	0.000	0.200
	Doc 64	0.000	0.000	0.000	...	0.000	0.000	0.100	0.000	0.000	...	0.100	0.000
	Doc 65	0.000	0.000	0.000	...	0.000	0.000	0.000	0.000	0.100	...	0.000	0.200

จากตาราง จะได้ค่าน้ำหนักคำ TF-IDF แทนค่าคุณลักษณะ เป็นตัวแทนของเอกสารฉบับนั้นๆ จากเอกสารสำคัญจะได้ค่าน้ำหนักในช่วงของคุณลักษณะสำคัญที่ดึงมากจากกลุ่มของเอกสารสำคัญ ส่วนค่าน้ำหนักในช่วงของคุณลักษณะที่ดึงมากจากกลุ่มของเอกสารทั่วไปนั้น จะให้

ค่าเป็น 0 และในส่วนของเอกสารทั่วไป จะได้ค่าน้ำหนักคำทั้งในส่วนของค่าคุณลักษณะที่ดึงมาจากกลุ่มเอกสารสำคัญ และกลุ่มเอกสารทั่วไป แสดงว่าเอกสารฉบับที่ 31 และ 32 มีคำสำคัญที่ปรากฏตรงกับคุณลักษณะของเอกสารสำคัญนั่นเอง

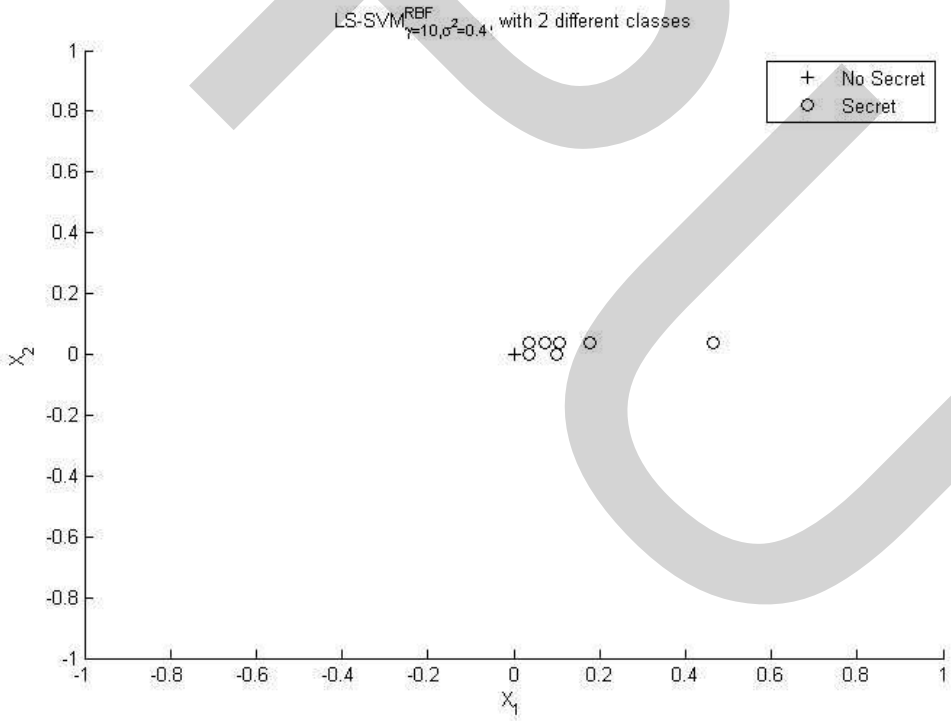
จากค่าคุณลักษณะที่ได้ หลังจากผ่านการลดค่าคุณลักษณะ(Normalize) ค่าที่ได้จะมีการกระจายตัวดีมากยิ่งขึ้น และค่าของข้อมูลไม่ห่างกันมากอยู่ระหว่างค่า 0 และ 1 นำข้อมูลนี้นำเข้ากระบวนการคัดแยกประเภทเอกสารด้วยวิธี LS-SVM ร่วมกับค่าผลลัพธ์ของประเภทเอกสารแต่ละฉบับ ซึ่งถูกกำหนดประเภทไว้แทนด้วยตัวเลข (-1) คือเอกสารทั่วไป และ (+1) คือเอกสารสำคัญ ทำการทดสอบการแยกประเภทเอกสาร ด้วยชุดข้อมูลฝึกสอน 145 เอกสาร ประกอบไปด้วยข้อมูลที่เป็นเอกสารสำคัญจำนวน 60 เอกสาร และเอกสารทั่วไปจำนวน 85 เอกสารนำมาทดสอบกับข้อมูลชุดทดสอบ โดยการทดสอบจะทำการทดสอบใน 2 กรณี คือทำการทดสอบกับชุดข้อมูลทดสอบที่ไม่อยู่ในกลุ่มเป้าหมาย และชุดทดสอบที่อยู่ในกลุ่มเป้าหมาย

4.3.1 ทำการทดสอบกับข้อมูลชุดทดสอบที่ไม่อยู่ในกลุ่มเป้าหมาย โดยนำชุดข้อมูลฝึกสอนทดสอบการคัดแยกประเภทเอกสารกับข้อมูลชุดทดสอบนอกกลุ่มเป้าหมาย นำเอกสารชุดทดสอบจำนวน 60 ฉบับ จากชุดข้อมูลทดสอบที่อยู่นอกเหนือกลุ่มเป้าหมาย และได้กำหนดประเภทเอกสารไว้ล่วงหน้า ให้ข้อมูลชุดทดสอบประกอบด้วยเอกสารสำคัญ 30 เอกสาร และเอกสารทั่วไป 30 เอกสาร โดยมีค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับ นำเข้าทดสอบกับเอกสารชุดฝึกสอนจำนวน 145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt นำมาคำนวณหาอัตราความถูกต้องตามจำนวนคุณลักษณะที่เลือกใช้ในการทดลองแต่ละครั้ง 10, 20, 30 และ 40 คุณลักษณะ ได้อัตราความถูกต้องเป็น 83.33%, 81.67%, 80.00% และ 83.33% ตามลำดับ จะเห็นได้ว่าค่าความถูกต้องมากที่สุด 83.33% โดยใช้จำนวนคุณลักษณะเป็น 10 และ 40 คุณลักษณะ แสดงให้เห็นว่าคำคุณลักษณะที่ปรากฏในเอกสารชุดทดสอบเป็นคุณลักษณะที่อยู่ในกลุ่มของคุณลักษณะลำดับที่ 1-10 และคุณลักษณะลำดับที่ 31-40

4.3.2 ทำการทดสอบกับข้อมูลชุดทดสอบที่อยู่ในกลุ่มเป้าหมาย โดยนำชุดข้อมูลฝึกสอนทดสอบการคัดแยกประเภทเอกสารกับข้อมูลชุดทดสอบที่อยู่ในกลุ่มเป้าหมาย นำเอกสารชุดทดสอบจำนวน 60 ฉบับจากชุดข้อมูลทดสอบที่อยู่ในกลุ่มเป้าหมาย ซึ่งได้กำหนดประเภทเอกสารไว้ล่วงหน้า ให้ข้อมูลชุดทดสอบประกอบด้วยเอกสารสำคัญ 30 เอกสาร และเอกสารทั่วไปจำนวน 30 เอกสาร โดยมีค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับ นำเข้าทดสอบกับเอกสารชุดฝึกสอนจำนวน 145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt และแบ่งการทดสอบออกเป็น 4 การทดสอบ ตามจำนวนคุณลักษณะที่เลือกใช้ในการทดสอบแต่ละครั้ง นำมาคำนวณหาอัตราความถูกต้องตามจำนวน

คุณลักษณะที่เลือกใช้ในการทดลองแต่ละครั้ง 10, 20, 30 และ 40 คุณลักษณะ ได้อัตราความถูกต้องเป็น 83.33%, 88.33%, 86.67% และ 86.67% ตามลำดับ ดังแสดงตามผลการทดสอบต่อไปนี้

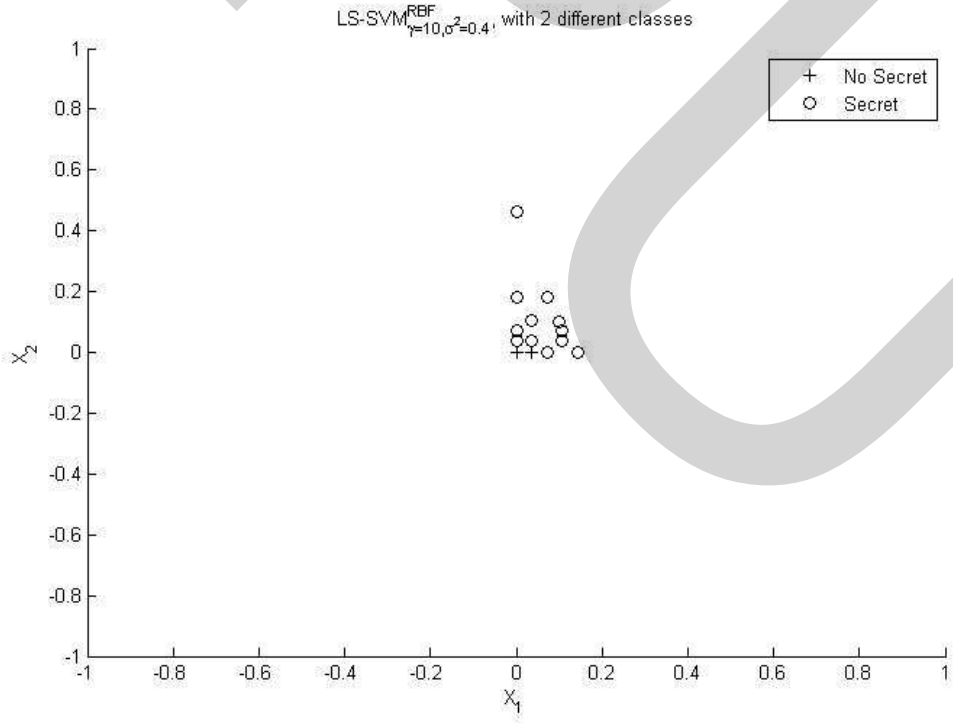
กราฟแสดงการแยกประเภทเอกสาร สามารถแยกได้ 2 ประเภท คือเอกสารสำคัญ แทนด้วยสัญลักษณ์ ○ และเอกสารทั่วไปแทนด้วยสัญลักษณ์ + ค่า $Xt = 60 \times 10$ แสดงจำนวนเอกสารชุดทดสอบจำนวน 60 ฉบับ คัดเลือกคุณลักษณะสำคัญ (Feature Selection) โดยเลือกจากสองส่วน จำนวน 10 คุณลักษณะ โดยแบ่งเป็นคุณลักษณะจากเอกสารสำคัญ จำนวน 5 คำ มีคุณลักษณะเรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [หน่วยงาน ข้อความ บันทึก กอง ผอ] และคุณลักษณะจากเอกสารทั่วไป จำนวน 5 คำ มีคุณลักษณะ เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากน้อย [ทำงาน ปัญหา วิธี มือ ไทย] ค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับ นำเข้าทดสอบกับเอกสารชุดฝึกสอนจำนวน 145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt ดังในรูปที่ 4.3



รูปที่ 4.3 กราฟแสดงผลการแยกประเภทเอกสารด้วยวิธี LS-SVM กับชุดข้อมูลทดสอบ จำนวน 60 เอกสาร โดยใช้คุณลักษณะจำนวน 10 คุณลักษณะ

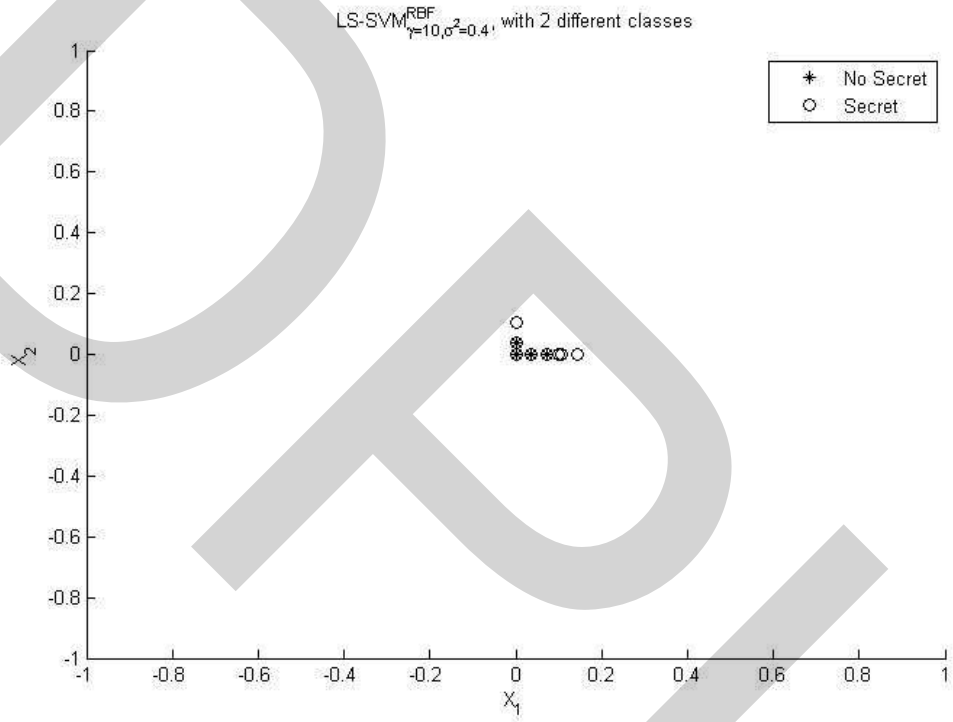
ผลลัพธ์ที่ได้จากการแยกประเภทเอกสารชุดทดสอบจำนวน 60 เอกสารที่อยู่ในกลุ่มเป้าหมาย ประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไปประเภทละ 30 เอกสาร ทดสอบกับข้อมูลชุดฝึกสอนจำนวน 145 เอกสาร โดยใช้ค่าคุณลักษณะจำนวน 10 คุณลักษณะ ได้ค่าความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสารอยู่ที่ 83.33%

ทดสอบกับเอกสาชุดทดสอบ โดยใช้จำนวนคุณลักษณะจำนวน 20 คุณลักษณะ เมื่อ $Xt = 60 \times 20$ แสดงจำนวนเอกสารชุดทดสอบจำนวน 60 ฉบับ คัดเลือกคุณลักษณะสำคัญ (Feature Selection) จำนวน 20 คุณลักษณะ โดยแบ่งเป็นคุณลักษณะจากเอกสารสำคัญ จำนวน 10 คำ มีคุณลักษณะเรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง] และคุณลักษณะจากเอกสารทั่วไป จำนวน 10 คำ มีคุณลักษณะเรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ระดับ พื้นที่ ป่วย] ค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับ ประกอบด้วยเอกสารสำคัญ และเอกสารทั่วไป นำเข้าทดสอบกับเอกสารชุดฝึกสอนจำนวน 145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt ดังในรูปที่ 4.4



รูปที่ 4.4 กราฟแสดงผลการแยกประเภทเอกสารด้วยวิธี LS-SVM กับชุดข้อมูลทดสอบ จำนวน 60 เอกสาร โดยใช้คุณลักษณะจำนวน 20 คุณลักษณะ

ผลลัพธ์ที่ได้จากการแยกประเภทเอกสารชุดทดสอบจำนวน 60 เอกสารที่ประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไป ทดสอบกับข้อมูลชุดฝึกสอนจำนวน 145 เอกสาร โดยใช้ค่าคุณลักษณะจำนวน 20 คุณลักษณะ ได้ค่าความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสารอยู่ที่ 88.33%

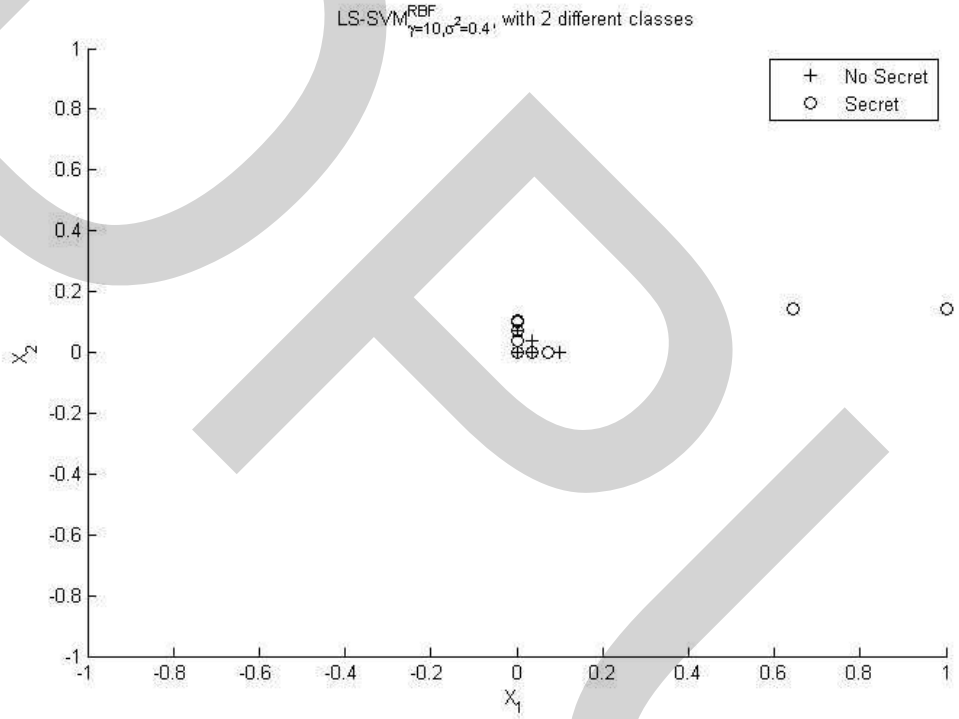


รูปที่ 4.5 กราฟแสดงผลการแยกประเภทเอกสารด้วยวิธี LS-SVM กับชุดข้อมูลทดสอบ จำนวน 60 เอกสาร โดยใช้คุณลักษณะจำนวน 30 คุณลักษณะ

ทดสอบกับเอกสาชุดทดสอบ โดยใช้จำนวนคุณลักษณะจำนวน 30 คุณลักษณะ เมื่อ $Xt = 60 \times 30$ แสดงจำนวนเอกสารชุดทดสอบจำนวน 60 ฉบับ คัดเลือกคุณลักษณะสำคัญ (Feature Selection) จำนวน 30 คุณลักษณะ โดยแบ่งเป็นคุณลักษณะจากเอกสารสำคัญ จำนวน 15 คำ มีคุณลักษณะเรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง งบประมาณ เบิก เอกสาร ฝบบ เห็นชอบ] และคุณลักษณะจากเอกสารทั่วไป จำนวน 15 คำ มีคุณลักษณะ เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อย [ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าวระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุน เทคนิค] ค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับที่ประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไป นำเข้าทดสอบกับเอกสารชุดฝึกสอนจำนวน

145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt ดังในรูปที่ 4.5

ผลลัพธ์ที่ได้จากการแยกประเภทเอกสารชุดทดสอบจำนวน 60 เอกสารที่ประกอบไปด้วยเอกสารสำคัญที่ และเอกสารทั่วไป ทดสอบกับข้อมูลชุดฝึกสอนจำนวน 145 เอกสาร โดยใช้คุณลักษณะจำนวน 30 คุณลักษณะ ได้ค่าความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสารอยู่ที่ 86.67%



รูปที่ 4.6 กราฟแสดงผลการแยกประเภทเอกสารด้วยวิธี LS-SVM กับชุดข้อมูลทดสอบ จำนวน 60 เอกสาร โดยใช้คุณลักษณะจำนวน 40 คุณลักษณะ

ทดสอบกับเอกสาชุดทดสอบ โดยใช้จำนวนคุณลักษณะจำนวน 40 คุณลักษณะ เมื่อ $Xt = 60 \times 40$ แสดงจำนวนเอกสารชุดทดสอบจำนวน 60 ฉบับ คัดเลือกคุณลักษณะสำคัญ (Feature Selection) จำนวน 40 คุณลักษณะ โดยแบ่งเป็นคุณลักษณะจากเอกสารสำคัญ จำนวน 20 คำ มีคุณลักษณะ เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไปน้อยเป็น [หน่วยงาน ข้อความ บันทึก กอง ผอ ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง งบประมาณ เบิก เอกสาร ฝบบ เห็นชอบ โปรแกรม สาขา SAP สารสนเทศ ศทส] และคุณลักษณะจากเอกสารทั่วไป จำนวน 20 คำ มีคุณลักษณะ เรียงตามลำดับความถี่ที่ปรากฏในเอกสารมากฉบับเรียงจากมากไป

น้อย [ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุนเทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข] ค่าคุณลักษณะแทนด้วยค่าน้ำหนักคำของเอกสารทั้ง 60 ฉบับ ที่ประกอบไปด้วยเอกสารสำคัญและเอกสารทั่วไป นำเข้าทดสอบกับเอกสารชุดฝึกสอน จำนวน 145 ฉบับ ตามทฤษฎีของ LS-SVM ค่า $Yt = 60 \times 1$ แสดงผลลัพธ์ที่ได้จากการแยกประเภทเอกสาร Yt ดังในรูปที่ 4.6

ผลลัพธ์ที่ได้จากการแยกประเภทเอกสารจำนวน 60 เอกสารที่ประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไป ทดสอบกับข้อมูลชุดฝึกสอนจำนวน 145 เอกสาร โดยใช้คุณลักษณะจำนวน 40 คุณลักษณะ ได้ค่าความถูกต้อง (Accuracy Rate) ของการคัดแยกประเภทเอกสารอยู่ที่ 86.67%

ตาราง 4.5 เปรียบเทียบค่าความถูกต้องของการคัดแยกประเภทเอกสารด้วยจำนวนคุณลักษณะในการทดสอบกับข้อมูลชุดทดสอบนอกกลุ่มเป้าหมาย

Feature	***Data Test***		***error***		***Accuracy Rate***
	Secret	***No Secret***	***Secret***	***No Secret***	
10	**30**	**30**	**9**	**1**	**83.33%**
20	**30**	**30**	**9**	**2**	**81.67%**
30	**30**	**30**	**11**	**1**	**80.00%**
40	**30**	**30**	**8**	**2**	**83.33%**

จากตารางเปรียบเทียบค่าความถูกต้องของการคัดแยกประเภทเอกสาร ด้วยจำนวนคุณลักษณะ (Feature) ที่แตกต่างกัน สำหรับข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบเดียวกัน ชุดฝึกสอนจำนวน 145 เอกสาร โดยแบ่งเป็นเอกสารสำคัญ 60 เอกสาร และเอกสารทั่วไปจำนวน 85 เอกสาร กับเอกสารชุดทดสอบที่อยู่นอกกลุ่มเป้าหมาย แบ่งเป็นเอกสารสำคัญ 30 เอกสาร และเอกสารทั่วไปจำนวน 30 เอกสาร จำนวนคุณลักษณะ ที่มีค่าคุณลักษณะ (Feature Vector) ถูกแทนด้วยค่าน้ำหนักคำในชุดเอกสาร ค่าความถูกต้องในการทดสอบดีที่สุดคือ 83.33% โดยใช้จำนวนคุณลักษณะ 10 และ 40 คุณลักษณะ ในขณะที่จำนวนคุณลักษณะ 20 และ 30 คุณลักษณะ ให้ค่าความถูกต้องอยู่ที่ 81.67% และ 80.00% ตามลำดับ

ตาราง 4.6 เปรียบเทียบค่าความถูกต้องของการคัดแยกประเภทเอกสารด้วยจำนวนคุณลักษณะในการทดสอบกับข้อมูลชุดทดสอบกลุ่มเป้าหมาย

Feature	*Data Test*		*error*		*Accuracy Rate*
	Secret	*No Secret*	*Secret*	*No Secret*	
10	**30**	**30**	**9**	**1**	**83.33%**
20	**30**	**30**	**6**	**1**	**88.33%**
30	**30**	**30**	**7**	**1**	**86.67%**
40	**30**	**30**	**7**	**1**	**86.67%**

จากตารางเปรียบเทียบค่าความถูกต้องของการคัดแยกประเภทเอกสาร ด้วยจำนวนคุณลักษณะ (Feature) ที่แตกต่างกัน สำหรับข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบเดียวกัน ชุดฝึกสอนจำนวน 145 เอกสาร โดยแบ่งเป็นเอกสาระสำคัญ 60 เอกสาร และเอกสารทั่วไปจำนวน 85 เอกสาร กับเอกสารชุดทดสอบที่มาจากกลุ่มเป้าหมาย แบ่งเป็นเอกสารสำคัญ 30 เอกสาร และเอกสารทั่วไปจำนวน 30 เอกสาร จำนวนคุณลักษณะ ที่มีค่าคุณลักษณะ (Feature Vector) ถูกแทนด้วยค่าน้ำหนักคำในชุดเอกสาร ค่าความถูกต้องในการทดสอบดีที่สุดคือ 88.33% โดยใช้จำนวนคุณลักษณะ 20 คุณลักษณะ ในขณะที่จำนวนคุณลักษณะ 30 และ 40 คุณลักษณะ ให้ค่าความถูกต้องต่ำที่สุดอยู่ที่ 86.6% และที่จำนวนคุณลักษณะ 10 คุณลักษณะให้ค่าความถูกต้อง 83.33%

สามารถวิเคราะห์ได้ว่าประสิทธิภาพในการคัดแยกประเภทเอกสารนั้นขึ้นอยู่กับคุณลักษณะที่เลือกใช้ และกลุ่มของข้อมูลชุดทดสอบว่าเป็นกลุ่มเป้าหมายเดียวกันกับกลุ่มข้อมูลชุดข้อมูลฝึกสอน หรือชุดข้อมูลที่นำมาคัดเลือกคุณลักษณะสำคัญหรือไม่ ซึ่งเป็นตัวแปรหลักในการการคัดแยกประเภทเอกสาร ถ้าหากคุณลักษณะที่เลือกใช้มีความคล้ายคลึง หรือเหมือนกับคำที่ปรากฏในเอกสารชุดทดสอบมากคำ ก็จะมีประสิทธิภาพ และให้ค่าความถูกต้องมาก ในขณะเดียวกันถ้าหากทำการทดสอบกับชุดทดสอบที่อยู่นอกเหนือจากกลุ่มเป้าหมาย ค่าความถูกต้องก็จะมีค่าน้อยกว่า ถึงแม้จะปรากฏคุณลักษณะที่ตรงกันบ้างแต่ก็ยังน้อยกว่าชุดทดสอบที่อยู่ในกลุ่มเป้าหมาย

ดังนั้นการคัดแยกประเภทเอกสารที่จะให้ประสิทธิภาพ หรือค่าความถูกต้องสูงนั้นจำเป็นต้องทำการคัดเลือกคุณลักษณะที่เป็นตัวแทนที่ดีของชุดเอกสารในกลุ่มเป้าหมาย รวมทั้งจำนวนคุณลักษณะที่เลือกใช้ว่าเหมาะสม และพอเหมาะสำหรับใช้เป็นตัวแทนที่ดีของชุดเอกสารในการคัดแยกสำหรับกลุ่มเป้าหมายที่ต้องการทดสอบ เพราะฉะนั้นจำเป็นต้องตั้งกลุ่มเป้าหมายก่อน

และทำการคัดเลือกชุดเอกสารฝึกสอนออกมา เพื่อหาคุณลักษณะที่เหมาะสมเพื่อเป็นตัวแทนของชุดเอกสารของกลุ่มเป้าหมาย คำที่ใช้เป็นคุณลักษณะ รวมทั้งจำนวนคุณลักษณะที่เหมาะสม ก็จะทำให้ได้ผลการคัดแยกที่มีความถูกต้อง และมีประสิทธิภาพ

บทที่ 5
สรุปผลการศึกษา

ในบทนี้จะเป็นการอภิปรายเพื่อสรุปผลที่ได้จากการทดสอบงานวิจัย รวมทั้งข้อจำกัดของระบบที่พบจากการทดสอบระบบ และข้อเสนอแนะสำหรับแนวทางในการพัฒนางานวิจัยให้มีประสิทธิภาพมากขึ้น

5.1 สรุปผลการศึกษา

5.1.1 สรุปผลตามวัตถุประสงค์ของงานวิจัย

5.1.1.1 ศึกษาโครงสร้างการทำงานเกี่ยวกับการป้องกันข้อมูลรั่วไหล เพื่อให้เกิดความชัดเจนในเรื่องของกระบวนการทำงานของระบบ

5.1.1.2 ศึกษาเทคนิคที่นำมาปรับใช้สำหรับการแยกประเภทเอกสาร เพื่อนำไปใช้งานเป็นโครงรูปในการคัดกรองของระบบการป้องกันเอกสารรั่วไหล โดยเลือกใช้เทคนิค LS-SVM เป็นการแยกประเภทแบบมีผู้สอน (Supervised Learning) ในรูปแบบของเทอมเวกเตอร์ สำหรับแทนเอกสารข้อความ เป็นเวกเตอร์ที่ระบุคุณลักษณะสำคัญ เพื่อเป็นข้อมูลที่ใช้ในการกรองการสืบค้นข้อมูลในการทำดัชนี และจัดอันดับความเกี่ยวข้องโดยใช้การกำหนดขอบเขตไว้ล่วงหน้าโดยทฤษฎีของ Support Vector Machine (SVM)

5.1.1.3 พัฒนาระบบทดสอบการแยกประเภทเอกสารสำคัญออกจากเอกสารทั่วไป โดยใช้โปรแกรม PHP ประกอบด้วยโปรแกรมย่อย Antiword , SWATH และ TF-IDF ในขั้นตอนการหาน้ำหนักคำ คัดเลือกคุณลักษณะสำคัญเพื่อเป็นตัวแทนของเอกสาร และนำคุณลักษณะสำคัญ (TF-IDF) เข้าสู่ขั้นตอนการแยกประเภทเอกสารผ่านโปรแกรม MATLAB โดยใช้ LS-SVM เพื่อแยกประเภทเอกสารสำคัญออกมา ก่อนนำเข้าระบบการป้องกันเอกสารรั่วไหล Symantec DLP

5.1.1.4 วิเคราะห์ความน่าเชื่อถือในการทดสอบ โดยการคำนวณค่าความถูกต้องของการแยก ประเภทเอกสาร เพื่อเปรียบเทียบจำนวนคุณลักษณะที่เลือกใช้ในการทดสอบแต่ละครั้ง

5.1.2 สรุปผลตามขอบเขตของงานวิจัย

5.1.2.1 ทำการทดสอบโดยใช้เทคนิคการให้น้ำหนักคำ TF-IDF และเทคนิคการแยก

ประเภทเอกสาร LS- SVM เพื่อแยกประเภทเอกสารสำคัญ ออกจากชุดเอกสารทดสอบที่ยังไม่ได้รับการคัดแยกประเภท นำผลลัพธ์(เอกสารสำคัญ) ที่ได้จากการแยกประเภทเอกสารเป็นโครงรูปในการคัดกรองของระบบการป้องกันเอกสารรั่วไหล

5.1.2.2 พัฒนาระบบการทดสอบ เพื่อแยกประเภทเอกสารสำคัญ ออกจากชุดเอกสารที่ยังไม่ได้มีการคัดแยก ซึ่งชุดเอกสารนี้จะประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไป มีการทดสอบประสิทธิภาพโดยคำนวณค่าความถูกต้อง (Accuracy Rate) ของการทดสอบ โดยมีตัวแปรอยู่ที่จำนวนคุณลักษณะที่เลือกใช้

5.2 ปัญหา และข้อเสนอแนะ

การทดสอบดังกล่าวเป็นการทำงานตามขั้นตอน ที่ไม่ต่อเนื่องกัน โดยจะประกอบไปด้วยขั้นตอนในส่วนของข้อมูลนำเข้า Text Preprocessing ขั้นตอนการหาค่าน้ำหนักคำ และขั้นตอนการคัดแยกเอกสาร ดังนั้นถ้าต้องการเพิ่มประสิทธิภาพการใช้งาน อาจจะต้องพัฒนาให้ระบบกลายเป็นโปรแกรมที่ทำงานอัตโนมัติ สามารถทำงานทุกขั้นตอนได้ในระบบเดียว

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

บทความ

จันทิมา พลพินิจ, ชมศักดิ์ สีบุญเรือง, รพีพร ช่ำชอง, อนิรุทธิ์ โชติถนอม และสมนึก พ่วงพรพิทักษ์. (2548). “Automated Obscenity Web Sites Filetering System”. **การประชุมวิชาการ สวทช . 2548 วิทยาศาสตร์ และเทคโนโลยีไทยสู่เศรษฐกิจโมเลกุล,** หน้า 325 - 330.

นิเวศ จิระวิชิตชัย ,ปริญญา สงวนสัตย์ และ พยุง มีสัจ. (2554). “การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ”. **วารสารพัฒนบริหารศาสตร์ – ปีที่ 51 ฉบับที่ 3/2554**, หน้า 187-206.

ปโยธร อุราธรรมกุล และกานดา รุณนะพงศา. (2549). “การปรับปรุงกฎสำหรับตัดคำในเอกสารภาษาไทย”. **The 3rd Joint Conference on Computer Science and Software Engineering (JCSSE 2006),** หน้า 34-40.

วิทยานิพนธ์

นนท์ บุญนิธิประเสริฐ.(2552). **การกรองข้อความภาษาไทยและภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่** .วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม บัณฑิตวิทยาลัย. กรุงเทพฯ: มหาวิทยาลัยธุรกิจบัณฑิต.

พรพล ธรรมรงค์รัตน์ . (2551). **การจำแนกประเภทเว็บเพจโดยวิธีการลดขนาดลักษณะเฉพาะและซัพพอร์ตเวกเตอร์แมชชีน.** วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์. สงขลา: มหาวิทยาลัยสงขลานครินทร์.

อภิชาต ขานทอง, วัลลภา ตันติประสงค์ชัย และ ชุลีรัตน์ จรัสกุลชัย. (2544). **การสรุปใจความสำคัญของเอกสาร.** โครงการวิทยาศาสตรบัณฑิต ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์. กรุงเทพฯ: มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน.

อาริยา เอื้ออภิสิทธิ์วงศ์ . (2549) . **การแบ่งประเภทลายนิ้วมือโดยใช้รหัสลายนิ้วมือ**. วิทยานิพนธ์คุรุศาสตร์อุตสาหกรรมมหาบัณฑิต สาขาวิชาเทคโนโลยีคอมพิวเตอร์ ภาควิชาคอมพิวเตอร์ศึกษา บัณฑิตวิทยาลัย. กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

สารสนเทศจากสื่ออิเล็กทรอนิกส์

กฎหมายกับความมั่นคงปลอดภัยสารสนเทศ. (2555). สืบค้นเมื่อ มีนาคม 2555, จาก **www.vrhris.com/klc/Article/HR/Manager/Computer Law2550.htm**.

ระบบมาตรฐานด้านความปลอดภัยของข้อมูล ISO 27001. (2555). สืบค้นเมื่อ มีนาคม 2555, จาก **www.tuv.com/th/_iso_27001.html**.

ระบบการจัดการความปลอดภัยทางข้อมูล ISO27001:2005.(2552). สืบค้นเมื่อ มีนาคม 2555, จาก **http://itm0052.blogspot.com/2009/02/iso27001bs7799.html.**

ภาษาต่างประเทศ

ARTICLES

DuraiPandian, N., Chellappan, C., Anna Univ. and Madras . (2006). “Dynamic information security level reclassification”. **Wireless and Optical Communications Networks, 2006 IFIP International Conference,** pp.1-3.

Daeseon Choi, Seunghun Jin and Hyunsoo Yoon . (2006). “A Personal Information Leakage Preventiion Method on the Internet”. **Consumer Electronics, 2006. ISCE '06. 2006 IEEE Tenth International Symposium,** pp. 646-650.

George Lawton. (2008). “New Technology Prevents Data Leakage”. **IEEE Computer Security September 2008,** pp.14-17.

Gilberto, Pedro, Edmo and Jayme. (2010). “A Security Framwork to Protect Against Social Network Services Threats.” **Fifty International Conference on Systems and Networks Communicatoins,** pp. 11-16.

Hua Zhang, Jun-feng Dial, and Qiao-yan Wen. (2008). "Secure files Management System in Intranet" . **2008 International Conference on Internet Computing in Science and Engineering**, pp 306-311.

Simon Liu, Rick Kuhn . (2010). " Data Loss Prevention". **IT Pro March/April 2010,Published by the IEEE Computer Society**, pp. 10-13.

Yuguo Wang . (2008). "A Tree-based Multi-class SVM Classifier for Digital Library." **International Conference on MultiMedia and Information Technology,** pp. 15-18.

Zhang Xiaosong , Liu Fei,Chen Ting and Li Hua . (2009). "Research and Application of the Transparent Data Encryption In Intranet Data Leakage Prevention". **2009 International conference on Computational Intelligence and Security,** pp. 376-379.

Zhijie Liu, Xueqiang Lv, Kun Liu and Shuicai Shi . (2010). "Study on SVM Compared with the other Text Classification Method". **Information Science and Technology University Beijing China 2010,** pp. 219-222.

ELECTRONIC SOURCES

7 Step to information Protection 2009. (2010). Symantec;White Paper Data Loss Prevention, Retrieved July 2010, from **www.symantec.com.**

Antiword Version 0.37. (2011). Retrieved April 2011, from **http://www.winfield.demon.n**

Arg max. (2012). Retrieved April 2012, from **http://en.wikipedia.org/wiki/Arg_max**.

Binary Search Pseudocode. (2012). Retrieved April 2012, From **http://www.cs.uiuc.edu/class/sp07/cs199lbp/lectures/CS199-Lectures17and18.pdf.**

Feature map approximation for RBF kernels. (2012). Retrieved March 2012, from **http://scikit-learn.sourceforge.net.**

ISO/IEC 27001. (2012). Retrieved March 2012, from **www.iso27001security.com/html/27001.html.**

Kristiaan Pelckmans , Johan A.K. Suykens. (2002). LS-SVMlab : a MATLAB/C toolbox for Least Squares Support Vector Machines, Retrieved July 2010, from **http://www.esat.kuleuven.be/sista/lssvmlab/.**

LS-SVMlab1.7. (2011). Retrieved April 2011, from **http://www.esat.kuleuven.ac.be/sista/lssvmlab.**

Software SWATH (Thai Word Segmentation). (2011). Retrieved April 2011, from **http://www.cs.cmu.edu/paisarn/software.html.**

Symantec Data Loss Prevention Administration Guide Version 10.5. (2010). Retrieved July 2010, from **www.sysmantec.com**.

ภาคผนวก

โปรแกรม PHP (กระบวนการหาค่าน้ำหนักคำ)

```
//ทำฟอร์ม Input ไฟล์เอกสาร โดยใส่เป็น Path
body{
}
#main{
        width:1000px;
        margin:0 auto 0 auto;
}
        #SendingForm {
                width:100%;
                margin:0 auto 0 auto;
                }
        #result {
                width:100%;
                margin:0 auto 0 auto;
                background-color:#0F0;
                }
        </style>
<body>
<div id="main">
<div id="SendingForm">
<form enctype="multipart/form-data" action="<?=$PHP_SELF?>" method="post">
<p>
        <input type="hidden" name="MAX_FILE_SIZE" value="3000000" />
          URL:
        <input type="text" name="url" size=50 />
        <br />
          <!--<iframe src="swf/upload" width="600" height="300" frameborder="0"></iframe>
-->
          <br />
```

```
            Send this file:<!--<br /> <input name="userfile" type="file"/>-->
            <br />
           <input type="submit" value="Send File" />
           </form>
</div>
<div id="result">
<?php
          $template = "";
          $url = str_replace(" ","",$_POST['url']);
          if ($url != ""){                    //Open images directory
                    $dir = opendir($url);       //List files in images directory
                    $i = 0;
                    $tmpdir = $url."\\tmp";
                    if (! file_exists($tmpdir) ) {
                              mkdir($tmpdir);
                    }
                    while ( ($file = readdir($dir) ) !== false ){ //
                              if ( strpos($file,".doc") ) {
                              echo  "filename[".$i++."]:  "  .  iconv("TIS-620",  "UTF-8",$file)  .  "<br
/>";
                                        $newfile =  Date("dmy_His");
                                        if ( !copy($url."\\".$file,$tmpdir."\\".$newfile) ) {
                                                  echo "failed to copy $file...\n";
                                                  }
                                        $newdata[$i-1] = antiword($tmpdir,$newfile,$i-1);
                                                            }
                              }
          $Tr = $i;
```

```
// กระบวนการหาค่าน้ำหนักคำ
        $num_vectors = sizeof($template[0]);
        echo "<table border=\"1\" width=\"100%\">";
        echo "<tr>";
        for ($k=0;$k<$num_vectors;$k++)
                {
                        echo "<td>".$template[0][$k]."</td>";
                }
                echo "</tr>";
                echo "<tr>";
                for ($k=0;$k<$num_vectors;$k++)
                {
                        echo "<td><ul>";
                         for ($j=0;$j<$Tr;$j++)
                         {
                                $position = array_search($template[0][$k],$newdata[$j][0]);
                                if ($position !== false)
                                {
                                        $Tr_tk = $template[1][$k];
                                        $tk_dj = $newdata[$j][1][$position];
                                        $cal_tmp = log10($Tr/$Tr_tk);
                                        $w[$k][$j] = $tk_dj * $cal_tmp;
                                }else
                                {
                                         $Tr_tk = 0;
                                         $tk_dj = 0;
                                         $w[$k][$j] = 0;
                                }
                                echo "<li>".round($w[$k][$j],2)."</li>";
                        }
```

```
                    echo "</ul></td>";
                }
                echo "</tr>";
                echo "</table>";
            }
            else {
              if (isset($_FILES['userfile']['name'])) {
                $uploaddir = 'tmp\\';
                $upload_file = $uploaddir . (Date("dmy_His"));
                $userfile =  $_FILES['userfile']['name'];
                if (move_uploaded_file($_FILES['userfile']['tmp_name'],$upload_file)) {
                    antiword($upload_file);
                }
              }
            }
              ?>
              </div>
              </div>
              </body>
            </html>
```

//เรียกใช้โปรแกรมย่อย Antiword (อ่านคำจากไฟล์เอกสาร Microsoft words)

```
            <?
                function antiword($dir,$upload_file,$round)
                {
                    $filedoc = $dir."\\".$upload_file;
                    $filetext = $dir."\\".$upload_file.".txt";
                    $command="C:/antiword/antiword -m UTF-8.txt $filedoc > $filetext";
                        exec($command,$error);
                        if (!$error)
                        {
```

```
                $output = swath($filetext,$dir,$round);
            }
            else die("<BR>ERROR Read Document<BR>");
            return $output;
        }
    ?>
    //เรียกใช้โปรแกรมย่อย SWATH ตัดคำภาษาไทย และภาษาอังกฤษ
    <?
    function swath($input_filename ,$dir, $round){
        echo "<h1>TEST</h1>";
        $input_filename = $input_filename;
      $output_filename= tempnam($dir, "swath");
        $input_text = file_get_contents($input_filename);
        $input_text = str_replace("]","",str_replace("[","",$input_text));
        $input_text = str_replace("~","",$input_text);
        $per = 100;
        $txt = split("\r\n",$input_text);
        $div = sizeof($txt);
        $input_text_tis620 = "";
        for ($i=0;$i<$div+1;$i++){
            $tmp =  $txt[$i];
            $input_text_tis620 .= iconv("UTF-8", "TIS-620",$tmp);
        }
    file_put_contents($input_filename, $input_text_tis620);
    $ourFileHandle = fopen($output_filename, 'w') or die("can't open file");
    fclose($ourFileHandle);   //if ($round == 0)
    $cmd = "cd\ && cd C:\AppServ\www\Poi\mail_swath\Swath-2.0 && swath -m bi
<$input_filename> $output_filename";
    $cmd = "cd\ && cd C:\AppServ\www\Poi\mail_swath\Swath-2.0 && swath -m bi
<$input_filename> $output_filename";
```

```
$rtn = null;
exec($cmd,$rtn);
if(!$rtn){
        /*$datatext = "";
        $f = fopen($output_filename,"r") or die("Error can not open file $file");
        while (!feof($f)){$datatext.=fgetc($f);}
        fclose($f);*/
                $raw = file_get_contents($output_filename);          //echo $raw;
                $raw_utf8 = iconv("TIS-620", "UTF-8", $raw);      //echo $raw_utf8;
        }else{
                die("error");
        }
   $data = preg_split('/\|/', $raw_utf8);
   $newdata = removeStopword($data,$round);
   return $newdata;
}
function removeStopword($data,$round)       //Remove Stop Words
{
        global $template;
        $stop_phrase = array  (_____);    //ใส่คำที่ต้องการให้ตัด
  $pass_phrase = "/[ก-ฮa-zA-Z]/";
 $counts = array_count_values($data);
  arsort($counts);
  $i =0;
  #$echo = "";
```

```
//แสดงความถี่ของคำ และตารางค่าน้ำหนักคำ
foreach($counts as $key => $value)
        if          (str_replace("","",$key)&&str_replace("\r\n","",$key)          &&
        preg_match($pass_phrase,$key)  && !in_array($key,$stop_phrase) ) {
                $echo .= ' '.$key.' --> '.$value.' <br>';
                $new_data[0][$i] = $key;
                $new_data[1][$i] = $value;
                if ($round != 0){
                        $position = array_search($key,$template[0]) ;
                        if ( $position !== false ){
                                $template[0][$position] = $template[0][$position]++;}
                                else{
                                        $template[0][sizeof($template[0])] = $key;
                                        $template[1][sizeof($template[0])] = 1;
                                        }
                                }

                                else{
                                        $template[0][$i] = $key;
                                        $template[1][$i] = 1;
                                }
                                $i++;
                        }
                }
                echo " มีจำนวนคำทั้งหมด ". $i."<br>".$echo;
                return $new_data;
        }
        ?>
```

แสดงคำหยุดที่ถูกใช้ในขั้นตอนกำจัดคำหยุด

{"แบบ","วัน","ทั้งสิ้น","ส่ง","จำกัด","รา","ย่อ","ศรี","ลม","กม","ลาด","พร้า","โศก","นุช",
"อ่อน","ยบี","ราช","เอส","แหง","ทร","นา","มแพล็กซ์","คำ","คอ","สา","รัก","สิต","อิน","วงศ์",
"รัง","วา","สี่","ชัย","พา","หลุยส์","หล่อ","มพินี","ไดซ์","ระ","ชิด","งธานี","ทอ","ท่า","แค",
"นนท์","บัว","พลี","ปรารถ","ชิต","สุร","สุ","นทร์","โรง","สำ","สุข","สวัสดิ์","สิน","บุ","ขาว",
"นนท","หัว","เตย","คะนอง","ตาก","แยก","สาม","เกล้า","โอ","ยู","บี","ราย","อัม","วาด",
"นทร์พ","ลา","ซ่า","ทรง","รัด","เยา","วาน","งาม","องแขม","จักร","โชค","สาธุ","บอน","วร",
"ษม","หลา","มุม","ปี","งาน","ขอ","เรียน","สง","ติดตาม","แนบ","เรียบร้อย","สำหรับ",
"เพิ่มเติม","โปรด","มุสิ","กู","ร่วมมือ","ดังกล่าว","ทุ","ตั้ง","ลง","พร้อมทั้ง","พบ","งขนิษฐา",
"ใช้","ต่อไป","ผล","เจริญ","ปน","พงศ์","ดำรง","อัฐ","นาย","นาง","นัก","จง","รบ","ขอบคุณ",
"ครั้ง","ตรง","๙.","๘.","๗.","๖.","๕.","๔.","๓.","๒.","๑.","โดย","ต้อง","เช่น","เป็น","ไว้",
"ฯ",".","0","9","8","7","6","5","4","3","2","1","๐","๙","๘","๗","๖","๕","๔","๓","๒","๑",
"ฬ","ห","อ","ษ","ส","ศ","ว","ล","ร","ย","ม","ภ","ฟ","พ","ฝ","ผ","ป","บ","น","ถ","ต","ด",
"ณ","ฑ","ฐ","ฏ","ญ","ซ","ช","ฉ","จ","ง","ค","ข","ก","ก็","คือ","เมื่อ","ๆ","นี่","นี้","กับ",
"จาก","ไป","ที่","ที","บาง","นำ","ฝ่าย","ถือ","จึง","หาก","ตาม","มา","เพื่อ","ส่วน","ทราบ",
"ความ","เมื่อ","การ","ความ","ที่","ซึ่ง","อัน","จะ","ใน","ว่า","และ","มี","ได้","ของ","ให้","กัน",
"แล้ว","อีก","ทั้ง","ด้วย","นั้น","ท","ดี","ติ","ตุ","มี","A","B","C","D","E","F","G","H","I","J",
"K","L","M","N","O","P","Q","R","S","T","U","V","W","X","Y","Z","a","b","c","d","e","f","g",
"h","i","j","k","l","m","n","o","p","q","r","s","t","u","v","w","x","y","z","ยช","ธ","ยุ","บิ","กร",
"รุ","อุ","นง","ลุ","ริ","นค","ขา","ยุ","หุ","หน","อยู่","คือ","ปะ","ตุ","ลิ","ร่","ยัง","หรือ","และ",
"ถ้า","แล้ว","บ้าง","อยาก","ไหม","อื่น","คแ","ซี","รี","อน","เผ","อ่ว","ธุ","ฯลฯ","าพ","แต่",
"ออ","ล์","ใ","ไ","า","ช้","ยู่","มิ","มาก","เ","รม","นทร","เอง","กว่า","รล์","นฯ","ติฯ","ถึง",
"รม","เก","ม่","วน","is","are","of","or","and","งก","เริ่ม","สูง","ช่วง","สิ้น","เนื่องจาก",
"ต้องการ","เดียวกัน","เห็น","เดือน","เรียก","ทาง","ผ่าน","ลด","เสร็จ","เลือก","เกี่ยวกับ","กลุ่ม",
"หา","เดิม","ก่อน","พล","ด้าน","เพื่อให้","คุณ","ชี","งจา","อก","งา","จัด","เก็บ","เสนอ",
"ให้แก่","เหลือ","รู้","ทำ","สามารถ","ดังนี้","ตัว","อย่าง","ออก","ทุก","สร","แก่","แรก","ผู้",
"หน้า","เข้า","สอ","หน่วย","ต่าง","ท่าน","โอกาส","ณัฐ","ย่าน","มเพล็กซ์","พหล","โยธิน","วิท",
"ย่อย","พร้อม","ฟัง","รับฟัง","รอง","เช่นกัน","ช่วย","ร่อง","กลาง","ลน","แทน","บน","ใบ","ดู",
"ขึ้น","ชื่อ","วก","ล่วง","คน","แจ้","ชนิด","โทร","ถาวร","ใหม่","pic","ประจำ","ประมาณ",
"เรื่อง","ย่าน","ซิตี้","ห้วย","ขวาง","ปิ่น","หมาก","ภิบาล","ปรารภ","พาร์ค","เซ็นต์","อุทิศ",

"พหล","รับ","รวม","หัก","งฝบบ","สนับสนุน","เหมาะสม","ไม่","ถูก","ตนเอง","เขียน","เปลี่ยน",
"รองรับ","ข้อ","ท้าย","ค่า","ต้อ","เลย","ตั้งแต่","ต้น","สาย","แต่ละ","เพียงพอ","ชอบ","กิจ",
"งฯ","ถูก","จ้า","สะดวก","ยินดี","จล","กก","ทยอย","ต่างๆ","อือ","เชื่อมต่อ","พริ้ง","บบฯ",
"ยัทฯ","กะ","เท","ลลิเนท","ง่าย","ดังนั้น","ซับซ้อน","ยุ่งยาก","บเต","เม","สน","เส","รด","โป",
"สะดวก","คร","บัง","กระ","เข","ศิ","เกร็ด","ตัด","ทา","คร","โก้","วเวอร์","คา","ขุน","กา","อา",
"ภู","ชื่น","ลำ","ไอ","โก","มอลล์","ภู","สนิท","ทิพย์","พี","ยซี","ยาน","กอ","วัตร","ยถ","คคโล",
"ยสุ","ยศิ","รฟร์","จุ","จาม","นั่ม","คโนฯ","แป","รทู","จอ","รจิวเวล","รี่","ไซ","เล่ย์","ษา",
"รซัน","นว","เวอร์","โฮ","มเวิร์ค","โค","ณิชย","นัด","รพ","มัน","ละ","อี","เอสพ","ปี้","แฮ",
"อิม","เบ้","ยโบ้","สาว","วุฒา","ยอ","กุม","อนุ","เวศ","สัน","ยโร","โน","โซ","ยยู","รัถ","นคร",
"ยจา","เข","นานา","ซอ","สาน","กฏา","รัตน์","พลัด","มัย","ราง","เลิ้ง","ยบิ๊ซี","อง","คอน","เท",
"ชื่น","เอก","ลูก","ทิพย์","พี่","ครุ","รัต","เพล","พัฒน์","โพ","ไท","นธร","สิ","ยถ","เข","รัถ",
"ยจา","กฏา","มัย","เห","ดุ","ยเยส","เจ้าฯ","ธิ","สรง","ธัญ","จุ","ลฯ","จอ","อง","ยม","ค่ะ",
"ครับ","คะ","คิด","มัว","เล่น","อาจ","อาย","เอาใจ","ขี้","กล้า","ล่ะ","มัน","ทธ","ยิ่ง","ชด","ต่อ",
"อาจ","กล่าว","น่า","กฯ","เท่า","ธรรม","อา","นอก","พทะ","ตัว","พันธุ์","ใดๆ","มัก","นาน",
"เพิ่ง","ทัน","คาด","ใหญ่","สิงห์","ระหว่าง","นางสาว","พก","บรรยาย","พร้อมกับ","กรอก",}

โปรแกรม MATLAB เรียกใช้โปรแกรมย่อม LS-SVM

```
X = load('_____.txt');          //เอกสารชุดฝึกสอน
Y = load('_____.txt');          //คลาสของเอกสารชุดฝึกสอน
type='classification';
plotlssvm({X,Y,type,gam,sig2,'RBF_kernel'},{alpha,b});
//พล็อตกราฟการแยกประเภทของชุดเอกสารฝึกสอน
Xt=load('_____.txt');                                   //เอกสารชุดทดสอบ
Yt=simlssvm({X,Y,type,gam,sig2,'RBF_kernel'},{alpha,b},Xt);     //ประมวลผลการแยกประเภท
plotlssvm({Xt,Yt,type,gam,sig2,'RBF_kernel'},{alpha,b});
//พล็อตกราฟการแยกประเภทของชุดทดสอบ
```

แสดงค่าคุณลักษณะสำคัญ (Feature)

จำนวนคุณลักษณะสำคัญ

10 Feature = {โปรแกรม สาขา SAP สารสนเทศ ศทส ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข}

20 Feature = {งบประมาณ เบิก เอกสาร ฝบบ เห็นชอบ โปรแกรม สาขา SAP สารสนเทศ ศทส ทีม อนาคต หมอ ลงทุน เทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข}

30 Feature = {ดำเนินการ พิจารณา ฝปค เทคโนโลยี แจ้ง โปรแกรม สาขา SAP สารสนเทศ ศทส เมือง ข้าว ระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุน เทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข}

40 Feature = {หน่วยงาน ข้อความ บันทึก กอง ผอ งบประมาณ เบิก เอกสาร ฝบบ เห็นชอบ โปรแกรม สาขา SAP สารสนเทศ ศทส ทำงาน ปัญหา วิธี มือ ไทย เมือง ข้าว ระดับ พื้นที่ ป่วย ทีม อนาคต หมอ ลงทุน เทคนิค ประวัติ พาณิชย์ หน้าจอ ส่งเสริม เงื่อนไข}

ประวัติผู้เขียน

ชื่อ-นามสกุล	อรทิพย์ เลื่อยงาม
ประวัติการศึกษา	วิศวกรรมศาสตรบัณฑิต ปีการศึกษา 2548 สาขาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยนเรศวร
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักคอมพิวเตอร์ 4 ส่วนระบบรักษาความมั่นคงปลอดภัยเทคโนโลยี กองบริหารจัดการสื่อสารและความมั่นคงปลอดภัย ฝ่ายเทคโนโลยีและสื่อสาร การประปานครหลวง