การจัดประเภทเอกสารด้วยวิธีเอสวีเอ็ม เพื่อการป้องกันเอกสารรั่วไหล

อรทิพย์ เลื่อยงาม[1,*] และ ชัยพร เขมะภาตะพันธ[2]

Document classification by SVM to document leakage prevention

Orathip Lueyngam[1,*] and Chaiyaporn Khemapatapan[2]

[1]สำนักเทคโนโลยีสารสนเทศ การประปานครหลวง กรุงเทพฯ ประเทศไทย 10210

Information Technology Department ,Metropolitan Waterworks Authority ,Bangkok Thailand 10210

[2]อาจารย์ประจำภาควิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม มหาวิทยาลัยธุรกิจบัณฑิตย์ กรุงเทพฯ ประเทศไทย 10210

Department of Computer and Telecommunication Engineering ,Dhurakij Pundit University, Bangkok Thailand 10210

*Corresponding author. E-mail: orathip@mwa.co.th

บทคัดย่อ

บทความนี้นำเสนอการออกแบบระบบการแยกประเภทเอกสารสำคัญออกจากเอกสารทั่วไป ที่มีใช้ภายในหน่วยงานของการประปานครหลวง เพื่อนำไปเป็นอินพุตของระบบการป้องกันข้อมูลรั่วไหล DLP ขณะนี้ยังไม่มีองค์กรหรือหน่วยงานใดออกมากำกับดูแลและให้มาตรฐานการทำงานของระบบ รวมทั้งยังไม่มีการกำหนดขั้นตอนการทำงานที่ชัดเจน ขั้นตอนการทำงานจึงมีหลากหลายตามแต่ละผลิตภัณฑ์ที่มีจำหน่ายออกมา Symantec DLP เป็นอุปกรณ์การป้องกันข้อมูลรั่วไหลที่ทางสำนักเทคโนโลยีสารสนเทศการประปานครหลวงนำมาปรับใช้งานเพื่อตอบสนองนโยบาย ของระบบการจัดการชั้นความลับข้อมูล และด้วยความสามารถของอุปกรณ์ จะทำงานในส่วนของการเฝ้าระวังข้อมูลสำคัญ และการบังคับใช้นโยบายที่ได้กำหนดขึ้นเท่านั้น ยังขาดส่วนของการแยกประเภทเอกสารก่อนการนำไปเป็นต้นแบบของการกรองข้อมูลสำคัญ ดังนั้นเพื่อลดเวลาในการคัดแยกเอกสารซึ่งมีปริมาณมาก และยากต่อการกำหนดระดับความสำคัญ จึงได้ทำการพัฒนาวิธีการแยกประเภทเอกสาร โดยใช้เทคนิคการแยกประเภทแบบมีผู้สอนตามทฤษฎี SVM และการนำเอกสารมาแปลงเป็นเวกเตอร์ค่าน้ำหนัก TFIDF และนำค่าน้ำหนักที่ได้เข้าสู่กระบวนการแยกประเภทเอกสาร LS-SVM โดยการแบ่งประเภทเอกสารด้วยคุณลักษณะสำคัญที่ได้กำหนดไว้ในขอบเขตของเอกสารที่ไม่มีการแก้ไขระดับความสำคัญอีกในภายหลัง ถ้าหากมีการเปลี่ยนแปลงระดับความสำคัญถือว่าเป็นอีกกรณี และจะไม่ถูกนำมาพิจารณาในขั้นตอนการแยกประเภทนี้ จากผลการทดสอบพบว่า การแยกประเภทเอกสารโดยใช้ข้อมูลชุดฝึกสอน และข้อมูลชุดทดสอบทั้งสองประเภท คือเอกสารความลับ (1) และเอกสารทั่วไป (-1) ทำการทดสอบกับระบบที่ได้นำเสนอ สามารถแยกประเภทเอกสารตามที่ได้กำหนดประเภทเอกสารไว้ ซึ่งได้ผลเป็นที่น่าพอใจ และนำเอกสารความลับที่แยกประเภทได้นำเข้าเป็นอินพุตของระบบการป้องกันข้อมูลรั่วไหล Symantec DLP เพื่อเป็นต้นแบบในการกรองของระบบต่อไป

Abstract

This paper present a system to classify important document from the common used documents in the Metropolitan Waterworks Authority (Thailand). To use as the input of data Leak prevention system. DLP does not have any organization or agency out supervision and the standard operation of this system. DLP Process has a wide range of each product . Symantec DLP is a device to prevent data leakage at the Office of Information Technology , Metropolitan Waterworks Authority (Thailand) to deploy applications to meet policy management system and data confidentiality. With the ability of the device that work in terms of monitoring critical data and implementing policies that are based only. The device can also be part of a classified document before adoption as a model of filtering information. In order to minimize the duration of the separation document, which is plenty. And difficult to set priorities. The classification used by an instructor theory Support Vector Machines. By the way, the document is converted is converted to a vector of TF-IDF weight and the weight. Through the process of classifying document, LS-SVM. The scope of the document has not been altered priorities later. If you have changed the priority is considered a special case. Will not be taken into consideration in the process of this sort. The results showed that to classify documents using training data set. And test two types. What is the secret document (1) and non secret documents (-1) to test the system at present. Users can sort the document according to the type of document. The result is satisfactory. The document, classified secret ,which has led to the input of data leak prevention system, Symantec DLP to the filters in the system.

บทนำ

ในปัจจุบันการทำงานร่วมกับเอกสาร และใช้งานเอกสารร่วมกันเป็นประจำในกลุ่มของพนักงาน ลูกค้า และคู่ค้า เช่น ข้อมูลการเงิน แผนธุรกิจ กลยุทธ์ทางการตลาด ข้อมูลลูกค้านั้น นำไปสู่การพิจารณาในเรื่องของการป้องกันข้อมูลสำคัญ และด้วยความก้าวหน้าทางเทคโนโลยีการจัดเก็บในรูปแบบของไฟล์ข้อมูลแบบดิจิตอล กลายเป็นช่องโหว่ที่ไฟล์ข้อมูลเหล่านี้ที่สามารถถ่ายโอนไปยังบุคคลที่ไม่ประสงค์ดีได้อย่างง่ายดายและรวดเร็ว เช่น การถ่ายโอนผ่านอุปกรณ์จัดเก็บข้อมูลแบบพกพา การส่งไฟล์แนบผ่านทางจดหมายอิเล็กทรอนิกส์ การนำไปโพสต์ไว้บนเว็บไซต์ต่าง ๆ โดยไม่ได้รับอนุญาต หรือแม้แต่การส่งข้อมูลออกไปนอกองค์กรอย่างไม่ตั้งใจ หรือโดยตั้งใจก็ตาม สิ่งเหล่านี้นำไปสู่การรั่วไหลของข้อมูลได้ทั้งนั้น และพบว่าสาเหตุของการรั่วไหลส่วนมากเกิดจากบุคคลภายในองค์กร

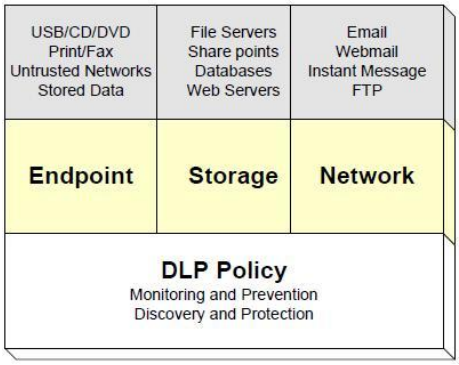
การจัดหมวดหมู่ของข้อมูล คือนโยบายการแยกประเภทข้อมูลตามคุณค่าที่มีต่อองค์กร และผลกระทบถ้าหากข้อมูลเหล่านั้นถูกนำไปเผยแพร่ แก้ไข การขโมยข้อมูลทรัพย์สินทางปัญญา หรือการนำข้อมูลส่วนตัวไปโพสต์กล่าวร้ายเป็นต้น ดังนั้นจำเป็นต้องเข้าใจคุณค่าของสิ่งที่จะปกป้อง ว่าจะปกป้องได้มากน้อยแค่ไหนและสิ่งที่จะปกป้องอยู่ที่ใดเสียก่อน

จากปัญหาข้างต้น จึงได้นำเสนอระบบการจัดประเภทเอกสาร เพื่อนำเข้าสู่กระบวนการทำงานร่วมกับโซลูชั่นด้านการรักษาความมั่นคงปลอดภัยของเอกสารในส่วนของการป้องกันข้อมูลรั่วไหล DLP (Data Leakage Prevention) เพื่อควบคุมการไหลของเอกสารสำคัญและบังคับใช้นโยบายการป้องกันข้อมูลสำคัญ เพื่อให้สอดคล้องกับการบริหารจัดการความมั่นคงปลอดภัย ISO27001 ซึ่งจัดเป็นมาตรฐานที่ได้รับการยอมรับจากหลายประเทศ ในการนำไปใช้บริหารจัดการระบบสารสนเทศขององค์กร และการควบคุมการบริหารความเสี่ยงของหน่วยงาน โดยศึกษาเทคนิคที่ใช้กันอยู่ในปัจจุบันคือการแยกประเภทเอกสารซึ่งเป็นเทคนิคการจำแนกประเภทเอกสาร โดยมีการกำหนดประเภทเอกสารไว้ล่วงหน้า เทคนิคนี้ต้องทำการสอนระบบให้รู้จำรูปแบบเอกสาร ในแต่ละประเภทก่อน หลังจากนั้นจึงนำเอกสารที่ต้องการจำแนกประเภทเข้าไปในระบบ ให้ระบบทำการแยกประเภทเอกสารเพื่อสร้างโครงรูป (template) ตามประเภทเอกสารที่กำหนดไว้ (เอกสารลับ ,เอกสารทั่วไป) โดยใช้ทฤษฎีที่เกี่ยวข้องคือ TF-IDF (Term Frequency and Inverse Document Frequency) เป็นวิธีการคำนวณค่าน้ำหนักจากความถี่ของคำ ที่ปรากฏในเอกสาร และ SVM (Support Vector Machine) เป็นขั้นตอนการใช้งานตัวจำแนก นำมาสนับสนุนการแยกประเภทเอกสาร ก่อนนำเข้าสู่โซลูชั่นของการป้องกันข้อมูลรั่วไหลเพื่อลดระยะเวลาในการแยกประเภทเอกสารและเพิ่มประสิทธิภาพในการทำงานของระบบการป้องกันข้อมูลรั่วไหลต่อไป

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1. เทคโนโลยีการป้องกันข้อมูลรั่วไหล DLP (Data Leakage Prevention)

เป็นกระบวนการทำงานและวิธีการป้องกันการรั่วไหลของข้อมูลสำคัญ ในการถ่ายโอนข้อมูลสำคัญโดยไม่ได้รับอนุญาต หรือเปิดเผยข้อมูลสำคัญ เพื่อลดความเสี่ยงจากการถูกขโมย และสูญเสียโอกาสทางธุรกิจ โดยการตรวจสอบและควบคุมการใช้งานข้อมูล รวมทั้งกำหนดนโยบายให้สอดคล้องกับกลยุทธ์และกระบวนการทางธุรกิจ พิจารณาในส่วนของการรั่วไหลของข้อมูล ,ช่องทางการรั่วไหล,วิธีการรั่วไหลและผลกระทบ การระบุและแยกประเภทข้อมูลที่เป็นความลับเพื่อกำหนดนโยบาย วิธีการในการป้องกันข้อมูล และปรับใช้เทคโนโลยีเข้ามาช่วยในการบังคับใช้เพื่อปฏิบัติตามนโยบาย พิจารณาพื้นฐานของการป้องกันการรั่วไหลของข้อมูล ซึ่งประกอบด้วย การค้นหาข้อมูล (Data Discover) เป็นกระบวนการค้นหาข้อมูลความลับ การแยกประเภทข้อมูล (Data Classification) เป็นกระบวนการจำแนกข้อมูลตามมูลค่า และผลกระทบที่มีต่อองค์กรเมื่อมีการรั่วไหล การเฝ้าระวังข้อมูล (Data Monitor) เป็นกระบวนการในการตรวจสอบช่องทางการสื่อสารต่าง ๆ ขององค์กร และการป้องกันข้อมูล (Data Protect) เป็นกระบวนการซึ่งดำเนินการตามนโยบายที่ได้กำหนดไว้ในเชิงป้องกัน ที่ระบุต้นทาง ปลายทาง และวิธีการที่อาจทำให้ข้อมูลรั่วไหลออกไปได้

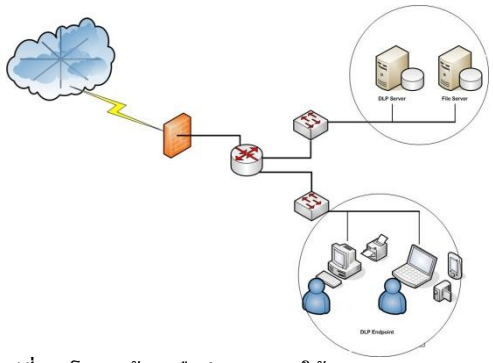


รูปที่ 1 รูปแบบของ DLP (Data Loss Prevention System Model)

จากรูปที่ 1 รูปแบบของเทคโนโลยีพื้นฐาน DLP ส่วนประกอบที่ใช้ในการพิจารณา คือ ข้อมูลเคลื่อนไหว (Network) ข้อมูลที่มีอยู่ (Storage) และข้อมูล ณ จุดปลายทาง (Endpoint) ซึ่งถูกควบคุมด้วยนโยบายรักษาความปลอดภัย (DLP Policy)

2. Symantec DLP Solution

เป็นระบบรักษาความปลอดภัยให้กับข้อมูลสำคัญขององค์การ มีขั้นตอนการทำงานหลักๆ คือ การค้นหาข้อมูล และการเฝ้าระวังข้อมูลสำคัญ ในขั้นตอนของการค้นหาข้อมูลสำคัญนั้นจะทำการตรวจหาข้อมูลสำคัญผ่านทางโปรแกรมย่อย (agent) ที่ติดตั้งไว้ที่เครื่องลูกข่ายต่างๆ และทำอยู่อย่างสม่ำเสมอโดยการตั้งเวลาการตรวจสอบ ผลที่ได้คือ รู้ตำแหน่งที่อยู่ของเอกสารสำคัญที่ตรงกับเอกสารต้นแบบที่ระบุไว้ในระบบ มีการรวบรวมตำแหน่งการจัดเก็บของเอกสารสำคัญไว้ และมีการป้องกันข้อมูลเหล่านั้น โดยวางโครงสร้างเครือข่าย ดังรูปที่ 2



รูปที่ 2 โครงสร้างเครือข่ายของการใช้งานระบบ DLP

รูปแบบของนโยบายเกี่ยวกับเอกสารสำคัญของอุปกรณ์ Symantec DLP อยู่ในรูปแบบของ IDM (Indexed document matching) Rule หรือการทำดัชนีเอกสารสำคัญ จากกฎจะมองเนื้อหาจากเอกสารเฉพาะที่มีการลงทะเบียนไว้ว่าสำคัญ (เอกสารสำคัญที่เป็นเอกสารต้นแบบ) และจะมีการตรวจสอบเทียบกับเอกสารที่ได้จากการแยกประเภทไว้ก่อนนำเข้าสู่ระบบ จะต้องมีความเหมือน 80% หรือมากกว่าเอกสารต้นฉบับ

3. การตัดคำในเอกสารภาษาไทย (Thai Word Segmentation)

เป็นกระบวนการแยกคำในเอกสารภาษาไทย ซึ่งอาจจะประกอบไปด้วยตัวหนังสือภาษาไทยตัวหนังสือภาษาอังกฤษ ตัวเลข และสัญลักษณ์พิเศษต่างๆ ออกมาเป็นแต่ละคำเพื่อนำไปใช้ในกระบวนการหาค่าน้ำหนักคำ สำหรับตัดคำในเอกสารภาษาไทยจะถูกพิจารณาเป็น 2 ขั้นตอนคือ 1.การใช้กฎ โดยใช้ไวยากรณ์ทางภาษา แบ่งตัวอักษรเป็นหมวดหมู่ตามลักษณะการใช้งาน ได้แก่ กลุ่มพยัญชนะ กลุ่มสระ กลุ่มวรรณยุกต์ กลุ่มตัวเลข และกลุ่มตัวอักษรพิเศษ ขั้นตอนการตัดพยางค์จะทำจากซ้ายไปขวา .2การใช้พจนานุกรม ผลลัพธ์ที่ได้จะอยู่ในระดับคำ โดยมีหลักการว่าให้ทำการตรวจสอบสายอักขระ (String) และนำไปค้นหาจากพจนานุกรม หากพบคำในพจนานุกรมที่

สามารถเป็นคำในสายอักขระนั้นได้มากกว่าหนึ่งคำ จะทำการเลือกคำที่ยาวที่สุด (Longest matching) หากอักษรตัวต่อมาไม่พบว่าเป็นคำที่มีอยู่ในพจนานุกรมที่มีอยู่ก็จะทำการย้อนกลับไปเลือกคำที่สั้นกว่าแทน

4. การให้น้ำหนักคำ (Word Weighting)

เป็นการสร้างเนื้อหาของเอกสาร ให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเรียนรู้ได้ สำหรับนำไปใช้ในกระบวนการเรียนรู้ลักษณะของตัวแทนเอกสาร จะอยู่ในรูปแบบเวกเตอร์น้ำหนักคำ TF-IDF (Term Frequency Inverse Document Frequency) เป็นวิธีคำนวณน้ำหนักจากความถี่ ของการปรากฏของคำ T_j ในเอกสาร D_i และพิจารณาความถี่ของคำ T_j ที่ปรากฏในเอกสารอื่นร่วมด้วย แสดงค่าน้ำหนักในตารางที่ 1

$$w_{ij} = tf_{ij} \times \log_2 \frac{N}{n_j}$$

โดยที่ w_{ij} = น้ำหนักของคำ T_j ในเอกสาร D_i

tf_{ij} = ความถี่ของคำ T_j ในเอกสาร D_i

N = จำนวนเอกสารทั้งหมดในระบบ

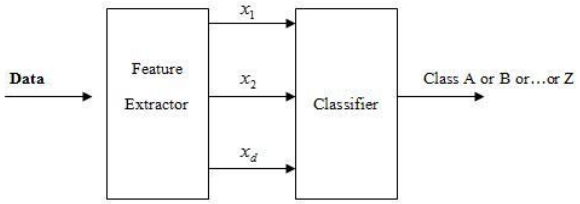
n_j = จำนวนเอกสารที่มีคำ T_j ปรากฏอยู่อย่างน้อยหนึ่งครั้ง

ตารางที่ 1 ผลของค่าน้ำหนักของแต่ละเอกสาร

	T_1	T_2	•••	T_j
D_1	w_{11}	w_{12}	•••	w_{1j}
D_2	w_{21}	w_{22}	•••	w_{2j}
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
D_i	w_{i1}	w_{i2}	•••	w_{ij}

5. การจัดหมวดหมู่

เป็นกระบวนการแบ่งแยกวัตถุหรือเหตุการณ์ออกเป็นกลุ่มโดยใช้ลักษณะสำคัญ เป็นวิธีที่ใช้ควบคุมค่าความแปรผันของข้อมูล เพื่อใช้เป็นเกณฑ์ในการแบ่งกลุ่ม การออกแบบจึงขึ้นกับผู้ออกแบบที่จะเลือกลักษณะสำคัญดังแสดงการสกัดลักษณะสำคัญ และการจัดกลุ่ม ในรูปที่ 3



รูปที่ 3 กระบวนการสกัดลักษณะสำคัญสำหรับแบ่งกลุ่มข้อมูล

โดยทั่วไปจะอยู่ในรูปของเวกเตอร์ ซึ่งจะเรียกเวกเตอร์นี้ว่าเวกเตอร์ลักษณะ (Feature Vector) มีรูปแบบเป็นคอลัมน์เวกเตอร์ขนาด d x 1 เมื่อ d คือ จำนวนที่นำมาใช้ในการจำแนก

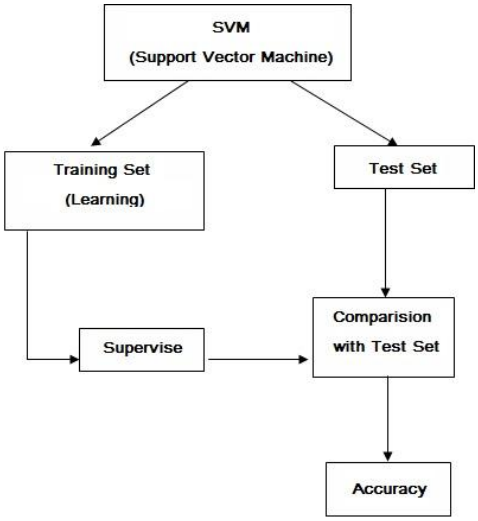
$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

เมื่อ x_1 คือ เอกสารที่เป็นความลับ

x_2 คือ เอกสารทั่วไป

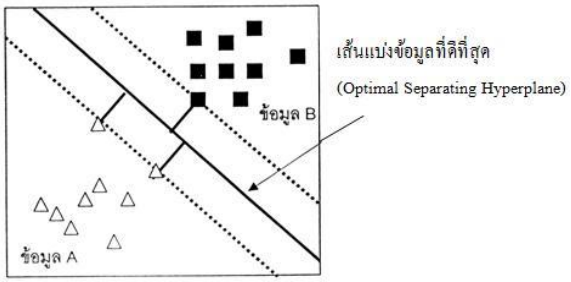
6. ทฤษฎี SVM (Support Vector Machines)

เป็นการแยกประเภทแบบมีผู้สอน (Supervised Learning) ทฤษฎีนี้ได้มาจากแนวความคิดของ Vapnik and Chervonenkis และได้มีการนำมาเสนอโดย Boser, Guyon, Vapnik in COLT-92 ดังรูปที่ 4



รูปที่ 4 แผนผังการทำงานของ Support Vector Machines

จัดเป็นเทคนิคที่ใช้ในการแก้ไขปัญหา ทางด้านการรู้จำรูปแบบข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลจากข้อมูลที่ถูกป้อนเพื่อใช้ในการสอนให้ระบบรู้จำ โดยเน้นไปยังเส้นแบ่งที่แยกแยะกลุ่มข้อมูลได้ดีที่สุด (Optimal separating hyperplane)



รูปที่ 5 เส้นแบ่งกลุ่มข้อมูลจาก SVM

$$f(x) = sign(w \cdot x + b)$$

เมื่อ w หมายถึง เวกเตอร์ค่าน้ำหนักของการรู้จำที่ได้จากกระบวนการการให้น้ำหนักคำ

b หมายถึง ค่าไบอัส (bias) สำหรับระบบรู้จำ

x หมายถึง เวกเตอร์ข้อมูลที่ใช้ในการสอนระบบ (Feature Vector)

sign หมายถึง ถ้าค่าที่ได้มากกว่า 0 จะเป็น +1 ถ้าน้อยกว่า 0 จะเป็น -1

ในกรณีที่ต้องทำงานกับข้อมูลที่ไม่เป็นเชิงเส้น สามารถแก้ปัญหาได้ด้วยการเปลี่ยนแปลงมิติของข้อมูลให้มีมิติที่สูงขึ้นซึ่งเรียกว่าฟีเจอร์สเปช (Feature Space) ในฟีเจอร์สเปชใหม่นี้ข้อมูลจะถูกวางตัวเป็นข้อมูลแบบเชิงเส้นและทำการสร้างสมการแบ่งข้อมูลในรูปแบบของเชิงเส้นบนมิติ ตามทฤษฎีของเมอร์เซอร์ (Mercer’s Theorem) โดยฟังก์ชัน $K(x, x_i)$ ซึ่งเรียกว่า เคอร์แนลฟังก์ชัน (Kernel Function) เป็นการสร้างเส้นแบ่งกลุ่มข้อมูลที่ดีที่สุดในฟีเจอร์สเปชใหม่ โดยเลือกใช้ชนิดของเคอร์เนลฟังก์ชันเป็น Radial Basis Function (RBF) ดังสมการ

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2)$$

ตามสมการในการแบ่งกลุ่ม

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i^* K(x, x_i) + b^*)$$

จากสมการ ค่า y_i คือคำตอบ (Class) ของ Support Vector ที่ x_i และ x คือข้อมูลชุดทดสอบ ค่า α_i และค่า b เป็นค่าสัมประสิทธิ์ที่ได้มาจากการคำนวณ (Optimization)

การดำเนินงานวิจัย

1. แนวทางการวิจัยและพัฒนา

เนื่องด้วยมีการนำระบบงานการป้องกันข้อมูลรั่วไหลเข้ามาใช้ในองค์กร ด้วยอุปกรณ์ Symantec DLP ดังได้อธิบายโครงสร้างการทำงานของระบบดังกล่าวแล้ว ด้วยความสามารถของอุปกรณ์ จะทำงานในส่วนของการเฝ้าระวังข้อมูลสำคัญและการบังคับใช้นโยบายที่ได้กำหนดขึ้นเท่านั้น ขั้นตอนก่อนหน้านั้นจำเป็นต้องมีการกำหนดข้อมูลที่ถือว่าสำคัญและเป็นความลับขององค์กรก่อนนำเข้าระบบ โดยมีการจัดเก็บในรูปแบบอิเล็กทรอนิกส์ในปริมาณมากและยากต่อการระบุและแยกประเภทของเอกสาร จึงมีวัตถุประสงค์ในการออกแบบวิธีการแยกประเภทเอกสาร ที่มีใช้งานในหน่วยงานภายในของการประปานครหลวงโดยการสร้างระบบการคัดแยกเอกสาร เพื่อคัดแยกเอกสารลับ ออกจากเอกสารทั่วไป จะนำเอกสารที่ถูกกำหนดว่าเป็นเอกสารลับ โดยหน่วยงานเจ้าของเอกสารนั้น ๆ และเอกสารที่ไม่ใช่เอกสารลับ(เอกสารทั่วไป) นำมาเป็นข้อมูลชุดทดสอบ ยกตัวอย่างเช่น ถ้ามีเอกสาร 1000 ฉบับ เรานำต้นแบบเอกสาร ฉบับ 100 ซึ่งในเอกสาร ฉบับ อาจจะประกอบไปด้วยเอกสารสำคัญ และเอกสารทั่วไป 100จากเอกสารทั้งหมดมาทำการฝึกสอนเพื่อจำแนกประเภทเอกสารที่เหลืออีก ฉบับที่จะนำมาทดสอบกับข้อมูลต้นแบบใน 900 ภายหลัง เพื่อแยกประเภทเอกสารทั้งหมด โดยมีข้อจำกัดว่าเอกสารเหล่านั้นจะต้องไม่มีการเปลี่ยนแปลง และแก้ไขระดับความสำคัญ เช่นการระบุว่าเป็นเอกสารลับทีหลัง หลังจากที่กำหนดเป็นเอกสารทั่วไปแล้ว และจะไม่ครอบคลุมถึงเอกสารที่มีการเปลี่ยนแปลงในภายหลังจากการแยกประเภทแล้ว ขั้นตอนต่อไปจะนำเอกสารไปทำการทดลองตามกระบวนการต่าง ๆ เพื่อให้สามารถแบ่งประเภทของเอกสารตามขั้นตอนในการดำเนินงานวิจัยดังตารางที่ 2 และรูปที่ 6

ตารางที่ 2 รายละเอียดกระบวนการดำเนินงาน

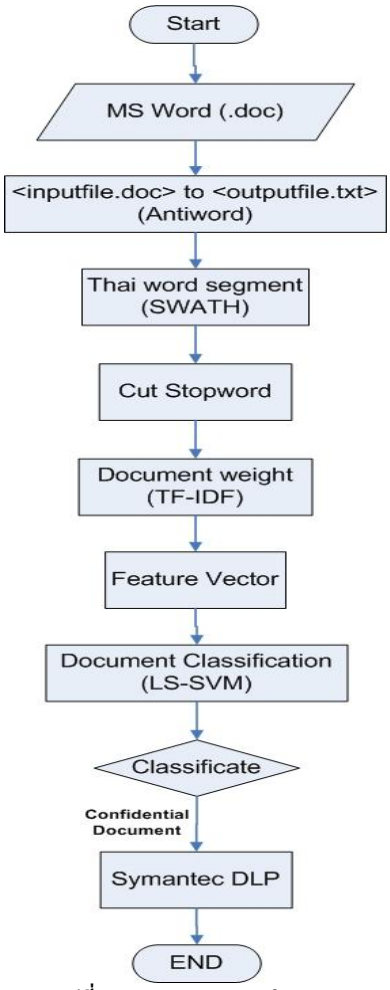
ขั้นตอน	วิธีการทำงาน
1	นำเอกสาร Ms word (.doc) เข้าสู่ระบบ
2	แปลงไฟล์เอกสารจาก .doc ไปเป็นไฟล์ .txt โดยใช้ Antiword
3	นำเข้าไฟล์ .txt เข้าสู่กระบวนการตัดคำภาษาไทยโดยใช้ SWATH
4	นำคำที่ได้จากการตัดคำมาเปรียบเทียบกับข้อมูลคำหยุด และลบคำหยุดทิ้ง
5	คำนวณค่าน้ำหนักของแต่ละเอกสารโดยใช้สมการ TFIDF
6	กำหนดคุณลักษณะสำคัญของการแยกประเภท (Feature Vector)
7	นำเข้าค่าน้ำหนักและคุณลักษณะสำคัญเพื่อคำนวณการแยกประเภท LS-SVM
8	นำเอกสารที่ทำการแยกประเภทและได้ผลเป็นเอกสารลับเข้าไปเป็นต้นแบบการกรองให้ระบบการป้องกันข้อมูลรั่วไหล Symantec DLP

2. **กระบวนการนำเข้าเอกสาร**

เอกสารที่นำเข้าระบบจะอยู่ในรูปแบบของ Microsoft Word (.doc) ต้องทำการแปลงไฟล์เอกสารให้อยู่ในรูปแบบของไฟล์ข้อความธรรมดา (.txt)โดยใช้โปรแกรม Antiword นำไฟล์ข้อความที่ได้นำเข้าสู่กระบวนการตัดคำภาษาไทย โดยใช้โปรแกรม SWATH ลักษณะการทำงานเป็นการเปรียบเทียบคำในเอกสารกับข้อมูลในพจนานุกรม ถ้าตรงกันในลักษณะการเปรียบเทียบจากซ้ายไปขวา และเลือกคำที่ยาวที่สุดที่สามารถเปรียบเทียบได้ (Longest Matching) เมื่อได้คำที่ตัดแล้ว จะเข้าสู่ขั้นตอนการตัดคำหยุด (Stop word) คือคำฟุ่มเฟือย คำเชื่อม ตัวเลข หรือคำที่ไม่มีความจำเป็นออก เช่น เป็น,อยู่,คือ,และ,มี, ...ฯลฯ นำความถี่ของคำที่เหลือเข้าสู่กระบวนการหาค่าน้ำหนักคำของแต่ละคำในแต่ละเอกสาร โดยใช้ TFIDF ก่อนนำไปเป็นอินพุตเข้าสู่กระบวนการแบ่งประเภทเอกสารต่อไป

3.กระบวนการแบ่งประเภท

นำค่าน้ำหนักเอกสาร TFIDF เป็นอินพุตเข้าสู่กระบวนการแบ่งประเภทเอกสาร โดยใช้วิธี LS-SVM วิธีนี้จะต้องให้คอมพิวเตอร์ทำการเรียนรู้ข้อมูลก่อนที่จะทำการเปรียบเทียบเพื่อแบ่งประเภทเอกสาร ได้กำหนดลักษณะสำคัญ (Feature Vector) ขึ้นมา 2 ประเภท คือ เอกสารลับ (Secret) และเอกสารทั่วไป (Non Secret) โดยแทนด้วยค่าตัวเลข 1 และ -1 ตามลำดับ



รูปที่ 6 กระบวนการดำเนินงาน

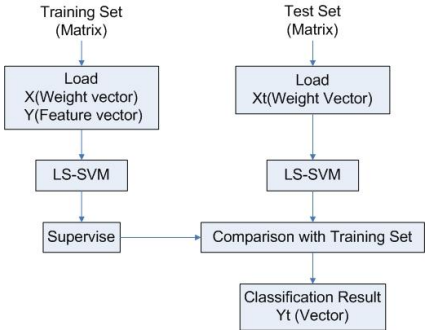
4. กระบวนการทดสอบ

นำข้อมูลทั้งหมดมาทำการฝึกสอน โดยกำหนดประเภทเอกสารให้ข้อมูลชุดฝึกสอนก่อน ว่าจัดอยู่ในประเภทเอกสารลับ (Class 1) หรือเอกสารทั่วไป (Class -1) และทำการทดสอบโดยนำข้อมูลชุดทดสอบ โดยนำเอาข้อมูลส่วนหนึ่งของข้อมูลฝึกสอน มาทำการทดสอบตามประเภทที่ได้กำหนดไว้ เริ่มจากนำข้อมูลเอกสารทั้งหมด 140 ฉบับซึ่งถือว่าเป็นข้อมูลชุดฝึกสอน แบ่งเป็น เอกสารลับ 40 ฉบับ และเอกสารทั่วไป 100 ฉบับ ดังแสดงในตารางที่ 3

ตารางที่ 3 เอกสารที่ใช้ในการฝึกสอนและทดสอบ

ชุดข้อมูล (Data Set)	ชุดข้อมูลตั้งต้น
ฝึกสอน (Training Set)	140 ฉบับ
ทดสอบ (Test Set)	25 ฉบับ/ประเภท

นำข้อมูลชุดทดสอบประเภทโดยเป็นเอกสารลับ 5 และเอกสารทั่วไป 20 นำเข้าสู่กระบวนการข้างต้น และดูผลการทดสอบว่าผลของการแยกประเภทจะตรงกับประเภทเอกสารที่ระบบได้เรียนรู้ไว้หรือไม่ โดยการทดสอบด้วยโปรแกรม MATLAB LS-SVM ขั้นตอนที่กล่าวมาแสดงขั้นตอนดังรูปที่ 7



รูปที่ 7 ขั้นตอนการฝึกสอนและทดสอบการแยกประเภทเอกสารโดยใช้ LS-SVM

ผลการทดสอบ

ผลลัพธ์ที่ได้จากการนำเวกเตอร์ข้อมูลชุดฝึกสอน $X(140\times 255)$ และเวกเตอร์ลักษณะสำคัญ $Y(140\times 1)$ โดยกำหนดให้เป็น $[-1,1]$ โดยค่าสัมประสิทธิ์ที่ได้จากการคำนวณเป็น $\alpha = -84.29 \quad 149.77$ และค่าไบอัส (bias) สำหรับให้ระบบรู้จำที่ได้จากการคำนวณของระบบ $b = -2.4214$ ส่วนค่า $gamma = 92.8812$ เป็นค่าเริ่มต้นของ LS-SVM ส่วน $sig2 = 5.0911$ ซึ่งได้จากการกำหนดสัมประสิทธิ์เพื่อให้ได้เส้นแบ่งข้อมูล (Optimal Sparating Hyperplane) สามารถทำการแบ่งเอกสารได้เป็น 2 ประเภท (Class 1,Class-1) แสดงดังตารางที่ 4

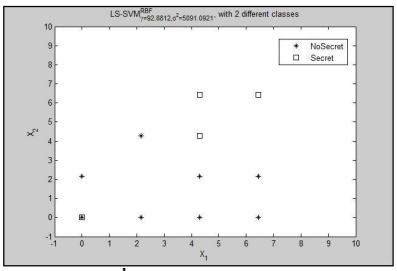
ตารางที่ 4 ผลที่ได้จากข้อมูลชุดทดสอบ

ตัวแปร	ค่า	ต่ำสุด	สูงสุด
ค่าน้ำหนักเอกสารชุดฝึกสอน (X)	140×255	0	154.52
คุณลักษณะสำคัญ (Y)	140×1	-1	1
ค่าสัมประสิทธิ์ที่ได้จากการคำนวณ	140×1	-84.29	149.77
ค่าไบอัส (b)	-2.4214	-2.4214	-2.4214
ค่าแกมม่า	92.8812	92.8812	92.8812
ค่า $sig2$	5.0911	5.0911	5.0911

ขั้นตอนการทดสอบโดยการนำข้อมูลชุดทดสอบในแต่ละประเภทเอกสารมาเปรียบเทียบกับข้อมูลชุดฝึกสอน ในที่นี้ได้ทำการทดสอบ โดยนำเอกสารจำนวน และเอกสารทั่วไป 5 เอกสารที่ประกอบไปด้วยเอกสารลับ 2520 เอกสาร ที่ถูกกำหนดให้เป็น Class 1 และ Class-1 มาทดสอบกับข้อมูลชุดฝึกสอน 140 เอกสาร ได้ผลเป็น Class 1 และ Class-1 ตามที่ได้กำหนดไว้ดังแสดงในตารางที่ 5 และแสดงกราฟ LS-SVM ที่ได้จากการแยกประเภทดังรูปที่ 8

ตารางที่ 5 ผลที่ได้จากข้อมูลชุดทดสอบประเภทข้อมูลสำคัญ

ตัวแปร	ค่า	ต่ำสุด	สูงสุด
ค่าน้ำหนักเอกสารชุดฝึกสอน	140×255	0	154.52
ค่าน้ำหนักชุดทดสอบ	25×255	0	154.52
คุณลักษณะสำคัญ	140×1	-1	1
ผลลัพธ์การแยกประเภทเอกสารลับ	5×1	1	1
ผลลัพธ์การแยกประเภทเอกสารทั่วไป	20×1	-1	-1



รูปที่ 8 กราฟ LS-ที่ได้จากการแยกประเภทเอกสาร SVM

สรุปผลการวิจัยและข้อเสนอแนะ

จากผลการทดลองแยกประเภทเอกสาร โดยใช้วิธี LS-SVM แสดงให้เห็นว่าการนำข้อมูลชุดทดสอบที่ประกอบไปด้วยเอกสารทั้งสองประเภททั้งหมด 25 เอกสาร มาทดลองกับกลุ่มข้อมูลชุดฝึกสอนจำนวน 140 เอกสาร ผลที่ได้สามารถจำแนกประเภทได้ตรงกับที่ได้กำหนดประเภทเอกสารไว้ล่วงหน้า และด้วยความหลากหลายของชุดเอกสารที่มีความจำเพาะในแต่ละหน่วยงานของการประปานครหลวง จำเป็นต้องทำการแยกการทดสอบเป็นข้อมูลคนละชุดต่อไป

เอกสารอ้างอิง

จันทิมา พลพินิจ, ชมศักดิ์ สีบุญเรือง, รพีพร ช่ำชอง, อนิรุทธิ์ โชติถนอม และ สมนึก พ่วงพรพิทักษ์. (2548). *Automated Obscenity Web Sites Filetering System*. คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม การประชุมวิชาการ สวทช.

นนท์ บุญนิธิประเสริฐ. (2552). *การกรองข้อความภาษาไทย และภาษาอังกฤษของบริการส่งข้อความสั้นบนเครือข่ายโทรศัพท์เคลื่อนที่*. วิทยานิพนธ์ วิศวกรรมศาสตร์มหาบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์และโทรคมนาคม บัณฑิตวิทยาลัย มหาวิทยาลัยธุรกิจบัณฑิต.

ปโยธร อุราธรรมกุล และกานดา รุณนะพงศา. (2005). *การปรับปรุงกฎสำหรับตัดคำในเอกสารภาษาไทย*. ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น.

พรพล ธรรมรงค์รัตน์. (2551). *การจำแนกประเภทเว็บเพจ โดยวิธีการลดขนาดลักษณะเฉพาะและซัพพอร์ตเวกเตอร์แมชชีน*. วิทยาศาสตร์มหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยสงขลานครินทร์.

อภิชาต ขานทอง ; วัลลภา ตันติประสงค์ชัย และชุลีรัตน์ จรัสกุลชัย. *การสรุปใจความสำคัญของเอกสาร*. ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตบางเขน

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

George Lawton, "*New Technology Prevents Data Leakage*", IEEE Computer Security September 2008, pp.14-17.

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

Daeseon Choi,Seunghun Jin and Hyunsoo Yoon : *A Personal Information Leakage Preventiion Method* on the Internet,2006

Johan A.K.Suykens, Tony Van Gestel, Jos De Brabanter, Bart De Moor and Joos Vandewalle. *Least Squares Support Vector Machines*,K.U. Leuven,Belgium

Zhijie Liu,Xueqiang Lv,Kun Liu,Shuicai Shi :*Study on SVM Compared with the other Text Classification Method*, Information Science and Technology University Beijing China,2010

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

Gilberto, Pedro, Edmo, Jayme. *A Security Framwork to Protect Against Social Networks Services Threats*. 2010 Filfth International Conference on Systems and Networks Communications.

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

Yuguo Wang. *A Tree-based Multi-class SVM Classifier for Digital Library*. International Conference

on MultiMedia and Information Technology,2008

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

7 step to information Protection in 2009 , Symantec ;White Paper Data Loss Prevention. Retrieved Jul2010, from www.symantec.com

Antiword Version 0.37. Retrieved Jul 2010, from http://www.winfield.demon.n

BitArmor. Retrieved Jul 2011, from http://www.bitarmor.com/solutions/enhance-dlp

Bluecoat. Retrieved Jul 2011, from http://www.bluecoat.com/products/dataloss-prevention

Check point. Retrieved Jul 2011, from http://www.checkpoint.com/products/index.html#endpoint

Code Green. Retrieved Jul 2011, from http://www.codegreennetworks.com/products/endpoint.htm

Endpoint Protector. Retrieved Jul 2011, from http://www.endpointprotector.com/en/index.php/products/my_endpoint_protector_Saa#

GTB Technology. Retrieved Jul 2011, from http://www.gtbtechnologies.com/products_end_point_protector.asp

Ironport. Retrieved Jul 2011, from http://www.ironport.com/technology/ironport_dlp_overview.html

LS-SVMlab1.7 . Retrieved Jul 2010, from http://www.esat.kuleuven.ac.be/sista/lssvmlab/

NexTier. Retrieved Jul 2011, from http://www.nextiernetworks.com/technology.php

NextLabs . Retrieved Jul 2011, from http://www.nextlabs.com/html/?q=endpoint-data-loss-prevention

Palisade Systems . Retrieved Jul 2011, from http://www.palisadesystems.com/products/packetsure_ad.aspx

RSA. Retrieved Jul 2011, from http://www.rsa.com/node.aspx?id=3426

Sophos. Retrieved Jul 2011, from http://www.sophos.com/products/enterprise/endpoint/

Software SWATH (Thai Word Segmentation). Retrieved Jul 2010, from http://www.cs.cmu.edu/paisarn/software.html

Symantec Data Loss Prevention AdministrationGuide Version 10.5. Retrieved Jul 2010, from www.symantec.com

Symantec. Retrieved Jul 2011, from http://www.symantec.com/en/uk/business/solutions/solutiondetail.jsp?solid=sol_security&solfid=sol_endpoint_security

Trus Wave. Retrieved Jul 2011, from https://www.trustwave.com/dlpoverview.php

Vericept. Retrieved Jul 2011, from https://www.vericept.com/index.php?id=58

WebSense. Retrieved Jul 2011, from http://www.websense.com/content/DataDiscover.aspx

Websense. Retrieved Jul 2011, from http://www.websense.com/content/DataSecuritySuite.aspx