PPA143

การจัดกลุ่มเอกสารข้อความภาษาไทยด้วยอัลกอริทึม SVMs

(Thai Text Classification By Support Vector Machines)

ได้รับทุนอุดหนุนจากโครงการวิจัย พัฒนาและวิศวกรรม
โครงการแข่งขันพัฒนาโปรแกรมคอมพิวเตอร์
แห่งประเทศไทย ครั้งที่ 7 ประจำปีงบประมาณ 2547

บทคัดย่อ

เนื่องจากความนิยมในการใช้งานระบบอินเตอร์เน็ตและความหลากหลายเกี่ยวกับบริการบนระบบอินเตอร์เน็ตที่เพิ่มมากขึ้น ส่งให้เกิดปัญหาตามมาในเรื่องของจำนวนสารสนเทศที่เพิ่มมากขึ้น การจัดกลุ่มเอกสารเป็นอีกหนึ่งทางเลือกที่สามารถช่วยให้ผู้ใช้ระบบสามารถค้นหาสารสนเทศที่ต้องการได้รวดเร็วมากขึ้น สำหรับงานวิจัยฉบับนี้ได้นำเสนอวิธีการจัดกลุ่มเอกสารด้วยอัลกอริทึม Support Vector Machines (SVMs) สำหรับการสร้างตัวแยกแยะเอกสารเพื่อจัดกลุ่มเอกสารข้อความภาษาไทยแบบอัตโนมัติ บนพื้นฐานของข้อมูลที่มีใช้งานอยู่จริง

หลักการและเหตุผล

เนื่องจากการเพิ่มขึ้นของจำนวนเอกสารซึ่งมีอยู่เป็นจำนวนมาก ได้เริ่มสร้างปัญหาให้กับผู้สืบค้นข้อมูลสารสนเทศเพื่อนำไปใช้งาน ดังนั้นงานทางด้านการค้นคืนเอกสารแบบอัตโนมัติซึ่งได้รับความสนใจกับนักวิจัยที่ทำงานในด้านการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งหนึ่งในงานวิจัยที่สำคัญ คือ การจำแนกประเภทของเอกสาร (Text Classification)

ในการจัดกลุ่มเอกสารกับข้อความภาษาไทยจะค่อนข้างยุ่งยากกว่าเอกสารที่เป็นภาษาอังกฤษ (ชุลีรัตน์และคณะ, 2000) เนื่องจากภาษาไทยมีลักษณะภาษาที่มีการเขียนในรูปแบบสายอักขระต่อเนื่องที่ไม่มีจุดสิ้นสุดของประโยคอย่างชัดเจน ทำให้ปัญหาในเรื่องของการสร้างคำขอ (Query) ที่เป็นส่วนสำคัญในการจัดกลุ่มเอกสาร

เนื่องจากการเพิ่มขึ้นของจำนวนเอกสารซึ่งมีอยู่เป็นจำนวนมาก ได้เริ่มสร้างปัญหาให้กับผู้สืบค้นข้อมูลสารสนเทศเพื่อนำไปใช้งาน ดังนั้นงานทางด้านการค้นคืนเอกสารแบบอัตโนมัติซึ่งได้รับความสนใจกับนักวิจัยที่ทำงานในด้านการค้นคืนสารสนเทศ (Information Retrieval) ซึ่งหนึ่งในงานวิจัยที่สำคัญ คือ การจำแนกประเภทของเอกสาร (Text Classification)

สำหรับโครงงานนี้นำเสนออัลกอริทึม Support Vector Machines (SVMs) มาประยุกต์ใช้ในการจัดกลุ่มเอกสารข้อความภาษาไทย เพื่อเป็นอีกแนวทางหนึ่งในการจัดกลุ่มเอกสารโดยเฉพาะเอกสารข้อความภาษาไทย

วัตถุประสงค์และเป้าหมายของการจัดทำโครงงาน

เพื่อพัฒนาระบบการจัดกลุ่มเอกสารข้อความภาษาไทย (Thai Text Classification) แบบอัตโนมัติ โดยการประยุกต์อัลกอริทึมทางด้าน Machine Learning ที่เรียกว่า Support Vector Machines (SVMs) และเพื่อเพิ่มประสิทธิภาพในการค้นคืนสารสนเทศ

ขอบเขตและข้อจำกัดของโครงงาน

พัฒนาระบบการจัดกลุ่มเอกสารข้อความภาษาไทย (Thai Text Classification) ด้วยอัลกอริทึมด้าน Machine Learning ที่เรียกว่า Support Vector Machines (SVMs) โดยเอกสารที่นำมาใช้ในการทดลองการจัดกลุ่มเอกสารจะเป็นเอกสารข้อความภาษาไทยเท่านั้น ซึ่งลักษณะของเอกสารจะจัดให้อยู่ในรูป HTML ที่ไม่รวมถึงภาพ อักขระพิเศษ สูตรหรือสมการใดๆ

สมการหลัก

$$wx + b$$

ถ้าค่าของ $wx + b > 0$ จะกำหนดให้ค่า $y = 1$ ซึ่งจะจัดอยู่ใน Class 1

ถ้าค่าของ $wx + b < 0$ จะกำหนดให้ค่า $y = -1$ ซึ่งจะจัดอยู่ใน Class 2

การวัดประสิทธิภาพ

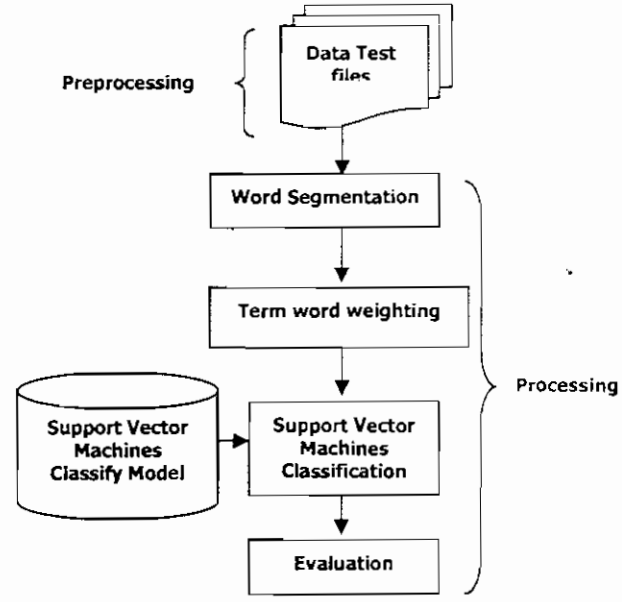
$$\text{ค่าความแม่นยำ} = \frac{\text{จำนวนเอกสารที่ถูกต้องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่ถูกต้องทั้งหมดที่ค้นคืนออกมาได้}}$$

$$\text{ค่าความระลึก} = \frac{\text{จำนวนเอกสารที่ถูกต้องที่ค้นคืนได้}}{\text{จำนวนเอกสารที่ถูกต้องทั้งหมดในฐานข้อมูล}}$$

การวัดคุณภาพซึ่งเป็นค่าผสมระหว่างค่าความแม่นยำและค่าความระลึก

$$F_{\beta,j} = \frac{(b^2 + 1)P_j R_j}{b^2 P_j + R_j}$$

ขั้นตอนการดำเนินการ

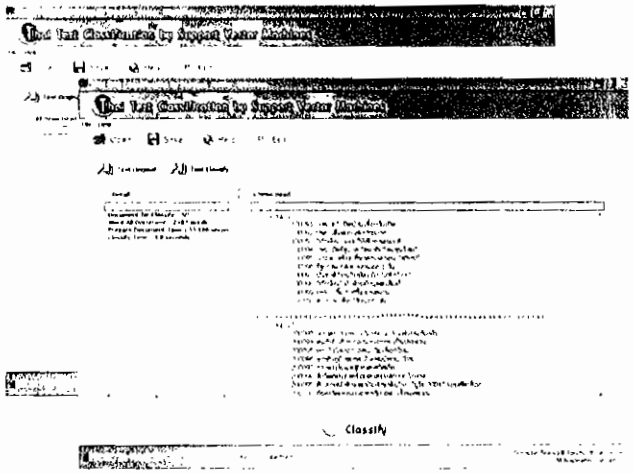


สรุปผลการทดลอง

ผลการทำงานโปรแกรมที่ได้พัฒนาขึ้น โดยได้นำไปทดสอบจัดกลุ่มเอกสารข่าว จำนวน 100 เอกสาร โดยแบ่งออกเป็น 5 ประเภท คือ ข่าวการเมือง , ข่าวกีฬา , ข่าวเศรษฐกิจ , ข่าวบันเทิง และข่าวต่างประเทศ หลังจากนั้นจะเป็นการประเมินผลความถูกต้องของผลการจัดกลุ่มเอกสารโดยโปรแกรมที่พัฒนาขึ้นดังกล่าว สามารถสรุปได้ว่า ระบบมีความน่าเชื่อถือในระดับที่น่าพอใจคือประมาณ 72 %

ตัวอย่างโปรแกรม

ตัวอย่างโปรแกรมที่ทำการจัดกลุ่มเอกสาร



ผู้จัดทำ

นายภาณุวัฒน์ บุญผาสุก

(phanuwat_cs@hotmail.com)

นางสาวสุรีย์พร สีดาคุณ

(Victoria_cs99@hotmail.com)

อาจารย์ที่ปรึกษา

อุมาภรณ์ สายแสงจันทร์

(umaporn.s@msu.ac.th)

จันทิมา พลพินิจ

(jantima.p@msu.ac.th)

สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ

มหาวิทยาลัยมหาสารคาม

อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150

โทรศัพท์ :: 0-4375-4359 หรือ

0-4375-4322-40 ต่อ2414 , 2453

แฟกซ์ :: 0-4375-4359