

บทที่ 5

สรุปผลการศึกษาวิจัยและข้อเสนอแนะ (Conclusions and Recommendations)

5.1 สรุปผลการศึกษาและวิจารณ์

การจัดทำดรรชนีหาความหมายแฝง (Latent Semantic Indexing) มีจุดมุ่งหมายเพื่อดึงความหมายจากชุดเอกสารมากกว่าการพิจารณาแค่ความหมายของคำหลัก โดยใช้การกระจายเมทริกซ์ (Matrix decomposition) ซึ่งเป็นเทคนิคที่ใช้ในระบบการค้นคืนสารสนเทศที่ใช้ในการจัดทำดรรชนีเพื่อหาความหมายแฝงในเอกสาร การกระจายเมทริกซ์ที่นิยมใช้ในการจัดทำดรรชนีเพื่อหาความหมายแฝง คือการกระจายเมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว (Singular Value Decomposition) ซึ่งได้รับการนำเสนอในงานวิจัยด้านการค้นคืนสารสนเทศจากข้อความภาษาไทยหลายงาน เช่น การจัดทำดรรชนีเพื่อหาความหมายแฝง การย่อความเอกสารภาษาไทย การประเมินคุณภาพการถอดความภาษาไทย แต่การกระจายเมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว (Singular Value Decomposition) มีจุดอ่อนคือ ใช้พื้นที่ในการประมวลผลระหว่างกระบวนการกระจายเมทริกซ์ (Decomposition storage) จำนวนมากจึงไม่เหมาะสมกับฐานข้อมูลเอกสารขนาดใหญ่ การกระจายเมทริกซ์ด้วยวิธีเซมิดีสครีต (Semidiscrete Matrix Decomposition) เป็นเทคนิคการกระจายเมทริกซ์วิธีหนึ่งซึ่งมีข้อดีในแง่ของการใช้พื้นที่ประมวลผลการกระจายเมทริกซ์น้อยกว่ามาก

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบผลการค้นคืนสารสนเทศระหว่างการกระจายเมทริกซ์ด้วยวิธีเซมิดีสครีต (Semidiscrete Matrix Decomposition) กับการกระจายเมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว (Singular Value Decomposition)

การวิจัยนี้ใช้ชุดเอกสารที่เป็นบทคัดย่อของงานวิจัย 4 สาขา คือ วิทยาการคอมพิวเตอร์ ประวัติศาสตร์ บรรณารักษศาสตร์ และรัฐศาสตร์ เอกสารทั้งหมดอยู่ในรูปแบบแฟ้มข้อมูลข้อความ (Text file format) เนื่องจากภาษาไทยไม่มีขอบเขตของคำที่ชัดเจนเหมือนในภาษาอังกฤษ จึงต้องใช้โปรแกรมตัดคำ SWATH เพื่อทำการตัดคำโดยแทรกช่องว่างเข้าไประหว่างคำเพื่อแยกคำให้ชัดเจน แล้วสร้างเมทริกซ์คำ-เอกสารจากชุดเอกสารที่ผ่านการตัดคำแล้ว สดมภ์ของเมทริกซ์แทนเอกสารที่มีในชุดเอกสาร ส่วนแถวของเมทริกซ์แทนคำที่มีอยู่ในชุดเอกสาร ส่วนสมาชิกของเมทริกซ์หมายถึงความถี่ของคำที่ปรากฏในเอกสารหนึ่งๆ ซึ่งอาจได้รับการให้ค่า

นำหน้าด้วยวิธีใดวิธีหนึ่ง หลังจากสร้างเมทริกซ์ค่า-เอกสารแล้วทำการกระจายเมทริกซ์ดังกล่าวด้วยวิธีเซมิดีสครีตและวิธีแยกค่าแบบเดี่ยวที่ค่าลำดับชั้น (Rank) ของเมทริกซ์ต่างๆโดยคำนวณจากความเปลี่ยนแปลงสัมพัทธ์ (Relative change) ร้อยละ 0 10 20 30 และ 40 เพื่อศึกษาเปรียบเทียบผลการค้นคืนสารสนเทศการวิจัยนี้ใช้ชุดข้อความ 30 คำต่อ 1 ชุดเอกสาร โดยเปรียบเทียบพื้นที่ที่ใช้ในการประมวลผลการกระจายเมทริกซ์ (Decomposition storage) เวลาที่ใช้ในการประมวลผลการกระจายเมทริกซ์ (Decomposition time) ความแม่นยำ (Precision) การเรียกกลับ (Recall) และเวลาที่ใช้ในการประมวลผลข้อความ (Query time) ระหว่างการกระจายเมทริกซ์ทั้งสองวิธี

ผลการวิจัยปรากฏว่า

(1) พื้นที่ที่ใช้ในการประมวลผลการกระจายเมทริกซ์ (Decomposition storage) ด้วยวิธีเซมิดีสครีตใช้น้อยกว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ ซึ่งสอดคล้องกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346) ทั้งนี้เนื่องจากการกระจายเมทริกซ์ด้วยวิธีเซมิดีสครีตเกิดเมทริกซ์ย่อย 2 เมทริกซ์ (X,Y) ซึ่งสมาชิกของเมทริกซ์เป็นจำนวนเต็มที่มีค่าที่เป็นไปได้ คือ -1, 0, 1 ทำให้ใช้พื้นที่ในการประมวลผลน้อยกว่าวิธีแยกค่าแบบเดี่ยว ขนาดพื้นที่ที่ใช้ในการประมวลผลการกระจายเมทริกซ์ของงานวิจัยนี้ที่ค่าลำดับชั้น (Rank) ใดๆ สามารถคำนวณได้ดังแสดงในตารางที่ 5.1

ตารางที่ 5.1

การคำนวณขนาดพื้นที่ที่ใช้ในการประมวลผลการกระจายเมทริกซ์

ชุดเอกสารสาขา	SDD	SVD
วิทยาการคอมพิวเตอร์	1,332.75k	42,528k
ประวัติศาสตร์	1,791.50k	57,208k
บรรณารักษศาสตร์	1,444.50k	46,104k
รัฐศาสตร์	1,400.25k	44,688k
ขนาดพื้นที่โดยเฉลี่ย	1,492.25k	47,632k

(2) เวลาที่ใช้ในการประมวลผลการกระจายแมทริกซ์ (Decomposition time) ด้วยวิธีเคมีดีสคริตีใช้น้อยกว่าวิธีแยกค่าแบบเดี่ยวอย่างมีนัยสำคัญทางสถิติ ซึ่งสอดคล้องกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346)

(3) ความแม่นยำ (Precision) และการเรียกกลับ (Recall) ของการกระจายแมทริกซ์ด้วยวิธีเคมีดีสคริตีต่ำกว่าวิธีแยกค่าแบบเดี่ยวในชุดเอกสารทุกสาขาอย่างมีนัยสำคัญทางสถิติ โดยปัจจัยแยกแมทริกซ์ (Splitting factor) ไม่มีผลต่อความแม่นยำ (Precision) และการเรียกกลับ (Recall) ส่วนค่าขีดแบ่งความเหมือนของเอกสาร (Similarity threshold value) มีผลต่อความแม่นยำ (Precision) และการเรียกกลับ (Recall) กล่าวคือมีค่าสูงขึ้นเมื่อค่าลำดับชั้นของแมทริกซ์สูงขึ้น

(4) เวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ของการกระจายแมทริกซ์ด้วยวิธีเคมีดีสคริตีไม่แตกต่างกับวิธีแยกค่าแบบเดี่ยว อย่างมีนัยสำคัญทางสถิติ ซึ่งขัดกับสมมุติฐานของงานวิจัย และงานของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346) ซึ่งรายงานเวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ของการกระจายแมทริกซ์ด้วยวิธีเคมีดีสคริตีใช้น้อยกว่าวิธีแยกค่าแบบเดี่ยว ทั้งๆที่การประมวลผลในสมการ 3.9 (Kolda and O'Leary, 1998, p. 322-346) มีการกระทำกับแมทริกซ์มากกว่าสมการของวิธีแยกค่าแบบเดี่ยวตามสมการ 3.7 ดังที่กล่าวไว้ในบทที่ 3 งานวิจัยนี้จึงเสนอสมการ 3.8 ซึ่งมีการกระทำกับแมทริกซ์เท่ากับสมการของวิธีแยกค่าแบบเดี่ยวตามสมการ 3.7 จึงทำให้เวลาที่ใช้ในการประมวลผลข้อคำถามใกล้เคียงกัน นอกจากนั้นเนื่องจากงานวิจัยนี้ใช้โปรแกรม MATLAB 2006a ในการประมวลผลแมทริกซ์ ซึ่งโปรแกรม MATLAB 2006a จะเก็บข้อมูลสมาชิกของแมทริกซ์เป็นเลขทศนิยม จึงทำให้ไม่สามารถใช้ข้อได้เปรียบของการกระจายแมทริกซ์ด้วยวิธีเคมีดีสคริตีซึ่งจะเกิดแมทริกซ์ย่อย 2 แมทริกซ์ (X, Y) ที่มีสมาชิกของแมทริกซ์เป็นจำนวนเต็มที่มีค่า $\{-1, 0, 1\}$ จึงทำให้ผลการประมวลผลข้อคำถามที่มีการแยกแมทริกซ์ด้วยวิธีเคมีดีสคริตีไม่สอดคล้องกับงานวิจัยของ Kolda และคณะ (Kolda and O'Leary, 1998, p. 322-346) ดังนั้นหากเลือกใช้เครื่องมืออื่นที่สามารถใช้ประโยชน์จากจุดนี้ได้ ก็จะทำให้เวลาในการประมวลผลข้อคำถาม (Query time) ของการกระจายแมทริกซ์ด้วยวิธีเคมีดีสคริตีดีกว่าวิธีแยกค่าแบบเดี่ยวมากขึ้นได้

(5) การทดลองในเอกสารทั้ง 4 สาขา ให้ผลการทดลองเกี่ยวกับ พื้นที่ที่ใช้ในการประมวลผลการกระจายแมทริกซ์ (Decomposition storage) เวลาที่ใช้ในการประมวลผลการกระจายแมทริกซ์ (Decomposition time) ความแม่นยำ (Precision) การเรียกกลับ (Recall) และ

เวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) ระหว่างการกระจายแมทริกซ์ทั้งสองวิธี ไม่แตกต่างกัน

5.2 ข้อเสนอแนะในการนำผลการวิจัยไปใช้

จากผลการวิจัยจะเห็นได้ว่าการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตเป็นเทคนิคหนึ่งที่น่าสนใจนอกเหนือจากวิธีแยกค่าแบบเดียวในการนำไปใช้ออกแบบระบบการค้นคืนสารสนเทศ ด้วยวิธีการจัดทำดรชนีเพื่อหาความหมายแฝง แต่ผู้พัฒนาระบบการค้นคืนสารสนเทศควรต้องพิจารณาถึงข้อดีข้อเสียของทั้งสองวิธีเพื่อเลือกนำไปใช้ให้ตรงกับความต้องการได้ เนื่องจากการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตนั้นมีข้อดีในแง่ของการใช้พื้นที่การประมวลผลการกระจายแมทริกซ์ที่ใช้พื้นที่น้อยกว่า แม้ว่าการประมวลผลการกระจายแมทริกซ์จะมากกว่าแต่การประมวลผลการกระจายแมทริกซ์จะเกิดขึ้นเพียงครั้งเดียวตอนเริ่มต้นเท่านั้น เวลาในการประมวลผลข้อคำถามนั้นก็สามารปรับปรุงให้ดีขึ้นได้ ส่วนความแม่นยำ (Precision) และการเรียกกลับ (Recall) นั้นสามารถปรับปรุงค่าขีดแบ่งความเหมือนของเอกสาร (Similarity threshold value) ให้ได้ค่าที่ต้องการ นอกจากนี้ถ้ามีการเปลี่ยนแปลงของชุดเอกสาร การกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตจะสามารถทำได้ง่ายวิธีแยกค่าแบบเดียว ซึ่งจะต้องคงความเป็นแมทริกซ์เชิงตั้งฉากเอาไว้ จนอาจต้องประมวลผลการกระจายแมทริกซ์โดยใช้แมทริกซ์ค่า-เอกสารใหม่ทั้งหมดซึ่งใช้พื้นที่และเวลามากจนไม่เหมาะสมที่จะนำมาใช้ในทางปฏิบัติจริง ดังนั้นวิธีเซมิดีสครีตจึงเหมาะสมที่ควรจะนำมาใช้ในระบบการค้นคืนสารสนเทศต่อไปได้ โดยเฉพาะในอนาคตฐานข้อมูลเอกสารจะมีเอกสารจำนวนมาก

5.3 ข้อเสนอแนะสำหรับการศึกษาวิจัยในอนาคต

อย่างไรก็ตาม การศึกษาวิจัยเพื่อศึกษาการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตในการค้นคืนสารสนเทศจากข้อความภาษาไทยควรมีต่อไปในประเด็นต่างๆดังนี้

(1) การปรับปรุงเวลาที่ใช้ในการประมวลผลข้อคำถาม (Query time) โดยเลือกใช้เครื่องมืออื่นที่สามารถใช้ประโยชน์จากคุณสมบัติของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตซึ่งมีการประมวลผลกับเขจำนวนเต็ม

(2) การใช้การกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตกับการค้นคืนสารสนเทศข้ามภาษา (Cross languages)

- (3) เทคนิคที่ใช้ในเมื่อมีเปลี่ยนแปลงชุดเอกสารแบบพลวัต (Updating technique)
- (4) ศึกษาแนวทางการปรับปรุงอัลกอริทึมในการกระจายแมทริกซ์ด้วยวิธีไฮบริดวิธีใดวิธีหนึ่งที่สามารถทำให้ใช้เวลาในการกระจายแมทริกซ์ได้หรือไม่
- (5) ควรจะนำเทคนิคการแยกแมทริกซ์ไปประยุกต์ใช้ในระบบการค้นคืนสารสนเทศจริง

