บทที่ 4
การลดมิติเอกสาร

บทนี้กล่าวถึงการลดมิติเอกสาร จากตัวแทนเอกสารที่ได้จากการสร้างดัชนีเอกสาร ในบทที่ 3 เนื้อหาในบทนี้ประกอบด้วย หลักการการลดมิติเอกสาร การลดมิติเอกสารด้วยวิธีการเลือกลักษณะเด่นด้วยกระบวนการรีเลแวนซี่สกอร์ การลดมิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีแอลเอสไอ การลดมิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีพีแอลเอส และวิจารณ์และสรุปผล

4.1 หลักการการลดมิติเอกสาร

จากกระบวนการการสร้างเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร พบว่าจำนวนคำสำคัญที่ได้จากการสร้างดัชนีเอกสารนั้นมีมากถึง 3,586 คำ ซึ่งคำสำคัญบางคำอาจทำให้ประสิทธิภาพในการจำแนกกลุ่มของเอกสารลดลง ดังนั้นจึงมีความจำเป็นในการกำจัดหรือลดปริมาณคำเหล่านี้ กระบวนการในการกำจัดคำสำคัญมีสองประเภท ประกอบด้วยวิธีการเลือกลักษณะเด่น ซึ่งเป็นวิธีการหาค่าน้ำหนักของแต่ละคำสำคัญ จากนั้น ทำการพิจารณาค่าน้ำหนักของคำสำคัญโดยเรียงค่าน้ำหนักจากมากไปน้อย ต่อไปคือการเลือกลำดับของคำสำคัญที่จะใช้ในการสร้างดัชนีเอกสาร อีกวิธีคือ วิธีการสกัดลักษณะเด่น คือการแปลงเวกเตอร์เอกสารให้อยู่ในรูปแบบอื่นที่มีมิติที่เล็กลง โดยกระบวนการนี้สามารถช่วยแก้ปัญหาด้านความหมายของคำสำคัญได้ในระดับที่ยอมรับได้

4.2 การลดมิติเอกสารด้วยวิธีการเลือกลักษณะเด่นด้วยกระบวนการรีเลแวนซี่สกอร์ (Relevancy Score)

การเลือกลักษณะเด่นมีองค์ประกอบที่สำคัญในการหาค่าน้ำหนัก ประกอบด้วยเมตริกซ์ความถี่ของคำสำคัญ-เอกสาร และ การระบุกลุ่มของเอกสารของข้อมูลสำหรับเรียนรู้ โดยการเลือกลักษณะเด่นมีหลายวิธีและถูกใช้ในงานทางด้านการจำแนกเอกสาร เช่นในงานของ Lam (Lam and Lee, 1999) และ Yang (Yang and Pedersen, 1997)

งานวิจัยนี้ได้ใช้กระบวนการรีเลแวนซี่สกอร์ ซึ่งถูกใช้โดย Wiener (Wiener, Pedersen and Weigend, 1995) โดย ซึ่งงานวิจัยของ Wiener ได้แสดงให้เห็นว่าการใช้กระบวนการนี้ให้ความถูกต้องในการจำแนกกลุ่มใกล้เคียงกับวิธีการเลือกลักษณะเด่นวิธีการอื่นๆ ขั้นตอนการเลือกลักษณะเด่นมีดังนี้

4.2.1 ขั้นตอนการเลือกลักษณะเด่นด้วยวิธีรีเลแวนซี่สกอร์

กำหนดให้ เมตริกซ์ความถี่ของคำสำคัญ-เอกสาร X มีขนาด k×d โดยมีกลุ่มทั้งหมด t กลุ่ม โดยแต่ละเอกสารต้องเป็นสมาชิกในกลุ่มใดกลุ่มหนึ่ง

1) ทำการคำนวณค่ารีเลแวนซี่สกอร์ r ของแต่ละคำสำคัญ k ในแต่ละกลุ่มเอกสาร t ดังสมการ

$$r_{tk} = \log \frac{w_{tk} / d_t + 1/6}{w_{\bar{t}k} / d_{\bar{t}} + 1/6}$$

โดยที่ w_{tk} คือจำนวนเอกสารที่มีคำสำคัญ k ในกลุ่มเอกสาร t และ d_t คือ จำนวนเอกสารทั้งหมดในกลุ่มเอกสาร t

2) ทำการหาค่าสูงสุดของค่ารีเลแวนซี่สกอร์ของแต่ละกลุ่ม r_i ดังสมการ

$$mr_k = \max(r_k)$$

3) ทำการเรียงลำดับคำสำคัญตามค่า mr_k จากมากไปน้อย

4) ทำการเลือกจำนวนคำสำคัญที่ต้องการ และเลือกคำสำคัญจากมากไปน้อย

5) ทำการสร้างดัชนีเอกสารจากคำสำคัญที่เลือกมา

6) จากขั้นตอนการเลือกลักษณะเด่น งานวิจัยนี้ได้ทำการเลือกคำสำคัญเพื่อสร้างเมตริกซ์คำสำคัญ-เอกสารเพื่อใช้ในการทดสอบการจำแนกกลุ่มในช่วงระหว่าง 1-500 คำสำคัญ โดยได้มาจากการเลือกมิติของแอลเอสไอในหัวข้อ 4.3

4.2.2 ผลของการเลือกลักษณะเด่นด้วยวิธีรีเลแวนซี่สกอร์

จากการหาค่ารีเลแวนซี่สกอร์เพื่อเลือกคำสำคัญเพื่อใช้สำหรับสร้างดัชนีเอกสารนั้นจะได้เมตริกซ์ที่มีขนาด k×d โดยที่ k มีค่าระหว่าง 1-500 ซึ่งจากการเลือกคำสำคัญมาเพื่อสร้างดัชนีเอกสารนั้นจะได้เมตริกซ์สำหรับเรียนรู้ k×1614 และเมตริกซ์สำหรับทดสอบ k×1614

4.3 การลดมิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีแอลเอสไอ

การสกัดลักษณะเด่น คือกระบวนการแปลงเวกเตอร์คำสำคัญ-เอกสาร ให้อยู่ในรูปแบบอื่นที่มีมิติที่เล็กลง ซึ่งวิธีที่เป็นที่นิยมและมีประสิทธิภาพสูงคือ แอลเอสไอ ซึ่งเป็นกระบวนการแปลงข้อมูลในรูปแบบเดิมให้อยู่ในรูปของพื้นที่ความหมาย (Semantic space) ที่มีขนาดเล็กลง การแปลงข้อมูลให้อยู่ในรูปแบบของแอลเอสไอสามารถกระทำโดยการนำเมตริกซ์ความถี่ของคำสำคัญ-เอกสารผ่านกระบวนการ การคำนวณเอสวีดี จากนั้นทำการตัดกลุ่มเมตริกซ์ที่สร้างจากเอสวีดีเพื่อใช้สร้างตัวแทนข้อมูล อย่างไรก็ตาม แอลเอสไอถูกเสนอขึ้นมาเพื่อใช้กับงานทางด้านการค้นคืน

เอกสาร แต่แอลเอสไอถูกใช้กับงานจำแนกเอกสารอย่างแพรหลายเช่นกัน เช่นงานของ Yang (Yang,1995), Wu (Wu, 2002), Zelikovitz (Zelikovitz and Hirsh, 2001)

4.3.1 ขั้นตอนการสกัดลักษณะเด่นด้วยวิธีแอลเอสไอ

ในการสร้างตัวแทนเอกสารเพื่อให้มีมิติที่เล็กลงด้วยแอลเอสไอมีขั้นตอนดังนี้

กำหนดให้ เมตริกซ์ A คือเมตริกซ์เอกสารสำหรับเรียนรู้ ที่ได้จากการสร้างตัวแทนเอกสารด้วยวิธีการสร้างดัชนีเอกสาร ที่มีขนาด t×d โดยที่ t มีค่าเท่ากับจำนวนคำสำคัญ และ d มีค่าเท่ากับจำนวนเอกสาร

1) ทำการคำนวณกระบวนการเอสวีดีบนเมตริกซ์ A

$$svd(A_{t\times d}) = T_{t\times n} \times S_{n\times n} \times (D_{d\times n})^T$$

2) ทำการสร้างตัวแทนเอกสารโดยใช้ S และ T โดยมีขั้นตนดังนี้

กำหนดให้ q คือเวกเตอร์เอกสารที่ต้องการสร้างตัวแทนเอกสาร ซึ่งในที่นี้ก็คือเอกสารสำหรับเรียนรู้ และ เอกสารสำหรับทดสอบ

2.1) ทำการเลือกจำนวนลักษณะเด่น k ที่ต้องการสร้างตัวแทนเอกสาร

2.2) ทำการสร้างตัวแทนเอกสาร จากสมการดังนี้

$$q' = 1/S_{k\times k} \times T'_{t\times k} \times q$$

โดยที่ q' คือตัวแทนของเอกสาร q

2.3) ทำการสร้างตัวแทนเอกสารทุกเอกสารจากขั้นตอน 2.2)

ในการสร้างตัวแทนเอกสารด้วยวิธีแอลเอสไอนี้ เราได้ทำการสร้างตัวแทนโดยมีมิติของตัวแทนเอกสาร k มีค่าระหว่าง 1-500 มิติ (Zeng, Wang and Nie, 2007) ซึ่งจากการทดสอบความถูกต้องของการจำแนกเอกสารพบว่าความถูกต้องของการจำแนกจะมีค่าเพิ่มขึ้นเรื่อยๆจนถึงค่า k อยู่ในช่วงระหว่าง 380-450 และลดลงเมื่อ k มีค่าเท่ากับ 500 โดยที่ชุดข้อมูลสำหรับเรียนรู้และทดสอบในแบบจำลองการจำแนกเดียวกันจำนวนมิติ k ที่เท่ากัน

4.3.2 ผลของการสกัดลักษณะเด่นด้วยวิธีแอลเอสไอ

จากการคำนวณเอสวีดีแล้วทำการสร้างตัวแทนเอกสาร ที่มีมิติของตัวแทนเอกสาร k มีค่าระหว่าง 1-500 มิติ ซึ่งจากการสร้างตัวแทนเอกสารด้วยวิธีแอลเอสไอนั้นจะได้เมตริกซ์สำหรับเรียนรู้ k×1614 และเมตริกซ์สำหรับทดสอบ k×1614

4.4 การลดมิติด้วยวิธีการสกัดลักษณะเด่นด้วยวิธีพีแอลเอส

พีแอลเอสเป็นวิธีหนึ่งที่ถูกใช้ในการสร้างตัวแทนเอกสาร ถูกนำเสนอโดย Wang (Wang and Nie, 2003) ซึ่งมีการทำงานคล้ายกับวิธีแอลเอสไอ โดยอยู่บนพื้นฐานการหาพื้นที่ความหมาย

เหมือนกัน แต่สิ่งที่แตกต่างจากวิธีแอลเอสไอ คือ วิธีพีแอลเอสมีการนำข้อมูลของกลุ่มเอกสารมาใช้ในกระบวนการหาพื้นที่ความหมายด้วย จึงทำให้สามารถรักษาความแตกต่างของกลุ่มข้อมูลที่มีขนาดเล็กไว้ได้

4.4.1 ขั้นตอนการสกัดลักษณะเด่นด้วยวิธีพีแอลเอส

สมมติให้เมตริกซ์ X คือเมตริกซ์ของเอกสาร-คำสำคัญ ของเอกสารสำหรับเรียนรู้

1) ทำการระบุกลุ่มของเอกสาร โดยการสร้างเมตริกซ์เอกสาร-กลุ่มของเอกสาร Y โดยที่ m แทนเอกสารและ r แทนกลุ่มเอกสาร จะได้ว่า

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1r} \\ y_{21} & y_{22} & \cdots & y_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mr} \end{bmatrix}$$

โดยกำหนด y_{ij} มีค่าเป็น 1 ก็ต่อเมื่อเอกสาร X_i เป็นสมาชิกของกลุ่มนั้น ๆ และมีค่าเป็น 0 ก็ต่อเมื่อเอกสาร X_i ไม่เป็นสมาชิกของกลุ่มนั้น

2) ทำการเลือกจำนวนลักษณะเด่น k ที่ต้องการสร้างตัวแทนเอกสาร

3) จากนั้นใช้ขั้นตอนวิธีในการสร้างพีแอลเอส (Wang and Nie,2003) โดยมีขั้นตอนวิธีดังนี้

ALGORITHM-1(LSR/PLS2) :

$E_0 = X; \quad F_0 = Y;$

FOR $k = 1, \ldots, s$ DO

$u_k =$ first column of F_{k-1};

Do until convergence in ξ_k// computing pairs of latent variables

$\xi_k = E_{k-1}^T u_k / u_k^T u_k$

$\xi_k = \frac{\xi_k}{||\xi_k||}$

$\xi t_k = E_{k-1} \xi_k$

$\omega_k = F_{k-1}^T t_k / t_k^T t_k$

$u_k = F_{k-1} \omega_k / \omega_k^T \omega_k$

ENDDO

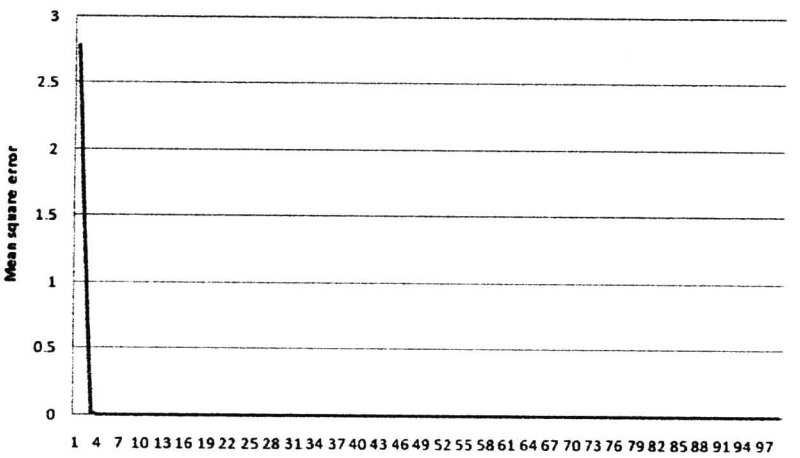
$P_k = F_{k-1}^T t_k / t_k^T t_k$

$E_k = E_{k-1} t_k P_k^T$ //remaining information of document - term matrix

$F_k = F_{k-1} t_k \omega_k^T$ //remaining information of document - class matrix

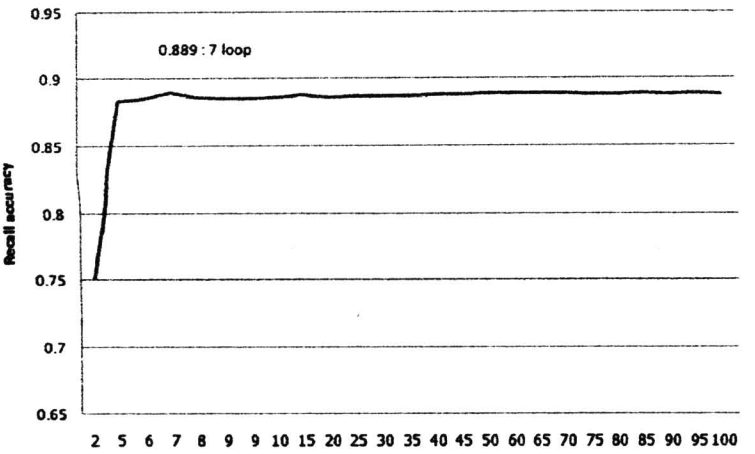
ENDFOR

จากขั้นตอนวิธีการหาพีแอลเอส งานวิจัยนี้ได้ปรับปรุงการทำงานของขั้นตอนวิธีการหาพีแอลเอส โดยกำหนดจำนวนรอบสูงสุดที่เหมาะสม กับเอกสารชีวสารสนเทศที่นำมาทดสอบ โดยการหาจำนวนรอบสูงสุดในการปรับค่า ξ_k มีขั้นตอนดังนี้ เริ่มจากทำการทดสอบการสร้างตัวแทนเอกสาร โดยเลือกมิติตัวแทนเอกสารมีค่าระหว่าง 1-500 มิติ และใช้รอบในการปรับค่าเท่าจนค่าไม่เปลี่ยนแปลง หรือ ไม่เกิน 100 รอบ ผลที่ได้คือการสร้างตัวแทนเอกสารจำนวนมิติเท่ากับ 25 มิติให้ความถูกต้องในการจำแนกสูงสุด จากนั้นนำมิติแรกมาใช้ในการเปรียบเทียบค่าความผิดพลาดเฉลี่ยกำลังสอง (Mean square error) ผลที่ได้ดังนี้



รูป 4.1 แสดงค่าความผิดพลาดเฉลี่ยกำลังสองของ ξ_k มิติแรก

จากรูป 4.1 พบว่าค่าผิดพลาดเฉลี่ยกำลังสองเริ่มลดลงเข้าใกล้ 0 ตั้งแต่รอบในการปรับค่าของ ξ_k เท่ากับ 4 และเมื่อทดสอบโดยการสร้างตัวแทนเอกสาร จากนั้นนำไปจำแนกโดยแบบจำลองโครงข่ายประสาท พบว่าการปรับค่า ξ_k จำนวน 7 รอบให้ความถูกต้องในการจำแนกสูงสุด ดังรูป 4.2



รูป 4.2 แสดงค่าความถูกต้องของการจำแนกที่สร้างตัวแทนเอกสาร 25 มิติ ในแต่ละรอบของการปรับค่า ξ_k ในมิติที่ 1

4) เมื่อได้เมตริกซ์ ξ ทำการสร้างตัวแทนเอกสารดังนี้

$$q' = q' \times \xi$$

โดยที่ q' คือตัวแทนเอกสารของเอกสาร q (Zeng, Wang and Nie, 2007)

5) ทำการสร้างตัวแทนเอกสารทุกเอกสารจากขั้นตอนที่ 4)

ในการสร้างตัวแทนเอกสารด้วยวิธีพีแอลเอส จึงทำการสร้างตัวแทนโดยมีมิติของตัวแทนเอกสาร k มีค่าระหว่าง 1-500 มิติ

4.4.2 ผลของการสกัดลักษณะเด่นด้วยวิธีพีแอลเอส

จากการสร้างตัวแทนเอกสารด้วยวิธีพีแอลเอส ที่มีมิติของตัวแทนเอกสาร k มีค่าระหว่าง 1-500 มิติ ซึ่งจากการสร้างตัวแทนเอกสารด้วยวิธีพีแอลเอส นั้นจะได้เมตริกซ์สำหรับเรียนรู้ k×1614 และเมตริกซ์สำหรับทดสอบ k×1614

4.5 วิจารณ์และสรุปผล

จากการลดมิติของเอกสารด้วยวิธีการเลือกลักษณะเด่นและวิธีการสกัดลักษณะเด่นผลที่ได้คือตัวแทนเอกสารที่มีมิติเล็กลง ซึ่งการลดมิติแบบต่างๆนั้นมีจุดเด่นที่ต่างกันคือ การเลือกลักษณะเด่นด้วยวิธีการรีเลแวนซี่สกอร์ เป็นวิธีที่คำนวณด้วยวิธีทางสถิติจึงทำให้คำนวณง่ายและใช้เวลาการคำนวณน้อย และยังสามารถช่วยกำจัดคำสำคัญที่ทำให้ความแตกต่างระหว่างกลุ่มไม่ชัดเจนออกไป ส่วนวิธีแอลเอสไอเป็นกระบวนการการหาพื้นที่ความหมายจากความสัมพันธ์ระหว่างคำสำคัญ โดยการหารูปแบบของความสัมพันธ์เองโดยไม่ใช้ข้อมูลของกลุ่มของเอกสารเข้ามาร่วมพิจารณาในการ

หาพื้นที่ความหมาย ส่วนวิธีพีแอลเอสเป็นกระบวนการหาพื้นที่ความหมายคล้ายกับวิธีแอลเอสไอ แต่สิ่งที่ต่างกันก็คือ พีแอลเอสมีการนำข้อมูลกลุ่มของเอกสารเข้ามาร่วมพิจารณาด้วย

แต่จากการลดมิติเหล่านี้ สิ่งที่ต้องพิจารณาต่อก็คือการหาจำนวนมิติที่เหมาะสมของแต่ละวิธีที่ให้ความถูกต้องสูงในการจำแนกกลุ่มเอกสารของแบบจำลองการจำแนกกลุ่ม และช่วยลดเวลาการคำนวณระหว่างการเรียนรู้ของแบบจำลองได้ดีไปพร้อมๆกัน ซึ่งจะกล่าวถึงในบทต่อไป