บทที่ 3

การทำดัชนีเอกสาร

บทนี้กล่าวถึงการสร้างดัชนีเอกสารเพื่อใช้เป็นตัวแทนเอกสารในคำนวณ โดยมีเนื้อหาประกอบด้วย หลักการการทำดัชนีเอกสาร การทดลอง ผลการทดลอง และวิจารณ์และสรุปผลการสร้างดัชนีเอกสาร

3.1 หลักการการทำดัชนีเอกสาร

ในงานทางด้านการค้นคืนเอกสารและงานทางด้านการจำแนกเอกสาร การทำดัชนีเอกสารเป็นกระบวนการตั้งต้นที่มีความสำคัญ เพราะเอกสารในรูปแบบของข้อความ (Text) นั้นไม่สามารถคำนวณได้ด้วยวิธีการทางคณิตศาสตร์ ดังนั้นจึงต้องมีวิธีการแปลงเอกสารให้อยู่ในรูปแบบที่สามารถนำไปคำนวณได้ด้วยวิธีการสร้างตัวแทนเอกสาร โดยใช้วิธีการสร้างดัชนีของเอกสารในรูปแบบเวกเตอร์คำสำคัญ ซึ่งได้อธิบายในบทที่ 2 หัวข้อ 2.2

3.2 การทดลองทำดัชนีเอกสาร

งานวิจัยนี้ใช้ข้อมูลสำหรับนำมาทดลอง จาก BMC (Bio Med Central, 2008) ซึ่งเป็นแหล่งรวบรวมงานวิจัยทางด้านชีววิทยาและทางการแพทย์ โดยเลือกเฉพาะบทคัดย่อ (Abstract) ของเอกสารงานวิจัยในกลุ่ม BMC จากนั้นทำการเลือกกลุ่มย่อยที่มีจำนวนเอกสารมากกว่า 200 เอกสาร จำนวน 7 กลุ่มประกอบด้วย Biotechnology, Cancer, Evolutionary Biology, Family Practice, Infection Diseases, Medical Genetics และ Musculoskeletal Disorders โดยแต่ละกลุ่มเอกสารจะถูกแบ่งออกเป็นชุดข้อมูลสำหรับเรียนรู้และเป็นข้อมูลสำหรับทดสอบ ในจำนวนที่เท่าๆกัน โดยมีเอกสารสำหรับใช้ในการเรียนรู้จำนวน 1,614 เอกสาร และเอกสารสำหรับทดสอบจำนวน 1,614 เอกสาร เพื่อแปลงเอกสารสำหรับเรียนรู้ให้อยู่ในรูปแบบที่สามารถนำไปใช้ในการคำนวณได้ มีขั้นตอนดังต่อไปนี้

1) ทำการกำจัดเครื่องหมายต่างๆ และเปลี่ยนอักษรตัวพิมพ์ใหญ่ให้เป็นตัวพิมพ์เล็ก
2) ทำการกำจัดคำศัพท์ที่ไม่มีผลกระทบต่อความหมายโดยทั่วไป (Remove Stopword)

ตัวอย่างการกำจัดคำเช่น ก่อนทำการกำจัดคำ “With the technique presented it is possible to produce, within a short timeframe, pure P34, suitable for further studies where an

example antigen is needed" ผลที่ได้จากการกำจัดคำ คือ "technique presented possible produce, short timeframe, pure P34, suitable studies example antigen needed"

3) ตัดคำเพื่อหารากศัพท์ (Word stemming) โดยใช้พอร์ทเตอร์อัลกอริทึม (Porter algorithms)

การตัดคำเพื่อหารากศัพท์โดยใช้พอร์ทเตอร์อัลกอริทึม คือกระบวนการลดรูปคำสำคัญที่แตกต่างกันให้อยู่ในรูปของรากศัพท์ แต่การทำงานของพอร์ทเตอร์อัลกอริทึมไม่ได้พยายามหารากศัพท์ที่ถูกต้องของคำสำคัญนั้นๆ เพียงแต่พยายามสร้างรูปแบบของคำสำคัญให้มีลักษณะทั่วไปที่เหมือนกัน โดยการตัดส่วนที่ไม่เหมือนกันออกไป ตัวอย่างเช่น

CONNECT

CONNECTED

CONNECTING

CONNECTION

CONNECTIONS

จากตัวอย่างของกลุ่มคำสำคัญนี้ เมื่อทำการตัดคำด้วยพอร์ทเตอร์อัลกอริทึม ผลที่ได้คือ CONNECT ซึ่งเป็นลักษณะโดยทั่วไปของกลุ่มคำสำคัญกลุ่มนี้ (Porter, 1980)

4) ทำการสร้างคำสำคัญจากชุดสำหรับเรียนรู้

5) ทำการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร และเลือกคำสำคัญที่ปรากฏในเอกสารทั้งหมดมากกว่า 4 เอกสาร โดยการเลือกจำนวนคำสำคัญนี้ได้จากการทดลองจำแนกเอกสารชุดข้อมูลของ BMC เพื่อให้สามารถนำไปคำนวณในคอมพิวเตอร์ส่วนบุคคล และให้ประสิทธิภาพการจำแนกที่ดีในระดับที่ยอมรับได้

6) เมื่อทำการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสารของข้อมูลสำหรับเรียนรู้แล้ว ขั้นต่อไปคือการสร้างเมตริกซ์คำสำคัญ-เอกสารของข้อมูลสำหรับทดสอบ โดยทำตามกระบวนการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสารของข้อมูลสำหรับเรียนรู้ ในขั้นตอนที่ 1) ถึง 3) จากนั้นใช้คำสำคัญในขั้นตอนที่ 4) ของกระบวนการสร้างชุดข้อมูลสำหรับเรียนรู้ในการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสารของข้อมูลสำหรับทดสอบ

3.3 ผลการทดลองทำดัชนีเอกสาร

ผลที่ได้จากกระบวนการสร้างดัชนีเอกสาร แบ่งออกเป็นสองเมตริกซ์คือ เมตริกซ์ความถี่ของคำสำคัญ-เอกสารของข้อมูลสำหรับเรียนรู้ โดยมีขนาด 3,586×1,614 และ เมตริกซ์ความถี่ของ

คำสำคัญ-เอกสารของข้อมูลทดสอบ โดยมีขนาด 3,586×1,614 และ กลุ่มของคำสำคัญที่ได้จากกระบวนการการสร้างดัชนีของเอกสาร

3.4 วิจารณ์และสรุปทำดัชนีเอกสาร

ในการสร้างดัชนีเอกสารเพื่อให้สามารถใช้ในการคำนวณในแบบจำลองการจำแนกได้ ซึ่งจากการสร้างดัชนีเอกสารนี้อาจจะไม่ได้ทำให้การจำแนกเอกสารได้อย่างแม่นยำเนื่องจากอาจจะมีคำสำคัญบางคำที่ไม่ได้สร้างความแตกต่างระหว่างกลุ่มที่ชัดเจน และ เมื่อใช้การสร้างตัวแทนเอกสารที่สร้างด้วยวิธีการนี้อาจจะทำให้การใช้เวลาการเรียนรู้และการจำแนกที่มากเนื่องจากปริมาณข้อมูลเข้า ซึ่งจะมีปริมาณเท่ากับจำนวนคำสำคัญหรือจำนวนคอลัมน์ (Column) ของเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร