การวิเคราะห์ความคล้ายคลึงของเอกสารด้วยการประยุกต์ใช้เทคนิคการจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์

The analysis of document similarity with Dewey Decimal Classification Multiple Relation Technique

จุฬาลักษณ์ วัฒนานนท์
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี
E-mail: watthananon@hotmail.com

บทคัดย่อ

บทความนี้มีวัตถุประสงค์ เพื่อวิเคราะห์ความคล้ายคลึงของเอกสาร โดยใช้เทคนิค DDC-MR ซึ่งข้อมูลกลุ่มตัวอย่างที่นำมาใช้ทดสอบเป็นข้อมูลบทความที่ได้มาจากฐานข้อมูลเว็บไซต์ Science Direct จำนวน 70 บทความ และเลือกใช้ตัวแทนข้อมูลสำหรับวิเคราะห์ จำนวน 4 ส่วน คือ ชื่อเรื่อง บทคัดย่อ คำสำคัญ และอภิปรายผล ด้วยสัดส่วนตัวแทนข้อมูลที่เหมาะสมคือ 4:2:2:2 จากการทดลองพบว่า เทคนิคการจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์ สามารถอธิบายถึงรายละเอียดความคล้ายคลึงระหว่างบทความได้อย่างมีประสิทธิภาพ และเมื่อพิจารณาถึงการวัดประสิทธิภาพของระบบจะพิจารณาจาก Recall, Precision และ F-measure โดยผลการสกัดคำเดี่ยวคือ 94.98%, 98.46% และ 96.69% ตามลำดับ และผลของการสกัดความสัมพันธ์ คือ 88.33%, 94.51 และ 85.02% ตามลำดับ

คำสำคัญ: ความคล้ายคลึง, ตัวแทนข้อมูล, การค้นคืนสารสนเทศ, ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์

Abstract

The main purpose of the paper is to analyze document similarity within a dataset (including 70 articles in total) collected from Science Direct database using DDC-MR technique. The paper provided 4 types of representative data (i.e., Title, Abstract, Keyword and Conclusion) in order to determine the appropriate proportion 4:2:2:2. The experimental results showed that, our proposed technique had better effectiveness to describe and analyze document similarity. Accordingly, the performances of the system were illustrated in terms of Recall, Precision, F-measure scores. The scores of extraction were 94.98%, 98.69% and 96.69%, respectively and the scores of relation extraction were 88.33%, 94.51% and 85.02%, respectively.

1. บทนำ

การจัดการความรู้ (KM: Knowledge Management) ในปัจจุบัน ถือได้ว่าเป็นประเด็น

สำคัญ ก่อให้เกิดการพัฒนาโดยการประยุกต์ใช้เทคโนโลยีสารสนเทศ ข้อมูล และความรู้ เข้ามารวมกัน เพื่อลดความเลื่อมล้ำของการเข้าถึงข้อมูล ส่งเสริมให้เกิดการพัฒนา และจัดเก็บเป็นองค์ความรู้ที่เป็นประโยชน์ต่อองค์กร และผู้ใช้เป็นจำนวนมาก จากรายงานการวิจัย (Oded, 2005; Sajjad, 2005) พบว่า ด้วยวิธีการจัดการและจัดเก็บสารสนเทศที่ทันสมัยและเป็นประโยชน์ต่อองค์กรนั้น มีการเพิ่มขึ้นของความรู้ตลอดเวลาอย่างต่อเนื่อง ทำให้ในปัจจุบันลักษณะของข้อมูลข่าวสาร มีปริมาณมากจนเกินไปและหลากหลายรูปแบบ ส่งผลกระทบให้เกิดปัญหาตามมามากมายหลายประการ อาทิ เกิดการหลงทางในความรู้ (Lost in space) เนื่องจากแม้จะมีคลังข้อมูลที่บรรจุความรู้ไว้มากมาย แต่ผู้ใช้ไม่สามารถเข้าถึงข้อมูล และไม่สามารถเข้าใจถึงความสัมพันธ์ของโครงสร้างความรู้ได้อย่างถ่องแท้ เป็นผลให้เกิดพฤติกรรมการสืบค้นข้อมูลในลักษณะการหลงทางของคลังความรู้ ที่ไม่สามารถย้อนกลับไป-มา ในสิ่งที่ผู้ใช้กำลังสนใจ หรือต้องการได้อย่างต่อเนื่อง นอกจากนี้ความนิยมในการใช้ Social Network และการทำธุรกิจหรือธุรกรรมออนไลน์ต่าง ๆ ทำให้มีข้อมูลเกิดขึ้นในระบบออนไลน์เหล่านี้เป็นจำนวนมาก และอยู่ในรูปแบบ Unstructured การจัดการกับข้อมูลจำนวนมากและเกิดขึ้นตลอดเวลาประเภทนี้ ไม่สามารถทำได้ด้วยวิธีการจัดเก็บไว้ใน Database รูปแบบเดิม ๆ ได้ดี หรือถ้าทำได้ก็ไม่สะดวกสบายนัก และการจะนำมาใช้ให้เกิดประโยชน์ก็ยากมากขึ้นตามไปด้วย

จะเห็นได้ว่าองค์กรต่าง ๆ ส่วนมากมักจะมีแต่การจัดเก็บสะสมไปเรื่อย ๆ แต่ไม่ได้มีการนำมาใช้งานใด ๆ สุดท้ายก็เป็นเหมือนขยะกองโตขององค์กรที่สิ้นเปลืองทรัพยากรในการเก็บรักษาหรือไม่ก็ถูกปล่อยให้สูญหายไปอย่างไร้ค่า ทั้งที่จริงแล้วถ้ามีการจัดเก็บและนำมาวิเคราะห์ให้ดี จะพบว่าข้อมูลเหล่านี้เปรียบเสมือนขุมทรัพย์ที่จะทำให้ธุรกิจเติบโตได้อย่างมาก เพราะอุดมไปด้วยข้อมูลที่หลากหลาย สามารถนำมาสร้างมูลค่าเพิ่มให้กับองค์กรได้ในหลาย ๆ มุมมอง รวมถึงสามารถนำข้อมูลที่ได้มาใช้เพื่อการป้องกันหรือปิดช่องโหว่เพื่อสร้างความได้เปรียบในการแข่งขันอีกด้วย หากระบบสามารถอธิบายความสัมพันธ์ หรือวิเคราะห์ความคล้ายคลึงของข้อมูลสารสนเทศที่จัดเก็บเหล่านั้นได้ ซึ่งจากการศึกษาและค้นคว้า ทำให้มีนักวิจัย Feng และคณะ (Feng et al, 1998) ได้นำเสนอวิธีสร้างความสัมพันธ์ด้วยการใช้กฎความสัมพันธ์ของข้อมูล (Association Rules) เพื่อพยากรณ์หาแนวโน้มของความสัมพันธ์ และในปี 2007 ได้มีการศึกษางานวิจัยเกี่ยวกับการใช้หัวข้อ (Topic Maps) (Redman et al. 2007; Kawtrakut et al., 2007) ของสารสนเทศเป็นตัวเชื่อมความสัมพันธ์ในรูปแบบของคำสำคัญที่คล้ายกัน (Keyword Matching) หรือชื่อเรื่องที่คล้ายกัน (Title Matching) ซึ่งสามารถอ้างถึงสารสนเทศที่เกี่ยวข้องกันได้ แต่วิธีการนี้เนื้อหาภายในของสารสนเทศไม่ได้ถูกนำมาวิเคราะห์ความสัมพันธ์และความหมายที่แอบแฝง

ดังนั้น บทความนี้มุ่งนำเสนอ เทคนิคการวิเคราะห์ความคล้ายคลึงเนื้อหาระหว่างเอกสาร เพื่อแก้ปัญหาจากข้อจำกัดที่กล่าวข้างต้น ด้วยการใช้กรอบมาตรฐานการจัดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์ (Lertmahakiat et al., 2008) และการ

จำแนกเนื้อหาสารสนเทศในระดับลึก (จุฬาลักษณ์และอนิราช, 2552) ซึ่งครอบคลุมถึงคุณลักษณะของคำ (Keywords) ด้วยการวิเคราะห์เนื้อหาสารสนเทศที่อยู่ในองค์กรอย่างเป็นระบบ และเชื่อมโยงความสัมพันธ์ด้วยการแจกแจงสัดส่วนความสัมพันธ์ระหว่างสารสนเทศ ซึ่งจะช่วยให้ทุกคนในองค์กรสามารถจำแนกเนื้อหาของสารสนเทศ เข้าใจถึงสิ่งที่คล้ายคลึงกันอย่างมีมิติสัมพันธ์ และเข้าถึงสารสนเทศเหล่านั้นได้อย่างมีประสิทธิภาพ อันจะส่งผลต่อประสิทธิภาพในการค้นคืนสารสนเทศที่เพิ่มขึ้น

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การวิเคราะห์คำ

บทความนี้เลือกใช้กระบวนการวิเคราะห์คำด้วยการสกัดคำแบบ *N-grams* โดยทำการแยกคำทีละคำออกจากกัน จากนั้นตัดคำที่ไม่มีผลต่อเนื้อหา เช่น a, an, the, to, and, of เป็นต้น และทำการประมาณค่าความน่าจะเป็นของชุดคำที่เกิดขึ้นร่วมกันว่ามีค่าเท่ากับผลคูณของความน่าจะเป็น ที่จะพบหน่วยคำ (Term) ตามค่า N เพื่อใช้แทนการตัดคำ โดยบทความนี้ใช้ความยาวของชุดอักขระ (อัษฎางค์และชุรีรัตน์, 2547) และคำที่เขียนเรียงกันที่มีความแตกต่างกัน (อัครพล, 2548) ตั้งแต่ *1-Gram, 2-Gram, 3-Gram, 4-Gram* ฯลฯ ซึ่งทำให้ลดเวลาในการค้นหาคำภายในเอกสาร และคำภายในพจนานุกรมได้รวดเร็วขึ้นเมื่อทำการเปรียบเทียบคำ 2 คำ และกลุ่มคำ 3 คำ ตามลำดับ สามารถทำการประมาณค่าด้วยสมการ 1 ดังนี้

$$P(w_1 w_2 K w_T) = \prod_{i=1}^{T} P(w_i \mid w_1 K w_{i-1}) \qquad (1)$$

โดยที่ w คือ คำ, n คือ จำนวนนับต่อไป, P คือ ค่าความน่าจะเป็น (Probability) ซึ่งประมาณได้จากคลังข้อมูล, T คือ จำนวนของคำ, i คือ ลำดับของคำโดยเริ่มต้นจากลำดับที่ 1 และ $(w_1, w_2, w_3, ..., w_n)$ คือ ชุดคำที่ประกอบด้วยคำที่มากกว่า 3 คำขึ้นไป

จากสมการที่ 1 จะเห็นได้ว่า $P(w_i \mid w_1, ..., w_{i-1})$ คือ ความน่าจะเป็นของคำ w_i หลังจากเกิดคำ $w_1, w_2, ..., w_{i-1}$ ก่อนหน้านี้ ดังนั้น ความน่าจะเป็นของประโยคโดยใช้วิธี *2-Gram* คือ $P(w_1, w_2, w_3, ..., w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)$, $..., P(w_n \mid w_{n-1})$ และความน่าจะเป็นของประโยคโดยใช้วิธี คือ *3-Gram* คือ $P(w_1, w_2, w_3, ..., w_n) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2), ..., P(w_n \mid w_{n-2} w_{n-1})$ ดังนั้น หากประมาณค่าความน่าจะเป็นของประโยคโดยใช้วิธี *2-Gram* จะสามารถปรับเปลี่ยนดังสมการ 2 ดังนี้

$$P(w_1 w_2 K w_T) = P(w_1 \mid < s >)P(w_2 \mid w_1) K P(w_i \mid w_{i-1}) \qquad (2)$$

โดยที่ $P(w_1 \mid ...s...)$ คือ ความน่าจะเป็นของคำที่หนึ่ง เมื่อเกิดเป็นคำแรกของประโยค ซึ่ง ในที่นี้คือ ช่องว่าง

$P(w_2 \mid w_1)$ คือ ความน่าจะเป็นของคำ w_2 หลังจากเกิดคำ w_1

$P(w_i \mid w_{i-1})$ คือ ความน่าจะเป็นของคำ w_i หลังจากเกิดคำ w_{i-1}

2.2 การคำนวณค่าความคล้ายคลึง

สำหรับการวิเคราะห์ความคล้ายคลึงระหว่างเอกสารได้มีการเก็บรวบรวมข้อมูลในรูปแบบของตาราง (Matrix) ซึ่งประกอบด้วยจำนวนความถี่คำ *n* คำ และจำนวนเอกสาร *m* เอกสาร โดยข้อมูลนี้จะถูกเก็บไว้ในรูปแบบของตาราง ดังนั้น การคำนวณหาค่าความคล้ายคลึงระหว่างเอกสาร 2 เอกสาร นิยมใช้มี 2 วิธี (Breese et al., 1998; Terveen and Hill, 2001; Sarwar et al., 2001) คือ Correlation-based และ Cosine-based ซึ่งในงานวิจัยของ Herlocker และคณะ (1999) ได้นำค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน (Pearson Correlation Coefficient: PCC) มาใช้ในการคำนวณค่าความคล้ายคลึงระหว่างเอกสาร โดยมีวิธีการคำนวณดังสมการ 3 ดังนี้

$$C_{a,u} = \frac{\text{cov}(r_a, r_u)}{\sigma_{r_a}\sigma_{r_u}} \quad (3)$$

โดยที่ r_a และ r_u คือ คะแนนของรายการข้อมูลที่ได้จากชุดข้อมูล a และชุดข้อมูล u

Covariance:

$$\text{cov}(r_a, r_u) = \frac{\sum_{i-1}^{m}(r_{a,i} - \overline{r_a})(r_{u,i} - \overline{r_u})}{m} \quad (4)$$

$$\overline{r_u} = \frac{\sum_{i=1}^{m} r_{u,i}}{m} \quad (5)$$

โดยที่ $r_{a,i}$ และ $r_{u,i}$ คือ ค่าคะแนนของรายการข้อมูล i ได้จากชุดข้อมูล a และชุดข้อมูล u

$\overline{r_u}$ คือ ค่าเฉลี่ยคะแนนของรายการข้อมูลของชุดข้อมูล u

m คือ จำนวน Co-Rated Items

Standard Deviation:

$$\sigma_{r_u} = \sqrt{\frac{\sum_{i=1}^{m}(r_{u,i} - \overline{r_u})^2}{m}} \quad (6)$$

จากงานวิจัยของ Herlocker และคณะ (1999) ได้กล่าวไว้ว่า การใช้ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันเพียงอย่างเดียว ไม่เพียงพอในการคำนวณหาความคล้ายคลึง ดังนั้น Herlocker และคณะ (1999) ได้ทำการคำนวณค่าน้ำหนัก (weight) เพื่อใช้ในการคำนวณหาค่าน้ำหนักของเอกสารที่มีจำนวนคำและความถี่เหมือนกัน โดยมีการคำนวณดังสมการที่ 7 และ 8 ดังนี้

$$S_{a,u} = \begin{cases} 1 & if \quad m > 50 \\ \dfrac{m}{50} & if \quad m \le 50 \end{cases} \quad (7)$$

โดยที่ m คือ จำนวน Co-Rated Items

$$sim(a,u) = S_{a,u}C_{a,u} \quad (8)$$

โดยที่ $S_{a,u}$ คือ ค่าน้ำหนัก

$C_{a,u}$ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ของเพียร์สัน ระหว่างชุดข้อมูล a และชุดข้อมูล u

2.3 การจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์

การจำแนกหมวดหมู่เอกสารในงานวิจัยโดยทั่วไปนิยมนำ Metadata มาเป็นตัวแทนข้อมูลของเอกสาร (Haung, 2007) ส่วนงานวิจัยของ Peery และคณะ (Peery et al., 2008) มีการใช้ตัวแทนข้อมูลเพิ่มมากขึ้นเพื่อเพิ่มประสิทธิภาพด้านความแม่นยำ แต่ส่งผลกระทบต่อเวลาในการประมวลผลที่เพิ่มมากขึ้น ซึ่งในบทความนี้มุ่งเน้นนำเสนอเนื้อหาในระดับลึกและครอบคลุมถึงเนื้อหาของหนังสือทั้งหมด จึงได้นำตัวแทนข้อมูล 4 ชนิด คือ ข้อมูลส่วนชื่อเรื่อง ข้อมูลส่วนบทคัดย่อ ข้อมูลส่วนคำสำคัญ และข้อมูลส่วน

อภิปรายผลมาเป็นตัวแทนข้อมูล และปรับเปลี่ยนเทคนิคการวิเคราะห์หมวดหมู่หนังสือจากหมวดหมู่เดียว (Only class collection) เป็นการวิเคราะห์หมวดหมู่หนังสือแบบสหความสัมพันธ์ (Multiple classes collection) โดยจัดเก็บเนื้อหาสารสนเทศในเอกสารทุกเรื่องแบบสหความสัมพันธ์ภายใต้กรอบหมวดหมู่ระบบทศนิยมดิวอี้ เพื่อแสดงให้เห็นว่าเอกสารแต่ละฉบับมีเนื้อหาเกี่ยวข้องกับหมวดหมู่ใดและมีรูปแบบความสัมพันธ์อย่างไรและทำให้สามารถแสดงสารสนเทศที่เคยมองไม่เห็น(Representation unseen information) ได้อย่างชัดเจนขึ้นจากการคำนวณน้ำหนักในแต่ละหมวดหมู่ (Calculate Weight of Each Class) ดังสมการ (9)

$$weight = \frac{c_i \times 100}{\sum_{i=1}^{n} c_i} \qquad (9)$$

โดยที่ c คือ หมวดหมู่ที่ปรากฏค่าความถี่, i คือ ลำดับของหมวดหมู่ และ n คือ จำนวนหมวดหมู่ในลำดับชั้นที่ 3

3. วิธีดำเนินงานวิจัย

งานวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อพัฒนาระบบวิเคราะห์การค้นค้นสารสนเทศที่มีความคล้ายคลึงกันด้วยการใช้กรอบการจัดหมู่ทศนิยมดิวอี้แบบสหความสัมพันธ์ โดยมีรายละเอียด ดังนี้

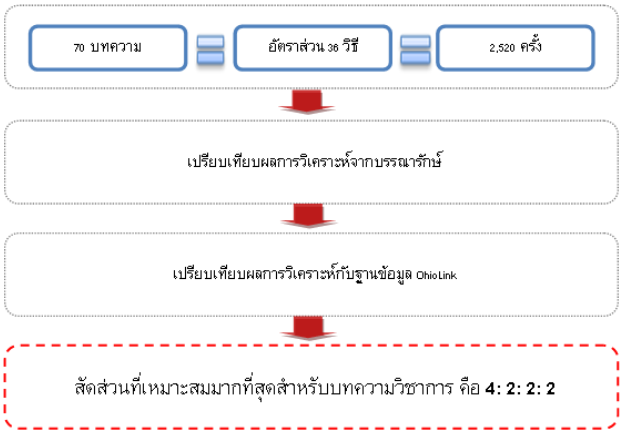
3.1 การคัดเลือกตัวแทนข้อมูล

การคัดเลือกตัวแทนข้อมูลสำหรับบทความนี้เป็นการคัดเลือกจากบทความวิชาการซึ่งมีองค์ประกอบหลักที่ใช้ในการนำเสนอ 10 ส่วน ได้แก่ ชื่อเรื่อง (Title) บทคัดย่อ (Abstract) คำสำคัญ (Keyword) วัน/เดือน/ปีที่เผยแพร่ (Date) บทนำ (Introduction) ทฤษฎีและงานวิจัยที่เกี่ยวข้อง (Related Works) วิธีการวิจัย (Method) ผลการวิจัย (Experimental Result) อภิปรายผล (Conclusion) และเอกสารอ้างอิง (References) โดยบทความนี้ได้ทำการคัดเลือกจำนวน 4 ส่วน คือ ชื่อเรื่อง บทคัดย่อ คำสำคัญ และอภิปรายผล จากทั้งหมด 10 ส่วน เพื่อเป็นตัวแทนของบทความวิชาการที่ใช้ในการทดลอง เนื่องจากผลการวิจัยที่ผ่านมาสรุปได้ว่า การคัดเลือกตัวแทนสำหรับการค้นคืนสารสนเทศสามารถใช้ตัวแทนข้อมูลเพียงบางส่วนได้ โดยไม่จำเป็นต้องเลือกใช้จากข้อมูลทุกส่วน (Peery et al., 2008; Fuller et al., 2008)

3.2 การคัดเลือกสัดส่วนที่เหมาะสมของตัวแทน

การคัดเลือกสัดส่วนที่เหมาะสมของตัวแทนเป็นการเปรียบเทียบหาสัดส่วนที่เหมาะสม เพื่อใช้กำหนดน้ำหนักให้กับแต่ละส่วนของตัวแทนบทความ โดยการวิเคราะห์เนื้อหาจากตัวแทนของบทความจำนวน 4 ส่วน คือ ชื่อเรื่อง (Title = T) บทคัดย่อ (Abstract = A) คำสำคัญ (Keyword = K) และอภิปรายผล (Conclusion = C) โดยมีรายละเอียดของขั้นตอนการคัดเลือกสัดส่วนที่เหมาะสมของตัวแทนบทความ ดังแสดงในรูปที่ 1

จากรูปที่ 1 แสดงขั้นตอนการคัดเลือกสัดส่วนที่เหมาะสมของตัวแทนบทความทั้งหมดที่ใช้ในการทดสอบ (70 บทความ) โดยขั้นตอนแรกทำการคำนวณอัตราส่วนวิธี (36 วิธี) (จุฬาลักษณ์, 2553) ที่สามารถปรากฏเหตุการณ์ได้ผลที่ได้ออกมาเท่ากับ



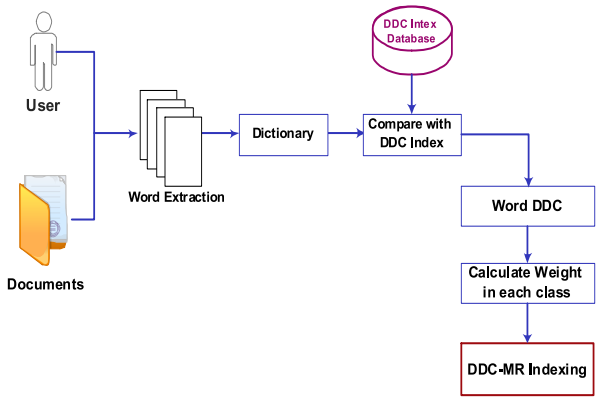
รูปที่ 1 ขั้นตอนการคัดเลือกสัดส่วนที่เหมาะสมของตัวแทน

2,520 ครั้ง จากนั้นทำการเปรียบเทียบผลการวิเคราะห์จากบรรณารักษ์ ที่มีประสบการณ์เกี่ยวกับการจัดหมู่ระบบทศนิยมดิวอี้ไม่ต่ำกว่า 10 ปี และเปรียบเทียบผลการวิเคราะห์กับฐานข้อมูล OhioLink ผลลัพธ์ที่ได้คือ 4:2:2:2 เป็นสัดส่วนที่เหมาะสมมากที่สุดสำหรับการกำหนดค่าน้ำหนักของตัวแทนข้อมูล ซึ่งสอดคล้องกับวิธีการของบรรณารักษ์ที่ให้ความสำคัญกับชื่อเรื่องเป็นลำดับแรก สำหรับการจัดหมวดหมู่สารสนเทศภายในห้องสมุด (วิไลพร, 2553)

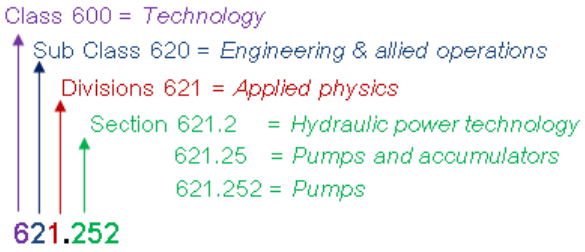
3.3 การวิเคราะห์และจัดเก็บด้วย DDC-MR

การวิเคราะห์ความคล้ายคลึงของเอกสารด้วยการประยุกต์ใช้เทคนิค DDC-MR นั้น ได้ทำการหาสัดส่วนสหความสัมพันธ์ของเนื้อหา เพื่อแสดงให้เห็นถึงกรอบเนื้อหาและความคล้ายคลึงของเอกสารได้อย่างชัดเจนว่ามีความสัมพันธ์กับหมวดใด หมู่ใดในสัดส่วนเท่าใด ซึ่งสามารถแสดงขั้นตอนการวิเคราะห์ ดังแสดงในรูปที่ 2

จากรูปที่ 2 แสดงขั้นตอนการวิเคราะห์หมวดหมู่ DDC-MR ที่เป็นลำดับขั้น โดยการวิเคราะห์บทความเริ่มต้นด้วยการนำเข้าข้อมูลที่ได้จากการคัดเลือก 4 ส่วน คือ ชื่อเรื่อง บทคัดย่อ คำสำคัญ และอภิปรายผล จากนั้นระบบจะทำการวิเคราะห์คำด้วยการสกัดคำแยกเป็นคำ และกลุ่มคำ จากนั้นทำการเปรียบเทียบคำสำคัญเพื่อค้นหาหมวดหมู่ DDC พร้อมค่าความถี่ในแต่ละหมวด ซึ่งเป็นขั้นตอนเริ่มต้นในการคำสำคัญที่ได้นำเข้าข้อมูลไปเปรียบเทียบกับคลังข้อมูลที่ได้จากการจัดหมวดหมู่หนังสือระบบทศนิยมดิวอี้ โดยทำการจำแนกคำและวิเคราะห์คำตามการแบ่ง 3 หมวดใหญ่ โดยการวิเคราะห์ครั้งที่ 1 (First Summary) เป็นการวิเคราะห์คำตามหมวดใหญ่จำนวน 10 หมวด การวิเคราะห์ครั้งที่ 2 (Second Summary) เป็นการวิเคราะห์คำจากหมวดใหญ่แต่ละหมวดออกเป็นหมวดย่อย 10 หมวด และการวิเคราะห์ครั้งที่ 3 (Third Summary) เป็นการวิเคราะห์คำในหมวดย่อยแต่ละหมวดออกเป็นหมู่ย่อย 10 หมู่ย่อย นอกจากนี้ระบบยังสามารถวิเคราะห์คำในระดับเฉพาะ หรือเจาะจงในระดับลึกที่มีการกระจายตัวเลขหลังจุดทศนิยมต่อจากเลข 3 หลักข้างต้น ในระดับที่ 4 และ 5 เพื่อจัดเก็บไว้ในฐานข้อมูล แสดงดังรูปที่ 3



รูปที่ 2 ขั้นตอนการวิเคราะห์ DDC-MR



รูปที่ 3 การจัดหมวดหมู่หนังสือระบบทศนิยมดิวอี้

จากรูปที่ 3 แสดงการจัดหมวดหมู่ด้วยการกำหนดสัญลักษณ์ที่มีความเชื่อมโยงเป็นลำดับชั้นที่ได้จากการคลังข้อมูล คือ หมายเลขหมวดหมู่ 621.252 หมายถึง หนังสือที่เกี่ยวกับปั๊ม (Pumps) ซึ่งจะเห็นได้ว่ามีความสัมพันธ์เป็นลำดับชั้นกับศาสตร์ในด้านการประยุกต์ใช้เทคโนโลยีกับงานทางวิศวกรรม

จากนั้นนำข้อมูลที่ได้ จากการจำแนกทุกคำของเอกสารไปเปรียบเทียบ และค้นหาคำในแต่ละหมวดหมู่ DDC มีคำเหล่านั้นปรากฏอยู่หรือไม่ และมากน้อยเพียงใด โดยทำการนับและคำนวณเป็นค่าความถี่ของแต่ละคำที่ปรากฏ ที่ถูกจัดเก็บในรูปแบบสัดส่วนสหความสัมพันธ์ของบทความ (Document DDC-MR) เป็นค่าร้อยละของแต่ละสัญลักษณ์ในลำดับชั้นที่ 3 ซึ่งมีทั้งหมด 1000 สัญลักษณ์ เพื่อใช้สำหรับการเปรียบเทียบความคล้ายคลึงระหว่างบทความ ส่วนค่าร้อยละของแต่ละสัญลักษณ์ในลำดับชั้นที่ 2 และ 1 ซึ่งเกิดจากการรวมค่าสัดส่วนในลำดับชั้นที่ 3 จะถูกจัดเก็บไว้ในฐานข้อมูล เพื่อใช้ประกอบการแสดงผลความสัมพันธ์ในเชิงลึกต่อไป ดังแสดงการวิเคราะห์หมวดหมู่จากความสำคัญในผลการวิเคราะห์ความถี่ของแต่ละหมวดหมู่คำ ในตารางที่ 1

ขั้นตอนการนับค่าความถี่ของหมวดหมู่ DDC ตามความสัมพันธ์ของเนื้อหาจากบทความ 4 ส่วน ด้วยขั้นตอนนี้จะทำให้ได้ค่าความถี่ของคำที่ปรากฏออกมา โดยนำค่าความถี่มาคำนวณค่าน้ำหนักของสี่ส่วนก่อน ว่าข้อมูลนั้นมาจากส่วนไหน ซึ่งมีการกำหนดค่าน้ำหนักที่แตกต่างกัน ดังนั้น ถ้าคำที่ปรากฏมาจากส่วนของชื่อเรื่องต้องนำมาคำนวณด้วย 0.3 ถ้ามาจากบทคัดย่อ คำสำคัญ และอภิปรายผลต้องนำมาคำนวณด้วย 0.2 ตามลำดับ เพื่อกำหนดค่าน้ำหนักตามสัดส่วนของข้อมูล

ขั้นตอนการรวมสัดส่วนความสัมพันธ์ของทุกคำ เพื่อหาสัดส่วนความสัมพันธ์ของเนื้อหาจากแต่ละบทความ ดังนั้น จากขั้นตอนที่ผ่านมาเมื่อได้ค่าตัวเลขทั้งหมดแล้ว ก็จะทำการรวบรวมสัดส่วนความสัมพันธ์ของทุกหมวดหมู่ ว่าแต่ละหมวดหมู่มีเลขหมู่อะไรบ้าง ซึ่งหากเกิดกรณีที่เลขหมวดหมู่ซ้ำกัน จะต้องทำการรวมค่าน้ำหนักนั้นเข้าไว้ด้วยกัน เนื่องจากอยู่ในเลขหมวดเดียวกัน นั่นหมายความว่ามีเนื้อหาเดียวกัน

ขั้นตอนสุดท้ายทำการคำนวณน้ำหนัก โดยนำค่าความถี่ทั้งหมดในแต่ละหมวดหมู่มาปรับเป็นค่าสัดส่วนความสัมพันธ์ เพื่อใช้เป็นดัชนีสำหรับสร้างความสัมพันธ์

4. การทดลองและการอภิปรายผล

4.1 ข้อมูลที่ใช้ในการทดลอง

- ข้อมูลที่ใช้ในการทดลองเป็นข้อมูลบทความวิจัยฉบับภาษาอังกฤษด้วยการสืบค้นสารสนเทศจากฐานข้อมูลเว็บไซต์ Science Direct จำนวนทั้งหมด 70 บทความ ซึ่งได้รับการตีพิมพ์และเผยแพร่ระหว่างปี 2010 – 2014 โดยแบ่งออกเป็น 7 หัวข้อเรื่อง ได้แก่

Art and Humanities จำนวน 10 บทความ, Economics, Econometrics and Finance จำนวน 10 บทความ, Social Science จำนวน 10 บทความ, Chemistry จำนวน 10 บทความ, Computer Science จำนวน 10 บทความ, Psychology จำนวน 10 บทความ และ Mathematics จำนวน 10 บทความ

- สำหรับบทความนี้ ผู้วิจัยได้แบ่งข้อมูลออกเป็น 2 ส่วน คือ ส่วนแรกจำนวน 90% ใช้ในการ Training (ฝึกฝนระบบ) และส่วนที่สองจำนวน 10% ใช้ในการ Test (ทดสอบจริง)

- ตัวแทนของข้อมูลเพื่อใช้ในการวิเคราะห์ข้อมูลชุดนี้เลือกใช้ 4 ส่วน คือ ชื่อเรื่อง (Title) บทคัดย่อ (Abstract) คำสำคัญ (Keyword) และอภิปรายผล (Conclusion)

4.2 การคำนวณประสิทธิภาพ

การทดลองครั้งนี้ เลือกใช้วิธีการวัด 3 วิธี ได้แก่ ค่าความแม่นยำ (Precision: P) ค่าความระลึก (Recall: R) และการวัดประสิทธิภาพโดยรวม (F1 measure) (Baeza and Ribeiro, 1999) ตามหลักการทดสอบประสิทธิภาพการค้นคืนสารสนเทศ ดังสมการ 10 และ 11 ดังนี้

$$\Pr ecision = \frac{A}{(A+M)} \qquad (10)$$

$$\mathrm{Re}\, call = \frac{A}{(A+N)} \qquad (11)$$

โดยที่ N คือ จำนวนบทความที่เกี่ยวข้องที่ไม่ถูกค้นคืน, A คือ จำนวนบทความที่เกี่ยวข้องและถูกค้นคืน และ M คือ จำนวนบทความที่ไม่เกี่ยวข้อง

การทดลองประสิทธิภาพของการค้นคืนสารสนเทศโดยพิจารณาเฉพาะค่าความแม่นยำ หรือค่าความระลึกเพียงอย่างเดียว อาจทำให้การประเมินประสิทธิภาพได้ไม่ถูกต้อง ทั้งนี้ เนื่องจากการค้นคืนแบบสหความสัมพันธ์นั้นคำหนึ่งคำสามารถเชื่อมโยงได้หลายหมวดหมู่ ซึ่งถ้าให้ค่าแม่นยำสูงอาจจะให้ค่าความระลึกต่ำได้ หรือการค้นคืนที่ให้ค่าความแม่นยำต่ำแต่ให้ค่าความระลึกสูง ดังนั้น จึงมีการนำวิธีการวัดประสิทธิภาพโดยรวม (F1 Measure) มาใช้ในการวัดเพิ่มเติม ดังสมการ 12 ดังนี้

$$F1 = \frac{2 \times P \times R}{P + R} \qquad (12)$$

4.3 การอภิปรายผล

จากการทดลองวิเคราะห์ความคล้ายคลึงของบทความหลากหลายหัวข้อเรื่อง จำนวน 70 บทความ สามารถอธิบายถึงรายละเอียดความคล้ายคลึงภายในระหว่างบทความด้วยวิธีการจำแนกเป็นหมู่หลัก หมู่รอง และหมู่ย่อย ตามมาตรฐานการจัดหมวดหมู่ระบบทศนิยมดิวอี้ ซึ่งบทความนี้ได้ให้ความสำคัญต่อการวิเคราะห์เนื้อหาของบทความแบบสหความสัมพันธ์ โดยกำหนดประเด็น 4 ข้อ คือ การวิเคราะห์คำในบทความ การวิเคราะห์สัดส่วนความสัมพันธ์ของแต่ละบทความ การวิเคราะห์ความคล้ายคลึงระหว่างบทความ และการวิเคราะห์ความคล้ายคลึงระหว่างบทความในระดับลึก ดังนี้

- *ประเด็น:* ผลการวิเคราะห์ความถี่คำที่ปรากฏในบทความ

ตารางที่ 1 ตัวอย่างผลการวิเคราะห์ความถี่ของแต่ละหมวดหมู่คำว่า “Technology”

Level 4	Freq	Level 3	Freq	Level 2	Freq	Level 1	Freq
215.7		215	2	210	2	200	6
261.56		261	1	260	1	300	13
291.6		291	3	290	3		
303.483		303	1				
…	…	…	…	…	…	…	…
604.7	2	604	2	600	2	600	77
613.26		613	2	610	2		
…	…	…	…	…	…	…	..
771.5		771	1	770	1	700	2

จากตาราง 1 แสดงผลจากการวิเคราะห์คำสำคัญแต่ละคำ ที่ได้จากการวิเคราะห์โดยระบบ ซึ่งหลังจากที่ได้ทำการทดสอบระบบ พบว่า คำว่า Technology มีความสัมพันธ์กับหลายหมวดหมู่ ในระดับชั้นที่ 4 พร้อมแสดงค่าความถี่ที่ปรากฏในหมู่ย่อย ความถี่ของหมู่ย่อยใดที่อยู่ในหมู่เดียวกันจะถูกรวมค่าความถี่ไว้ด้วยกัน ตัวอย่างเช่น ในระดับชั้นที่ 4 ค่าของ หมู่ย่อย 633.5 และ 633.72 มีค่าความถี่เป็น 1 จะถูกรวมกัน กลายเป็นค่าความถี่ 2 ในระดับชั้นที่ 3 ของหมู่ 633 และในระดับชั้นที่ 2 ค่าความถี่ ในหมวดย่อย 630 ยังคงมีค่าเป็น 2 เนื่องจากว่ายังไม่มีค่าเพิ่มขึ้น จนกระทั่งมาถึง ระดับชั้นที่ 1 หมวด 600 จะมีค่าความถี่เพิ่มขึ้นเป็น 6 เนื่องจากมีการรวมกันของค่าความถี่ในหมวดย่อย 630 และ 660

ตารางที่ 2 ผลการวัดประสิทธิภาพการสกัดคำ

	Recall (%)	Precision (%)	F-Measure (%)
Training	88.82	98.48	93.40
Test	94.98	98.46	96.69

จากตารางที่ 2 แสดงผลการวัดประสิทธิภาพของการสกัดคำจากบทความ โดยข้อมูลสำหรับ Training ระบบสามารถสกัดได้ถูกต้องจำนวน 1,359 คำ สกัดได้แต่ผิดจำนวน 21 คำ และสกัดคำไม่ได้ 171 คำ ในขณะที่ข้อมูลสำหรับ Test ระบบสามารถสกัดได้ถูกต้องจำนวน 833 คำ สกัดได้แต่ผิดจำนวน 13 คำ และสกัดไม่ได้ 44 คำ อธิบายได้ว่า วิธีการสกัดคำเมื่อนำไปตรวจสอบกับพจนานุกรมเพื่อนำไปวิเคราะห์ให้ผลลัพธ์ที่ดี ซึ่งสังเกตได้จากค่า Precision ที่ค่อนข้างสูงทั้งใน Training และ Test (98.48% และ 98.46% ตามลำดับ) ซึ่งหมายความว่า เมื่อระบบวิเคราะห์คำออกมาแล้วมีความผิดพลาดต่ำนั่นเอง

ตารางที่ 3 ผลการวัดประสิทธิภาพการสกัดความสัมพันธ์

	Recall (%)	Precision (%)	F-Measure (%)
Training	81.33	89.05	91.25
Test	88.21	94.51	85.02

จากตารางที่ 3 แสดงผลการวัดประสิทธิภาพของการสกัดความสัมพันธ์ของคำจากบทความ เนื่องจากหนึ่งคำสามารถปรากฏได้มากกว่าหนึ่งหมวด ดังนั้น ข้อมูลสำหรับ Training สามารถสกัดความสัมพันธ์ได้ถูกต้อง จำนวน 244 ความสัมพันธ์ สกัดได้แต่ผิด 30 ความสัมพันธ์ และสกัดไม่ได้ 56 ความสัมพันธ์ ขณะที่ข้อมูลสำหรับ Test สามารถสกัดความสัมพันธ์ได้ถูกต้อง จำนวน 172 ความสัมพันธ์ สกัดได้แต่ผิด 10 ความสัมพันธ์ และสกัดไม่ได้ 23 ความสัมพันธ์ อธิบายได้ว่า การสกัดความสัมพันธ์ของคำนั้น มีความซับซ้อนมากกว่า และมีลักษณะที่ไม่ตายตัวมากกว่าการสกัดคำเดี่ยว ซึ่งสังเกตได้จากค่า Recall สำหรับ Test เท่ากับ 88.21%

- *ประเด็น:* ผลการวิเคราะห์สัดส่วนความสัมพันธ์ของบทความ

จากตารางที่ 4 แสดงผลการคำนวณค่าสัดส่วนความสัมพันธ์ (อ้างอิงการจัดหมวดหมู่ระบบทศนิยมดิวอี้) ของบทความ “The role of

knowledge-oriented leadership in knowledge management practices and innovation" จำนวน 10 ลำดับ สามารถอธิบายได้ว่า บทความดังกล่าวมีความสัมพันธ์แบบสหความสัมพันธ์นั่นคือมากกว่าหนึ่งหมวด ซึ่งประกอบด้วยหมวดหมู่ 658: General Management มีค่าความสัมพันธ์ในลำดับแรกเท่ากับ 8.0597% หมวดหมู่ 371: School & Activities มีค่าความสัมพันธ์เป็นลำดับที่สองเท่ากับ 3.4328% และหมวดหมู่ 338: Production มีค่าความสัมพันธ์เป็นลำดับที่สามเท่ากับ 3.1343% เป็นต้น

ตารางที่ 4 สัดส่วนความสัมพันธ์แยกตามหมวดหมู่ของบทความ "The role of knowledge-oriented leadership in knowledge management practices and innovation"

ลำดับ	หมวดหมู่*	ค่าสัดส่วนความสัมพันธ์ (%)
1	658	8.0597
2	371	3.4328
3	338	3.1343
4	271	2.1641
5	352	2.0149
6	005	1.9403
7	355	1.9403
8	676	1.5671
9	333	1.4925
10	004	1.4925

* อ้างอิงจากหมวดหมู่ระบบทศนิยมดิวอี้

● *ประเด็น: ผลการวิเคราะห์ความคล้ายคลึงของบทความ*

จากตารางที่ 5 แสดงผลการเปรียบเทียบความคล้ายคลึงของบทความ "A scalable communication middleware for real-time data collection of dangerous goods vehicle activities" โดยผ่านกระบวนการวิเคราะห์คำ จากนั้นจำแนกความสัมพันธ์และนำมาเปรียบเทียบความคล้ายคลึงกับบทความที่จัดเก็บในฐานข้อมูล พบว่า บทความที่จัดเก็บในฐานข้อมูล เมื่อทำการค้นคืนออกมามีความเกี่ยวข้องกับบทความหลักในระดับสูง (ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่าง 0.7 – 0.9) ซึ่งมีเพียง 6 บทความเท่านั้น จากจำนวน 69 บทความที่อยู่ในเกณฑ์ยอมรับได้

ตารางที่ 5 ผลการวิเคราะห์ความคล้ายคลึงของบทความ "A scalable communication middleware for real-time data collection of dangerous goods vehicle activities"

ลำดับ	ชื่อบทความ	สัมประสิทธิ์สหสัมพันธ์
1	A new wireless asynchronous data communications module for industrial applications	0.90
2	Synchronization of networked Euler-Lagrange systems by sampled data communication with time-varying transmission delays under directed topology	0.82
3	Streamlining life cycle inventory data generation in agriculture using traceability data and information and communication technologies – part II: application to viticulture	0.81
4	Security awareness of computer users: A phishing threat avoidance perspective	0.80
5	Ontology for knowledge management in software maintenance	0.70
6	Relationship among Soft Skills, Hard Skills, and Innovativeness of Knowledge Workers in the Knowledge Economy Era	0.70

ตารางที่ 6 ผลการวิเคราะห์ตามลำดับชั้นที่ 1

บทความ #38			จำนวนหมวดที่สัมพันธ์	บทความ #11		
หมวดหลัก	จำนวน	ค่า (%)		หมวดหลัก	จำนวน	ค่า (%)
000	10	7.143	10	000	33	6.776
100	12	8.571	12	100	27	5.544
200	10	7.143	10	200	44	9.035
300	34	24.286	34	300	74	15.195
400	11	7.857	11	400	54	11.088
500	13	9.286	12	500	65	13.347
600	33	23.571	33	600	76	15.606
700	15	10.714	15	700	58	11.910
800	1	0.714	1	800	19	3.901
900	1	0.714	1	900	37	7.589
รวม	**140**	**100**	**139**	**รวม**	**487**	**100**

● *ประเด็น*: ผลการวิเคราะห์ความคล้ายคลึงร่วมระหว่างบทความในระดับลึก

จากตารางที่ 6 แสดงผลการวิเคราะห์ความคล้ายคลึงร่วมระหว่างบทความ #38 (A scalable communication middleware for real-time data collection of dangerous goods vehicle activities) และบทความ #11 (A new wireless asynchronous data communications module for industrial applications) ตามลำดับชั้นที่ 1 ด้วยจำนวนหมวดหมู่ซึ่งปรากฏหมวดหมู่ที่มีความคล้ายคลึงร่วมกันภายในจำนวน 139 หมวดหมู่ จากตารางอธิบายได้ว่า ทั้งสองบทความมีความคล้ายคลึงกันตามหมวดหมู่ ด้วยค่าความสัมพันธ์ของหมวดหมู่ตามลำดับชั้นความสัมพันธ์ ซึ่งเป็นการเชื่อมโยงในระดับลึกของตัวเนื้อหาของแต่ละบทความ

ตารางที่ 7 ผลการวิเคราะห์ตามลำดับชั้นที่ 2

บทความ #38 (10)			จำนวนหมวดที่สัมพันธ์	บทความ #11 (33)		
หมวดย่อย	จำนวน	ค่า (%)		หมวดย่อย	จำนวน	ค่า (%)
000	4	2.857	4	000	5	1.027
010	1	0.714	1	010	5	1.027
020	4	2.857	4	020	8	1.643
030	-	-	-	030	9	1.848
040	-	-	-	040	-	-
050	-	-	-	050	1	0.205
060	1	0.714	1	060	2	0.411
070	-	-	-	070	1	0.205
080	-	-	-	080	1	0.205
090	-	-	-	090	1	0.205
รวม	**10**	**7.143**	**10**	**รวม**	**33**	**6.776**

จากตารางที่ 7 แสดงผลการวิเคราะห์ความคล้ายคลึงร่วมระหว่างบทความ #38 และบทความ #11 ตามลำดับชั้นที่ 2 ของหมวดหลัก 000-General ซึ่งสามารถจำแนกเป็นหมวดย่อยภายในได้ 10 หมวดย่อย จากตารางพบว่าหมวดย่อยที่มีความคล้ายคลึงกับทั้งสองบทความมีจำนวน 4 หมวดย่อย ประกอบไปด้วย 10 หมวดหมู่ โดยที่บทความ #38 มีความคล้ายคลึงเพียง 3 หมวดย่อย ซึ่งมีค่าความสัมพันธ์โดยรวมเท่ากับ 7.143% ขณะที่บทความ #11 ปรากฏหมวดย่อยที่มีความคล้ายคลึงเท่ากับ 4 หมวดย่อย และมีค่าความสัมพันธ์โดยรวมเท่ากับ 6.776% สอดคล้องกับความคล้ายคลึงในลำดับชั้นที่ 1 ของหมวดหลัก 000-General แสดงดังตารางที่ 4 ข้างต้น

ตารางที่ 8 ผลการวิเคราะห์ตามลำดับชั้นที่ 3

บทความ #38 (2.857%)			จำนวนหมวดที่สัมพันธ์	บทความ #11 (1.027%)		
หมวดย่อย	จำนวน	ค่า (%)		หมวดย่อย	จำนวน	ค่า (%)
000	-	-	-	000	-	-
001	-	-	-	001	1	0.325
002	-	-	-	002	-	-
003	1	0.282	1	003	1	0.075
004	1	1.695	1	004	1	1.426
005	1	9.322	1	005	1	3.228
006	1	1.129	1	006	1	0.601
007	-	-	-	007	-	-
008	-	-	-	008	-	-
009	-	-	-	009	-	-
รวม	**4**	**-**	**4**	**รวม**	**5**	**-**

จากตารางที่ 8 แสดงผลการวิเคราะห์ความคล้ายคลึงร่วมระหว่างบทความ #38 และบทความ #11 ตามลำดับชั้นที่ 3 ของหมวดย่อย 000-General (จากหมวดหลักสู่หมวดย่อย และจากหมวดย่อยสู่หมวดหมู่) ซึ่งสามารถจำแนกเป็นหมวดหมู่ภายในได้ 10 หมวดหมู่ จากตารางพบว่าทั้งสองบทความปรากฏหมวดหมู่ที่คล้ายคลึงกันจำนวน 4 หมวดหมู่ ได้แก่ หมวดหมู่ 003-Systems, 004-Computer Science, 005-Computer Programming และ 006-Special Computer Methods โดยที่บทความ #38 มีค่าความคล้ายคลึง ที่ได้จากการวิเคราะห์น้ำหนักเท่ากับ 0.282%, 1.696%, 9.322% และ 1.129% ตามลำดับ ในขณะที่บทความ #11 มีจำนวนหมวดหมู่ที่ปรากฏความคล้ายคลึงจำนวน 5 หมวดหมู่ คือ หมวดหมู่ 001-Knowledge, 003-Systems, 004-Computer Science, 005-Computer Programming และ 006-Special Computer Methods ซึ่งมีค่าความสัมพันธ์ที่ได้จากการวิเคราะห์

น้ำหนักเท่ากับ 0.243%, 0,041%, 0.081%, 0.203% และ 0.081% ตามลำดับ

5. สรุป

บทความนี้ได้นำเสนอวิธีการวิเคราะห์ความคล้ายคลึงของเอกสารด้วยการประยุกต์ใช้เทคนิคการจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์เพื่อเพิ่มประสิทธิภาพในการเข้าถึง และค้นคืนสารสนเทศตรงกับความต้องการของผู้ใช้ และสามารถแนะนำเอกสารที่มีความคล้ายคลึงกันได้อย่างมีประสิทธิภาพ ซึ่งได้แบ่งการทดลองเพื่อพัฒนาออกเป็น 2 ส่วน คือ ส่วนที่ 1 การคัดเลือกสัดส่วนที่เหมาะสมของตัวแทนบทความ พบว่า สัดส่วนที่เหมาะสมสำหรับบทความนี้ คือ 4:2:2:2 จากตัวแทนข้อมูล 4 ส่วน คือ ชื่อเรื่อง บทคัดย่อ คำสำคัญ และอภิปรายผล ส่วนที่ 2 การออกแบบและวิเคราะห์ความคล้ายคลึงของเอกสาร จากผลการทดลองพบว่า วิธีการนี้สามารถอธิบายความคล้ายคลึงระหว่างเอกสารได้อย่างชัดเจนขึ้นว่า เอกสารหนึ่ง ๆ มีความสัมพันธ์กับเรื่องใดบ้างในปริมาณเท่าใดและมีรูปแบบความสัมพันธ์อย่างไร รวมทั้งสามารถนำเสนอเอกสารที่มีรูปแบบมิติความสัมพันธ์ของเนื้อหาที่ตรงกัน หรือมีความคล้ายคลึงกับรูปแบบมิติความสัมพันธ์ของเอกสารหลักในระดับลึก ด้วยหลักเกณฑ์การจำแนกหมวดหมู่ของระบบทศนิยมดิวอี้แบบสหความสัมพันธ์

ดังนั้น ด้วยเทคนิคนี้ผู้วิจัยเชื่อว่า จะทำให้สารสนเทศที่เคยมองไม่เห็นหรือไม่สามารถแสดงผลได้ ให้สามารถนำกลับออกมาเป็นสารสนเทศและนำเสนอต่อผู้ใช้ในรูปแบบความสัมพันธ์ที่ชัดเจนยิ่งขึ้น และมุ่งหวังว่าเทคนิคนี้จะเป็นประโยชน์ต่อการจัดการกับปริมาณข้อมูลจำนวนมากที่เกิดขึ้นในโลก Social Network ให้มีประสิทธิภาพมากยิ่งขึ้น

เอกสารอ้างอิง

[1] จุฬาลักษณ์ วัฒนานนท์ และอนิราช มิ่งขวัญ. “เทคนิคการสร้างภาพขยายความสัมพันธ์ของกลุ่มความรู้แบบสหความสัมพันธ์ระบบทศนิยมดิวอี้.” ใน เอกสารประกอบการประชุมทางวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 4 (NCCIT 09). กรุงเทพฯ : มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2552.

[2] จุฬาลักษณ์ วัฒนานนท์. “การหาความสัมพันธ์ของความรู้ โดยใช้กรอบความรู้การจัดหมวดหมู่ระบบทศนิยมดิวอี้แบบสหความสัมพันธ์.” วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ. 2553.

[3] วิไลพร เลิศมหาเกียรติ และอนิราช มิ่งขวัญ. “นวัตกรรมการค้นคืนสารสนเทศแบบสหความสัมพันธ์.” วารสารวิชาการพระจอมเกล้าพระนครเหนือ, ปีที่ 20, ฉบับที่ 3 ก.ย. – ธ.ค., 2553.

[4] วิไลพร เลิศมหาเกียรติ. “การค้นคืนสารสนเทศแบบสหความสัมพันธ์ตามหมวดหมู่ระบบทศนิยมดิวอี้.” วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ภาควิชา

เทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ. 2553.

[5] อัครพล เอกวงศ์อนันต์. การระบุคำไทยและคำทับศัพท์ด้วยแบบจำลองเอ็นแกรม. วิทยานิพนธ์อักษรศาสตร์มหาบัณฑิตสาขาภาษาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2548.

[6] อัษฎางค์ แตงไทย และชุลีรัตน์ จรัสกุลชัย. “การย่อความภาษาไทยโดยกรรมวิธีการแยกค่าแบบดี่ยว.” การประชุมทางวิชาการ NCSEC, 2004.

[7] Baeza-Yates, R., and B. Ribeiro-Neto. Modern Information Retrieval. New York : Addison-Wesley, 2005.

[8] Beall.J., “Representation of DDC in MARC21.,” In New Perspectives on Subject Indexing and Classification: International Symposium in Honour of Magda Heiner-Freiling; Deutsche Nationalbibliothek: Leipzig, Frankfurt am Main Berlin, pp. 131 – 137, 2008.

[9] Breese, J. S., Heckerman, D., and Kadie, C. “Empirical Analysis of Predictive Algorithms for Collaborative Filtering.” Technical Report MSR-TR-98-12. [n.p. : n.p.], 1998.

[10] Feng, Y., and Feng, J. “Incremental updating algorithms for mining association rules.” Journal of Software, 1998 : 301-306.

[11] Fuller, M. C., Biros, P. D. and Delen, D. “Exploration of Feature Selection and Advanced Classification Models for High-Stakes Deception Detection.” Proceedings of the 41st Hawaii International Conference on System Sciences, IEEE, 2008.

[12] Herlocker, J. L., et al. “An Algorithmic Framework for Performing Collaborative Filtering.” Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [n.p. : n.p.], 1999 : 230 – 237.

[13] Huang, C. T., Yonghong Zhou, Z. and Huang, T. “Towards Multi-Granularity Multi-Facet E-book Retrieval.” WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM, 2007.

[14] Kawtrakut, A., Yingsaeree, C., and Andres, F. “A Framework of NLP Based Information Tracking and Related Knowledge Organizing with Topic Maps”. Natural Languages Processing and Information Systems. 2007 : 272-283.

[15] Lertmahakiat, W., and Mingkwan, A. “Information Retrieval by Multiple DDC Relational Classification.” Proceedings NCCIT 2008. The 4th National Conference on Computing and Information Technology. 22 – 23 May 2008, Bangkok, Thailand.

[16] Oded, N. “Creativity, Knowledge and IS: A Critical View,” Proceedings of the 38th

Hawaii International Conference on System Sciences, 2005.

[17] Peery, C., Wei Wang, A. M. and Nguyen, Thu D. "Multi-Dimensional Search for Personal Information Management Systems." EDBT'08. Nantes, France, 2008.

[18] Redmann, T., and Thomas, H. "The wiki way of knowledge management with topic maps." Proceedings of the International Conference on Information Society (i-Society 2007), October 7 – 11, 2007.

[19] Sajjad M. Jasimuddin. "Storage of Transferred Knowledge or Transfer of Stored Knowledge: Which Direction? If both, then how?," Proceedings of the 38th Hawaii International Conference on System Sciences, 2005.

[20] Sarwar, B., et al. "Item-based Collaborative Filtering Recommendation Algorithms." Proceedings of the WWW10 Conference. [n.p.: n.p.], 2001.

[21] Scott, M. L., "Dewey decimals classification, 22nd edition: A study manual and number building guide.," Westport, Conn: Libraries Unlimited, 2005.

[22] Terveen, L. and Hill, W. "Beyond Recommender Systems: Helping People Help Each Other." HCI In The New Millennium, Addison-Wesley. 2001.