บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง (Review of Literature)

2.1 คุณลักษณะของภาษาไทย

ภาษาไทยจัดเป็นภาษาคำโดด (Isolative language) (Vongvipanond, 1993, p. 519-545) ภาษาในกลุ่มนี้เป็นภาษาที่ไม่มีการเปลี่ยนรูปเพื่อบอกไวยากรณ์ ลักษณะการเขียนของภาษาไทยจะถูกเขียนเป็นบท (Script) ด้วยตัวอักษรต่างๆ แบ่งออกได้เป็นพยัญชนะ สระ วรรณยุกต์ และอักษรพิเศษ ดังแสดงในภาพที่ 2.1

ภาพที่ 2.1
ตัวอักษรในภาษาไทย

พยัญชนะ (44)
ก ข ฃ ค ฅ ฆ ง จ ฉ ช ซ ฌ ญ ฎ ฏ ฐ ฑ ฒ ณ ฒ ณ
ด ต ถ ท ธ น บ ป ผ ฝ พ ฟ ภ ม ย ร ล ว ศ ษ ส ห ฬ อ ฮ

สระ (16)
เ แ โ ไ ใ ะ า ย ว อ
อิ อี อึ อื อุ อู

อักษรพิเศษ (4) ฯ ฯๆ ฤ ฤๅ

วรรณยุกต์ (4) เอก โท ตรี จัตวา

อักษรอื่นๆ (4)
เครื่องหมายแสดงการซ้ำ ๆ
เครื่องหมายแสดงการย่อ ฯ
เครื่องหมายแสดงรายการตัวอย่าง ฯลฯ

ที่มา: Linguistic Problems in Computer Processing of the Thai Language (Vongvipanond, 1993, p. 519-545)

ตัวอักษรในภาพที่ 2.1 ถูกเขียนเป็น 4 ระดับ กล่าวคือระดับ 1 เป็นตำแหน่งของวรรณยุกต์ ระดับ 2 เป็นตำแหน่งของวรรณยุกต์และสระที่เขียนด้านบน ระดับ 3 เป็นตำแหน่งของพยัญชนะและสระส่วนใหญ่ ส่วนระดับ 4 เป็นตำแหน่งของสระที่อยู่ด้านล่าง สระอาจถูกเขียนได้ทั้งก่อนหน้า ตามหลัง ข้างบน และด้านล่างของพยัญชนะ

การรู้จำคำภาษาไทยเป็นปัญหาหนึ่งในการประมวลผลภาษาไทย เนื่องจากภาษาไทยไม่มีช่องระหว่างคำในการเขียนเหมือนภาษาอังกฤษ ดังนั้นจึงอาจเกิดความกำกวมในการสื่อความหมายขึ้นได้ดังแสดงในภาพที่ 2.2

ภาพที่ 2.2
ความเป็นไปได้ในการตัดคำภาษาไทย

ชุดของตัวอักษร: กรรมการกรมพลศึกษารอยกร่าง
การแบ่งพยางค์แบบที่ 1: กร รม กา รก รม พล ศึก ษา รอย กร่าง
การแบ่งพยางค์แบบที่ 2: กรรม การ กรม พล ศึก ษา รอ ยก ร่าง

การตัดคำแบบที่ 1: กรรม การ กรม พล ศึกษา รอ ยก ร่าง
การตัดคำแบบที่ 2: กรรมการ กรม พล ศึกษา รอย กร่าง
การตัดคำแบบที่ 3: กรรมการ กรม พลศึกษา รอย กร่าง

การตัดคำทางขวาแบบที่ 1: กรรม การ กรม พลศึกษา รอ ยก ร่าง
การตัดคำทางขวาแบบที่ 2: กรรมการ กรมพลศึกษา รอ ยก ร่าง

ที่มา: Linguistic Problems in Computer Processing of the Thai Language (Vongvipanond, 1993, p. 519-545)

ภาพที่ 2.2 เห็นได้ว่าการแบ่งพยางค์และการตัดคำไม่จำเป็นต้องได้คำที่มีความหมายเสมอไป สิ่งที่ระบบคอมพิวเตอร์ต้องการคือความรู้ในความหมายของคำเหมือนมนุษย์ ดังนั้นการ

ตัดคำ (Word segmentation) จึงเป็นปัญหาหนึ่งที่สำคัญในการจัดการกับภาษาไทยในระบบคอมพิวเตอร์

2.2 การตัดคำ
(Word Segmentation)

ดังที่กล่าวแล้วข้างต้นการเขียนภาษาไทยจะเป็นลักษณะการเขียนที่ต่อเนื่องกันไป ไม่สามารถหาขอบเขตของคำได้ ซึ่งแตกต่างกับภาษาอังกฤษที่มีช่องว่างในการแบ่งคำแต่ละคำอย่างชัดเจน ดังนั้นงานวิจัยในด้านการตัดคำภาษาไทยจึงมีเป้าหมายที่ต้องหาขอบเขตของคำในภาษาไทยเพื่อให้มีความถูกต้องมากที่สุด มีงานวิจัยเกี่ยวกับการตัดคำอยู่หลายเรื่อง เช่น การตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) (วิรัช, 2536), วิธีการตัดคำแบบคุณลักษณะ (Feature - Based Approach) (Meknavin, Charoenpornsawat and Kijsirikul, 1997) เป็นต้น ตารางที่ 2.1 จะแสดงให้เห็นความเป็นไปได้หากใช้โปรแกรมที่มีอัลกอริธึมการตัดคำที่แตกต่างกันจะส่งผลให้ผลลัพธ์ของการตัดคำนั้นแตกต่างกันไป

ตารางที่ 2.1
ตัวอย่างผลลัพธ์ของการตัดคำที่ใช้อัลกอริธึมที่แตกต่างกัน

โปรแกรม	ผลลัพธ์ของการตัดคำ
A	ดวง | อาทิตย์ | ขึ้น | ทาง | ทิศ | ตะวัน | ออก
B	ดวง | อาทิตย์ | ขึ้น |ทาง | ทิศตะวันออก
C	ดวงอาทิตย์ | ขึ้น | ทาง | ทิศ | ตะวัน | ออก
D	ดวงอาทิตย์ | ขึ้น | ทาง | ทิศตะวันออก
…	...

ที่มา : Thai Automatic Indexing using Latent Semantic Indexing (Tunpaiboon, 2000)

ในงานวิจัยนี้ได้นำโปรแกรมตัดคำที่ชื่อว่า Smart Word Analysis for Thai - SWATH โดยทางศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) โปรแกรมตัดคำ

SWATH ใช้อัลกอริธึมในการตัดคำภาษาไทย 3 แบบ กล่าวคือ Longest Matching, Maximal Matching และ Part-Of-Speech Bigram โดยที่สามารถสนับสนุนแฟ้มข้อมูลได้ทั้งแบบ html, rtf, LaTeX และ plain text สามารถหาโปรแกรมนี้ได้ที่ http://www.links.nectec.or.th และ http://www.cs.cmu.edu/~paisarn/software.html)

เนื่องจากงานวิจัยนี้เลือกใช้การตัดคำวิธีการเทียบคำที่ยาวที่สุด (Longest Matching) (Paisarn,1999, Meknavin, Charoenpornsawat and Kijsirikul,1997) จึงจะขอกล่าวถึงเฉพาะการทำงานของอัลกอริธึมการเทียบคำที่ยาวที่สุดที่ประมวลผลโดย เริ่มทำการอ่านข้อความเข้ามาแล้วนำไปเทียบกับพจนานุกรม ถ้าไม่พบข้อความทั้งหมดในพจนานุกรมจะทำการลดอักษรของข้อความลงทีละตัวอักษรและทำเครื่องหมายเพื่อเป็นจุดย้อนกลับ (Backtracking) แล้วนำข้อความที่ลดตัวอักษรไปค้นหาในพจนานุกรม ถ้าพบก็ให้เริ่มทำงานใหม่จากจุดย้อนกลับเพื่อตรวจสอบตัวอักษรที่เหลือในข้อความนั้นว่าสามารถตัดเป็นคำที่มีอยู่ในพจนานุกรมได้หรือไม่ และทำงานเช่นนี้ไปเรื่อยๆ จนจบข้อความที่อ่านเข้ามา ตัวอย่างเช่น การแบ่งคำในข้อความ “ฉันนั่งตากลมที่หน้าบ้าน” ข้อความแรกที่นำไปเทียบคือ “ฉันนั่งตากลมที่หน้าบ้าน” ปรากฏว่าไม่มีในพจนานุกรม ให้ทำการลดตัวอักษรของข้อความ จะได้ “ฉันนั่งตากลมที่หน้าบ้า” แล้วนำไปเทียบอีกครั้ง ก็ไม่มี ก็ให้ลดตัวอักษรของข้อความไปเรื่อย ๆ จนข้อความเหลือคำ “ฉัน” ซึ่งเป็นข้อความที่มีในพจนานุกรม จากจุดนี้จะเป็นจุดย้อนกลับ ก็เริ่มทำการเทียบข้อความที่เหลือ คือ “นั่งตากลมที่หน้าบ้าน” ทำเหมือนเดิมจนกว่าจะจบข้อความ จะได้คำดังต่อไปนี้ “นั่ง” “ตาก” “ลม” “ที่” “หน้า” “บ้าน” สำหรับข้อด้อยของอัลกอริธึมนี้เกิดจากลักษณะความพยายามที่จะตัดคำให้ได้ยาวที่สุด(Poowarawan, 1986) เช่น “ไปหามเหสี” จะตัดได้ “ไป หาม เห สี” เพราะอัลกอริธึมนี้จะเลือกคำว่า “หาม” เพราะเลือกคำที่ยาวที่สุด ทำให้มีผลทำให้การตัดคำที่ตามมาผิดเพี้ยน

2.3 การค้นคืนสารสนเทศ
(Information Retrieval)

การค้นหาสารสนเทศให้ตรงกับความต้องการของผู้ใช้เป็นปัญหาที่สำคัญปัญหาหนึ่งในงานวิจัยการค้นคืนสารสนเทศเนื่องจากขนาดของฐานข้อมูลเอกสารมีมากขึ้น โดยมีการเก็บข้อมูลทั้งในรูปแบบ ข้อความ ภาพ และวิดีโอ นอกจากนั้นการค้นคืนสารสนเทศยังไม่สามารถค้นหาสารสนเทศได้ในเชิงความหมาย งานวิจัยทางด้านการค้นคืนสารสนเทศได้มีการศึกษาพัฒนารูปแบบการค้นคืนหลายรูปแบบแต่สามสารถจัดเป็นกลุ่มใหญ่ได้ 2 กลุ่ม ดังนี้ รูปแบบการค้นคืนสารสนเทศเชิงคำหลัก (Keyword-oriented technique) และรูปแบบการค้นคืนสารสนเทศ

เชิงแมทริกซ์ (Matrix-oriented technique) (Ye, 2000) รูปแบบการค้นคืนสารสนเทศที่นิยมใช้กันมาก่อน คือ เทคนิคเชิงคำหลัก จะทำการสร้างคำดรรชนี (Index term) เพื่อการค้นคืนเอกสาร คำดรรชนีหมายถึงคำหลักหรือกลุ่มของคำหลักที่สัมพันธ์กันที่มีความหมายในตัวของมันเอง คำดรรชนีเป็นคำซึ่งปรากฏในเอกสารในชุดเอกสาร การค้นคืนสารสนเทศโดยใช้เทคนิคเชิงคำหลักสามารถทำได้ง่าย แต่ก่อให้เกิดปัญหาสำคัญในการค้นคืน (Baeza-Yates and Reibeiro-Neto, 1999) เนื่องจากคำหนึ่งคำอาจมีได้หลายความหมาย (Polysemy) ในขณะเดียวกัน คำหลายคำที่ต่างกันอาจมีความหมายเหมือนกัน (Synonymy) ก็ได้

ตัวอย่างสำหรับคำภาษาไทย เช่น คำว่า “ที่” สามารถเป็นได้ทั้งคำนาม คำแยกประเภท และคำบุพบท ถ้าเป็นคำนามหมายถึงพื้นที่ ในฐานะคำแยกประเภทหมายถึงส่วนย่อย และในฐานะคำบุพบทหมายถึงตำแหน่ง นอกจากนี้ยังปรากฏในฐานะคำเชื่อมประโยคหรือเครื่องหมายของประโยคย่อย (Vongvipanond, 1993, p. 519-545)

คำที่สามารถมีได้หลายความหมาย (Polysemy) และคำหลายคำที่สามารถใช้แทนความหมายเดียวกัน (Synonymy) เป็นสองปัญหาหลักที่การทำดรรชนีเพื่อหาความหมายต้องพยายามแก้ไข Polysemyทำให้เราค้นคืนได้เอกสารที่ไม่เกี่ยวข้องออกมา ในขณะที่ Synonymy ทำให้เราเสียเอกสารที่เกี่ยวข้องไป ตัวอย่างเช่น การใช้คำว่า ‘‘บ้าน’’ เป็นข้อคำถามอาจทำให้พลาดเอกสารอื่นที่มีคำว่า ‘‘เรือน’’ ‘‘ที่พัก’’ ‘‘ที่อาศัย’’ ไป หรือ การใช้คำว่า ‘‘ที่’’ เป็นข้อคำถามอาจค้นคืนได้เอกสารที่มีคำนี้ทั้งในความหมายของ ‘‘ที่ดิน’’ และเอกสารที่มีคำนี้เป็นคำเชื่อมประโยคหรือบุพบทอยู่ก็ได้ เป็นต้น

2.4 แบบจำลองเวกเตอร์ปริภูมิ (Vector Space Model)

แบบจำลองเวกเตอร์ปริภูมิ (Vector Space Model หรือ VSM) เป็นหนึ่งในเทคนิคการค้นคืนสารสนเทศด้วยวิธีการเชิงแมทริกซ์ ในแบบจำลองนี้จะประกอบไปด้วยเวกเตอร์ซึ่งเวกเตอร์แต่ละเวกเตอร์จะแทนเอกสารแต่ละเอกสารในชุดเอกสาร สมาชิกแต่ละตัวในเวกเตอร์แสดงถึงคำหลักหรือคำที่สัมพันธ์กับเอกสาร ค่าของสมาชิกที่กำหนดให้กับเอกสารจะบ่งถึงความสำคัญของคำในการแสดงความหมายของเอกสารนั้น โดยปกติค่าของสมาชิกในเวกเตอร์นั้นเป็นฟังก์ชันของความถี่ของคำที่ปรากฏในเอกสารหรือในชุดเอกสารทั้งหมด

ดังที่กล่าวข้างต้น ฐานข้อมูลของชุดเอกสารที่มีเอกสาร *d* ชุด และมีคำ *t* คำถูกสร้างเป็นแมทริกซ์คำ-เอกสาร *(t x d term-by-document matrix)* ชุดของเวกเตอร์ *d* แทนเอกสารด้วยสดมภ์ของแมทริกซ์ สมาชิกของแมทริกซ์ (a_{ij}) หมายถึงค่าความถี่ของคำซึ่งเป็นค่าน้ำหนักที่คำ *i* ปรากฏในเอกสาร *j* (Berry , Dumais and O'Brien 1995, p. 573-595) กล่าวโดยสรุป สดมภ์ของแมทริกซ์คำ-เอกสารเป็นเวกเตอร์ของเอกสาร ส่วนแถวของแมทริกซ์คำ-เอกสารเป็นเวกเตอร์ของคำที่ปรากฏในชุดเอกสาร บางเวกเตอร์ที่อยู่ในสดมภ์ของแมทริกซ์อาจมีความหมายเฉพาะในชุดเอกสารนั้น แมทริกซ์ที่เป็นตัวแทนของเอกสารจะสามารถถูกนำไปประมวลผลเพื่อหาความสัมพันธ์ระหว่างเวกเตอร์ของเอกสารในแง่ของความหมายหรือความแตกต่างของเนื้อหา นอกจากนั้นยังสามารถนำไปคำนวณเพื่อทำการเปรียบเทียบเวกเตอร์ของคำด้วยวิธีการทางเรขาคณิตเพื่อแสดงความเหมือนหรือความต่างในการใช้คำในเอกสารได้อีกด้วย

ผู้ใช้สามารถค้นคืนเอกสารจากระบบการค้นคืนสารสนเทศโดยการป้อนข้อคำถาม (Query) เข้าไปในระบบ ข้อคำถามของผู้ใช้ประกอบด้วยชุดของคำซึ่งอาจพิจารณาให้ค่าน้ำหนักไว้เสมือนเป็นเวกเตอร์ของเอกสารฉบับหนึ่ง สมาชิกส่วนใหญ่ของเวกเตอร์ข้อคำถามมีค่าเท่ากับศูนย์ ในการค้นหาเอกสารที่เกี่ยวข้องจะสามารถกระทำได้ด้วยวิธีการจับคู่ข้อคำถาม (Query matching) ซึ่งเป็นวิธีการการค้นหาเอกสารที่เกี่ยวข้องที่เหมือนกับเวกเตอร์ข้อคำถามมากที่สุด ในแบบจำลองเวกเตอร์ปริภูมิสามารถคำนวณหาค่าความเหมือน (Similarity) ได้ด้วยวิธีการทางเรขาคณิตเพื่อหาเอกสารที่ใกล้เคียงกับข้อคำถามมากที่สุด

วิธีการหาค่าความเหมือน (Similarity) ที่นิยมมากวิธีหนึ่งคือ การหาค่าโคไซน์ของมุมระหว่างเวกเตอร์ของเอกสารและเวกเตอร์ของข้อคำถาม ถ้าแมทริกซ์คำ-เอกสาร (A) มีสดมภ์ a_j โดยที่ j =1,…,d ค่าโคไซน์สามารถหาได้จากสมการ 2.1 :

$$\cos\theta_j = \frac{a^T q}{\|a_j\|_2 \|q\|_2} = \frac{\sum_{i=1}^{t} a_{ij} q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^2}\sqrt{\sum_{i=1}^{t} q_i^2}} \qquad (2.1)$$

โดยที่

$\cos\theta_j$ หมายถึง ค่าโคไซน์ของมุมระหว่างเวกเตอร์ของเอกสารและเวกเตอร์ของข้อคำถาม

a หมายถึง เวกเตอร์ของเอกสาร

q หมายถึง เวกเตอร์ของข้อคำถาม

i หมายถึง ตำแหน่งของแถวในแมทริกซ์เอกสาร

j หมายถึง ตำแหน่งของสดมภ์ในแมทริกซ์เอกสาร

$\|x\|_2$ หมายถึง ค่าประจำเวกเตอร์ ซึ่งสามารถคำนวณได้จากสมการ 2.2

$$\|x\|_2 = \sqrt{x^T x} = \sum_{i=1}^{t} x_i^2 \qquad (2.2)$$

ค่าความเหมือนที่คำนวณได้จะถูกนำไปเปรียบเทียบกับค่าขีดแบ่งความเหมือน (Threshold value) ที่กำหนดไว้ ถ้าค่าความเหมือนของเอกสารมากกว่าค่าขีดแบ่งความเหมือนจะถือว่าเอกสารนั้นถูกค้นคืนขึ้นมาได้

2.5 การจัดทำดรรชนีหาความหมายแอบแฝง (Latent Semantic Indexing)

การจัดทำดรรชนีหาความหมายแอบแฝง (Latent Semantic Indexing หรือ LSI) เป็นเทคนิคที่พัฒนาต่อมาจากแบบจำลองเวกเตอร์ปริภูมิโดยการใช้การประมาณการเพื่อลดค่าลำดับชั้นของแมทริกซ์สำหรับเวกเตอร์ปริภูมิของเอกสารทั้งหมด นั่นคือการกระจายแมทริกซ์เดิมเป็นแมทริกซ์ย่อยหลายแมทริกซ์ที่ใกล้เคียงกับแมทริกซ์เดิมเท่าที่จะเป็นไปได้ การจัดทำดรรชนีหาความหมายแอบแฝงถูกนำเสนอครั้งแรกโดย S.T. Dumains ในปี 1998 เพื่อแก้ปัญหา Polysemy และ Synonymy

วิธีการที่นิยมใช้ในการจัดทำดรรชนีหาความหมายแอบแฝงคือ การกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว (Singular Value Decomposition หรือ SVD) ซึ่งถูกใช้เพื่อประมาณค่าลำดับชั้น (Rank) ขนาด k ที่ดีที่สุดของแมทริกซ์คำ-เอกสารเพื่อให้ทำการดึงความหมายของคำที่แฝงอยู่ในเอกสารออกมาได้ดีที่สุด สดมภ์ของแมทริกซ์คำเอกสารใหม่ที่ได้ทำการลดค่าลำดับชั้น (Rank) แล้วจะแสดงความหมายแอบแฝง ซึ่งเชื่อว่าเก็บความหมายแอบแฝงของเอกสารไว้มากกว่าแมทริกซ์ของแบบจำลองเวกเตอร์ปริภูมิ

2.6 การกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Value Decomposition)

การแยกค่าแบบเดี่ยว (Singular Value Decomposition) เป็นกระบวนการทางคณิตศาสตร์ สำหรับทำการกระจายแมทริกซ์ออกเป็นผลคูณของแมทริกซ์ย่อย โดยที่แมทริกซ์ย่อย

เหล่านี้สามารถแสดงให้เห็นถึงความสัมพันธ์หรือเนื้อหาที่มีอยู่ในแมทริกซ์เดิมและความสัมพันธ์หรือเนื้อหาเหล่านี้ยังสามารถแบ่งตามระดับความสำคัญได้ จากความสามารถดังกล่าวนี้จึงนิยมใช้วิธีการแยกค่าแบบเดี่ยวเป็นตัวลดขนาดข้อมูลหรือใช้เป็นตัวที่ดึงความสัมพันธ์ของแมทริกซ์ออกมา

ส่วนนิยามทางคณิตศาสตร์ของการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวในการกระจายแมทริกซ์คำ-เอกสาร (A) เป็นสามแมทริกซ์ย่อย (Berry, Drmac and Jessup, 1999) สามารถทำได้โดยใช้สมการ 2.3

$$A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \qquad (2.3)$$

โดยที่ $U \in \Re^{mxn}$ และ $V \in \Re^{nxn}$ มีสดมภ์เชิงตั้งฉาก (Orthogonal column) และ Σ เป็นแมทริกซ์เอกฐานที่ประกอบด้วยสมาชิก $\sigma_1, \sigma_2, ...\sigma_n$ และมีค่าของ $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r > 0 = \sigma_{r+1} = ... = \sigma_n$ ภาพที่ 2.3 แสดงให้ถึงลักษณะของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยว

ภาพที่ 2.3
ผลลัพธ์ของการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว

ส่วนภาพที่ 2.4 แสดงให้เห็นตัวอย่างของการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวของแมทริกซ์เอกสารเป็นแมทริกซ์ย่อย สมาชิกในแมทริกซ์ย่อยทั้งสามแมทริกซ์มีค่าเป็นเลข

ทศนิยมที่มีความเที่ยงสองเท่า (double precision number) (Kolda and O'Leary, 1998, p. 322-346)

ภาพที่ 2.4

ตัวอย่างผลลัพธ์ของการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว

$$\begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 0 & 2 \end{bmatrix}$$

$$\begin{bmatrix} -0.8213 & -0.5704 \\ -0.2551 & 0.3673 \\ -0.5102 & 0.7346 \end{bmatrix} \begin{bmatrix} 3.8287 & 0 \\ 0 & 0.5840 \end{bmatrix} \begin{bmatrix} -0.2145 & -0.9767 \\ -0.9767 & 0.2145 \end{bmatrix}$$

ถึงแม้ว่าการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวจะมีความแน่นอนและคุณสมบัติในการประมาณค่าที่ดีจนเป็นที่นิยมในการจัดทำดรรชนีหาความหมายแอบแฝง แต่จุดอ่อนของวิธีการนี้คือ ความซับซ้อนในการประมวลผลสูง ทำให้การประมวลผลในแมทริกซ์ที่ขนาดใหญ่มากเป็นไปได้ยาก และที่สำคัญที่สุดก็คือการเพิ่มเติมเปลี่ยนแปลงเอกสาร การกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวทำได้ยากมาก เนื่องจากต้องการคำนวณต้องรักษาความเป็นแมทริกซ์เชิงตั้งฉาก (Orthogonal matrix) ของแมทริกซ์ย่อยไว้ จนต้องการการประมวลผลแมทริกซ์คำ-เอกสารใหม่ทั้งหมด

A

2.7 การกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีต (Semidiscrete Matrix Decomposition)

การกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีต (Semidiscrete Matrix Decomposition หรือ SDD) ถูกนำเสนอครั้งแรกโดย O'Leary และ Peleg (O'Leary and Peleg, 1983, p.441-444) ในการบีบอัดภาพดิจิตอล ต่อมา Kolda และ O'Leary (Kolda and O'Leary, 1998, p. 322-346) ได้นำเทคนิคนี้มาใช้สำหรับการจัดทำดรรชนีเพื่อหาความหมายแอบแฝงในการค้นคืนสารสนเทศ

A =

U

การกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตกระจายแมทริกซ์ (Ak) ขนาด $m \; x \; n$ ตามสมการที่ 2.4 (Kolda and O'Leary, 1998, p. 322-346) สามารถแสดงได้ดังนี้

$$A_k = \underbrace{\begin{bmatrix} x_1 & x_2 & . & . & . & x_k \end{bmatrix}}_{X_k} \underbrace{\begin{bmatrix} d_1 & 0 & . & . & . & 0 \\ 0 & d_2 & & & & 0 \\ . & & . & & & . \\ . & & & . & & . \\ . & & & & . & . \\ 0 & 0 & . & . & . & d_k \end{bmatrix}}_{D_k} \underbrace{\begin{bmatrix} y_1^T \\ y_2^T \\ . \\ . \\ . \\ y_k^T \end{bmatrix}}_{Y_k^T} = \sum_{i=1}^{k} d_i x_i y_i^T \qquad (2.4)$$

เมื่อ x_i เป็นเวกเตอร์ขนาด m (m-vector) ซึ่งสมาชิกเป็นเลขจำนวนเต็มที่มีค่าในเซต {-1, 0, 1} , y_i เป็นเวกเตอร์ขนาด n (n-vector) ที่มีสมาชิกเป็นเลขจำนวนเต็มที่มีค่าในเซต {-1, 0, 1} เช่นเดียวกัน .ในขณะที่ d_i เป็นจำนวนที่เป็นบวก สมการที่ 2.4 แสดงการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตที่มีค่าลำดับชั้น (Rank) ขนาด k (k-term SDD)

ภาพที่ 2.5 แสดงตัวอย่างผลลัพธ์ของการการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีต สมาชิกในแมทริกซ์ X และ Y เป็นจำนวนเต็มที่มีค่าในเซต {-1, 0, 1} ส่วนแมทริกซ์ D มีสมาชิกที่เป็นจำนวนทศนิยมที่มีความเที่ยงหนึ่งเท่า (single precision number) (Kolda and O'Leary, 1998, p. 322-346)

ภาพที่ 2.5

ตัวอย่างผลลัพธ์ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีต

$$A \quad \begin{bmatrix} 1 & 3 \\ 0 & 1 \\ 0 & 2 \end{bmatrix}$$

$$A = \underbrace{\begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} 2.5000 & 0 \\ 0 & 0.6250 \end{bmatrix}}_{D} \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}}_{Y}$$

ตารางที่ 2.2

เปรียบเทียบพื้นที่เก็บข้อมูลระหว่างการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตและการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวที่มีค่าลำดับชั้น (Rank) ขนาด *k* ของแมทริกซ์ขนาด *m x n*

วิธีการ	ส่วนประกอบ		ขนาด (Bytes)
SVD	*U*	km double-precision numbers	
	V	kn double-precision numbers	8k(m+n+1)
	Σ	k double-precision numbers	
SDD	*X*	km numbers from {-1,0,1}	
	Y	kn numbers from {-1,0,1}	4k+1/4(m+n)
	D	k single precision numbers	

ที่มา: A Semidiscrete Matrix Decomposition for Latent Semantic Indexing in Information Retrieval (Kolda and O'Leary, 1998, p. 322-346)

การคำนวณความต้องการพื้นที่สำหรับการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตที่มีค่าลำดับชั้น (Rank) ขนาด k (k-term SDD) สามารถทำได้ (Kolda and O'Leary, 1998, p. 322-346) โดยคำนวณ k(m+n) จากเลขจำนวนเต็มที่มีค่า -1, 0, 1 ในแมทริกซ์ X และ Y รวมกับจำนวนสเกลาร์ k จำนวน จากแมทริกซ์ D เนื่องจากจำนวนสเกลาร์ของการกระจายแมทริกซ์ด้วยวิธีเซมิดีสครีตใช้ข้อมูลชนิดจำนวนทศนิยมความเที่ยงหนึ่งเท่า (Single precision) ส่วนการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวใช้ข้อมูลจำนวนทศนิยมความเที่ยงสองเท่า (Double precision) ถ้าข้อมูลชนิดจำนวนทศนิยมความเที่ยงสองเท่าใช้พื้นที่เก็บข้อมูลขนาด 8 ไบต์ และข้อมูลชนิดจำนวนทศนิยมความเที่ยงหนึ่งเท่าใช้พื้นที่เก็บข้อมูลขนาด 4 ไบต์ ส่วนข้อมูลเลขจำนวนเต็มใช้พื้นที่เก็บข้อมูลขนาด 1 ไบต์ โดยให้ 8 บิต เท่ากับ 1 ไบต์แล้ว ตารางที่ 2.2 แสดงการเปรียบเทียบพื้นที่เก็บข้อมูลระหว่างการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตและการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวที่มีค่าลำดับชั้น (Rank) ขนาด *k* โดยวิธีการดังกล่าวข้างต้น

ความแตกต่างที่สำคัญระหว่างการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตและการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวคือการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตจะไม่ใช่แมทริกซ์เชิงตั้งฉาก (Orthogonal matrix) อีกต่อไป จึงทำให้การกระจายแมทริกซ์เมื่อมีการ

เปลี่ยนแปลงชุดเอกสารทำได้ง่ายกว่ากระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยวมาก (Ye, 2000)

เนื่องจากการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตมีแมทริกซ์ย่อยถึงสองแมทริกซ์ ซึ่งมีสมาชิกเป็นเลขจำนวนเต็มจึงทำให้ใช้พื้นที่ในการเก็บข้อมูลน้อยกว่า แมทริกซ์จึงมีขนาดเล็ก ทำให้ใช้เวลาน้อยลงในการค้นคืนสารสนเทศ (Query time) ในขณะเดียวกันผลการค้นคืนเอกสารทำได้ใกล้เคียงกับระบบที่มีดรรชนีหาความหมายแอบแฝงที่ใช้การกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Kolda and O'Leary, 1998, p. 322-346) ดังนั้นทำให้การกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตจึงเหมาะสมสำหรับการค้นคืนเอกสารข้อมูลที่มีขนาดใหญ่ อย่างไรก็ตามการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีตก็มีจุดอ่อนเหมือนกันคือ การกระจายแมทริกซ์ด้วยวิธีนี้ใช้เวลาประมวลผลในการกระจายแมทริกซ์มากกว่าการกระจายแมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว เนื่องจากอัลกอริธึมในการกระจายแมทริกซ์ของวิธีเซมิดีสครีตใช้เวลามากกว่า

2.8 ผลงานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้อความภาษาไทย

งานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศเชิงหาความหมายในข้อความภาษาไทยมีผู้เสนอเทคนิคหลายวิธี ได้แก่ Universal Word (UW) (Sornlertlamvanich, Potipiti and Charoenporn, 2000), Singular Value Decomposition (SVD) (Tunpaiboon, 2000; Pongpinigpinyo and Rivepiboon, 2005; Tangthai , 2003; Premkusolchai, 2005) ตารางที่ 2.3 แสดงการสรุปงานวิจัยต่างๆที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้อความภาษาไทยเชิงหาความหมายด้วยวิธีการกระจายแมทริกซ์

ถึงแม้จะมีงานวิจัยที่ได้ทำการวิเคราะห์เพื่อหาความหมายแอบแฝง และการจัดทำดรรชนีเพื่อหาความหมายแอบแฝงในภาษาไทย แต่ยังไม่มีงานวิจัยที่ได้นำการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีต (SDD) มาใช้ประยุกต์ ซึ่งเป็นเทคนิคการกระจายแมทริกซ์ที่มีข้อดีในแง่ของการพื้นที่ในการประมวลผล เวลาในการค้นคืน ประสิทธิภาพของการค้นคืน และความง่ายในการเปลี่ยนแปลงเอกสาร ดังนั้นงานวิจัยนี้จึงเป็นการเริ่มต้นที่จะนำการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีต (SDD) มาใช้ในการค้นคืนสารสนเทศข้อความภาษาไทยเพื่อเป็นการทดสอบหาประสิทธิภาพในการค้นคืนสารสนเทศเชิงความหมายที่มีการกระจายแมทริกซ์ด้วยวิธีการเซมิดีสครีต จะให้ผลการค้นคืนข้อความภาษาไทยได้แตกต่างการกระจายแมทริกซ์ด้วยวิธีแยกค่าแบบเดี่ยวหรือไม่

ตารางที่ 2.3

งานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้อความภาษาไทยเชิงหาความหมายด้วยวิธีการกระจายแมทริกซ์

ผู้วิจัย	เรื่อง/ปี	ประเด็นในการศึกษา	ข้อมูลทดสอบ	การวัดผลการทดลอง	ผลการทดลอง
ธีระเดช ตันไพบูลย์	Thai Automatic Indexing using Latent Semantic Indexing/ 2000	การค้นคืนโดยไม่ได้เอกสารตามต้องการ	สารานุกรมไทยสำหรับเยาวชน	-ความแม่นยำ (Precision) -การเรียกกลับ (Recall) -ค่าความเหมือน (Similarity)	การเรียกกลับ เพิ่มขึ้นเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์เพิ่มขึ้นแต่ลดลงเมื่อค่าขีดแบ่งความเหมือนเพิ่มขึ้น ในขณะที่ความแม่นยำเพิ่มขึ้นเมื่อค่าขีดแบ่งความเหมือนเพิ่มขึ้นแต่ลดลงเมื่อค่าลำดับชั้น (Rank) ของแมทริกซ์เพิ่มขึ้น
ศุภวรรณ เปรมกุศลชัย	การประเมินคุณภาพการถอดความภาษาไทย	การถอดความภาษาไทย	บทคัดย่อของเอกสารเกี่ยวกับ E-Learning	สัมประสิทธิ์สหสัมพันธ์ด้วย Spearman’s rank ของค่าความเหมือนระหว่างเอกสาร	การวิเคราะห์เพื่อหาความหมายแอบแฝงสามารถใช้ในการประเมินคุณภาพการถอดความภาษาไทยได้

ตารางที่ 2.3 (ต่อ)

งานวิจัยที่เกี่ยวข้องกับการค้นคืนสารสนเทศข้อความภาษาไทยเชิงหาความหมายด้วยวิธีการกระจายแมทริกซ์

ผู้วิจัย	เรื่อง/ปี	ประเด็นในการศึกษา	ข้อมูลทดสอบ	การวัดผลการทดลอง	ผลการทดลอง
อัษฎางค์ แตงไทย	การย่อความเอกสารภาษาไทยโดยการวิเคราะห์หาความหมายที่แอบแฝง	การย่อความเอกสารภาษาไทย	เอกสารข่าวภาษาไทย	-ความแม่นยำ (Precision) -การเรียกกลับ (Recall) -ค่าความเหมือน (Similarity)	ระบบสามารถวิเคราะห์ความสามารถในการย่อความเอกสารภาษาไทยของเอกสารที่ใช้ในการทดสอบได้