

บทที่ 6

สรุปผลการอภิปรายผลของการค้นคว้าแบบอิสระ

ในบทนี้กล่าวถึง การสรุปผลการทดลองในการจัดกลุ่มเอกสารโดยใช้ซอฟต์แวร์คอมพิวเตอร์แมชชีน และข้อเสนอแนะการทำงานค้นคว้าแบบอิสระ

6.1 บทสรุป

งานค้นคว้าแบบอิสระฉบับนี้เป็นงานค้นคว้าแบบอิสระเกี่ยวกับการจัดกลุ่มเอกสารโดยการประยุกต์ใช้อัลกอริทึม ซอฟต์แวร์คอมพิวเตอร์แมชชีน ในขั้นก่อนการประมวลผลของงานค้นคว้าแบบอิสระฉบับนี้ จะทำการจัดเก็บเอกสารจากการครอว์เลอร์ของเว็บไซต์โดยกำหนดคยูอาร์แอลเป็น “http://www.cnn.com” ในรูปแบบโครงสร้างเป็น richtext, html, htm, asp, aspx, php และ mspcx จากการทดสอบการทำงานของเว็บไซต์เป็นที่น่าสนใจ ซึ่งสามารถค้นหายูอาร์แอลได้ถึงห้าแสนยูอาร์แอลและสามารถดาวน์โหลดไฟล์ได้หกพันกว่าไฟล์ภายในเวลาหนึ่งชั่วโมง โดยมีไฟล์ที่ไม่สามารถดาวน์โหลด (Errors) เนื่องจากการส่งและรับ Timeout ทุกข้อผิดพลาดของครอว์เลอร์มากกว่าที่ได้กำหนดไว้คือ 20 Timeout นำเอกสารจากการครอว์เลอร์ของเว็บไซต์มา สร้างโมเดลจำนวน 1,000 เอกสาร และทดสอบจำนวน 500 เอกสารและจะนำเอกสารทำการสกัดสัญลักษณ์ สคริปต์ แท็ก (Tag) และแอททริบิวต์ (Attribute) ที่ไม่ต้องการเพื่อต้องการออกมาในรูปแบบของ Text file โดยภาษา Perl โดยทดสอบข้อมูลข่าว 5 ประเภท คือ ข่าวธุรกิจ ข่าวสุขภาพ ข่าวกีฬา ข่าวท่องเที่ยว และข่าวคอมพิวเตอร์

ส่วนในขั้นการประมวลผลจะแบ่งออกเป็นสองกระบวนการคือกระบวนการเรียนรู้และกระบวนการทดสอบ โดยทั้งสองกระบวนการจะมีขั้นตอนการทำงาน คือ จะนำเอาเอกสารแต่ละกลุ่มมาทำการตัดคำและการสกัดคำหลัก (การตัดคำจากช่องว่าง การเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก การกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป การตัดคำเพื่อหารากศัพท์) หลังจากได้คำทั้งหมดในเอกสารแล้วจึงนำคำเหล่านั้นมาหาความถี่และนำหนักจากเอกสารทั้งหมด เพื่อนำไปสร้างตัวแทนเวกเตอร์ของเอกสาร แล้วนำข้อมูลที่ผ่านกระบวนการดังกล่าวเข้าสู่การสร้างการจัดกลุ่มเอกสารด้วยอัลกอริทึม ซอฟต์แวร์คอมพิวเตอร์แมชชีน โดยอาศัยโปรแกรมอาร์ (R Language) แพกเกจที่ใช้ทดลองคือ แพกเกจทีเอ็มและแพกเกจอี1071 ในการประมวลผล เมื่อโมเดลการจัดกลุ่มเอกสารด้วยอัลกอริทึม ซอฟต์แวร์คอมพิวเตอร์แมชชีนสร้างสำเร็จ จะมีการวัดประสิทธิภาพของระบบโดยใช้วิธีการวัดค่าความระลึก (Precision) ค่าความแม่นยำ (Recall)

และค่า ประสิทธิภาพของการจัดกลุ่มเอกสาร (F-measure) โดยค่าความแม่นยำ เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนทั้งหมดที่ทำการค้นหาได้ ส่วนค่าความระลึกลับ เป็นอัตราส่วนของการค้นพบเอกสารที่ถูกต้องจากจำนวนเอกสารที่ถูกต้องทั้งหมด ค่าประสิทธิภาพของการจัดกลุ่มเอกสารเป็นค่าความสัมพันธ์ระหว่าง ค่าความแม่นยำและค่าความระลึกลับ ในเชิงฮาร์โมนิก ซึ่งผลการทดสอบสามารถจำแนกได้ 3 กลุ่มคือ 1.กลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารที่ดีที่สุดที่สุดคือกลุ่มข่าวกีฬา 2.กลุ่มวัดประสิทธิภาพการจัดกลุ่มเอกสารที่ได้ปานกลางคือกลุ่มข่าว สุขภาพ และกลุ่มข่าวท่องเที่ยว 3.กลุ่มวัดประสิทธิภาพการจัดกลุ่มที่ได้ต่ำคือกลุ่มข่าว ธุรกิจและกลุ่มข่าว คอมพิวเตอร์

ดังนั้นจากผลการทดสอบระบบมีความน่าเชื่อถือในระบบอยู่ในระดับที่ดี ทั้งนี้ความน่าเชื่อถือของระบบนั้นขึ้นอยู่กับเอกสารและคำหลักที่นำมาใช้ในการเรียนรู้เพื่อสร้างแบบของเอกสาร ซึ่งถ้าเอกสารที่นำมาใช้ในการเรียนรู้เพื่อสร้างรูปแบบเอกสารนั้นมีความชัดเจนในแต่ละด้าน คือมีการใช้คำที่สื่อความหมายในการระบุกลุ่มเอกสารจึงจะทำให้ระบบมีความถูกต้องและความน่าเชื่อถือมากขึ้นและเนื่องจากคำในภาษาอังกฤษมีความหลากหลายในการใช้คำจึงไม่สามารถที่จะระบุประเภทของกลุ่มเอกสาร ได้ชัดเจน และนอกจากนั้นในการเรียนรู้ การที่ใช้จำนวนเอกสารเป็นจำนวนมากเท่าใดก็จะยิ่งทำให้การสร้างโมเดลของการจัดกลุ่มเอกสารมีขอบเขตในการจัดกลุ่มอย่างชัดเจน นั่นหมายความว่าโอกาสที่ความถูกต้องในการจัดกลุ่มเอกสารก็จะยิ่งมากขึ้นตามไปด้วย

6.2 ข้อเสนอแนะสำหรับการค้นคว้าแบบอิสระ

งานค้นคว้าแบบอิสระนี้ ได้ใช้เว็บไซต์ไปเดอร์ในการเก็บหน้าแรกของเอกสารเว็บเพจ ซึ่งในปัจจุบันในหน้าเว็บเพจมีการใช้ การเปลี่ยนทิศทางใหม่ของ ยูอาร์แอล (URL Redirect) มากยิ่งขึ้น ทำให้ขั้นตอนการเตรียมเอกสารในหัวข้อ 3.2 ในบทที่ 3 ได้เอกสารที่มีรายละเอียดของเอกสารน้อย และเนื้อหาของเอกสารเว็บเพจที่มีการเปลี่ยนแปลงได้ตลอดเวลา (Dynamic Content) มีผลทำให้การจัดกลุ่มของเอกสารมีความไม่ชัดเจนตามกลุ่มเอกสาร

ส่วนในขั้นตอนการเรียนรู้ของระบบ หรือขั้นตอนของการสร้างโมเดลการจัดกลุ่มเอกสาร ถ้าเอกสารที่นำมาใช้ในการเรียนรู้มีจำนวนน้อย จะทำให้ได้คำที่นำมาสร้าง โมเดลมีจำนวนน้อย ผลที่ตามมาคือทำให้การจัดกลุ่มเอกสารมีประสิทธิภาพที่ไม่ดี นอกจากนี้ยังมีการพบคำที่มีเป็นจำนวนมากแต่คำนั้นๆ ไม่สัมพันธ์ต่อการระบุกลุ่มของเอกสาร ก็จะทำให้ระบบมีประสิทธิภาพลดลง เช่นเดียวกัน

6.3 แนวทางการพัฒนาต่อในอนาคต

งานค้นคว้าแบบอิสระนี้ยังมีประเด็นที่จะนำไปพัฒนาต่อเพิ่มเติมเพื่อให้การจัดกลุ่มของเอกสารมีประสิทธิภาพดีขึ้นโดยพัฒนาให้สามารถจัดกลุ่มเอกสารข่าวอื่นๆ ได้ เช่น เอกสารข่าวการศึกษา เอกสารข่าวเกษตรกรรม เป็นต้น ส่วนของเว็บสไปเดอร์สามารถพัฒนาให้เว็บสไปเดอร์ครอว์เลอร์ ยูอาร์แอลที่มีการเปลี่ยนทิศทางใหม่ของ ยูอาร์แอล (URL Redirect) ได้ ซึ่งจะทำให้เว็บสไปเดอร์เพิ่มประสิทธิภาพดียิ่งขึ้น

The logo of Chiang Mai University is a circular emblem. In the center is a stylized elephant standing and facing left. Above the elephant's head is a traditional Thai oil lamp (diya) with a flame, emitting rays of light. The entire emblem is surrounded by a circular border containing the university's name in Thai script at the top and 'CHIANG MAI UNIVERSITY 1964' in English at the bottom. There are decorative floral motifs on either side of the elephant.

ลิขสิทธิ์มหาวิทยาลัยเชียงใหม่
Copyright© by Chiang Mai University
All rights reserved