บทที่ 1

บทนำ

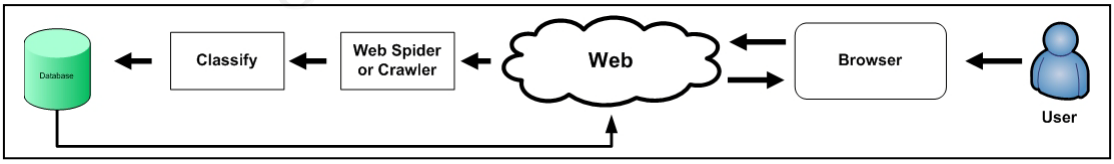
บทนี้จะกล่าวถึงหลักการและเหตุผล วัตถุประสงค์ของการศึกษา ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา ขอบเขตของการศึกษา และคำนิยามศัพท์ เพื่อให้ทราบถึงมูลเหตุจูงใจและขอบเขตในการจัดทำการค้นคว้าแบบอิสระนี้

1.1 หลักการและเหตุผล

ปัจจุบันเว็บไซต์ส่วนใหญ่เป็นไดนามิกเว็บเพจ (Dynamic Web Page) ซึ่งมีการเปลี่ยนแปลงอยู่ตลอดเวลาและปริมาณเนื้อหาข้อมูลที่เพิ่มขึ้นอย่างรวดเร็วบนเว็บไซต์ต่างๆ เป็นปัจจัยหลักที่ทำให้ผู้ใช้เกิดความลำบากในการค้นหาข้อมูล ซึ่งมีส่วนข้อมูลอื่นๆ ที่ไม่ต้องการเรียกว่า Web Spam หรือเมื่อเปรียบเทียบกับอีเมล์แล้ว Web Spam ก็คือ ข้อมูลที่ไม่ต้องการและมีส่วนข้อมูลที่ต้องการของกลุ่มเป้าหมาย เช่น กลุ่มวิศวกร กลุ่มการแพทย์ กลุ่มคอมพิวเตอร์ ฯลฯ ที่ต้องการความแม่นยำของการค้นหาข้อมูล

เสิร์ชเอนจิน (Search Engine) เป็นโปรแกรมที่ช่วยในการสืบค้นหาข้อมูล โดยเฉพาะข้อมูลบนอินเทอร์เน็ต โดยครอบคลุมทั้งข้อความ รูปภาพ ภาพเคลื่อนไหว เพลง ซอฟต์แวร์ แผนที่ ข้อมูลบุคคล กลุ่มข่าว และอื่น ๆ ซึ่งแตกต่างกันไป แล้วแต่โปรแกรมหรือผู้ให้บริการแต่ละราย เสิร์ชเอนจินส่วนใหญ่จะค้นหาข้อมูลจาก คำสำคัญ -เอกสาร ที่ผู้ใช้ป้อนเข้าไป จากนั้นก็จะแสดงรายการผลลัพธ์ที่คิดว่าผู้ใช้น่าจะต้องการขึ้นมา การทำงานของเสิร์ชเอนจินโดยทั่วไปประกอบด้วย 3 ส่วนหลัก คือ 1) โปรแกรม โรบอต (Robot) เว็บสไปเดอร์ (Web Spider) หรือ ครอว์เลอร์ (Crawler) 2) อินเด็กเซอร์ (Indexer) 3) ระบบค้นหา (Query)

ในส่วนของเว็บ สไปเดอร์ (Web Spider) จะเป็นตัวที่ทำหน้าที่เข้าสำรวจเว็บไซต์ต่างๆ แล้วดึงข้อมูลเหล่านั้นมาอัพเดตใส่ในรายการฐานข้อมูล (Database) ซึ่งเป็นส่วนที่เก็บรายการเว็บไซต์ ฐานข้อมูลที่ดีควรจะมีขนาดใหญ่เพียงพอและวิธีในการจัดเก็บที่เหมาะสมเพื่อรองรับกับการเติบโตของเว็บไซต์ในปัจจุบัน ดังรูป 1.1



รูป 1.1 การทำงานของเว็บสไปเดอร์

การเพิ่มขึ้นของ เว็บไซต์ ในงานด้านต่างๆ ทำให้เกิดปัญหาในการจัดการกับ เว็บไซต์ ที่เพิ่มขึ้นซึ่งรวมไปถึงการ คัดกรอง เว็บไซต์ ให้เป็นหมวดหมู่ด้วยเช่นกัน ถ้าเราสามารถ คัดกรอง เว็บไซต์เป็นหมวดหมู่ ก็จะช่วยให้สามารถจัดการ เว็บไซต์ได้อย่างมีประสิทธิภาพ เช่น การค้นคืนสารสนเทศ (Information Retrieval) สามารถค้นคืนเอกสารได้ตรงกับความต้องการและรวดเร็ว ด้วยเหตุนี้จึงมีการนำเทคโนโลยีสารสนเทศมาประยุกต์ใช้กับการ คัดกรอง เว็บไซต์ หลักการที่นำมาใช้ในการคัดกรองเว็บไซต์แต่ละกลุ่มนั้นคือการจัดกลุ่มของเอกสาร (Text Classification)

การจัดกลุ่มเอกสารข้อความ (Text Classification) คือ เป็นกระบวนการในจัดกลุ่มเอกสารข้อความออกเป็นหมวดหมู่ ซึ่งกระบวนการจัดกลุ่มเอกสารนี้จะมีการกำหนดเป้าหมายชุดของเอกสารเอาไว้ล่วงหน้า ซึ่งในแต่ละกลุ่มเอกสารที่จัดได้อาจจะมีเอกสารภายในเท่ากับศูนย์ฉบับหรือมีเอกสารอยู่เป็นจำนวนมากก็ได้ ดังนั้นอาจจะกล่าวได้ว่าการจัดกลุ่มข้อความเป็นกระบวนการจัดระเบียบชุดของเอกสาร เช่น จัดกลุ่มของเอกสารประเภทเว็บ บทความข่าว บทคัดย่อ และที่คั่นหนังสือ (Bookmarks) เพื่อให้ง่ายต่อการสืบค้นข้อมูล

ปัจจุบันได้มีการศึกษาและประยุกต์นำเอาอัลกอริทึมกลไกเชิงเรียนรู้ (Machine Learning) มาใช้ในการจัดกลุ่มเอกสารอย่างกว้างขวาง เนื่องจากให้ประสิทธิภาพที่น่าพอใจ ตัวอย่างอัลกอริทึมที่ได้รับความสนใจและถูกนำมาใช้ส่วนใหญ่เป็นอัลกอริทึมที่ต้องมีกระบวนการของการสอนเพื่อให้เกิดการรู้จำ (Recognition) และระบบสามารถทำการจัดกลุ่มเอกสารได้อย่างอัตโนมัติโดยอาศัยการรู้จำจากการสอนระบบ อัลกอริทึมในกลุ่มนี้เรียกว่า Supervised Learning ซึ่งมีมากมายที่สามารถประยุกต์มาสู่การสร้างโมเดลเพื่อการจัดกลุ่มเอกสารแบบอัตโนมัติ เช่น Support Vector Machines (Corinna and Vladimir, 1995), Decision Trees (Yuan and M.J. Shaw, 1995), Artificial Neural Networks (Han and Kamber, 2001), Instance based learning (Mitchel and McGrawHil, 1997), Rocchio relevance feedback (Jordan and others., 2004), Naïve bayes (Thomas Bayes, 1702), K-nearest Neighbour (Fix and Hodges, 1951) และ Regression model (Hardle, 1990)

ด้วยเหตุนี้ผู้วิจัยจึง มีแนวคิดที่จะสร้างเว็บสไปเดอร์ที่สามารถคัดกรอง (Filter) จำเพาะกลุ่มเป้าหมาย โดยใช้หลักการของซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) โดยมีวัตถุประสงค์ คือการคัดกรองเฉพาะกลุ่มเป้าหมาย ข้อมูลที่ใช้คือเอกสารที่ได้มาจากการ ครอว์เลอร์จากเว็บไซต์ เพื่อทำการสร้างแบบจำลองการจำแนกและการทดสอบประสิทธิภาพ

1.2 วัตถุประสงค์ของการศึกษา

(1) เพื่อพัฒนาเว็บสไปเดอร์แบบจำเพาะกลุ่มเป้าหมาย กรณีศึกษาของกลุ่มคอมพิวเตอร์

(2) เพื่อจัดการข้อมูลรายการเว็บไซต์ ตามกลุ่มเป้าหมาย

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา

(1) ได้เว็บสไปเดอร์ที่สามารถจำแนกกลุ่มเป้าหมาย

(2) ได้ศึกษาหลักจัดกลุ่มของข้อมูลโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน

(3) สามารถนำข้อมูลที่ได้สร้างดัชนีรายการเว็บเพจสำหรับเสิร์ชเอนจินต่อไป

1.4 แผนการดำเนินการ ขอบเขตและวิธีการศึกษา

1.4.1 แผนการดำเนินงาน

(1) ศึกษาหลักการของเสิร์ชเอนจินในส่วนของเว็บสไปเดอร์

(2) ศึกษาหลักการจัดกลุ่มของข้อมูลโดยอาศัยซัพพอร์ตเวกเตอร์แมชชีน

(3) ทำการทดสอบข้อมูล เพื่อให้ได้มาของคำสำคัญ-เอกสารตามกลุ่มเป้าหมาย

(4) ออกแบบและสร้างเว็บสไปเดอร์

(5) ทดสอบและปรับปรุงแก้ไข

(6) จัดทำเอกสารประกอบและคู่มือการใช้งาน

(7) นำเสนอรายงานการค้นคว้าแบบอิสระ

1.4.2 ขอบเขตการวิจัย

ขอบเขตการค้นคว้าคือ เว็บสไปเดอร์ที่สร้างขึ้น โดยการพัฒนาโปรแกรมที่ใช้บนเครื่องไมโครคอมพิวเตอร์ ที่ใช้หลักการเสิร์ชเอนจินในส่วนของตัวเว็บสไปเดอร์ ซึ่งสามารถคัดกรองประเภทของชื่อโดเมน (Domain Name) และจำแนกรายการเอกสารที่เป็นภาษาอังกฤษเท่านั้น กลุ่มเป้าหมายที่ต้องการศึกษาคือ กลุ่มของคอมพิวเตอร์ ซึ่งตัวเว็บสไปเดอร์สามารถเก็บรายการเว็บไซต์ ของหน้าแรก (Index Page) ของเว็บไซต์ ตามกลุ่มเป้าหมายที่ได้กำหนดไว้

1.4.3 วิธีการวิจัย

(1) ศึกษาหลักการของเสิร์ชเอนจินในส่วนของเว็บสไปเดอร์

(2) สร้างข้อมูลฝึกหัด (TrainingData) เพื่อให้ได้ของคำสำคัญ-เอกสารของกลุ่มเป้าหมาย โดยใช้หลักการ TF*IDF (Term Frequency/Inverse Document Frequency) เป็นภาษาอังกฤษเท่านั้น

(3) สร้างเว็บสไปเดอร์แบบจำเพาะกลุ่มเป้าหมาย

(4) ศึกษาหลักการคัดกรองข้อมูลโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน

(5) ทดสอบโปรแกรมและติดตามผล

(6) จัดทำรายงานฉบับสมบูรณ์

1.5 อุปกรณ์ที่ใช้ในการวิจัย

1.5.1 ฮาร์ดแวร์ (Hardware)

(1) หน่วยประมวลผลกลางทำงานด้วยความเร็ว 1.73 GHz

(2) หน่วยความจำหลัก (RAM) ขนาด 1.24 GB

(3) หน่วยความจำสำรอง (Hard Disk) ความจุ 80 GB

1.5.2 ซอฟต์แวร์ (Software)

(1) ระบบปฏิบัติการไมโครซอฟท์ วินโดวส์ เอ็กซ์พี (Microsoft Windows XP)

(2) โปรแกรมภาษาอาร์ (R Language) เวอร์ชั่น 2.8.1

(3) แพกเก็จทีเอ็ม (Text Mining: TM) ของภาษาอาร์

(4) แพกเก็จอี1071 (e1071) ของภาษาอาร์

(5) โปรแกรมไมโครซอฟท์ วิชวลสตูดิโอดอทเน็ต 2005 (Microsoft Visual Studio .NET 2005)

1.6 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล

(1) ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

(2) สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

1.7 นิยามศัพท์

กลุ่มเป้าหมาย หมายถึงกลุ่มที่เราสนใจ กลุ่มจุดมุ่งหมาย เช่น กลุ่มคอมพิวเตอร์ กลุ่มการแพทย์ กลุ่มวิศวกรรมศาสตร์ เป็นต้น

Training data หมายถึงการนำข้อมูลหรือเอกสารมาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใด อยู่ในกลุ่มเดียวกันและจำแนกข้อมูลออกเป็นกลุ่มตามที่ได้กำหนดไว้

เอกสาร หมายถึง กลุ่มของเอกสารที่เราสนใจหรือเป็นกลุ่มเป้าหมาย ซึ่งอาจจะอยู่ในรูปของเอกสารที่เป็นเวิร์ด (Word) หรือเป็นเอกสารสำหรับเว็บไซต์ (Web document)