



รายงานการวิจัยฉบับสมบูรณ์

โครงการ “ระบบแนะนำหนังสืออ่านเพิ่มเติมเพื่อส่งเสริมการอ่านและการศึกษา
ด้วยตนเอง (Supplementary books suggestion system for promoting
self- reading and self-learning)”

คณะผู้วิจัย

ผู้ช่วยศาสตราจารย์ ดร.โกเมศ อัมพวัน	หัวหน้าโครงการ
รองศาสตราจารย์ ดร.อนุชิต จิตพัฒนกุล	ผู้ร่วมวิจัย
ดร. วรารุณี ผ้าเจริญ	ผู้ร่วมวิจัย

โครงการวิจัยประเภทงบประมาณเงินรายได้

(เงินอุดหนุนจากรัฐบาล) ประจำปีงบประมาณ พ.ศ. 2561

มหาวิทยาลัยบูรพา

รหัสโครงการ 256107A1080013

สัญญาเลขที่ 16/2561

รายงานการวิจัยฉบับสมบูรณ์

โครงการ “ระบบแนะนำหนังสืออ่านเพิ่มเติมเพื่อส่งเสริมการอ่านและการศึกษา
ด้วยตนเอง (Supplementary books suggestion system for promoting
self- reading and self-learning)”

คณะผู้วิจัย

ผู้ช่วยศาสตราจารย์ ดร.โกเมศ อัมพวัน	หัวหน้าโครงการ*
รองศาสตราจารย์ ดร.อนุชิต จิตพัฒนกุล	ผู้ร่วมวิจัย**
ดร. วราวุฒิ ฟ้าเจริญ	ผู้ร่วมวิจัย***

*ห้องปฏิบัติการวิจัยนวัตกรรมการประมวลผล คณะวิทยาการสารสนเทศ

มหาวิทยาลัยบูรพา

**คณะวิทยาศาสตร์ประยุกต์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

***คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชธานี

กันยายน 2562

บทคัดย่อ

ในปัจจุบันมีระบบค้นหาหนังสือและระบบแนะนำหนังสือต่าง ๆ มากมาย เพื่อช่วยให้ผู้อ่านมีความสะดวกในการค้นหาและเลือกหนังสือที่ตนเองสนใจได้ดียิ่งขึ้น อาทิเช่น การค้นหาหนังสือจากคุณสมบัติของคำค้นหนึ่ง ๆ กับชื่อหนังสือหรือสารบัญหนังสือ การแนะนำหนังสือจากความนิยมของหนังสือ เป็นต้น แต่อย่างไรก็ตามแนวคิดทั้งหมดไม่ได้มุ่งเน้นที่การแนะนำหนังสือจากคุณสมบัติของหนังสือที่เกี่ยวข้องกับการศึกษาในชั้นเรียน และไม่ได้มุ่งเน้นในการแนะนำหนังสือที่เกี่ยวข้องกับศาสตร์ที่มีคำศัพท์เฉพาะทางค่อนข้างมากกับคำค้น ที่ซึ่งช่วยให้ผู้อ่านได้รับการแนะนำที่สอดคล้องกับเนื้อหาที่ตนเองสนใจมากยิ่งขึ้น

งานวิจัยนี้ได้นำเสนอระบบแนะนำหนังสืออ่านเพิ่มเติม (Supplementary Books Suggestion system หรือ SBS system) เพื่อทำการแนะนำหนังสือหนังสืออ่านเพิ่มเติมที่มีความเกี่ยวข้องกับเนื้อหาของคำอธิบายรายวิชาในหลักสูตรวิทยาการคอมพิวเตอร์ ประกอบไปด้วย 5 ขั้นตอนคือ 1) การรวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ 2) การรวบรวมข้อมูลคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ 3) การสกัดคำสำคัญจากคำอธิบายรายวิชาและการสกัดคำสำคัญจากหนังสือ 4) การค้นหาหนังสือที่มีความเกี่ยวข้องกับคำสำคัญในคำอธิบายรายวิชา และ 5) การคำนวณความเกี่ยวข้องของหนังสือที่ครอบคลุมคำสำคัญในคำอธิบายรายวิชา การจัดอันดับหนังสือที่มีความเกี่ยวข้องมากไปยังหนังสือที่มีความเกี่ยวข้องน้อย และการสร้างรายการแนะนำหนังสืออ่านเพิ่มเติมที่เกี่ยวข้องกับคำอธิบายรายวิชา โดยพิจารณารายการแนะนำหนังสือจำนวน N เล่ม โดยในงานวิจัยนี้พิจารณาค่า N ที่น่าสนใจที่ 10, 30, 50 และ 100 ตามลำดับ

ในการประเมินประสิทธิภาพของระบบแนะนำหนังสืออ่านเพิ่มเติมมี 2 ส่วน ได้แก่ 1) การวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ และ 2) การประเมินประสิทธิภาพของระบบแนะนำหนังสืออ่านเพิ่มเติม ที่ซึ่งได้เปรียบเทียบกับวิธีการสกัดคำสำคัญที่ใช้ในระบบอีก 2 วิธีการคือ “Termine” และ “RAKE” โดยจากผลการประเมินสามารถสรุปได้ว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบมีประสิทธิภาพมากกว่าทั้ง 2 ระบบแนะนำหนังสืออ่านเพิ่มเติมที่นำมาเปรียบเทียบทั้งด้านประสิทธิภาพของการสกัดคำสำคัญและประสิทธิภาพของระบบแนะนำหนังสืออ่านเพิ่มเติม

Abstract

At present, the current book-selling websites usually apply recommender systems to recommend a book or a set of books to customers. Based on relevance, clicks, reviews and so on. However, the recommender systems mostly focus on recommending books that users usually view or buy together and also on books having high review rates. This may cause failure to recommend books that cover most required contents, for example, books related to a course description of a course students have registered.

To address on this issue, we here introduce an alternative recommender system called *Supplementary Books Suggestion system* (also called *SBS system* for short) to create a list of supplementary books related/relevance to a course description of a course in computer science domain by regarding relevance between a book and a course description. The *SBS system* consists of five computational steps: 1) Terminology gathering, 2) Course descriptions input, 3) Keyword extraction, 4) Relevance books searching 5) Coverage calculation and ranking, respectively. To create a list of supplementary books related to a course description, the number of recommended books is set to be 10, 30, 50 and 100, respectively. This can help students easily find supplementary books to read and also may help to encourage the students doing self-learning.

Experiments on real course descriptions were conducted to investigate the effectiveness of the SBS system in the terms of precision, recall, F-measure and average coverage, total coverage, average uncoverage and total uncoverage of contents between a list of supplementary books and a course description.

สารบัญ

หน้า

บทคัดย่อ.....	ก
ABSTRACT	ข
สารบัญ.....	ค
สารบัญภาพ.....	จ
สารบัญตาราง.....	ช
บทที่ 1 บทนำ	1
1.1 ที่มาและความสำคัญ.....	1
1.2 วัตถุประสงค์ของงานวิจัย.....	2
1.3 ประโยชน์ที่คาดว่าจะได้รับ.....	3
1.4 ขอบเขตของงานวิจัย.....	4
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	5
2.1 ระบบแนะนำ (RECOMMENDER SYSTEM).....	5
2.1.1 วิธีการกรองเนื้อหา (Content based filtering).....	6
2.1.2 วิธีการกรองข้อมูลร่วม (Collaborative filtering)	9
2.1.3 วิธีการแบบผสมผสาน (Hybrid filtering)	11
2.2 การประมวลผลข้อความเบื้องต้น (TEXT PREPROCESSING)	11
2.2.1 การตัดคำ/ประโยค (Word/Sentence tokenization).....	11
2.2.2 การกำจัดคำหยุด (Stop word removal).....	11
2.2.3 การวิเคราะห์หน้าที่ของคำ (Part of speech tagging).....	12
2.2.4 การหารากศัพท์ (Stemming and lemmatization)	12
2.3 การสกัดคำสำคัญ (KEYWORD EXTRACTION).....	13
2.3.1 วิธีการทางภาษาศาสตร์ (Linguistic approach)	13
2.3.2 วิธีการทางสถิติ (Statistical approach).....	13

สารบัญ (ต่อ)

หน้า

2.3.3 วิธีการทางการเรียนรู้ของเครื่องจักร (Machine learning approach).....	14
2.3.4 วิธีการทางความเชี่ยวชาญเฉพาะด้าน (Domain specific approach).....	14
2.4 งานวิจัยที่เกี่ยวข้อง	15
2.4.1 งานวิจัยที่เกี่ยวข้องทางการแนะนำหนังสือ (Recommender system).....	15
2.4.2 งานวิจัยที่เกี่ยวข้องทางการสกัดคำสำคัญ (Keyword extraction).....	22
บทที่ 3 วิธีการที่นำเสนอ	25
3.1 รวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์.....	26
3.2 รวบรวมข้อมูลคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์.....	28
3.3 การสกัดคำสำคัญจากคำอธิบายรายวิชา	29
3.3.1 การประมวลผลข้อความเบื้องต้น (Text preprocessing).....	29
3.3.2 การระบุคำศัพท์เฉพาะ	31
3.3.3 การระบุกฎทางภาษาศาสตร์และการประยุกต์ใช้ในการสกัดคำสำคัญ.....	32
3.4 การค้นหาหนังสือที่มีความเกี่ยวข้องกับคำสำคัญในคำอธิบายรายวิชา	50
3.5 การคำนวณความครอบคลุมของหนังสือ และการจัดอันดับ	52
บทที่ 4 ผลการดำเนินงาน	65
4.1 การวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ	65
4.2 การประเมินระบบแนะนำหนังสืออ่านเพิ่มเติม	69
บทที่ 5 สรุปและอภิปรายผล	76
บรรณานุกรม.....	79
ภาคผนวก.....	81

สารบัญภาพ

หน้า

ภาพที่ 2-1 ระบบแนะนำประกอบไปด้วยขั้นตอนวิธีการต่าง ๆ	6
ภาพที่ 2-2 ขั้นตอนวิธีการกรองเนื้อหา	6
ภาพที่ 2-3 ขั้นตอนวิธีการกรองข้อมูลร่วม	9
ภาพที่ 2-4 ระบบแนะนำหนังสือแบบไฮบริดรวมถึงการพิจารณาข้อมูลผู้อ่านหนังสือ	17
ภาพที่ 2-5 โครงสร้างข้อมูลต้นไม้ 3 ระดับ	18
ภาพที่ 2-6 ขั้นตอนการแนะนำแบบ top-n items	19
ภาพที่ 2-7 การแบ่งปัญหาออกเป็นปัญหาย่อย	20
ภาพที่ 2-8 การแสดงผลหนังสือที่มีความเกี่ยวเนื่องกันซึ่งหาความสำคัญจาก FAHP	21
ภาพที่ 3-1 ขั้นตอนวิธีระบบแนะนำหนังสืออ่านเพิ่มเติม	25
ภาพที่ 3-2 ตัวอย่างคำศัพท์เฉพาะที่ต้องการรวบรวมจากพจนานุกรมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่จัดทำขึ้นโดยมหาวิทยาลัยออกซฟอร์ด (Oxford University)	27
ภาพที่ 3-3 ตัวอย่างคำศัพท์เฉพาะที่ต้องการรวบรวมจากรายการคำศัพท์เฉพาะทางการเขียนโปรแกรมและวิทยาการคอมพิวเตอร์ ข้อมูลจาก Labautopedia	27
ภาพที่ 3-4 การจัดเก็บข้อมูลคำอธิบายรายวิชาลงในฐานข้อมูลของคลังข้อมูลคำอธิบายรายวิชา	28
ภาพที่ 3-5 ตัวอย่างข้อมูลรับเข้าคำอธิบายรายวิชา “Data Mining”	29
ภาพที่ 3-6 การประมวลผลข้อความเบื้องต้นจากข้อมูลคำอธิบายรายวิชา “Data Mining”	30
ภาพที่ 3-7 การระบุคำศัพท์เฉพาะจากข้อมูลรับเข้าคำอธิบายรายวิชา “Data Mining”	32
ภาพที่ 3-8 การประยุกต์ใช้กฎทางภาษาศาสตร์ที่สร้างขึ้นกับข้อมูลคำอธิบายรายวิชา “Data Mining”	49
ภาพที่ 3-9 ตัวอย่างข้อมูลหนังสือที่ทำการเก็บรวบรวมไว้ในฐานข้อมูลของหนังสือ	51
ภาพที่ 3-10 คำอธิบายรายวิชาของวิชา “Data mining”	53
ภาพที่ 3-11 การประมวลผลคำอธิบายรายวิชาของวิชา “Data Mining” เบื้องต้น	54

สารบัญภาพ (ต่อ)

หน้า

ภาพที่ 3-12 การระบุถึงคำศัพท์เฉพาะที่ปรากฏในแต่ละประโยคของคำอธิบายรายวิชาของวิชา “Data Mining”	55
ภาพที่ 3-13 การสกัดคำสำคัญจากคำอธิบายรายวิชาของวิชา “Data Mining”	56
ภาพที่ 3-14 ระบุถึงความถูกต้องของคำสำคัญของคำอธิบายรายวิชาของวิชา “Data Mining” โดยผู้เชี่ยวชาญ	57
ภาพที่ 3-15 หน้าเว็บไซต์ Barnes & Noble จากการค้นหาหนังสือด้วยคำสำคัญคือ “data mining” โดยระบุหมวดหนังสือที่ทำการค้นหาเป็นหมวด “Computers”	58
ภาพที่ 4-1 การเปรียบเทียบผลลัพธ์ค่าความครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญ	71
ภาพที่ 4-2 การเปรียบเทียบผลลัพธ์ค่าความครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญ	71
ภาพที่ 4-3 การเปรียบเทียบผลลัพธ์ค่าความไม่ครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีสกัดความสำคัญ	73
ภาพที่ 4-4 การเปรียบเทียบผลลัพธ์ค่าความไม่ครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีสกัดความสำคัญ	74

สารบัญตาราง

หน้า

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ	33
ตารางที่ 3-2 ลักษณะของกฎทางภาษาศาสตร์ที่ไม่พิจารณาร่วมกับคำศัพท์เฉพาะ	47
ตารางที่ 3-3 การเปรียบเทียบค่าสำคัญระหว่างเซตของค่าสำคัญของคำอธิบายรายวิชาของวิชา “Data Mining” และเซตของค่าสำคัญของหนังสือชื่อ “Introduction to Data Mining” รหัสหนังสือ “9780133128901” ที่เกี่ยวข้องกับคำอธิบายรายวิชา	59
ตารางที่ 3-4 การคำนวณค่าความครอบคลุมระหว่างคำอธิบายรายวิชาของวิชา “Data Mining” และหนังสือทั้งหมดที่เกี่ยวข้องเกี่ยวกับคำอธิบายรายวิชา	60
ตารางที่ 3-5 การคำนวณค่าความไม่ครอบคลุมระหว่างคำอธิบายรายวิชาของวิชา “Data Mining” และหนังสือทั้งหมดที่เกี่ยวข้องเกี่ยวกับคำอธิบายรายวิชา	61
ตารางที่ 3-6 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความครอบคลุม	62
ตารางที่ 3-7 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความไม่ครอบคลุม	63
ตารางที่ 4-1 ตาราง Confusion matrix ระหว่างข้อมูลความเป็นจริงจริงและข้อมูลทำนาย	66
ตารางที่ 4-2 ตาราง Confusion matrix ระหว่างข้อมูลสำคัญที่ถูกต้องและคำสำคัญที่สกัดได้ของรายวิชา “Data Mining”	67
ตารางที่ 4-3 เปรียบเทียบการวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ, ขั้นตอนวิธี “Termine” และขั้นตอนวิธี “RAKE”	68

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญ

ในยุคปัจจุบันเป็นยุคของข้อมูลข่าวสาร โดยเราสามารถค้นหาข้อมูลข่าวสารได้จากเว็บไซต์ต่าง ๆ และจากสังคมออนไลน์เป็นจำนวนมาก แต่อย่างไรก็ตาม การอ่านหนังสือก็ยังคงเป็นปัจจัยสำคัญต่อการพัฒนาตนเองให้รอบรู้ รวมทั้งพัฒนาด้านกระบวนการทางความคิดได้ โดยในปัจจุบัน รูปแบบของหนังสือได้เปลี่ยนแปลงไปอย่างมาก โดยมีทั้งรูปเล่มที่ถูกพิมพ์จากสำนักพิมพ์ และในรูปแบบอิเล็กทรอนิกส์ ซึ่งจะช่วยเพิ่มความสะดวกในการเพิ่มพูนความรู้ที่มากขึ้น จากความหลากหลายในรูปแบบของหนังสือจึงได้มีระบบค้นหาหนังสือและระบบแนะนำหนังสือต่าง ๆ มากมาย เพื่อช่วยให้ผู้อ่านสามารถมีความสะดวกในการค้นหาและเลือกหนังสือที่ตนเองสนใจได้ดียิ่งขึ้น อาทิเช่น การค้นหาหนังสือจากคุณสมบัติของคำค้นหนึ่ง ๆ กับชื่อหนังสือ กับสารบัญหนังสือ (Garrido et al., 2014a) โดยดูจากความเหมือนของคำและความหมายของคำเป็นหลัก (Garrido et al., 2014b) การแนะนำหนังสือจากความนิยมของหนังสือ (Linden et al., 2003, Chandak et al., 2015) การแนะนำหนังสือจากข้อคิดเห็นที่ได้รับจากผู้อ่าน (Lu & Zhang, 2015) การแนะนำหนังสือจากพฤติกรรมการยืมหนังสือในห้องสมุด (Saxena et al., 2014, Teng et al., 2015, Chen et al., 2015) เป็นต้น ซึ่งจากระบบดังกล่าวจะสามารถช่วยให้ผู้อ่านหนังสือมีข้อมูลพื้นฐานในการเลือกอ่านหนังสือได้เหมาะสมมากยิ่งขึ้น และยังสามารถเพิ่มคุณภาพการให้บริการของห้องสมุดต่อผู้ใช้บริการ และยังสามารถรวมถึงการบริหารจัดการหนังสือในห้องสมุดได้อีกด้วย

แต่อย่างไรก็ตาม แนวคิดทั้งหมดไม่ได้มุ่งเน้นที่การแนะนำหนังสือจากคุณสมบัติของหนังสือที่เกี่ยวข้องกับการศึกษาในชั้นเรียน และไม่ได้มุ่งเน้นในการแนะนำหนังสือที่เกี่ยวข้องกับศาสตร์ที่มีคำศัพท์เฉพาะทางค่อนข้างมากกับคำค้น (คำค้นหนึ่งๆสามารถมีได้หลายรูปแบบ อาทิเช่น คำหนึ่งๆ คำหลายๆคำ คำอธิบายรายวิชาที่ประกอบไปด้วยเนื้อหาต่างๆ ประมวลผลรายวิชา) ที่ซึ่งช่วยให้ผู้อ่านได้รับการแนะนำที่สอดคล้องกับเนื้อหาที่ตนเองสนใจมากยิ่งขึ้น ตัวอย่างเช่น ถ้าผู้อ่านหนังสือต้องการค้นหาหนังสือเกี่ยวกับระบบแนะนำ (Recommendation system) ที่จะประกอบไปด้วย “นิยาม” “ขั้นตอนวิธีในการแนะนำ” “การวัดและประเมินประสิทธิภาพของการแนะนำ” “ตัวอย่าง/กรณีศึกษาของการประยุกต์ใช้ระบบแนะนำ” และอื่นๆ หรืออาจเป็นการค้นหาหนังสือที่เกี่ยวข้องกับคำอธิบายรายวิชา

หนึ่งๆ อาทิเช่น คำอธิบายของรายวิชา การทำเหมืองข้อมูล (Data mining) ที่จะมีคำอธิบายรายวิชา “Fundamental concepts of data mining. Types and techniques of data mining. Procedures of pattern discovery. Anomaly detection in large database. Association-rule mining. Classification and clustering algorithms. Data mining tools. Data warehousing and its applications using tool” ซึ่งมีความต้องการเนื้อหาในหนังสือที่ค่อนข้างมาก มีคำศัพท์เฉพาะ ค่อนข้างมากและหลากหลาย โดยจากความต้องการหนังสือที่มีความเฉพาะเจาะจงค่อนข้างสูง ระบบการแนะนำที่คิดค้นจากงานวิจัยข้างต้นอาจไม่สามารถแนะนำหนังสือที่ตรงกับความต้องการของผู้อ่านเท่าที่ควร ด้วยเหตุนี้ งานวิจัยนี้มุ่งเน้นที่การแนะนำหนังสือโดยพิจารณาจากความสอดคล้องของหนังสือที่เพิ่มเติมจากการศึกษาในชั้นเรียน ที่ซึ่งเป็นหนังสือที่เกี่ยวข้องกับคำค้นที่เฉพาะเจาะจงเป็นหลัก โดยจะมุ่งเน้นที่จะทำการแนะนำหนังสืออ่านเพิ่มเติมจากการศึกษาในโรงเรียน หรือ มหาวิทยาลัยผ่านการพิจารณาคำอธิบายรายวิชาและ/หรือประมวลรายวิชาเพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเอง อันนำมาซึ่งการพัฒนาตนเอง พัฒนาการเรียนรู้อย่างยั่งยืน และรวมถึงการพัฒนาประเทศชาติสืบไป โดยรูปแบบการแนะนำหนังสืออ่านเพิ่มเติมจะเป็นการสร้างลิสต์ของหนังสือที่มีความเกี่ยวข้องกับคำค้นหรือกลุ่มของคำค้นที่ผู้อ่านกำหนด (คำค้นอาจเป็นคำหนึ่ง ๆ กลุ่มของคำ คำอธิบายรายวิชาและ/หรือประมวลรายวิชา) โดยหนังสือในลิสต์ที่ถูกแนะนำจะถูกเรียงลำดับตามความเกี่ยวข้อง (เรียงจากมากไปน้อยและแสดงถึงความเกี่ยวข้องของหนังสือกับคำค้น) โดยจากลิสต์ที่ถูกแนะนำจะช่วยให้ผู้อ่านสามารถตัดสินใจที่จะเลือกหนังสืออ่านเพิ่มเติมได้ง่ายและสะดวกยิ่งขึ้น อันนำมาซึ่งการส่งเสริมการอ่านและการพัฒนาตนเองของคนในประเทศสืบไป

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาเกี่ยวกับการแนะนำหนังสือเพื่ออ่านเพิ่มเติม การประยุกต์ใช้การแนะนำหนังสือในแง่มุมต่าง ๆ และประโยชน์ที่ได้รับจากการประยุกต์ใช้การแนะนำหนังสือ โดยศึกษาจากงานวิจัยก่อนหน้า และเว็บไซต์ที่เกี่ยวข้อง

2. เพื่อศึกษาและค้นหาแนวทางการปรับปรุงพัฒนาขั้นตอนวิธีในการแนะนำหนังสือเพื่ออ่านเพิ่มเติมทางด้านวิทยาการคอมพิวเตอร์เพื่อเป็นต้นแบบในการสร้างระบบแนะนำหนังสือเพิ่มเติมจากการศึกษาในชั้นเรียน และการเรียนรู้ด้วยตนเอง อันนำมาซึ่งการก้าวทันเทคโนโลยีของบุคลากรทางด้านวิทยาการคอมพิวเตอร์เพื่อเพิ่มพูนประโยชน์ที่มากขึ้นในองค์กรและการสร้างนวัตกรรมสืบไป

3. เพื่อสร้างระบบการแนะนำหนังสือเพื่ออ่านเพิ่มเติมทางด้านวิทยาการคอมพิวเตอร์แบบอัตโนมัติที่สามารถใช้ได้จริง ที่ซึ่งสามารถอำนวยความสะดวกให้นิสิต/นักศึกษา/บุคคลทั่วไปใช้ในการเลือกหนังสือเพื่ออ่านเพิ่มเติมได้ อันนำมาซึ่งการส่งเสริมให้นิสิต/นักศึกษา/บุคคลทั่วไปอ่านหนังสือมากขึ้น

4. เพื่อสร้างต้นแบบงานวิจัยทางการสร้างระบบแนะนำหนังสือที่จะสามารถเปิดโอกาสให้ผู้สนใจสามารถนำความคิดที่นำเสนอไปศึกษาเพิ่มเติม พัฒนาเป็นผลผลิตเพื่อใช้ในองค์กรของตนเองและเพื่อประยุกต์ใช้ในงานวิจัยของตนเองต่อไป

1.3 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้โครงสร้างพื้นฐานของระบบแนะนำหนังสือเพิ่มเติม ขั้นตอนวิธีในการประมวลผลแบบอัตโนมัติที่สามารถนำไปพัฒนาต่อยอดให้เข้ากับองค์กรต่าง ๆ ได้

2. ได้ระบบการแนะนำหนังสืออ่านเพิ่มเติมที่สามารถใช้งานได้จริง โดยสามารถประยุกต์ใช้กับเว็บไซต์ของรายวิชาหนึ่งในระดับอุดมศึกษา เพื่อแนะนำหนังสืออ่านเพิ่มเติมให้กับนิสิต/นักศึกษาที่ลงทะเบียนเรียนในวิชานั้น ๆ

3. สามารถนำแนวคิดและระบบการแนะนำหนังสืออ่านเพิ่มเติมไปวิเคราะห์เพิ่มเติม เพื่อพัฒนาการให้บริการของห้องสมุดในมหาวิทยาลัยต่าง ๆ โดยจะสามารถส่งเสริมประสิทธิภาพการสั่งซื้อหนังสือเข้าห้องสมุดที่ซึ่งจะเป็นหนังสือที่ผ่านการคัดกรองและจัดอันดับจากระบบ ซึ่งการดำเนินการดังกล่าวจะช่วยส่งเสริมการยืมหนังสือห้องสมุดได้มากขึ้น

4. สามารถนำแนวคิดและระบบการแนะนำหนังสืออ่านเพิ่มเติมไปใช้ในภาคธุรกิจสำนักพิมพ์หนังสือ เพื่อให้ทราบถึงข้อมูลความเกี่ยวเนื่องกันระหว่างคำค้นและเนื้อหาในหนังสือที่เกี่ยวข้องกับคำค้น อันนำมาซึ่งการจัดทำแผนการตลาด/โปรโมชั่น และการคาดการณ์ความต้องการของลูกค้าได้

5. สามารถนำแนวคิดและระบบการแนะนำหนังสืออ่านเพิ่มเติมเพื่อประกอบการตัดสินใจในการวางโครงเนื้อหาของผู้เขียนก่อนเขียนหนังสือเล่มหนึ่ง ๆ

6. สามารถนำแนวคิดและระบบการแนะนำหนังสืออ่านเพิ่มเติมไปพัฒนาต่อยอดเพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเองของนักเรียนและนักศึกษาในประเทศไทยได้

7. ได้ผลงานวิจัยที่จะสามารถตีพิมพ์ในงานประชุมวิชาการหรือวารสารวิชาการ

1.4 ขอบเขตของงานวิจัย

1. ระบบแนะนำหนังสือจะทำการแนะนำหนังสือที่เกี่ยวกับศาสตร์วิทยาการคอมพิวเตอร์ที่ซึ่งเป็นศาสตร์ที่มีคำศัพท์เฉพาะค่อนข้างมาก และเป็นศาสตร์ที่เกี่ยวข้องกับผู้วิจัย

2. การแนะนำหนังสือจะทำงานจากคำค้นของผู้ใช้ที่เป็นภาษาอังกฤษเท่านั้น โดยคำค้นอาจเป็นคำหนึ่ง ๆ หรือ อาจเป็นกลุ่มของคำที่ได้จากคำอธิบายรายวิชาหนึ่ง ๆ หรือประมวลผลรายวิชาต่าง ๆ ในสาขาวิทยาการคอมพิวเตอร์

3. การพิจารณาความน่าสนใจของหนังสือที่จะแนะนำจะพิจารณาจากเนื้อหาในหนังสือที่เกี่ยวข้องกับคำค้นของผู้ใช้

4. หนังสือเล่มหนึ่ง ๆ ที่จะถูกพิจารณาหาความน่าสนใจ และถูกแนะนำให้แก่ผู้ใช้จะเป็นหนังสือที่สามารถสืบค้นได้จากเว็บไซต์ www.barnesandnoble.com และเว็บไซต์ของสำนักพิมพ์ต่าง ๆ

5. ผลลัพธ์ที่ได้จากการแนะนำหนังสือจะถูกเรียงลำดับตามความน่าสนใจโดยเรียงลำดับจากความน่าสนใจมากไปยังความน่าสนใจน้อย

6. การประเมินผลของการแนะนำหนังสือจะสามารถวัดได้จากประสิทธิภาพในการประมวลผลและ/หรือ ความถูกต้องของผลลัพธ์

บทที่ 2

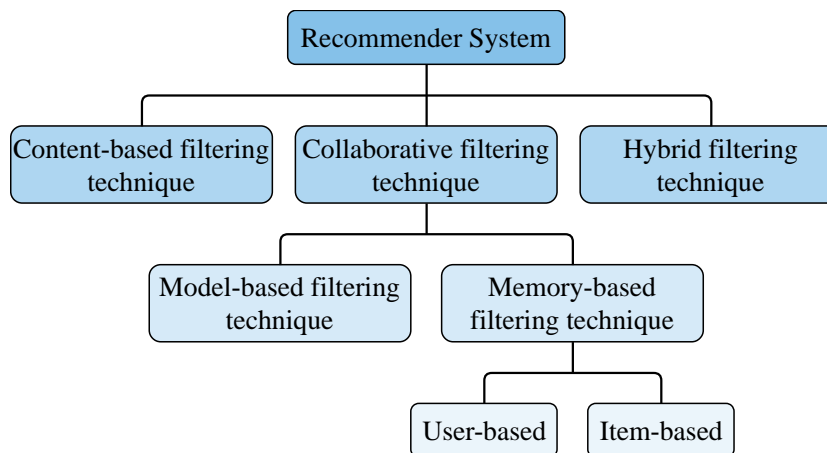
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้มีการนำขั้นตอนวิธีของระบบแนะนำ (Recommender system) การประมวลผลข้อความเบื้องต้น (Text preprocessing) และการสกัดคำสำคัญ (Keyword extraction) มาใช้ในการสร้างรายการแนะนำหนังสืออ่านเพิ่มเติม ซึ่งในบทนี้จะอธิบายถึงความเป็นมาและความสำคัญ ขั้นตอนวิธี รวมทั้งอธิบายถึงงานวิจัยที่เกี่ยวข้อง รายละเอียดต่าง ๆ มีดังนี้

- 2.1 ระบบแนะนำ (Recommender system)
- 2.2 การประมวลผลข้อความเบื้องต้น (Text preprocessing)
- 2.3 การสกัดคำสำคัญ (Keyword extraction)
- 2.4 งานวิจัยที่เกี่ยวข้อง

2.1 ระบบแนะนำ (Recommender system)

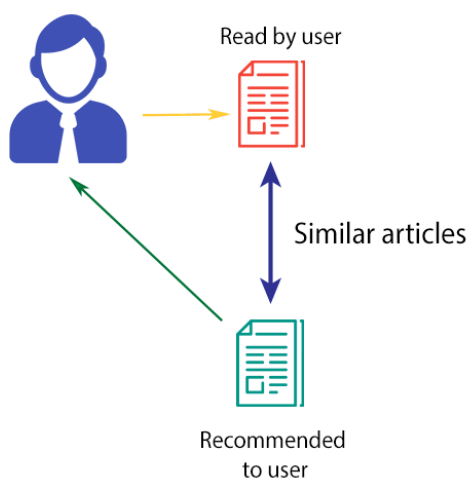
ระบบแนะนำ (Recommender system) เป็นระบบที่ถูกสร้างขึ้นมาเพื่อช่วยเหลือผู้ใช้ในการคัดกรองสิ่งที่สนใจ หรือชื่นชอบจากสิ่งต่าง ๆ ที่มีอยู่เป็นจำนวนมาก มีบทบาทสำคัญในการสร้างรายการแนะนำข้อมูลที่คาดเดาว่าผู้ใช้ชื่นชอบหรือให้ความสนใจจากฐานข้อมูลที่มีขนาดใหญ่ ซึ่งช่วยให้ผู้ใช้สามารถเข้าถึงข้อมูลสินค้าหรือบริการที่ชื่นชอบได้อย่างมีประสิทธิภาพ และสามารถเพิ่มผลกำไรและรักษาลูกค้าให้กับระบบธุรกิจพาณิชย์อิเล็กทรอนิกส์ อาทิเช่น ระบบแนะนำหนังสือ (www.goodreads.com, www.amazon.com) ระบบแนะนำภาพยนตร์ (www.netflix.com, www.imdb.com) ระบบแนะนำเพลง (www.lastfm.com www.pandora.com) และระบบแนะนำวิดีโอ (www.youtube.com) โดยวิธีการที่นิยมใช้ในการสร้างรายการแนะนำข้อมูลสามารถแบ่งได้ 3 วิธีการ ได้แก่ วิธีการกรองเนื้อหา (Content based filtering), วิธีการกรองข้อมูลร่วม (Collaborative filtering) และ วิธีการแบบผสมผสาน (Hybrid filtering) แสดงดังภาพที่ 2-1



ภาพที่ 2-1 ระบบแนะนำประกอบไปด้วยขั้นตอนวิธีการต่าง ๆ

2.1.1 วิธีการกรองเนื้อหา (Content based filtering)

วิธีการกรองเนื้อหาเป็นวิธีการสร้างรายการแนะนำข้อมูล โดยพิจารณาจากความคล้ายคลึงของเนื้อหาหรือคุณสมบัติต่าง ๆ ของข้อมูลที่คล้ายกันกับข้อมูลของผู้ใช้ในอดีต การทำงานของวิธีการนี้จะให้ความสนใจกับเนื้อหาของข้อมูลเป็นสำคัญ เช่น การสร้างรายการแนะนำหนังสือนั้น วิธีการนี้จะทำการพิจารณาเนื้อหาของหนังสือ, ชื่อหนังสือ, ชื่อผู้แต่ง, หมวดหมู่ของหนังสือ และสำนักพิมพ์ที่มีความคล้ายกันของหนังสือกับหนังสือที่ผู้ใช้เคยอ่านในอดีต แสดงดังภาพที่ 2-2



ภาพที่ 2-2 ขั้นตอนวิธีการกรองเนื้อหา

วิธีการกรองเนื้อหาที่นิยมในการสร้างรายการแนะนำ คือ วิธีการหาค่าน้ำหนักของคำสำคัญ หรือ Term Frequency/Inverse Document Frequency (TF-IDF) เป็นวิธีการทางสถิติในการสืบค้นข้อมูล โดยพิจารณาคำสำคัญของเอกสาร (Keyword) ความสำคัญของคำเป็นสัดส่วนโดยตรงกับความถี่ที่คำนั้น ๆ ปรากฏในเอกสารหนึ่ง ๆ แต่ถูกลดความสำคัญโดยความถี่ของคำนั้น ๆ ที่ปรากฏในเอกสารอื่น ๆ ซึ่งหาค่าน้ำหนักของคำสำคัญ เริ่มจากการคำนวณหาค่าความถี่ของคำ หรือ Term Frequency ($TF_{i,j}$) แสดงดังสมการที่ 1

$$TF_{i,j} = \frac{f_{i,j}}{\sum_{i=0}^k f_{k,j}} \quad (1)$$

กำหนดให้	$TF_{i,j}$	คือ ค่าความถี่ของคำ k_i ที่ปรากฏในเอกสาร d_j
	$f_{i,j}$	คือ จำนวนครั้งของคำ k_i ที่ปรากฏในเอกสาร d_j
	$\sum_{i=0}^k f_{k,j}$	คือ จำนวนครั้งของทุกคำที่ปรากฏในเอกสาร d_j

ต่อมาการคำนวณหาค่าความทั่วไปของคำ Inverse Document Frequency (IDF_i) เพื่อลดค่าความสำคัญของคำที่ปรากฏทั่วไปในหลาย ๆ เอกสาร ซึ่งคำนวณได้ดังสมการที่ 2

$$IDF_i = \log \frac{N}{n_i} \quad (2)$$

กำหนดให้	IDF_i	คือ ค่าความทั่วไปของคำ k_i
	N	คือ จำนวนเอกสารทั้งหมด
	n_i	คือ จำนวนเอกสารทั้งหมดที่ปรากฏคำ k_i

และการวิธีการคำนวณหาค่าน้ำหนักของ TF-IDF ($W_{i,j}$) ของคีย์เวิร์ด i ในเอกสาร j แสดงดังสมการที่ 3

$$W_{i,j} = TF_{i,j} \times IDF_i \quad (3)$$

กำหนดให้	$W_{i,j}$	คือ ค่าน้ำหนักของคำ k_i ในเอกสาร d_j
	$TF_{i,j}$	คือ ค่าความถี่ของคำ k_i ที่ปรากฏในเอกสาร d_j
	IDF_i	คือ ค่าความทั่วไปของคำ k_i

ดังนั้นวิธีการนี้จะต้องมีการคำนวณหาค่าความคล้ายคลึงกัน (Similarity measure) ระหว่างข้อมูลที่มีอยู่ในฐานข้อมูล ซึ่งขั้นตอนวิธีที่นิยมใช้ก็มีอย่างหลากหลาย เช่น การหาค่าความคล้ายคลึงแบบโคไซน์ (Cosine Similarity) แสดงดังสมการที่ 4

$$\cos(w_c, w_s) = \frac{\sum_{i=1}^n w_{c,i} w_{s,i}}{\sqrt{\sum_{i=1}^n w_{c,i}^2} \sqrt{\sum_{i=1}^n w_{s,i}^2}} \quad (4)$$

กำหนดให้ $w_{c,i}$ คือ ค่าน้ำหนักของผู้ใช้ C กับคำสำคัญ i
 $w_{s,i}$ คือ ค่าน้ำหนักของเอกสาร S กับคำสำคัญ i
 n คือ จำนวนคำสำคัญทั้งหมด

การหาค่าระยะห่างระหว่างข้อมูลแบบยูคลิด (Euclidean Distance) ซึ่งเหมาะกับชนิดข้อมูลที่เป็นตัวเลข แสดงดังสมการที่ 5

$$D_{Euclidean}(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2} \quad (5)$$

กำหนดให้ X_i คือ ค่าของข้อมูล X ที่ตำแหน่ง i
 Y_i คือ ค่าของข้อมูล Y ที่ตำแหน่ง i
 n คือ จำนวนตำแหน่งของทั้งหมด

การหาค่าความคล้ายคลึงกับแบบ Jaccard coefficient ซึ่งเหมาะกับชนิดข้อมูลที่เป็นไบนารี (Binary) แสดงดังสมการที่ 6

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} \quad (6)$$

การหาค่าความครอบคลุม (Coverage) แสดงดังสมการที่ 7

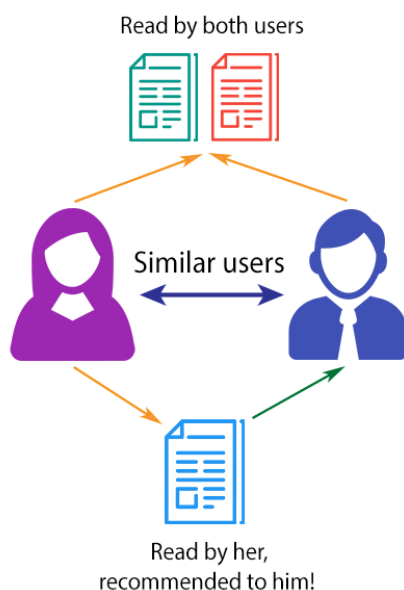
$$C(X, Y) = \frac{|X \cap Y|}{|X|} \quad (7)$$

กำหนดให้ X คือ เซตของคำสำคัญที่เกี่ยวข้องกับข้อมูล X
 Y คือ เซตของคำสำคัญที่เกี่ยวข้องกับข้อมูล Y

ข้อดีของวิธีการกรองเนื้อหา คือการให้ความสำคัญกับเนื้อหาของข้อมูลเป็นสำคัญ เพื่อค้นหาข้อมูลที่ผู้ใช้สนใจ ซึ่งวิธีการนี้จะไม่ประสบกับปัญหาการให้คะแนนความชอบ (Rating) กับข้อมูลที่ไม่ทั่วถึง และปัญหาข้อมูลที่ยังไม่ได้ให้คะแนนความชอบ (Rating) จึงสามารถแนะนำข้อมูลที่ตรงกับความชื่นชอบของผู้ใช้ได้ อย่างไรก็ตามข้อเสียของวิธีการนี้คือข้อมูลที่แนะนำให้กับผู้ใช้มีความไม่หลากหลาย (Overspecialization)

2.1.2 วิธีการกรองข้อมูลร่วม (Collaborative filtering)

วิธีการกรองข้อมูลร่วมเป็นวิธีการสร้างรายการแนะนำข้อมูลที่นิยมมากที่สุด โดยการพิจารณาความคล้ายของข้อมูลหรือผู้ใช้ที่มีพฤติกรรมในการเลือกซื้อหรือบริโภคข้อมูลที่คล้ายกันจากคะแนนความชอบที่ผู้ใช้ให้กับข้อมูลโดยตรง (Rating) หรือประวัติในการเลือกซื้อหรือรับชมข้อมูลนั้น ๆ แสดงดังภาพที่ 2-3



ภาพที่ 2-3 ขั้นตอนวิธีการกรองข้อมูลร่วม

โดยวิธีการกรองข้อมูลร่วมแบ่งออกเป็น 2 ประเภท คือ 1) วิธีการกรองข้อมูลร่วมแบบจำลอง (Model-based filtering) เป็นวิธีการนำข้อมูลของผู้ใช้ทั้งหมดที่มีในฐานข้อมูล มาสร้างเป็นแบบจำลองสำหรับสร้างรายการแนะนำให้กับผู้ใช้ที่มีลักษณะคล้ายคลึงกัน 2) วิธีการกรองข้อมูลร่วมแบบจดจำ (Memory-based filtering) เป็นวิธีการนำข้อมูลของผู้ใช้แต่ละคนมาเปรียบเทียบว่ามีความคล้ายคลึงกับ

ข้อมูลผู้ใช้งานอื่น ๆ ที่มีอยู่ในฐานข้อมูล ซึ่งถ้าพิจารณาจากความคล้ายคลึงหรือใกล้เคียงกันระหว่างผู้ใช้งานในฐานข้อมูล จะเรียกว่าวิธีการกรองข้อมูลร่วมโดยพิจารณาผู้ใช้งานเป็นหลัก (User-based Collaborative filtering) แต่ถ้าพิจารณาความคล้ายคลึงหรือใกล้เคียงกันระหว่างข้อมูลในฐานข้อมูล จะเรียกว่าวิธีการกรองข้อมูลร่วมโดยพิจารณาข้อมูลเป็นหลัก (Item-based Collaborative filtering) ดังต่อไปนี้

2.1.2.1 วิธีการกรองข้อมูลร่วมโดยพิจารณาผู้ใช้งานเป็นหลัก (User-based)

พิจารณาความชอบของผู้ใช้งานในการเลือกบริโภคหรือบริการข้อมูลที่คล้ายกันกับผู้ใช้งานอื่นในฐานข้อมูล เช่น การรับชม ความชอบ การเลือกซื้อ จากนั้นจะทำการเลือกผู้ใช้งานที่มีพฤติกรรมใกล้เคียงกับผู้ใช้งานเป้าหมายมากที่สุดแล้วทำการสร้างการทำนายค่าคะแนนความชอบของข้อมูลต่าง ๆ โดยพิจารณาจากผู้ใช้งานที่มีพฤติกรรมที่ใกล้เคียงกัน สุดท้ายทำการสร้างรายการแนะนำข้อมูลที่คาดว่าผู้ใช้งานจะชื่นชอบโดยเรียงลำดับค่าการทำนายจากมากไปหาน้อยตามลำดับ

2.1.2.2 วิธีการกรองข้อมูลร่วมโดยพิจารณาข้อมูลเป็นหลัก (Item-based)

พิจารณาความชอบของข้อมูลที่ผู้ใช้งานทำการเลือกบริโภคที่คล้ายกันกับข้อมูลอื่นในฐานข้อมูล จากนั้นพิจารณาข้อมูลที่มีลักษณะในการเลือกบริโภคที่คล้ายกันมากที่สุดแล้วสร้างรายการแนะนำข้อมูลโดยพิจารณาจากค่าการทำนายคะแนนความชอบของข้อมูลข้างเคียงจากมากไปหาน้อยตามลำดับ

ข้อดีของวิธีการกรองข้อมูลร่วม คือสามารถแนะนำข้อมูลให้กับผู้ใช้งานได้หลากหลายมากขึ้น แต่อย่างไรก็ตามข้อเสียของวิธีการนี้คือ 1) เมื่อเกิดเหตุการณ์ที่มีผู้ใช้งานใหม่ หรือข้อมูลใหม่ ผู้ใช้งานใหม่ที่เข้าสู่ระบบครั้งแรกจะยังไม่ได้รับการแนะนำและข้อมูลใหม่ก็จะไม่ถูกแนะนำเช่นกัน ปัญหานี้เรียกว่า “Cold start problem” 2) เมื่อเกิดเหตุการณ์ที่ข้อมูลมีมากกว่าผู้ใช้งานในฐานข้อมูล ทำให้เกิดความเบาบางของข้อมูลขึ้น ปัญหานี้เรียกว่า “Data sparsity problem”

2.1.3 วิธีการแบบผสมผสาน (Hybrid filtering)

วิธีการแบบผสมผสานเป็นการรวมข้อดีของทั้งวิธีการกรองเนื้อหาและวิธีการกรองข้อมูลรวมไว้ด้วยกัน ซึ่งสามารถช่วยในการหลีกเลี่ยงข้อจำกัดของทั้งสองวิธีการได้ ทำให้สามารถสร้างรายการแนะนำสินค้าคาดว่าผู้ใช้ชื่นชอบได้อย่างมีประสิทธิภาพ อย่างไรก็ตามวิธีการนี้ต้องใช้ข้อมูลของสินค้าจำนวนมากในการสร้างรายการแนะนำ อาจทำให้มีการใช้ทรัพยากรและมีความซับซ้อนในการแนะนำที่สูงขึ้น ซึ่งอาจไม่เหมาะสมกับระบบธุรกิจพาณิชย์อิเล็กทรอนิกส์ที่ไม่มีการจัดเก็บคุณสมบัติของข้อมูล

2.2 การประมวลผลข้อความเบื้องต้น (Text preprocessing)

การประมวลผลข้อความเป็นการแปลงรูปแบบข้อความให้อยู่ในรูปแบบที่ถูกต้อง หรือมีความพร้อมใช้งานในการประมวลผลเชิงลึกมากยิ่งขึ้น ซึ่งจะประกอบไปด้วยการตัดคำ/ประโยค, การกำจัดคำหยุด, การวิเคราะห์หน้าที่ของคำ, การหารากศัพท์ เป็นต้น

2.2.1 การตัดคำ/ประโยค (Word/Sentence tokenization)

การตัดประโยค เป็นขั้นตอนการแยกข้อความออกจากกัน ซึ่งการวิเคราะห์แบ่งประโยคเป็นประโยคความเดียว จะทำการวิเคราะห์ตามหลักไวยากรณ์ของประโยค และความสมบูรณ์ของความหมายในส่วนของคำตัดคำ เป็นขั้นตอนการแบ่งองค์ประกอบของประโยค ออกมาเป็นหน่วยคำ

2.2.2 การกำจัดคำหยุด (Stop word removal)

เป็นขั้นตอนการกำจัดคำที่ไม่มีนัยสำคัญออกจากประโยค โดยที่ความหมายของประโยคหรือข้อความไม่ได้เปลี่ยนแปลง คำหยุดมักจะปรากฏบ่อยในทุก ๆ ข้อความหรือเอกสาร ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องในการประมวลผลข้อความ นอกจากนี้การกำจัดคำหยุดยังช่วยลดขนาดพื้นที่และเวลาในการประมวลผลข้อความอีกด้วย

2.2.3 การวิเคราะห์หน้าที่ของคำ (Part of speech tagging)

การวิเคราะห์หน้าที่ของคำหน้าที่ของคำ คือการระบุ/กำกับหน้าที่ของคำที่ประกอบอยู่ในประโยค ว่าคำนั้น ๆ มีหน้าที่ชนิดใดในทางไวยากรณ์ ซึ่งหน้าที่ของคำภาษาอังกฤษมีทั้งหมด 8 ชนิด ดังนี้

- | | |
|----------------------------|---|
| 1. คำนาม (Noun) | คือ คำที่ใช้เรียกแทนชื่อคน, สัตว์, สิ่งของ, สถานที่ |
| 2. คำสรรพนาม (Pronoun) | คือ คำที่ใช้แทนคำนาม |
| 3. คำกริยา (Verb) | คือ คำที่แสดงอาการหรือการกระทำในประโยค |
| 4. คำคุณศัพท์ (Adjective) | คือ คำที่ทำหน้าที่ขยายคำนาม โดยตำแหน่งจะอยู่หน้าคำนามเสมอ |
| 5. คำกริยาวิเศษณ์ (Adverb) | คือ คำที่มีหน้าที่ขยายกริยา ขยายคุณศัพท์ และขยายกริยาวิเศษณ์ด้วยตัวเอง |
| 6. คำบุพบท (Preposition) | คือ คำที่ใช้บอกตำแหน่ง วันเวลา ทิศทาง สถานที่ หรือบอกความสัมพันธ์ระหว่างนามหรือสรรพนามกับคำอื่น ๆ |
| 7. คำสันธาน (Conjunction) | คือ คำที่ใช้เชื่อมคำกับคำ กลุ่มคำกับกลุ่มคำ หรือประโยคกับประโยค |
| 8. คำอุทาน (Interjection) | คือ คำที่ใช้บ่งบอกอารมณ์ ความรู้สึกที่เกิดขึ้นในตอนนั้น |

2.2.4 การหารากศัพท์ (Stemming and lemmatization)

เป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์นั้นจะใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกันหรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ แล้วใช้ในการเปรียบเทียบหารากศัพท์ต่อไป ซึ่งในตอนเริ่มแรกอาจจะต้องใช้ผู้เชี่ยวชาญในการจัดทำรายการรากศัพท์

นอกจากนี้ในการทำการประมวลผลข้อความเบื้องต้นอาจจะมีวิธีการอื่น ๆ เข้ามาร่วมด้วย ขึ้นอยู่กับรูปแบบข้อความที่ต้องการนำไปใช้ในการประมวลผลต่อไปเป็นอย่างไร

2.3 การสกัดคำสำคัญ (Keyword extraction)

คำสำคัญเป็นตัวแทนของเอกสารหรือบทความหนึ่ง ๆ โดยย่อเหลือเพียงคำที่แสดงความสำคัญที่กระชับรัด และชัดเจน โดยเป็นตัวแทนว่าเอกสารหรือบทความนั้น ๆ เกี่ยวข้องกับเรื่องใด ซึ่งการระบุคำสำคัญของเอกสารหรือบทความจะช่วยลดระยะเวลาในการอ่าน และช่วยพิจารณาได้ว่าเอกสารหรือบทความนั้นตรงตามความต้องการของผู้อ่านหรือไม่ โดยวิธีการสกัดคำสำคัญแบบอัตโนมัติสามารถแบ่งได้ 4 วิธี ได้แก่

2.3.1 วิธีการทางภาษาศาสตร์ (Linguistic approach)

เป็นวิธีการที่ใช้องค์ความรู้ หรือคุณสมบัติทางด้านภาษาศาสตร์เข้ามาช่วยในการสกัดคำสำคัญ โดยการวิเคราะห์คำในประโยคหรือเอกสารต่าง ๆ โดยใช้หลักไวยากรณ์ของภาษา รวมถึงการวิเคราะห์คำศัพท์, วิเคราะห์ประโยค เป็นต้น ในส่วนของกฎที่ได้สร้างขึ้นนั้นจะถูกกำหนดไว้ตั้งแต่ก่อนเริ่มทำการสกัดคำและจะไม่มีเปลี่ยนแปลงตามข้อมูลที่เข้ามาใหม่ ซึ่งวิธีการนี้อาจมีค่าความถูกต้องเป็นอย่างมาก แต่จะมีการคำนวณที่ละเอียดและจำเป็นจะต้องมีความเชี่ยวชาญในด้านภาษาศาสตร์

นอกจากนี้วิธีการทางภาษาศาสตร์สามารถนำไปใช้ร่วมกับวิธีการอื่น ๆ เพื่อเพิ่มประสิทธิภาพในการสกัดคำสำคัญมากยิ่งขึ้น

2.3.2 วิธีการทางสถิติ (Statistical approach)

เป็นวิธีการอย่างง่ายที่สามารถใช้สถิติของข้อมูลคำ ซึ่งอาจใช้สถิติทางด้านความถี่ หรือความน่าจะเป็น ในการระบุคำสำคัญของเอกสารต่าง ๆ ได้ โดยข้อดีที่สำคัญของวิธีการนี้คือการประยุกต์ใช้ได้ อย่างอิสระ สามารถนำมาใช้งานได้หลากหลายภาษา วิธีการนี้อาจจะไม่ได้ให้ค่าความถูกต้องที่ดีเมื่อเปรียบเทียบกับวิธีการทางภาษาศาสตร์ แต่วิธีการนี้เหมาะสมกับฐานข้อมูลที่มีขนาดใหญ่ ทฤษฎีที่นิยมนำมาใช้ ได้แก่ 1) TF-IDF เป็นค่าถ่วงน้ำหนักทางด้านสถิติที่ใช้ในการประเมินค่าความสำคัญของคำในเอกสาร 2) Word co-occurrences เป็นค่าทางด้านสถิติที่พิจารณาจากการเกิดขึ้นร่วมกันของสองคำที่อยู่ติดกันในประโยคเดียวกันในเอกสารต่าง ๆ เป็นต้น

2.3.3 วิธีการทางการเรียนรู้ของเครื่องจักร (Machine learning approach)

เป็นวิธีการที่ใช้การเรียนรู้ของเครื่องจักรเข้ามาช่วย โดยเรียนรู้จากเอกสารตัวอย่างที่นำมาใช้ในการสอนและสร้างโมเดลสำหรับสกัดคำสำคัญจากเอกสารต่าง ๆ ซึ่งวิธีการทางการเรียนรู้ของเครื่องจักรสามารถแบ่งออกเป็น 2 ประเภทคือ

2.3.3.1 วิธีการเรียนรู้แบบมีผู้สอน (Supervised learning)

เป็นการเรียนรู้ที่สามารถจำแนกข้อมูลภายในชุดข้อมูลว่ามีผลลัพธ์ที่ผิดหรือถูกต้องได้ ที่ซึ่งในชุดข้อมูลจะประกอบไปด้วยข้อมูลฝึกสอนและข้อมูลเป้าหมาย สำหรับข้อมูลเหล่านี้จะถูกนำไปพยากรณ์ค่าข้อมูล โดยพื้นฐานการพยากรณ์จะพิจารณาจากข้อมูลทั้งหมดในชุดข้อมูล โดยเทคนิคที่ใช้วิธีการเรียนรู้แบบมีผู้สอน ได้แก่ Naïve Bayes, Support Vector Machine เป็นต้น

2.3.3.2 วิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning)

เป็นการเรียนรู้ที่ไม่มีการกำหนดข้อมูลเป้าหมาย จึงไม่มีการจำแนกข้อมูลว่ามีผลลัพธ์อย่างไร แต่เป็นการเรียนรู้ที่เกี่ยวข้องกับความสัมพันธ์ของข้อมูล จะนำไปใช้สำหรับจำแนกหรือแบ่งข้อมูลออกเป็นกลุ่ม ๆ

2.3.4 วิธีการทางความเชี่ยวชาญเฉพาะด้าน (Domain specific approach)

เป็นวิธีการที่นำคลังคำศัพท์ในศาสตร์ด้านใดด้านหนึ่งมาประยุกต์ใช้กับวิธีการต่าง ๆ ซึ่งจะใช้ประโยชน์ในด้านขององค์ความรู้เชิงลึกที่เกี่ยวข้องกับศาสตร์ด้านนั้น ๆ และคลังคำศัพท์เหล่านั้นยังช่วยประสิทธิภาพของผลลัพธ์ในการการระบุ หรือสกัดคำสำคัญนั้นดียิ่งขึ้น

2.4 งานวิจัยที่เกี่ยวข้อง

ในส่วนนี้จะอธิบายถึงงานวิจัยที่เกี่ยวข้องทางการแนะนำหนังสือ (Recommender system) โดยแบ่งการพิจารณาในการแนะนำออกเป็น 5 แง่มุม ได้แก่ การแนะนำหนังสือจากการพิจารณาเนื้อหาสาระในหนังสือ, การแนะนำหนังสือจากการให้คะแนนความนิยมของผู้อ่าน, การแนะนำหนังสือที่สืบเนื่องจากการแนะนำผู้เขียน, การแนะนำหนังสือจากประวัติการยืมหนังสือของผู้ใช้ห้องสมุด และการแนะนำหนังสือโดยพิจารณาจากระดับทักษะการอ่าน นอกจากนี้ยังอธิบายถึงงานวิจัยที่เกี่ยวข้องทางการสกัดคำสำคัญ (Keyword extraction) ดังต่อไปนี้

2.4.1 งานวิจัยที่เกี่ยวข้องทางการแนะนำหนังสือ (Recommender system)

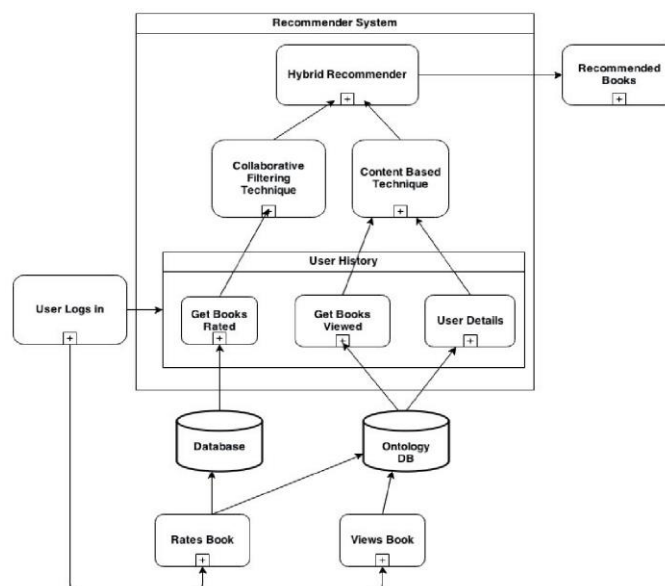
2.4.1.1 การแนะนำหนังสือจากการพิจารณาเนื้อหาสาระในหนังสือ

Garrido et al., 2014a นำเสนอวิธีการสร้างรายการแนะนำหนังสือให้กับผู้อ่านแต่ละบุคคล จากสมมติฐานที่ว่าคำอธิบายหนังสือและบทความวิจารณ์ของผู้อ่านคนอื่นของหนังสือสามารถช่วยให้ผู้อ่านได้รับรายการหนังสือที่ตรงกับความชื่นชอบของผู้อ่าน จากแนวความคิดข้างต้น A. L. Garrido et al. ได้นำเสนอวิธีการสร้างรายการแนะนำหนังสือที่เรียกว่า “Topic Map Recommender (TMR)” โดยการพิจารณาความสัมพันธ์ของหนังสือ (Topic map) ซึ่งสกัดจากคำอธิบายหนังสือโดยการใช้เครื่องมือที่เรียกว่า “TM-Gen” ที่จะพิจารณาคำสำคัญของคำอธิบายแล้วทำการวิเคราะห์ความสัมพันธ์ของข้อมูลร่วมกับการพิจารณาบทความวิจารณ์ของผู้อ่านคนอื่น โดยใช้วิธีการให้ค่าน้ำหนักของคำสำคัญ (TF-IDF) ในการวิเคราะห์บทความวิจารณ์ จากนั้นรายการแนะนำหนังสือของผู้อ่านจะถูกพิจารณาจากความคล้ายของความสัมพันธ์ของหนังสือทั้งในเชิงบวกและเชิงลบที่คาดว่าผู้อ่านน่าจะชื่นชอบ นอกเหนือจากระบบ TMR Garrido et al., 2014b นำเสนอการสร้างระบบแนะนำหนังสือที่มีชื่อว่า “SOLE-R” ที่ทำการพิจารณาความหมาย (Semantic) ออนโทโลยี (Ontology) และคุณลักษณะทางด้านภาษา (Linguistic) เพื่อให้สามารถแนะนำหนังสือได้อย่างมีประสิทธิภาพ นอกจากนี้ยังทำการพิจารณาความสัมพันธ์ของหนังสือ (Topic map) เพื่อทำการแนะนำหนังสือให้กับผู้อ่านหนังสือคนหนึ่ง ๆ

2.4.1.2 การแนะนำหนังสือจากการให้คะแนนความนิยมของผู้อ่าน

Linden et al., 2003 นำเสนอการแนะนำหนังสือที่คาดว่าจะถูกซื้อพร้อมกันของผู้ใช้เว็บไซต์ Amazon โดยงานวิจัยนี้ได้เสนอการแนะนำหนังสือด้วยวิธี “Item-to-Item Collaborative Filtering” ที่จะพิจารณาข้อมูลหนังสือที่ถูกซื้อ หรือถูกให้คะแนนความนิยมจากผู้ใช้นั้น ๆ โดยพิจารณาร่วมกับข้อมูลหนังสือที่ได้รับคะแนนความนิยมคล้าย ๆ กัน (ข้อมูลหนังสือที่พิจารณาอาจประกอบไปด้วย ชื่อผู้เขียน คำสำคัญ ประเภทหรือหมวดหมู่ของหนังสือ และอื่น ๆ) ซึ่งจากการพิจารณาดังกล่าวจะทำให้สามารถสร้างตารางหรือเมทริกซ์ที่แสดงถึงความเหมือนและความเกี่ยวเนื่องกันของหนังสือที่ซึ่งจะช่วยให้สามารถค้นหารายการหนังสือที่คาดว่าน่าจะถูกซื้อพร้อมกันได้

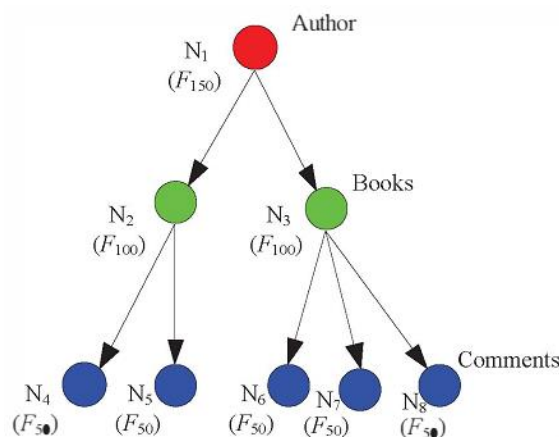
Chandak et al., 2015 นำเสนอการแนะนำหนังสือให้แก่ผู้อ่านด้วยการพิจารณาข้อมูลอินพุต 3 ส่วนหลัก คือ 1) ข้อมูลผู้อ่าน (เช่น อายุ เพศ และอื่น ๆ) 2) ข้อมูลหนังสือ (เช่น คำสำคัญ ประเภทหนังสือ และอื่น ๆ) และ 3) คะแนนความนิยมของผู้อ่าน ตามลำดับ โดยในงานวิจัยนี้ได้นำเสนอระบบการแนะนำหนังสือแบบไฮบริด (Hybrid book recommendation system) ที่จะผสมผสานการทำงานระหว่างขั้นตอนวิธี “Collaborative Filtering” และ “Content-based recommendation” รวมถึงการพิจารณาข้อมูลผู้อ่านเป็นส่วนประกอบหนึ่งของการแนะนำหนังสือ (ดังแสดงแผนผังการทำงานในรูปที่ 1) ที่จะสามารถเพิ่มมิติการพิจารณาและลดทอนข้อเสียของการประยุกต์ใช้วิธีการ “Collaborative Filtering” หรือ “Content-based recommendation” เพียงอย่างเดียว โดยจากอินพุตทั้ง 3 ส่วนข้างต้นจะถูกแยกการประมวลผลออกเป็นสองส่วนหลักและแยกเก็บไว้ในฐานข้อมูล และฐานข้อมูลออนไลน์ เพื่อส่งต่อให้กับการคำนวณแบบ “Collaborative Filtering” และ “Content-based recommendation” โดยจากการทดลองเพื่อทดสอบประสิทธิภาพจะทำให้เราทราบได้ว่าวิธีการดังกล่าวสามารถแนะนำหนังสือได้อย่างมีประสิทธิภาพ และลดปัญหาที่เกิดขึ้นจากการประยุกต์ใช้วิธีการคำนวณหนึ่ง ๆ ได้



ภาพที่ 2-4 ระบบแนะนำหนังสือแบบไฮบริดรวมถึงการพิจารณาข้อมูลผู้อ่านหนังสือ

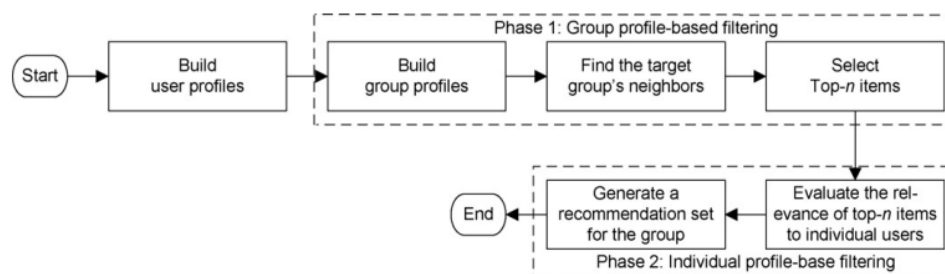
2.4.1.3 การแนะนำหนังสือที่สืบเนื่องจากการแนะนำผู้เขียน

Lu & Zhang, 2015 นำเสนอการแนะนำผู้เขียนหนังสือให้แก่ผู้อ่านหนังสือที่ซึ่งจะทำให้ผู้อ่านสามารถค้นหาหนังสือของผู้เขียนที่ถูกแนะนำได้ โดยงานวิจัยนี้มีการตั้งสมมติฐานที่ว่า “บุคคลหนึ่ง ๆ ที่ชอบหรือรักในการอ่านหนังสือมักจะชื่นชอบและติดตามหนังสือของผู้เขียนที่ตนเองชื่นชอบ บุคคลเหล่านี้มักจะมีคามยินดีที่จะอ่านหนังสือจากผู้เขียนอื่น ๆ ที่มีแนวทางการเขียนหนังสือที่คล้ายคลึงกับนักเขียนที่พวกเขาชอบ” ซึ่งจากสมมติฐานดังกล่าว Lu & Zhang ได้เสนอการแนะนำผู้เขียนให้แก่ผู้อ่านหนังสือพร้อมกับหนังสือที่ถูกเขียนโดยผู้เขียนที่ถูกแนะนำ และข้อคิดเห็นของผู้อ่านอื่น ๆ ที่เคยอ่านหนังสือของผู้เขียนที่ถูกแนะนำ ด้วยการประยุกต์ใช้โครงสร้างข้อมูลแบบต้นไม้ 3 ระดับ (ดังแสดงตัวอย่างในรูปที่ 2 โดยระดับแรกใช้สำหรับจัดเก็บข้อมูลที่เกี่ยวข้องกับผู้เขียน ระดับที่สองใช้สำหรับจัดเก็บข้อมูลหนังสือที่ถูกเขียนโดยผู้เขียนนั้น ๆ และ ระดับที่ 3 ใช้สำหรับจัดเก็บข้อมูลข้อคิดเห็น (50 ข้อคิดเห็นแรก) จากผู้ที่ทำการอ่านหนังสือเล่มนั้น ๆ ของผู้เขียน ตามลำดับ) และทำการประยุกต์ใช้เทคนิค Multi-Layer Self-Organizing Map (ML-SOM) เพื่อทำการจัดกลุ่มความคล้ายคลึงกันระหว่างผู้เขียนหนังสือที่ซึ่งจะสามารถทำให้ทราบถึงผู้เขียนที่มีสไตล์ที่คล้ายคลึงกัน อันนำมาซึ่งการแนะนำหนังสือจากผู้เขียนที่มีความคล้ายคลึงกันได้



ภาพที่ 2-5 โครงสร้างข้อมูลต้นไม้ 3 ระดับ

Kim et al, 2010 นำเสนอวิธีการแนะนำหนังสือให้กับกลุ่มผู้อ่านที่มีลักษณะความชอบหรือพฤติกรรมในการอ่านหนังสือที่คล้ายกัน จากสมมติฐานที่ว่าผู้อ่านหนังสือที่มีความชอบในการเลือกอ่านหนังสือที่คล้ายกันจะมีการแบ่งปันข้อมูลหนังสือที่มีความน่าสนใจกับผู้อ่านคนอื่น ๆ ภายในกลุ่ม ส่งผลให้ผู้อ่านในกลุ่มเดียวกันสามารถได้รับการแนะนำหนังสือที่ชื่นชอบโดยการพิจารณาการเลือกอ่านหนังสือของผู้อ่านคนอื่นภายในกลุ่ม ดังนั้น Kim et al. จึงได้นำเสนอวิธีการสร้างรายการแนะนำหนังสือด้วย 2 ขั้นตอนวิธีการแสดงดังภาพ โดยขั้นตอนที่ 1) พิจารณาความชอบของกลุ่มผู้อ่านที่มีลักษณะความชอบในการอ่านหนังสือที่คล้ายกัน ด้วยวิธีการกรองข้อมูลร่วม (Collaborative Filtering) จากขั้นตอนนี้รายการหนังสือของกลุ่มผู้อ่านที่มีลักษณะความชอบคล้ายกันจะถูกแนะนำร่วมกับรายการหนังสือที่ถูกเลือกซื้อ มากภายในกลุ่มผู้อ่าน และขั้นตอนที่ 2) วิธีการที่นำเสนอนี้ได้มีการพิจารณาความชอบของการอ่านหนังสือของผู้อ่านแต่ละบุคคล โดยหนังสือที่ผู้อ่านไม่ชื่นชอบจะไม่ถูกแนะนำ โดยการพิจารณาคีย์เวิร์ด (keyword) ของหนังสือจากประวัติการอ่านหนังสือของผู้อ่านแต่ละบุคคล



ภาพที่ 2-6 ขั้นตอนการแนะนำแบบ top-n items

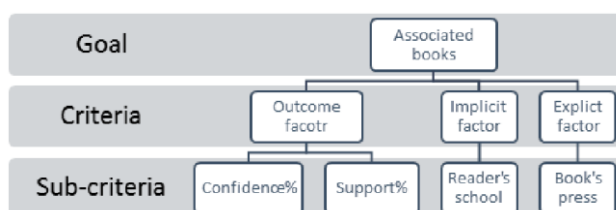
Vaz et al., 2012 นำเสนอวิธีการสร้างรายการแนะนำหนังสือให้กับผู้อ่าน เนื่องจากหนังสือภายในระบบมีจำนวนมากและมีความหลากหลายของเนื้อหาที่แตกต่างกัน ส่งผลให้ผู้อ่านมีความยุ่งยากในการเข้าถึงหนังสือที่ผู้อ่านชื่นชอบ ดังนั้น Vaz et al. ได้นำเสนอวิธีการสร้างรายการแนะนำหนังสือที่คาดว่าผู้อ่านชื่นชอบ จากสมมติฐานที่ว่าผู้อ่านน่าจะมีความชอบในการเลือกอ่านหนังสือจากสไตล์การเขียนหนังสือของนักเขียนที่ผู้อ่านเคยอ่านในอดีต วิธีการที่นำเสนอนี้จึงได้สร้างรายการหนังสือโดยการรวมค่าการทำนายคะแนนความชอบจาก 2 ขั้นตอนวิธีการ คือ 1) วิธีการกรองข้อมูลร่วมโดยพิจารณาความชอบของหนังสือเป็นหลัก (item based CF) ซึ่งพิจารณาจากประวัติของผู้อ่านที่ให้คะแนนความชอบกับหนังสือในระบบ และ 2) วิธีการกรองเนื้อหา (Content-based filtering) ซึ่งพิจารณาสไตล์การเขียนหนังสือจากคำศัพท์ที่ใช้ในการเขียน, ขนาดของเนื้อหา, คำศัพท์ที่ใช้ร่วมกัน และคำศัพท์ที่ใช้บ่อยร่วมกับค่าคะแนนความชอบของหนังสือในเชิงบวกและเชิงลบจากคะแนนความชอบของหนังสือ โดยทั้ง 2 วิธีการจะถูกรวมเป็นค่าการทำนายความชอบของหนังสือโดยการใช้ค่าเฉลี่ยเพื่อสร้างรายการแนะนำหนังสือต่อไป

2.4.1.4 การแนะนำหนังสือจากประวัติการยืมหนังสือของผู้ใช้ห้องสมุด

Saxena et al., 2014 เสนอการแนะนำหนังสือให้แก่ผู้ใช้ห้องสมุดโดยตั้งสมมติฐานที่ว่าในการค้นหนังสือเล่มหนึ่ง ๆ ที่เกี่ยวกับเรื่องหนึ่ง ๆ ที่ผู้ใช้ห้องสมุดสนใจจะมีหนังสือหลายเล่มที่เกี่ยวข้องกับเรื่องนั้น ๆ โดยจากการที่มีหนังสือหลายเล่มทำให้ผู้ใช้ห้องสมุดไม่สามารถตัดสินใจได้ว่าจะเลือกยืมหนังสือเพื่ออ่านเพิ่มเติมเนื่องจากไม่มีทราบข้อมูลเบื้องต้นว่าหนังสือแต่ละเล่มดีหรือไม่ดีอย่างไร ดังนั้น Saxena et al. ได้เสนอการประยุกต์ใช้แนวคิดของการค้นหารูปแบบปรากฏบ่อย (Frequent itemsets mining) เพื่อทำการค้นหาค่าความถี่ของจำนวนครั้งที่หนังสือเล่มหนึ่ง ๆ ถูกค้นหา หรือถูกยืมจากผู้ใช้ห้องสมุดทั้งหมด

โดยผลลัพธ์ที่สามารถค้นคืนได้จะประกอบไปด้วยชื่อผู้แต่งหนังสือ ชื่อหนังสือ สำนักพิมพ์ และจำนวนครั้งที่หนังสือเล่มนั้น ๆ ถูกค้นหาหรือถูกยืมจากผู้ใช้ห้องสมุดทั้งหมด ซึ่งจากผลลัพธ์ดังกล่าวจะทำให้ผู้ใช้ห้องสมุดมีข้อมูลเบื้องต้นเกี่ยวกับความนิยมหนังสือ อันนำมาซึ่งการตัดสินใจในการเลือกยืมหนังสือที่ดีขึ้นได้

Teng et al., 2015 เสนอการแนะนำหนังสือให้แก่ผู้ใช้ห้องสมุดจากประวัติการยืมหนังสือในอดีตของผู้ใช้ห้องสมุดด้วยการประยุกต์ใช้แนวคิด Fuzzy Analytical Hierarchy Process (FAHP) ร่วมกับอัลกอริทึม Apriori โดยขั้นตอนเริ่มแรก FAHP จะทำการแบ่งปัญหาออกเป็นปัญหาย่อย ๆ โดยแต่ละปัญหาย่อยสามารถวิเคราะห์ได้โดยเป็นอิสระต่อกัน (ดังแสดงลำดับชั้นของการแบ่งปัญหาย่อยในรูปที่ 4) ซึ่งจากการแนวคิดข้างต้น จะทำให้สามารถแบ่งส่วนของปัญหาออกเป็น 3 ส่วนย่อย คือ 1) Outcome factor จะเกี่ยวข้องกับการค้นหาค่าความถี่และค่าความเชื่อมั่นของหนังสือที่ถูกยืมร่วมกัน 2) Implicit factor จะเกี่ยวเนื่องกับข้อมูลพื้นฐานของผู้ใช้ห้องสมุด (อาทิเช่น ชื่อโรงเรียนของผู้ใช้ห้องสมุด โดยนักเรียนในโรงเรียนเดียวกันน่าจะมีความสนใจเหมือนกัน) และ 3) Explicit factor จะเกี่ยวเนื่องกับข้อมูลพื้นฐานของหนังสือ (อาทิเช่น สำนักพิมพ์ของหนังสือ) ตามลำดับ โดยจากการรวมกันของส่วนย่อยทั้งหมดและการกำหนดค่าความสำคัญของแต่ละส่วนย่อยโดยผู้ที่ต้องการวิเคราะห์ข้อมูลจะทำให้เราสามารถทราบถึงหนังสือที่มีความเกี่ยวเนื่องกันได้



ภาพที่ 2-7 การแบ่งปัญหาออกเป็นปัญหาย่อย



ภาพที่ 2-8 การแสดงผลหนังสือที่มีความเกี่ยวเนื่องกันซึ่งหาความสำคัญจาก FAHP

Chen et al., 2015 นำเสนอวิธีการแนะนำหนังสือให้กับผู้อ่านในห้องสมุดอิเล็กทรอนิกส์ ด้วยสมมติฐานที่ว่าผู้อ่านน่าจะมีความชอบในการเลือกอ่านหนังสือกับผู้อ่านคนอื่นที่มีพฤติกรรมการเลือกอ่านหนังสือที่คล้ายกันและหนังสือมีคุณลักษณะและคุณสมบัติที่หลากหลาย แต่อย่างไรก็ตามในการคำนวณค่าความคล้ายของลักษณะการอ่านหนังสือของผู้อ่านในระบบนั้นใช้เวลาในการคำนวณที่นาน เนื่องจากในระบบมีหนังสือจำนวนมากร่วมกับผู้อ่านมีการให้คะแนนกับหนังสือจำนวนน้อย ดังนั้น Chen et al. จึงได้นำเสนอการสร้างรายการแนะนำหนังสือ โดยใช้วิธีการกรองข้อมูลร่วมซึ่งพิจารณาความชอบในการอ่านหนังสือเพื่อแนะนำหนังสือให้กับผู้อ่านจากประวัติการยืมหนังสือของผู้อ่านที่มีพฤติกรรมการเลือกอ่านหนังสือที่คล้ายกัน ร่วมกับพิจารณาความสัมพันธ์ของคุณสมบัติของหนังสือของผู้อ่านโดยใช้วิธีการออนโทโลยี (Ontology) เพื่อสร้างรายการแนะนำหนังสือ โดยวิธีการที่นำเสนอมีการประมวลผลอยู่บนสถาปัตยกรรม Map-reduce โมเดลในเฟรมเวิร์คของ Hadoop ซึ่งสามารถใช้เวลาในการคำนวณที่รวดเร็วกว่าวิธีการแบบเดิม จากวิธีการที่นำเสนอนี้สามารถสร้างรายการแนะนำหนังสือได้อย่างมีประสิทธิภาพในด้านการประมวลผลข้อมูลและความถูกต้องของรายการแนะนำหนังสือ

2.4.1.5 การแนะนำหนังสือโดยพิจารณาจากระดับทักษะการอ่าน

Ng, 2016 ได้เสนอระบบ CRec สำหรับแนะนำหนังสือให้กับเด็กที่ทำการศึกษาในระบบ K-12 ของสหรัฐอเมริกา โดยตั้งสมมติฐานที่ว่าเด็กยังไม่มีความรู้พื้นฐานในการเลือกหนังสือที่ชอบหรือที่สนใจ ซึ่งจากสมมติฐานดังกล่าวจะทำให้เด็กใช้เวลาเลือกค่อนข้างมากและอาจต้องพึ่งพาคำแนะนำของผู้ปกครอง หรือครู/อาจารย์ที่อาจมีความชอบไม่เหมือนกัน ด้วยเหตุนี้ ระบบ CRec จึงถูกพัฒนาขึ้นเพื่อ

แก้ไขปัญหาลำต้น CRec จะประกอบด้วย 2 ขั้นตอนการทำงานหลัก คือ 1) การวิเคราะห์ระดับทักษะการอ่านของเด็ก ด้วยการวิเคราะห์ความยากง่ายของหนังสือที่เด็ก ๆ เคยอ่านไว้ และ 2) การระบุถึงหนังสือต่าง ๆ ที่ถูกจัดเก็บไว้ในเว็บไซต์ โดยหนังสือที่ถูกระบุจะมีความสอดคล้องกับระดับทักษะการอ่านของเด็กด้วยการผสมผสานระหว่างเทคนิค “user-based collaborative filtering” และ “content-based” ที่จะช่วยคัดกรองหนังสือที่จะแสดงให้เห็นแก่แต่ละคนเห็นด้วยการพิจารณาจากการให้คะแนนความนิยมของเด็กๆกับหนังสือ รวมถึงการวิเคราะห์คำอธิบาย (หรือเนื้อหา) ของหนังสือที่เด็กคนหนึ่ง ๆ เคยให้คะแนนความนิยมไว้ในอดีต ที่ซึ่งจะสามารถช่วยสร้างโปรไฟล์และสามารถทราบถึงความชื่นชอบของเด็กคนนั้น ๆ ได้ ซึ่งจากการดำเนินการดังกล่าวจะสามารถช่วยเพิ่มประสิทธิภาพของการแนะนำหนังสือได้

2.4.2 งานวิจัยที่เกี่ยวข้องทางด้านสกัดคำสำคัญ (Keyword extraction)

Frantzi et al., 2000 นำเสนอวิธีการ C-value และ NC-value สำหรับสกัดกลุ่มคำสำคัญแบบอัตโนมัติ โดยวิธีการทั้งสองนี้ประยุกต์ใช้วิธีการทางภาษาศาสตร์ (Linguistic approach) และวิธีการทางสถิติ (Statistical approach) ในการสกัดกลุ่มคำสำคัญ ดังนี้ 1) วิธีการ C-value เป็นขั้นตอนสร้างรายการแคนดิเดต (Candidate) ของกลุ่มคำสำคัญที่ถูกสกัดออกมาจากประโยค โดยจะพิจารณาด้วยวิธีการทางภาษาศาสตร์ ซึ่งกลุ่มคำสำคัญที่สนใจคือนามวลี (Noun phrase) และการนับความถี่ของกลุ่มคำสำคัญภายในรายการแคนดิเดตของกลุ่มคำสำคัญที่ถูกสกัด โดยใช้วิธีการทางสถิติ จากนั้นนำผู้เชี่ยวชาญเข้ามาประเมินผลลัพธ์ในรายการแคนดิเดตของกลุ่มคำสำคัญที่ถูกสกัด เมื่อผู้เชี่ยวชาญได้ระบุถึงคำสำคัญที่ต้องการแล้ว จะนำไปสู่วิธีการที่ 2) วิธีการ NC-value เป็นการพิจารณาคำอื่น ๆ ที่เกิดขึ้นร่วมกับคำสำคัญที่ผู้เชี่ยวชาญกำหนด รวมถึงความถี่ในการเกิดขึ้นร่วมกัน และระบุค่าที่เกิดขึ้นร่วมกับคำสำคัญ เป็นกลุ่มคำสำคัญที่สามารถสกัดได้ ซึ่งผลลัพธ์ในการสกัดกลุ่มคำสำคัญด้วยวิธีการ NC-value มีประสิทธิภาพมากกว่า C-value และ Frantzi et al. ยังทำการสร้างซอฟต์แวร์ชื่อว่า “Termine” จากวิธีการดังกล่าวข้างต้น เพื่อให้บุคคลทั่วไปสามารถใช้ในการสกัดคำสำคัญได้อีกด้วย

Hulth, 2003 ได้เสนอการสกัดคำสำคัญโดยอัตโนมัติจากบทคัดย่อภาษาอังกฤษ และนำวิธีการทางภาษาศาสตร์ (Linguistic approach) มาใช้ร่วมด้วยในการสกัดคำสำคัญ แทนที่จะใช้วิธีการทางสถิติเพียงอย่างเดียว โดยในการสกัดคำสำคัญแบ่งเป็น 3 วิธีการ ได้แก่ 1) วิธีการ n-gram โดยจะกรอง

ข้อความที่ไม่ใช่ตัวอักษรและข้อความที่เป็นตัวเลขออกไป แต่จะสนใจเก็บคำนามเฉพาะ เช่น ชื่อผู้คน, ชื่อเรื่อง เป็นต้น 2) วิธีการ NP-chunks จะสกัดคำสำคัญโดยสนใจเฉพาะนามวลี (Noun phrase) 3) POS Tag Patterns จะสกัดคำสำคัญ โดยพิจารณาหน้าที่ของคำในประโยค ซึ่งจะสนใจเฉพาะคำนามที่เกิดขึ้นร่วมกัน และคำนามที่เกิดขึ้นร่วมกับคำคุณศัพท์เท่านั้น และผลลัพธ์ในการสกัดคำสำคัญโดยวิธีการ NP-chunks ให้ค่าความถูกต้อง (Precision) สูงที่สุด

Mihalcea et al., 2004 นำเสนอวิธีการ TextRank ที่ใช้วิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) โดยทำการสร้าง Graph-based ranking ซึ่งเป็นโมเดลกราฟสำหรับการจัดลำดับคำสำคัญที่ได้จากข้อความ โดยจะนำคำทุกคำในเอกสารมาสร้างกราฟเชื่อมโยงคำที่อยู่ติดกันไปเรื่อย ๆ ต่อมาจะพิจารณาการเชื่อมโยงของคำที่เกิดขึ้นร่วมกัน โดยพิจารณาคำภายในโครงสร้างกราฟที่มีเส้นเชื่อมมากที่สุดแล้วจึงนำคำมารวมกัน ในการรวมกันระหว่างคำหนึ่ง ๆ กับคำอื่น ๆ จะใช้แนวความคิดของการ “voting” ระหว่างคำ เพื่อรวมคำเข้าด้วยกัน จากนั้นจะคำนวณน้ำหนักของคำที่นำมารวมกัน และสร้างรายการคำสำคัญจากค่าน้ำหนัก สุดท้ายจะสกัดคำสำคัญจากการจัดอันดับค่าน้ำหนัก และทำการคัดกรองคำสำคัญที่ค่าน้ำหนักผ่านเกณฑ์

Rose et al., 2010 ได้นำเสนอวิธีการสกัดคำโดยอัตโนมัติจากบทความภาษาอังกฤษ และได้นำวิธีการทางภาษาศาสตร์ (Linguistic approach) เข้ามาใช้ในการพิจารณาคำสำคัญ วิธีการนี้เรียกว่า Rapid automatic keyword extraction หรือ RAKE โดยวิธีการมีดังนี้ 1) พิจารณาคำที่อยู่ติดกันไปเรื่อย ๆ จนเจอคำหยุด (Stop word) แล้วจึงนำคำเหล่านั้นมาสร้างรายการแคนดิเดทของคำสำคัญ 2) ทำการสร้างเมทริกซ์ (Matrix) ของแต่ละคำที่ปรากฏในรายการแคนดิเดทของคำสำคัญ เพื่อคำนวณคะแนนความถี่ของคำที่เกิดขึ้นร่วมกัน 3) รวมคะแนนของคำสำคัญที่ปรากฏในรายการแคนดิเดทของคำสำคัญ และทำการเรียงลำดับรายการแคนดิเดทของคำสำคัญโดยพิจารณาคะแนนของคำสำคัญแต่ละคำจากมากไปหาน้อย 4) พิจารณา top-T scoring โดยจะสนใจเพียง 1 ใน 3 คำสำคัญจากรายการแคนดิเดทของคำสำคัญ วิธีการ RAKE ได้ทำการเปรียบเทียบกับวิธีการ n-gram, NP-chunks และ POS Tag Patterns ของ Hulth, 2003 และวิธีการ TextRank ของ Mihalcea et al., 2004 ซึ่งผลลัพธ์ในการสกัดคำสำคัญโดยวิธีการ RAKE ให้ค่าความถูกต้อง (Precision) มากที่สุด

จากงานวิจัยที่เกี่ยวข้องทางการแนะนำหนังสือที่ได้นำเสนอมาทั้ง 5 แง่มุมข้างต้น ได้แก่ การแนะนำหนังสือจากการพิจารณาเนื้อหาสาระในหนังสือ, การแนะนำหนังสือจากการให้คะแนนความนิยมของผู้อ่าน, การแนะนำหนังสือที่สืบเนื่องจากการแนะนำผู้เขียน, การแนะนำหนังสือจากประวัติการยืมหนังสือของผู้ใช้ห้องสมุด และการแนะนำหนังสือโดยพิจารณาจากระดับทักษะการอ่าน แต่อย่างไรก็ตามแนวคิดทั้งหมดไม่ได้มุ่งเน้นที่การแนะนำหนังสือจากความสอดคล้องของหนังสือกับ คำอธิบายรายวิชาที่ประกอบไปด้วยเนื้อหาต่าง ๆ และ/หรือประมวลผลรายวิชา ที่ซึ่งช่วยให้ผู้อ่านได้รับการแนะนำที่สอดคล้องกับเนื้อหาที่ตนเองสนใจมากยิ่งขึ้น

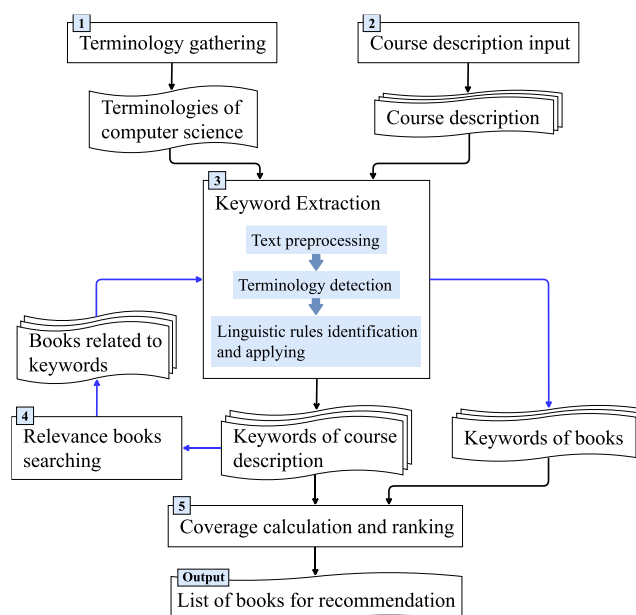
ด้วยเหตุนี้งานวิจัยนี้มุ่งเน้นที่การแนะนำหนังสือโดยพิจารณาจากความสอดคล้องของหนังสือกับ คำค้นที่เฉพาะเจาะจงเป็นหลัก โดยจะมุ่งเน้นที่จะทำการแนะนำหนังสืออ่านเพิ่มเติมจากการศึกษาในโรงเรียน หรือ มหาวิทยาลัยผ่านการพิจารณาคำอธิบายรายวิชาและ/หรือประมวลผลรายวิชาเพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเอง อันนำมาซึ่งการพัฒนาตนเอง พัฒนาการเรียนรู้อย่างยั่งยืน ผู้วิจัยจึงประยุกต์ใช้วิธีการกรองเนื้อหาที่จะทำการพิจารณาถึงคำสำคัญระหว่างคำอธิบายรายวิชาและหนังสือเป็นสำคัญ และใช้วิธีการหาค่าความครอบคลุม (Coverage) ระหว่างคำอธิบายรายวิชาและหนังสือ เพื่อแสดงให้เห็นถึงคำสำคัญที่ครอบคลุมระหว่างในหนังสือเล่มหนึ่ง ๆ ต่อคำอธิบายรายวิชาหนึ่ง ๆ และทำการสร้างรายการแนะนำหนังสืออ่านเพิ่มเติม

นอกจากนี้การแนะนำหนังสืออ่านเพิ่มเติมที่มีความสอดคล้องกับคำอธิบายรายวิชาและ/หรือประมวลผลรายวิชานั้น จะต้องทำการสกัดคำสำคัญระหว่างคำอธิบายรายวิชาและ/หรือประมวลผลรายวิชาและหนังสืออ่านเพิ่มเติมที่ต้องการแนะนำ วิธีการสกัดคำสำคัญจึงมีบทบาทสำคัญในขั้นตอนวิธีการแนะนำหนังสือ จากงานวิจัยที่เกี่ยวข้องทางการสกัดคำสำคัญข้างต้น จะเห็นได้ว่าวิธีการทางภาษาศาสตร์นั้นมีส่วนช่วยให้ผลลัพธ์ในการสกัดคำสำคัญมีความถูกต้องมากยิ่งขึ้น งานวิจัยนี้จึงประยุกต์ใช้วิธีการทางภาษาศาสตร์ร่วมกับวิธีการทางความเชี่ยวชาญเฉพาะด้านซึ่งเป็นการนำคลังคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ เพื่อสกัดคำสำคัญได้อย่างมีประสิทธิภาพ

บทที่ 3

วิธีการที่นำเสนอ

งานวิจัยนี้ได้นำเสนอระบบแนะนำหนังสืออ่านเพิ่มเติม (Supplementary book suggestion system) หรือเรียกว่า “SBS system” เพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเองภายใต้คำค้นของผู้ใช้ สำหรับงานวิจัยนี้จะมุ่งเน้นที่การแนะนำหนังสือจากคำอธิบายรายวิชาและ/หรือประมวลรายวิชา มาสร้างความเชื่อมโยงกับเนื้อหาในหนังสือเล่มหนึ่ง ๆ โดยขั้นตอนการดำเนินการจะประกอบไปด้วย 5 ขั้นตอนหลัก 1) รวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ 2) รวบรวมข้อมูลคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ 3) การสกัดคำสำคัญจากคำอธิบายรายวิชา (รวมทั้งการสกัดคำสำคัญจากหนังสือ) 4) การค้นหาหนังสือที่มีความเกี่ยวข้องกับคำสำคัญในคำอธิบายรายวิชา และ 5) การคำนวณความเกี่ยวข้องของหนังสือที่ครอบคลุมคำสำคัญในคำอธิบายรายวิชาและการจัดอันดับหนังสือที่มีความครอบคลุมคำสำคัญมากไปยังหนังสือที่มีความครอบคลุมคำสำคัญน้อย จากขั้นตอนข้างต้นแสดงดังภาพที่ 3-1



ภาพที่ 3-1 ขั้นตอนวิธีระบบแนะนำหนังสืออ่านเพิ่มเติม

3.1 รวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์

ศาสตร์ทางด้านวิทยาการคอมพิวเตอร์เป็นศาสตร์ที่มีคำศัพท์เฉพาะที่มีความหมายเฉพาะค่อนข้างมาก เช่น “Natural Language Processing”, “Data Mining”, “Data Structure”, “Data Science”, “Big Data” เป็นต้น ซึ่งจะช่วยในการระบุถึงคำสำคัญที่เฉพาะเจาะจงได้ นอกจากนี้การระบุถึงคำศัพท์เฉพาะในชื่อวิชา, คำอธิบายรายวิชา, ชื่อหนังสือ และสารบัญของหนังสือ สามารถช่วยในการทำความเข้าใจเกี่ยวกับหัวข้อสำคัญ ๆ ที่ควรศึกษาในแต่ละหลักสูตร ดังนั้นในการสกัดคำสำคัญจึงควรเริ่มต้นจากการรวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์เป็นอันดับแรก ซึ่งระบบแนะนำหนังสืออ่านเพิ่มเติมนั้นได้รวบรวมคำศัพท์เฉพาะจากพจนานุกรมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่จัดทำขึ้นโดยมหาวิทยาลัยออกซฟอร์ด (Oxford University)¹ และรายการคำศัพท์เฉพาะทางด้าน การเขียนโปรแกรมและวิทยาการคอมพิวเตอร์ ข้อมูลจากเว็บไซต์ Labautopedia²

โดยการได้มาซึ่งคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์จะใช้วิธีการสกัดข้อความบนเว็บไซต์ (Web Scraping) ซึ่งเป็นเทคนิคในการรวบรวมข้อมูลจากเว็บไซต์โดยอัตโนมัติ ภาพที่ 3-2 แสดงเว็บไซต์ที่มีข้อมูลพจนานุกรมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่จัดทำขึ้นโดยมหาวิทยาลัยออกซฟอร์ด (Oxford University) และภาพที่ 3-3 แสดงเว็บไซต์ที่มีข้อมูลรายการคำศัพท์เฉพาะทางด้าน การเขียนโปรแกรมและวิทยาการคอมพิวเตอร์ ข้อมูลจากเว็บไซต์ Labautopedia โดยได้ระบุถึงส่วนของข้อความที่ต้องการบนเว็บไซต์ภายในกรอบสี่ฟ้าที่ปรากฏในภาพที่ 3-2 และภาพที่ 3-3

หลังจากดำเนินการสกัดข้อความบนเว็บไซต์ (Web Scraping) เป็นที่เรียบร้อยแล้วนั้น จะนำคำศัพท์เฉพาะเหล่านั้นเก็บรวบรวมไว้ในฐานข้อมูลของคลังคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ (Terminology corpus)

¹ <http://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975>

² http://www.labautopedia.org/mw/List_of_programming_and_computer_science_terms

Oxford Reference

Search

Search within my subject specializations: Select ...

A Dictionary of Computer Science (7 ed.)

Edited by **Andrew Butterfield** and **Gerard Ekembe Ngondi**

Previous Edition (6 ed.)

Over 6,500 entries

Previously named *A Dictionary of Computing*, this bestselling dictionary has been fully revised by a team of computer specialists, making it the most up-to-date and authoritative guide to computing available. Including expanded coverage of multimedia, computer applications, networking, and personal computing, it encompasses all aspects of the subject. Terms are defined in a jargon-free and concise manner, with helpful examples where relevant, and the appendices include useful resources such as generic domain names, file extensions, and the Greek alphabet. This dictionary is s ... [More](#)

Andrew Butterfield, editor
Gerard Ekembe Ngondi, editor

Andrew Butterfield holds an honours degree in Engineering and a PhD in Computer Science and is currently Head of the Foundation and Methods Group at Trinity College Dublin, as well as Director of Undergraduate Teaching and Learning in the School of Computer [More](#)

BURAPHA UNIVERSITY

Sign in

Search within work

Subject(s) in Oxford Reference

- Mathematics and Computer Science
- Science and technology

Publishing Information

ALL CONTENTS | ENTRIES

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Items per page: 10 Starting with: Page 1 2 3 4 5 6 7 8 9 10 11 ... 651 652

A* algorithm

abduction

abelian group

ABI

ablative

ภาพที่ 3-2 ตัวอย่างคำศัพท์เฉพาะที่ต้องการรวบรวมจากพจนานุกรมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่จัดทำขึ้นโดยมหาวิทยาลัยออกซฟอร์ด (Oxford University)

LABAUTOPEDIA™
The SLAS Knowledge Network

Search

History View source Discussion Page

Personal tools

- Log in / create account

Navigation

- Main Page
- Contents
- Random page
- Create page
- Help
- Collaborative sites
- Interaction
- Community portal
- Current events
- Recent changes

List of programming and computer science terms

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Numerical

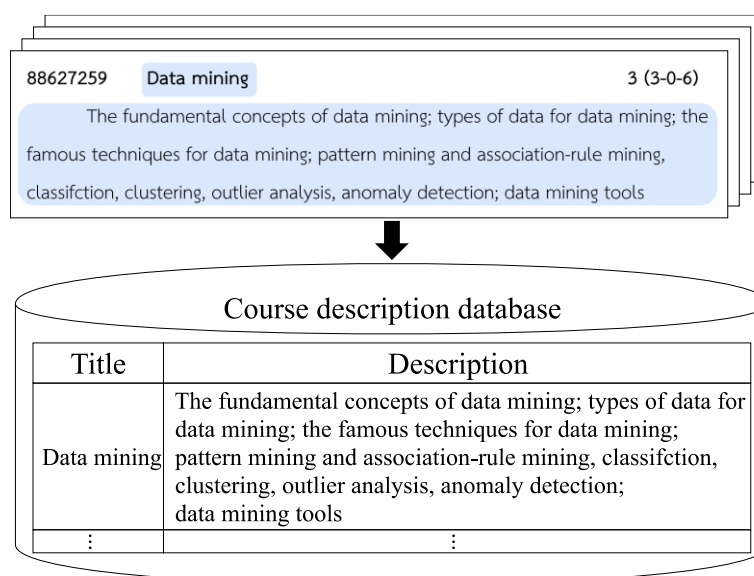
A

- AARON** - AARON is a screensaver program written by artist Harold Cohen who created the original artistic images. AARON utilizes artificial intelligence to continuously create original paintings on PCs.
- Abandonware** - Abandonware refers to software that is no longer being sold or supported by its publisher. Most abandonware is still considered illegal to sell unless the publisher has re-released the software as freeware.
- Abstract Base Class** - An abstract base class, a term used in the object oriented programming language, is meant to be used as the base class from which other classes are derived.
- Abstract Type** - In software engineering, an abstract type is a type in a nominative type system which is declared by the programmer, and which has the property that it contains no members which also are not members of some declared subtype. In many object-oriented programming languages, abstract types are known as abstract base classes, interfaces, traits, mixins, flavors, or roles. Note that these names refer to different language constructs which are (or may be) used to implement abstract types.
- ACD Canvas** - ACD Canvas is a graphics, publishing, and imaging product for personal computers. Canvas provides tools for creating and editing vector graphics and pixel-based images. It is used for illustration, page

ภาพที่ 3-3 ตัวอย่างคำศัพท์เฉพาะที่ต้องการรวบรวมจากรายการคำศัพท์เฉพาะทางการเขียนโปรแกรมและวิทยาการคอมพิวเตอร์ ข้อมูลจาก Labautopedia

3.2 รวบรวมข้อมูลคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์

ในขั้นตอนนี้ได้ทำการรวบรวมคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ ซึ่งได้รวบรวมคำอธิบายรายวิชาจากสาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา เพื่อใช้เป็นตัวอย่างข้อมูลรับเข้าของระบบแนะนำหนังสืออ่านเพิ่มเติม โดยข้อมูลคำอธิบายรายวิชาที่สนใจพิจารณาคือ ชื่อรายวิชา และรายละเอียดของคำอธิบายรายวิชาในรูปแบบภาษาอังกฤษ ดังภาพที่ 3-4 แสดงการจัดเก็บข้อมูลคำอธิบายรายวิชา โดยส่วนที่เน้นสีฟ้าคือส่วนที่ต้องการจัดเก็บข้อมูลเหล่านั้นลงในฐานข้อมูลของคลังข้อมูลคำอธิบายรายวิชา



ภาพที่ 3-4 การจัดเก็บข้อมูลคำอธิบายรายวิชาลงในฐานข้อมูลของคลังข้อมูลคำอธิบายรายวิชา

3.3 การสกัดคำสำคัญจากคำอธิบายรายวิชา

ในขั้นตอนนี้จะอธิบายถึงวิธีการสกัดคำสำคัญที่ได้นำวิธีการทางภาษาศาสตร์ (Linguistic approach) ร่วมกับวิธีการทางความเชี่ยวชาญเฉพาะด้าน (Domain specific approach) ซึ่งเป็นการนำคลังคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่ได้สร้างขึ้นจากขั้นตอนที่ 3.1 มาประยุกต์ใช้เพื่อสกัดคำสำคัญได้อย่างมีประสิทธิภาพ

เมื่อใดก็ตามที่มีการรับข้อมูลรับเข้า (คำอธิบายรายวิชา) เข้ามาในระบบ กระบวนการสกัดคำสำคัญจะดำเนินการกับข้อมูลรับเข้า (คำอธิบายรายวิชา) เพื่อให้ทราบถึงหัวข้อสำคัญที่ซ่อนอยู่ในคำอธิบายรายวิชา ภาพที่ 3-5 แสดงตัวอย่างข้อมูลรับเข้าคำอธิบายรายวิชา “Data Mining” ซึ่งวิธีการในการสกัดคำสำคัญนี้ประกอบไปด้วย 3 ขั้นตอนที่เชื่อมต่อกัน ดังนี้

Course description	
Data mining	The fundamental concepts of data mining; types of data for data mining; the famous techniques for data mining; pattern mining and association-rule mining, classification, clustering, outlier analysis, anomaly detection; data mining tools

ภาพที่ 3-5 ตัวอย่างข้อมูลรับเข้าคำอธิบายรายวิชา “Data Mining”

3.3.1 การประมวลผลข้อความเบื้องต้น (Text preprocessing)

การประมวลผลข้อความเบื้องต้นนี้จะทำให้ข้อมูลคำอธิบายรายวิชาอยู่ในรูปแบบที่มีความพร้อมใช้งานในการสกัดคำสำคัญมากยิ่งขึ้น การประมวลผลข้อความเบื้องต้นประกอบไปด้วย 4 ขั้นตอนที่เชื่อมต่อกัน ภาพที่ 3-6 แสดงการประมวลผลข้อความเบื้องต้นจากข้อมูลคำอธิบายรายวิชา “Data Mining” ผ่านขั้นตอนทั้ง 4 ขั้นตอนที่เชื่อมต่อกัน ดังนี้

Text preprocessing	
Stop word removal	The fundamental concepts of data mining; types of data for data mining; the famous techniques for data mining; pattern mining and association-rule mining, classification, clustering, outlier analysis, anomaly detection; data mining tools
Word error correction	Before fundamental concepts of data mining; types of data for data mining; famous techniques for data mining; pattern mining and association-rule mining, classifction , clustering, outlier analysis, anomaly detection; data mining tools
	After fundamental concepts of data mining; types of data for data mining; famous techniques for data mining; pattern mining and association-rule mining, classification , clustering, outlier analysis, anomaly detection; data mining tools
Sentence tokenization	fundamental concepts of data mining. types of data for data mining. famous techniques for data mining. pattern mining and association-rule mining. classification. clustering. outlier analysis. anomaly detection. data mining tools.
Part of speech tagging	('fundamental', 'JJ'), ('concepts', 'NNS'), ('of', 'IN'), ('data', 'NNS'), ('mining', 'NN') ('types', 'NNS'), ('of', 'IN'), ('data', 'NNS'), ('for', 'IN'), ('data', 'NN'), ('mining', 'NN') ('famous', 'JJ'), ('techniques', 'NNS'), ('for', 'IN'), ('data', 'NNS'), ('mining', 'NN') ('pattern', 'NN'), ('mining', 'NN'), ('and', 'CC'), ('association-rule', 'NN'), ('mining', 'NN') ('classification', 'NN') ('clustering', 'NN') ('outlier', 'JJR'), ('analysis', 'NN') ('anomaly', 'JJ'), ('detection', 'NN') ('data', 'NNS'), ('mining', 'NN'), ('tools', 'NNS')

ภาพที่ 3-6 การประมวลผลข้อความเบื้องต้นจากข้อมูลคำอธิบายรายวิชา “Data Mining”

3.3.1.1 การกำจัดคำหยุด (Stop word removal)

ขั้นตอนนี้จะเป็นการตัดคำหยุดต่าง ๆ ที่เกิดขึ้นในประโยค เช่น “the”, “a”, “an” เป็นต้น ออกไปจากประโยค โดยที่ความหมายของประโยคหรือข้อความไม่ได้เปลี่ยนแปลง แต่คำหยุดที่น่าสนใจ บางคำจะมีส่วนช่วยในการได้มาซึ่งคำสำคัญที่ซ่อนอยู่ในประโยค คำหยุดที่น่าสนใจเหล่านั้นก็จะพิจารณาให้ไม่ถูกกำจัดออกไป เช่น คำบุพบท (Prepositions), คำสันธาน (Conjunction) เป็นต้น ตัวอย่างประโยคเช่น “unstructured and structured data” เมื่อสกัดคำสำคัญจะได้ “unstructured data” และ “structured data” จะเห็นได้ว่าสันธาน “and” ที่เกิดขึ้นในประโยค จะสามารถช่วยเชื่อมโยงคำที่ซ่อนอยู่ในประโยค เพื่อให้ได้เนื้อหาหรือข้อมูลที่แม่นยำมากขึ้น

3.3.1.2 การแก้ไขคำผิด (Word error correction)

ในคำอธิบายรายวิชาอาจจะมีคำผิดเกิดขึ้นได้ ดังนั้นคำผิดที่พบในคำอธิบายรายวิชาจึงจะต้องได้รับการแก้ไขให้ถูกต้อง ขั้นตอนการแก้ไขคำผิดสามารถทำได้โดยวิธีการเปรียบเทียบคำกับคำศัพท์ในพจนานุกรมภาษาอังกฤษ (Dictionary) ด้วยวิธีการ n-gram เพื่อทำการเปรียบเทียบอักขระที่ปรากฏในคำกับคำศัพท์ในพจนานุกรม จากนั้นจึงดำเนินการแก้ไขคำผิด

3.3.1.3 การตัดประโยค (Sentence tokenization)

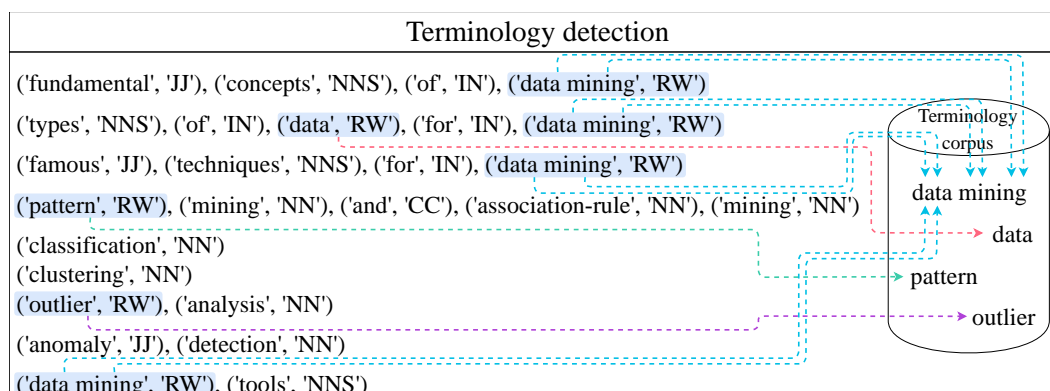
ในขั้นตอนนี้ได้ทำการแยก/ตัดประโยคจากคำอธิบายรายวิชาให้อยู่ในรูปแบบประโยคความเดียว ซึ่งจะทำให้การวิเคราะห์ตามหลักไวยากรณ์ของประโยค เช่น เมื่อพบ “.” (Full stop), “;” (Semi colon) เป็นต้น เมื่อพบคำบ่งบอกถึงการจบประโยค ก็ดำเนินการแยก/ตัดประโยคเหล่านั้นออกจากข้อความ ดังแสดงในภาพที่ 3-6 ในส่วนของ Sentence tokenization

3.3.1.4 การวิเคราะห์หน้าที่ของคำ (Part of speech tagging)

ขั้นตอนการวิเคราะห์หน้าที่ของคำแต่ละคำในประโยคหนึ่ง ๆ ทำได้โดยการตัดคำ (Word tokenization) ออกมาเป็นหน่วยคำต่าง ๆ และทำการระบุ/กำกับหน้าที่ของคำแต่ละชนิดไว้ในคำเหล่านั้น โดยหน้าที่ของคำแต่ละชนิดที่กำกับไว้ในคำหนึ่ง ๆ โดยประยุกต์เช่น “JJ”, “JJR”, “JJS” หมายถึง คำคุณศัพท์ (Adjective), “NN”, “NNS”, “NNP”, “NNPS” หมายถึง คำนาม (Noun), “IN” หมายถึง คำบุพบท (Preposition), “CC” หมายถึง คำสันธาน (Conjunction) เป็นต้น ตัวอย่างการระบุ/กำกับหน้าที่ของคำ แสดงดังภาพที่ 3-6 ในส่วนของ Part of speech tagging

3.3.2 การระบุคำศัพท์เฉพาะ

การค้นหาคำศัพท์เฉพาะในคำอธิบายรายวิชาด้วยวิธีการ n-gram เพื่อเปรียบเทียบคำแต่ละคำในประโยคและคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่ได้จากขั้นตอนที่ 3.1 แสดงดังภาพที่ 3-7 ตัวอย่างเช่น ประโยคแรกของคำอธิบายรายวิชา “Data Mining” คือ (‘fundamental’, ‘JJ’), (‘concepts’, ‘NNS’), (‘of’, ‘IN’), (‘data’, ‘NNS’), (‘mining’, ‘NN’). จากประโยคข้างต้นคำว่า “Data” และ “Mining” จะถูกรวมกลุ่มกันเป็นคำว่า “Data Mining” โดยวิธีการ n-gram จากนั้นจะนำมาเปรียบเทียบกับคำศัพท์ที่มีอยู่ในคลังคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์และระบุว่าเป็นคำศัพท์เฉพาะ ในที่นี้จะระบุหน้าที่ของคำเป็น “RW” (Reserved word) เป็นต้น



ภาพที่ 3-7 การระบุคำศัพท์เฉพาะจากข้อมูลรับเข้าคำอธิบายรายวิชา “Data Mining”

3.3.3 การระบุกฎทางภาษาศาสตร์และการประยุกต์ใช้ในการสกัดคำสำคัญ

ขั้นตอนการสร้างกฎทางภาษาศาสตร์ ที่ซึ่งประยุกต์ใช้การสกัดคำสำคัญโดยวิธีการทางภาษาศาสตร์ (Linguistic approach) เพื่อระบุถึงคำสำคัญที่ซ่อนอยู่ในแต่ละประโยคในแต่ละคำอธิบายรายวิชา ซึ่งในการสร้างกฎจะต้องผ่านการพิจารณาลักษณะประโยคจนทราบว่ารูปแบบของประโยคในคำอธิบายรายวิชา มีลักษณะอย่างไรบ้าง เช่น รูปแบบประโยคที่มีคำบุพบท (Preposition) เกิดขึ้นร่วมกับคำอื่น ๆ ในประโยคจะต้องมีการรวมคำเกิดขึ้น เพื่อสกัดคำสำคัญที่ซ่อนอยู่ในประโยค เป็นต้น เงื่อนไขเหล่านี้สามารถสร้างเป็นกฎบังคับไว้ได้ ซึ่งในการสร้างกฎจะแบ่งเป็น 2 ประเภท คือ 1) กฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ รูปแบบของกฎนี้จะพิจารณาคำศัพท์เฉพาะที่ปรากฏขึ้นในประโยคเป็นลำดับแรก และรวมคำศัพท์เฉพาะกับคำนาม (Noun) หรือคำคุณศัพท์ (Adjective) ที่เกิดขึ้นกับคำศัพท์เฉพาะในประโยคเดียวกัน และทำการระบุคำสำคัญ ลักษณะของกฎทางภาษาศาสตร์ร่วมกับคำศัพท์เฉพาะแสดงดังตารางที่ 3-1 และ 2) กฎทางภาษาศาสตร์ที่ไม่พิจารณาร่วมกับคำศัพท์เฉพาะ รูปแบบของกฎนี้จะพิจารณานามวลี (Noun phrase) คือพิจารณาคำนาม (Noun) ที่เกิดขึ้นร่วมกับคำคุณศัพท์ (Adjective) หรือคำบุพบท (Preposition) ในแต่ละประโยค และทำการระบุคำสำคัญ ลักษณะของกฎทางภาษาศาสตร์ไม่ร่วมกับคำศัพท์เฉพาะแสดงดังตารางที่ 3-2

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ

No.	Sentence	Result
1	$(W_0, \dots, W_n = 'RW')$	<ul style="list-style-type: none"> W_0, \dots, W_n
2	$(W_0 = 'RW') + (W_1 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP')$	<ul style="list-style-type: none"> $W_0 + W_1$
3	$(W_0 = 'RW') + (W_1, \dots, W_n = 'NN' \text{ or } 'NNS' \text{ or } 'NNP' \text{ or } 'NNPS')$	<ul style="list-style-type: none"> $W_0 + W_1, \dots, W_n$
4	$(W_0 = 'RW') + (W_1 = 'RW') + (W_2 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP')$	<ul style="list-style-type: none"> $W_0 + W_1 + W_2$
5	$(W_0 = 'RW') + (W_1 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_2 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_3 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP')$	<ul style="list-style-type: none"> $W_0 + W_1$ $W_2 + W_3$
6	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'RW') + (W_2 = 'RW')$	<ul style="list-style-type: none"> $W_1 + W_2$ $W_0 + W_1 + W_2$
7	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'RW') + (W_2 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP')$	<ul style="list-style-type: none"> $W_1 + W_2$ $W_0 + W_1 + W_2$
8	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_2 = 'RW') + (W_3 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP')$	<ul style="list-style-type: none"> $W_2 + W_3$ $W_0 + W_1 + W_2 + W_3$
9	$(W_0 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_1 = 'RW')$	<ul style="list-style-type: none"> $W_0 + W_1$
10	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'RW')$	<ul style="list-style-type: none"> $W_0 + W_1$
11	$(W_0 = 'RW') + (W_1 = 'IN') + (W_2 = 'RW')$	<ul style="list-style-type: none"> $W_0 + W_1 + W_2$
12	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_2 = 'IN') + (W_3 = 'RW')$	<ul style="list-style-type: none"> W_3 $W_0 + W_3$ $W_3 + W_1$
13	$(W_0 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_1 = 'IN') + (W_2 = 'RW') + (W_3 = 'IN') + (W_4 = 'RW')$	<ul style="list-style-type: none"> $W_2 + W_0$ $W_2 + W_0 + W_1 + W_4$
14	$(W_0 = 'JJ' \text{ or } 'JJR' \text{ or } 'JJS') + (W_1 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_2 = 'IN') + (W_3 = 'RW') + (W_4 = 'RW')$	<ul style="list-style-type: none"> $W_3 + W_4$ $W_0 + W_3 + W_4$ $W_3 + W_4 + W_1$
15	$(W_0 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_1 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_2 = 'IN') + (W_3 = 'RW') + (W_4 = 'RW')$	<ul style="list-style-type: none"> $W_3 + W_4$ $W_3 + W_4 + W_0$ $W_3 + W_4 + W_1$
16	$(W_0 = 'NN' \text{ or } 'NNS' \text{ or } 'NNP') + (W_1 = 'IN') + (W_2 = 'RW') + (W_3 = 'RW') + (W_4 = 'RW')$	<ul style="list-style-type: none"> $W_2 + W_3 + W_4$ $W_2 + W_3 + W_4 + W_0$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
17	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_1 + W_2 + W_3$
18	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$
19	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_3 + W_4 + W_5$ • $W_3 + W_4 + W_5 + W_1$ • $W_0 + W_3 + W_4 + W_5$
20	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • W_3
21	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
22	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	if $W_1 = W_4$: <ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
23	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	if $W_0 = W_2$: <ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
24	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
25	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
26	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • W_2
27	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • W_2
28	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
29	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
30	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
31	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
32	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
33	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
34	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_2 • $W_2 + W_0$ • $W_2 + W_4$
35	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_6 = \text{'RW'}) + (W_7 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_5 + W_6 + W_7$ • $W_2 + W_3$ • $W_5 + W_6 + W_7 + W_0$ • $W_2 + W_3 + W_0$
36	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3$ • $W_5 + W_6$ • $W_2 + W_3 + W_0$ • $W_5 + W_6 + W_0$
37	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_4 + W_0$
38	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_4 + W_5 + W_6$ • $W_3 + W_1 + W_4 + W_5 + W_6$
39	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
40	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
41	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
42	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
43	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
44	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
45	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$
46	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
47	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$ • $W_2 + W_3 + W_0 + W_4 + W_5 + W_6$
48	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
49	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4$ • $W_0 + W_3 + W_4$
50	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$ • $W_0 + W_1 + W_2 + W_3 + W_4$
51	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
52	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4 + W_5$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
53	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
54	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_4 + W_5$
55	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4$ • $W_0 + W_3 + W_4$
56	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
57	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_4 + W_5$ • $W_2 + W_3 + W_4 + W_5$
58	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_2 + W_0 + W_3 + W_4 + W_5$
59	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$ • $W_5 + W_6 + W_0$
60	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
61	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
62	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
63	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$
64	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
65	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
66	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4 + W_0$
67	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
68	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
69	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
70	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$ • $W_5 + W_0$
71	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
72	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
73	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'IN'}) + (W_6 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_5 + W_6$ • $W_3 + W_4 + W_5 + W_6$
74	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
75	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
76	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
77	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4$ • $W_0 + W_3 + W_4$
78	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_4 + W_5 + W_6 + W_0$
79	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • W_2

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
80	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • W_2
81	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
82	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	if $W_0 = W_2$: <ul style="list-style-type: none"> • W_0 • $W_2 + W_3 + W_4$
83	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$ • $W_0 + W_1 + W_2 + W_5$
84	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3 + W_4$
85	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_4 + W_0$
86	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_3 + W_1$ • $W_0 + W_3$
87	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
88	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_4 + W_5 + W_6 + W_0$
89	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
90	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$
91	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
92	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
93	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3 + W_4$
94	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3 + W_4$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
95	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
96	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
97	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
98	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
99	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
100	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3 + W_4$
101	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3 + W_4$
102	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_4 + W_5 + W_6$ • $W_3 + W_4 + W_5 + W_6$
103	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
104	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
105	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
106	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
107	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_0 • W_2
108	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
109	$(W_0 = \text{'RW'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'JJ' or 'JJR' or 'JJS'})$	<ul style="list-style-type: none"> • $W_1 + W_2 + W_3$ • $W_0 + W_1 + W_2 + W_3 + W_4 + W_5$
110	$(W_0 = \text{'RW'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'IN'}) + (W_6 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_1 + W_2 + W_3 + W_4$ • $W_0 + W_1 + W_2 + W_3 + W_4 + W_5 + W_6$
111	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$ • $W_5 + W_6 + W_0$
112	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
113	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
114	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • W_2
115	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $+ W_4 + W_5$
116	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
117	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'IN'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_5 + W_3$ • $W_5 + W_1$ • $W_0 + W_5$
118	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_4$ • $W_2 + W_3 + W_4$
119	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
120	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
121	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • W_2 • $W_2 + W_0$ • $W_2 + W_4$
122	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_2 + W_3$
123	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_3 + W_4$ • $W_0 + W_3 + W_4$ • $W_3 + W_4 + W_1$
124	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
125	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$ • $W_2 + W_0 + W_3 + W_4$
126	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<p>if $W_2 = W_5$:</p> <ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_4 + W_5$
127	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • W_3
128	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4$ • $W_2 + W_3 + W_4 + W_0$
129	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
130	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
131	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
132	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_6 = \text{'RW'}) + (W_7 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_5 + W_6 + W_7$ • $W_2 + W_3$ • $W_5 + W_6 + W_7 + W_0$ • $W_2 + W_3 + W_0$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
133	$(W_0 = \text{'RW'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'RW'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3$ • $W_5 + W_6$ • $W_2 + W_3 + W_0$ • $W_5 + W_6 + W_0$
134	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
135	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
136	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	if $W_2 = W_4$: <ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
137	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4 + W_5$
138	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'}) + (W_5 = \text{'CC'}) + (W_6 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$ • $W_0 + W_1 + W_6$ • $W_3 + W_4 + W_6$
139	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'CC'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$ • $W_4 + W_0$
140	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'}) + (W_5 = \text{'IN'}) + (W_6 = \text{'RW'}) + (W_7 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_5 + W_6 + W_7$ • $W_3 + W_4 + W_5 + W_6 + W_7$
141	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_4$ • $W_2 + W_3 + W_4$
142	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
143	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_4 + W_5 + W_6$
144	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4 + W_5$ • $W_2 + W_3 + W_4 + W_5$
145	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$
146	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_4 + W_5$
147	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
148	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
149	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
150	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
151	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_2 + W_3$
152	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$
153	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
154	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$
155	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_0$
156	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3$ • $W_0 + W_1 + W_2 + W_3$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
157	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4 + W_0$
158	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
159	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_2 + W_3$
160	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4 + W_0$
161	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_2$
162	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • W_0 • $W_2 + W_3$
163	$(W_0 = \text{'RW'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
164	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_3 + W_4 + W_1$ • $W_0 + W_3 + W_4$
165	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_4$ • $W_0 + W_1 + W_2 + W_3 + W_4$
166	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3$ • $W_0 + W_1 + W_2 + W_3$
167	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_4 + W_5$
168	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
169	$(W_0 = \text{'RW'}) + (W_1 = \text{'RW'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'RW'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_4 + W_5$

ตารางที่ 3-1 ลักษณะของกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
170	$(W_0 = \text{'RW'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'}) + (W_5 = \text{'IN'}) + (W_6 = \text{'RW'}) + (W_7 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_5 + W_6 + W_7$ • $W_3 + W_4 + W_5 + W_6 + W_7$
171	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'RW'}) + (W_3 = \text{'RW'})$	<ul style="list-style-type: none"> • $W_2 + W_3 + W_0$

ตารางที่ 3-2 ลักษณะของกฎทางภาษาศาสตร์ที่ไม่พิจารณาร่วมกับคำศัพท์เฉพาะ

No.	Sentence	Result
1	$(W_0, \dots, W_n = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • W_0, \dots, W_n
2	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1, \dots, W_n = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • $W_0 + W_1, \dots, W_n$
3	$(W_0 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_2 + W_3 + W_4$
4	$(W_0 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
5	$(W_0 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • $W_0 + W_3$ • $W_2 + W_3$
6	$(W_0 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	<ul style="list-style-type: none"> • W_0 • W_2
7	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP' or 'NNPS'})$	if $W_1 = W_4$: <ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
8	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
9	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
10	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$
11	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4 + W_5$ • $W_2 + W_3 + W_4 + W_5$
12	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'IN'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3$

ตารางที่ 3-2 ลักษณะของกฎทางภาษาศาสตร์ที่ไม่พิจารณาร่วมกับคำศัพท์เฉพาะ (ต่อ)

No.	Sentence	Result
13	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'CC'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_0 + W_1 + W_4$
14	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_2 + W_3$
15	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3$
16	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_3 + W_4$ • $W_2 + W_3 + W_4$
17	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$
18	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_2 + W_3$
19	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'}) + (W_3 = \text{'IN'}) + (W_4 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_5 = \text{'NN' or 'NNS' or 'NNP'}) + (W_6 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$ • $W_4 + W_5 + W_6$ • $W_0 + W_1 + W_2 + W_3 + W_4 + W_5 + W_6$
20	$(W_0 = \text{'NN' or 'NNS' or 'NNP'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'IN'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2 + W_3 + W_4$
21	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'CC'}) + (W_2 = \text{'JJ' or 'JJR' or 'JJS'})$	<ul style="list-style-type: none"> • W_0 • W_2
22	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_2 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1 + W_2$
23	$(W_0 = \text{'JJ' or 'JJR' or 'JJS'}) + (W_1 = \text{'NN' or 'NNS' or 'NNP'}) + (W_2 = \text{'CC'}) + (W_3 = \text{'NN' or 'NNS' or 'NNP'}) + (W_4 = \text{'NN' or 'NNS' or 'NNP'})$	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_0 + W_3 + W_4$

ตัวอย่างกฎที่สร้างขึ้น “Adjective (JJ) + Noun (NN or NNS or NNP) + Preposition (IN) + Terminology (RW)” สามารถเปลี่ยนแปลงลำดับของคำดังกล่าวให้เป็นคำสำคัญได้ดังต่อไปนี้ 1) “Terminology (RW) + Noun (NN or NNS or NNP)” และ 2) “Adjective (JJ) + Terminology (RW)” ตามลำดับ ภาพที่ 3-8 แสดงการประยุกต์ใช้กฎทางภาษาศาสตร์ที่สร้างขึ้นกับข้อมูลคำอธิบายรายวิชา “Data Mining” และแสดงคำสำคัญที่สามารถสกัดได้จากการประยุกต์ใช้กฎที่สร้างขึ้น

Linguistic rules identification and applying		
Sentence	Rules	Result
('fundamental', 'JJ'), ('concepts', 'NNS'), ('of', 'IN'), ('data mining', 'RW')	$(W_0 = \text{JJ/JJR/JJS}) +$ $(W_1 = \text{NN/ NNS/NNP}) +$ $(W_2 = \text{IN}) + (W_3 = \text{RW})$	<ul style="list-style-type: none"> ● data mining ● fundamental concepts ● fundamental data mining ● data mining concepts
('types', 'NNS'), ('of', 'IN'), ('data', 'RW'), ('for', 'IN'), ('data mining', 'RW')	<ul style="list-style-type: none"> ● $W_3 + W_1$ ● $W_0 + W_3$ 	<ul style="list-style-type: none"> ● data mining ● data types ● data types for data mining
('famous', 'JJ'), ('techniques', 'NNS'), ('for', 'IN'), ('data mining', 'RW')	$(W_0 = \text{RW}) + (W_1 = \text{NN/NNS/ NNP}) + (W_2 = \text{CC}) +$ $(W_3 = \text{NN/NNS/NNP}) +$ $(W_4 = \text{NN/NNS/ NNP})$ if $W_1 = W_4$:	<ul style="list-style-type: none"> ● data mining ● famous techniques ● data mining techniques ● famous data mining
('pattern', 'RW'), ('mining', 'NN'), ('and', 'CC'), ('association-rule', 'NN'), ('mining', 'NN')	<ul style="list-style-type: none"> ● $W_0 + W_1$ ● $W_3 + W_4$ 	<ul style="list-style-type: none"> ● pattern mining ● association-rule mining
('classification', 'NN')	(NN/NNS/NNP)	<ul style="list-style-type: none"> ● classification
('clustering', 'NN')	$(\text{RW}) + (\text{NN/NNS/NNP})$	<ul style="list-style-type: none"> ● clustering
('outlier', 'RW'), ('analysis', 'NN')		<ul style="list-style-type: none"> ● outlier analysis
('anomaly', 'JJ'), ('detection', 'NN')	$(\text{JJ/JJR/JJS}) + (\text{NN/NNS/NNP})$	<ul style="list-style-type: none"> ● anomaly detection
('data mining', 'RW'), ('tools', 'NNS')		<ul style="list-style-type: none"> ● data mining ● data mining tools

ภาพที่ 3-8 การประยุกต์ใช้กฎทางภาษาศาสตร์ที่สร้างขึ้นกับข้อมูลคำอธิบายรายวิชา “Data Mining”

3.4 การค้นหาหนังสือที่มีความเกี่ยวเนื่องกับคำสำคัญในคำอธิบายรายวิชา

หลังจากทำการสกัดคำสำคัญในขั้นตอนที่ 3.3 ตามวิธีการที่ได้กล่าวมาข้างต้นแล้วนั้น จะได้เซตของคำสำคัญ $K_{c_i} = \{k_{c_i,1}, k_{c_i,2}, \dots, k_{c_i,n}\}$ ของคำอธิบายรายวิชา c_i จากนั้นทำการค้นหาหนังสือที่มีความเกี่ยวเนื่องกับคำสำคัญแต่ละคำ $k_{c_i,x} \in K_{c_i}$ เพื่อค้นหาหนังสือที่เกี่ยวข้องเนื่องจากเว็บไซต์ Barnes & Noble³ ซึ่งเป็นเว็บไซต์ขายหนังสือออนไลน์ ที่ตั้งแต่ปี พ.ศ. 2556 ก็ได้มุ่งเน้นไปทางด้านหนังสือกลุ่มการศึกษา อีกทั้งข้อมูลหนังสือที่สามารถเข้าถึงได้จากหน้าเว็บไซต์ยังประกอบไปด้วยข้อมูลสารบัญหนังสือ ซึ่งเป็นข้อมูลที่งานวิจัยนี้จำเป็นต้องใช้สำหรับหาความเกี่ยวเนื่องระหว่างหนังสือและคำอธิบายรายวิชา ซึ่งข้อมูลสำคัญของหนังสือแต่ละเล่มที่ต้องการรวบรวมประกอบไปด้วย ชื่อหนังสือ, ชื่อผู้แต่ง, คำอธิบายหนังสือ, สารบัญหนังสือ เป็นต้น

โดยที่การได้มาซึ่งข้อมูลของหนังสือจะใช้วิธีการสกัดข้อความบนเว็บไซต์ (Web Scraping) เช่นเดียวกันกับรวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ในขั้นตอนที่ 3.1 ภาพที่ 3-9 แสดงข้อมูลหน้าเว็บไซต์ของหนังสือ “Introduction to Data Mining” จากการค้นหาด้วยคำสำคัญว่า “Data Mining” โดยได้ระบุถึงส่วนของข้อมูลที่ต้องการบนเว็บไซต์ภายในกรอบสี่เหลี่ยมที่ปรากฏในภาพที่ 3-9

หลังจากดำเนินการสกัดข้อความบนเว็บไซต์ (Web Scraping) เป็นที่เรียบร้อยแล้วนั้น จะข้อมูลหนังสือเหล่านั้นเก็บรวบรวมไว้ในฐานข้อมูลของหนังสือและกำหนดความเกี่ยวเนื่องกับคำสำคัญของคำอธิบายรายวิชา ดังนั้นสามารถอธิบายได้ว่าแต่ละคำอธิบายรายวิชา c_i จะมีรายการหนังสือที่เกี่ยวข้อง $B_{c_i} = \{b_j, \dots, b_k\}$

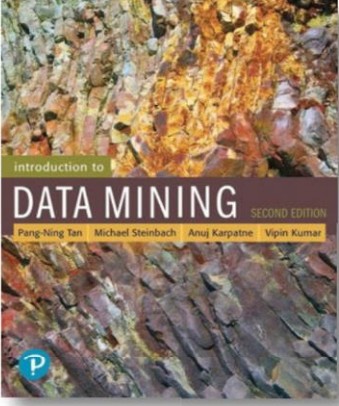
จากข้อมูลคำอธิบายหนังสือ และสารบัญหนังสือ ข้อมูลเหล่านี้จะต้องทำการสกัดคำสำคัญในหัวข้อที่ 3.3 เช่นเดียวกันกับคำอธิบายรายวิชา หลังจากนั้นคำสำคัญที่สกัดได้จากคำอธิบายหนังสือ และสารบัญหนังสือ จะถูกรวมเข้าด้วยกัน เป็นเซตของคำสำคัญ $K_{b_j} = \{k_{b_j,1}, k_{b_j,2}, \dots, k_{b_j,m}\}$ ของหนังสือ b_j

³ <https://www.barnesandnoble.com>

BARNES & NOBLE All Search Hello, Sign In My Account

Books NOOK Books NOOK Textbooks Newsstand Teens Kids Toys & Games Hobbies & Collectibles Home & Gifts Movies & TV Music Sale

Author Introduction to Data Mining
by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar



ISBN-10: 0133128903
ISBN-13: 9780133128901
Pub. Date: 01/18/2018
Publisher: Pearson

Hardcover
Select a Purchase Option (New Edition)

Buy New **Price**
\$108.50 (Save 12%)

Available for Buy Online, Pick up in Store
See Details

Marketplace - from \$114.37

ADD TO CART Instant Purchase

Add to Wishlist

Table of content

Overview Product Details About the Author Table of Contents

Book description

ISBN13 Product Details Publisher

Publication date Edition

Table of Contents

ISBN-13:	9780133128901
Publisher:	Pearson
Publication date:	01/18/2018
Series:	Publication date
Edition description:	2
Pages:	864
Sales rank:	203,541
Product dimensions:	7.70(w) x 9.20(h) x 1.40(d)

Table of Contents
1. Introduction
2. Data
3. Classification: Basic Concepts and Techniques
4. Classification: Alternative Techniques
5. Association Analysis: Basic Concepts and Algorithms
6. Association Analysis: Advanced Concepts
7. Cluster Analysis: Basic Concepts and Algorithms
8. Cluster Analysis: Additional Issues and Algorithms

ภาพที่ 3-9 ตัวอย่างข้อมูลหนังสือที่ทำการเก็บรวบรวมไว้ในฐานข้อมูลของหนังสือ

3.5 การคำนวณความครอบคลุมของหนังสือ และการจัดอันดับ

ขั้นตอนสุดท้ายของระบบแนะนำหนังสืออ่านเพิ่มเติมคือการคำนวณค่าความครอบคลุม (ความคล้ายคลึงกัน) ระหว่างคำสำคัญที่สกัดได้จากคำอธิบายรายวิชา และคำสำคัญที่สกัดได้จากหนังสือ และจัดอันดับหนังสือที่เกี่ยวข้องกับคำอธิบายรายวิชาตามค่าความครอบคลุม (เรียงลำดับจากหนังสือที่มีความเกี่ยวข้องมากที่สุดไปยังหนังสือที่มีความเกี่ยวข้องน้อยที่สุด)

จากหัวข้อที่ 3.4 ได้อธิบายไว้ว่าแต่ละคำอธิบายรายวิชา c_i จะมีรายการหนังสือที่เกี่ยวข้องเนื่อง $B_{c_i} = \{b_j, \dots, b_k\}$ การคำนวณหาค่าความครอบคลุมระหว่างคำอธิบายรายวิชา c_i ด้วยเซตของคำสำคัญ $K_{c_i} = \{k_{c_i,1}, k_{c_i,2}, \dots, k_{c_i,n}\}$ และหนังสือ b_j ที่เกี่ยวข้องกับคำอธิบายรายวิชา c_i ด้วยเซตของคำสำคัญ $K_{b_j} = \{k_{b_j,1}, k_{b_j,2}, \dots, k_{b_j,m}\}$ ดังนั้นการคำนวณค่าความครอบคลุม (Coverage) และค่าความไม่ครอบคลุม (Uncoverage) ระหว่างหนังสือ b_j และคำอธิบายรายวิชา c_i สามารถคำนวณได้จากสมการที่ 7 และ 8 ตามลำดับต่อไปนี้

$$C(c_i, b_j) = \frac{|K_{c_i} \cap K_{b_j}|}{|K_{c_i}|} \quad (8)$$

$$U(c_i, b_j) = 1 - C(c_i, b_j) \quad (9)$$

หลังจากคำนวณค่าความครอบคลุม และค่าความไม่ครอบคลุมจากสมการทั้งสองข้างต้น จะได้รายการค่าความครอบคลุม $C_{c_i} = \{C(c_i, b_j), \dots, C(c_i, b_k)\}$ และรายการค่าความไม่ครอบคลุม $U_{c_i} = \{U(c_i, b_j), \dots, U(c_i, b_k)\}$ แสดงถึงความคล้ายคลึงกันและความไม่คล้ายคลึงกันระหว่างคำอธิบายรายวิชา c_i และรายการหนังสือที่เกี่ยวข้องเนื่อง $B_{c_i} = \{b_j, \dots, b_k\}$ ดังนั้นรายการแนะนำหนังสือสามารถสร้างขึ้นได้จากการจัดอันดับค่าความครอบคลุม โดยพิจารณาค่าความครอบคลุมที่มีค่ามากที่สุดไปยังค่าความครอบคลุมที่มีค่าน้อยที่สุด และพิจารณาการแนะนำหนังสือจากค่าความครอบคลุมที่มีค่ามากที่สุด N เล่ม (Top-N recommendation)

โดยในงานวิจัยนี้พิจารณาค่า N ที่น่าสนใจที่ 10, 30, 50 และ 100 ตามลำดับ ในทางกลับกัน รายการแนะนำหนังสือสามารถสร้างขึ้นได้จากการจัดอันดับค่าความไม่ครอบคลุมได้เช่นกัน โดยสร้างจากการพิจารณาค่าความไม่ครอบคลุม เพื่อให้ผู้ใช้ได้พิจารณาหนังสือที่มีค่าสำคัญที่ขยาย หรือแตกต่างออกไป จากคำสำคัญของคำอธิบายรายวิชา

ตัวอย่างการแนะนำหนังสือจากคำอธิบายรายวิชาของวิชา “Data mining” รายละเอียดคำอธิบายรายวิชาแสดงดังภาพที่ 3-10 จะเห็นได้ว่าคำอธิบายรายวิชายังคงมีคำหยุดที่ซึ่งไม่มีนัยสำคัญในประโยค

Course description of Data Mining
The fundamental concepts of data mining; types of data for data mining; the famous techniques for data mining; pattern mining and association-rule mining, classification, clustering, outlier analysis, anomaly detection; data mining tools

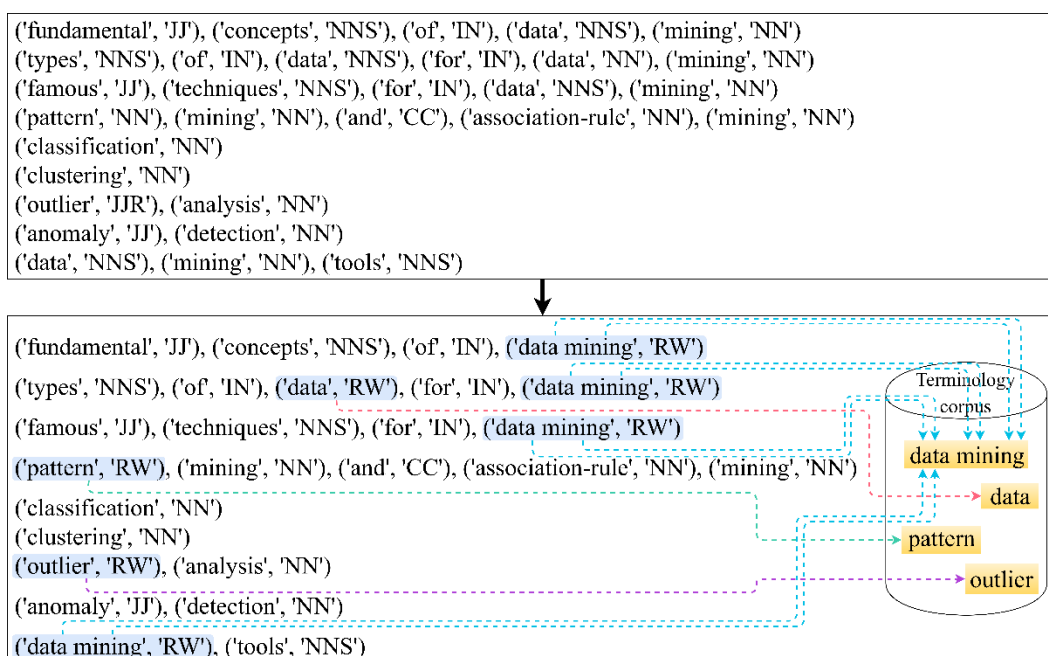
ภาพที่ 3-10 คำอธิบายรายวิชาของวิชา “Data mining”

ดังนั้นจึงทำการสกัดคำสำคัญจากคำอธิบายรายวิชา โดยผ่านการประมวลผลเบื้องต้นคือการกำจัดคำหยุด แก๊ไขคำผิด ตัดประโยค และวิเคราะห์หน้าที่ของคำ จะได้ผลลัพธ์ดังภาพที่ 3-11 หลังจากคำอธิบายรายวิชาผ่านการประมวลผลเบื้องต้นจะได้คำอธิบายรายวิชาที่ไม่มีคำหยุด แก๊ไขคำผิด ตัดประโยค วิเคราะห์และระบุหน้าที่ของคำเรียบร้อยแล้ว

Course description of Data Mining	The fundamental concepts of data mining; types of data for data mining; the famous techniques for data mining; pattern mining and association-rule mining, classification, clustering, outlier analysis, anomaly detection; data mining tools
Stop word removal	The fundamental concepts of data mining; types of data for data mining; the famous techniques for data mining; pattern mining and association-rule mining, classification, clustering, outlier analysis, anomaly detection; data mining tools
Word error correction	Before fundamental concepts of data mining; types of data for data mining; famous techniques for data mining; pattern mining and association-rule mining, classification , clustering, outlier analysis, anomaly detection; data mining tools
	After fundamental concepts of data mining; types of data for data mining; famous techniques for data mining; pattern mining and association-rule mining, classification , clustering, outlier analysis, anomaly detection; data mining tools
Sentence tokenization	fundamental concepts of data mining. types of data for data mining. famous techniques for data mining. pattern mining and association-rule mining. classification. clustering. outlier analysis. anomaly detection. data mining tools.
Part of speech tagging	('fundamental', 'JJ'), ('concepts', 'NNS'), ('of', 'IN'), ('data', 'NNS'), ('mining', 'NN') ('types', 'NNS'), ('of', 'IN'), ('data', 'NNS'), ('for', 'IN'), ('data', 'NN'), ('mining', 'NN') ('famous', 'JJ'), ('techniques', 'NNS'), ('for', 'IN'), ('data', 'NNS'), ('mining', 'NN') ('pattern', 'NN'), ('mining', 'NN'), ('and', 'CC'), ('association-rule', 'NN'), ('mining', 'NN') ('classification', 'NN') ('clustering', 'NN') ('outlier', 'JJR'), ('analysis', 'NN') ('anomaly', 'JJ'), ('detection', 'NN') ('data', 'NNS'), ('mining', 'NN'), ('tools', 'NNS')

ภาพที่ 3-11 การประมวลผลคำอธิบายรายวิชาของวิชา “Data Mining” เบื้องต้น

เมื่อคำอธิบายรายวิชาได้ผ่านการระบุหน้าที่ของคำที่ปรากฏในคำอธิบายรายวิชาแล้ว จะทำการระบุถึงคำศัพท์เฉพาะที่ปรากฏในแต่ละประโยค โดยนำคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ที่ได้รวบรวมไว้มาเปรียบเทียบกับคำที่ปรากฏในคำอธิบายรายวิชา จะได้ผลลัพธ์ดังภาพที่ 3-12



ภาพที่ 3-12 การระบุถึงคำศัพท์เฉพาะที่ปรากฏในแต่ละประโยคของคำอธิบายรายวิชาของวิชา “Data Mining”

หลังจากการระบุถึงคำศัพท์เฉพาะในคำอธิบายรายวิชา ต่อมาทำการสกัดคำสำคัญโดยใช้กฎทางภาษาศาสตร์ที่สร้างขึ้นในหัวข้อที่ 3.3.3 ข้างต้น การสกัดคำสำคัญจะพิจารณาว่าภายในประโยคที่มีการระบุถึงคำศัพท์เฉพาะหรือไม่ ถ้ามีการระบุถึงคำศัพท์เฉพาะจะทำการสกัดคำสำคัญจากกฎทางภาษาศาสตร์ที่พิจารณาร่วมกับคำศัพท์เฉพาะดังข้อมูลในตารางที่ 3-1 แต่ถ้าไม่มีการระบุถึงคำศัพท์เฉพาะจะทำการสกัดคำสำคัญจากกฎทางภาษาศาสตร์ที่ไม่พิจารณาร่วมกับคำศัพท์เฉพาะดังข้อมูลในตารางที่ 3-2 เมื่อคำอธิบายรายวิชาผ่านขั้นตอนการสกัดคำสำคัญเรียบร้อยแล้ว จะทำการตัดคำสำคัญที่ปรากฏซ้ำกันออกไปด้วย จะได้ผลลัพธ์ที่แสดงดังภาพที่ 3-13

Sentence	Rules	Result
('fundamental', 'JJ'), ('concepts', 'NNS'), ('of', 'IN'), ('data mining', 'RW')	$(W_0 = JJ/JJR/JJS) +$ $(W_1 = NN/ NNS/NNP) +$ $(W_2 = IN) + (W_3 = RW)$	<ul style="list-style-type: none"> • data mining • fundamental concepts • fundamental data mining • data mining concepts
('types', 'NNS'), ('of', 'IN'), ('data', 'RW'), ('for', 'IN'), ('data mining', 'RW')	<ul style="list-style-type: none"> • $W_3 + W_1$ • $W_0 + W_3$ 	<ul style="list-style-type: none"> • data mining • data types • data types for data mining
('famous', 'JJ'), ('techniques', 'NNS'), ('for', 'IN'), ('data mining', 'RW')	$(W_0 - RW) + (W_1 - NN/NNS/ NNP) + (W_2 - CC) +$ $(W_3 = NN/NNS/NNP) +$ $(W_4 = NN/NNS/ NNP)$ if $W_1 = W_4$:	<ul style="list-style-type: none"> • data mining • famous techniques • data mining techniques • famous data mining
('pattern', 'RW'), ('mining', 'NN'), ('and', 'CC'), ('association-rule', 'NN'), ('mining', 'NN')	<ul style="list-style-type: none"> • $W_0 + W_1$ • $W_3 + W_4$ 	<ul style="list-style-type: none"> • pattern mining • association-rule mining
('classification', 'NN')	$(NN/NNS/NNP)$	• classification
('clustering', 'NN')	$(RW) - (NN/NNS/NNP)$	• clustering
('outlier', 'RW'), ('analysis', 'NN')		• outlier analysis
('anomaly', 'JJ'), ('detection', 'NN')		• anomaly detection
('data mining', 'RW'), ('tools', 'NNS')	$(JJ/JJR/JJS) + (NN/NNS/NNP)$	<ul style="list-style-type: none"> • data mining • data mining tools

<ul style="list-style-type: none"> • data mining • fundamental concepts • fundamental data mining • data mining concepts • data types • data types for data mining • famous techniques • data mining techniques 	<ul style="list-style-type: none"> • famous data mining • pattern mining • association-rule mining • classification • clustering • outlier analysis • anomaly detection • data mining tools
---	---

ภาพที่ 3-13 การสกัดคำสำคัญจากคำอธิบายรายวิชาของวิชา “Data Mining”


ผลลัพธ์จากการสกัดคำสำคัญนั้นจะได้คำสำคัญที่เกี่ยวข้องกับคำอธิบายรายวิชาของวิชา “Data Mining” และคำสำคัญเหล่านั้นจะถูกนำไปให้ผู้เชี่ยวชาญ (อาจารย์ผู้สอนรายวิชา) ระบุถึงความถูกต้อง ภาพที่ 3-14 แสดงผลลัพธ์คำสำคัญของคำอธิบายรายวิชาที่ผู้เชี่ยวชาญระบุว่าถูกต้อง โดยเส้นทึบสีแดงบ่งบอกว่าคำสำคัญนั้นไม่ถูกต้อง

- | | |
|-----------------------------------|---------------------------------|
| ● data mining | ● famous data mining |
| ● fundamental concepts | ● pattern mining |
| ● fundamental data mining | ● association-rule mining |
| ● data mining concepts | ● classification |
| ● data types | ● clustering |
| ● data types for data mining | ● outlier analysis |
| ● famous techniques | ● anomaly detection |
| ● data mining techniques | ● data mining tools |


ภาพที่ 3-14 ระบุถึงความถูกต้องของคำสำคัญของคำอธิบายรายวิชาของวิชา “Data Mining” โดยผู้เชี่ยวชาญ

จากนั้นนำคำสำคัญมาค้นหาหนังสือที่มีความเกี่ยวเนื่องกับคำสำคัญจากเว็บไซต์ Barnes & Noble ดังภาพที่ 3-15 แสดงตัวอย่างหน้าเว็บไซต์จากการค้นหาหนังสือด้วยคำสำคัญคือ “data mining” โดยระบุหมวดหนังสือที่ทำการค้นหาเป็นหมวด “Computers”



จากการค้นหาหนังสือที่มีความเกี่ยวเนื่องกับคำสำคัญแล้ว ทำการเก็บรวบรวมข้อมูลหนังสือที่สำคัญประกอบไปด้วย ชื่อหนังสือ, ชื่อผู้แต่ง, ชื่อสำนักพิมพ์, คำอธิบายหนังสือ, สารบัญหนังสือ เป็นต้น และนำข้อมูลคำอธิบายหนังสือและสารบัญหนังสือมาทำการสกัดคำสำคัญเช่นเดียวกับคำอธิบายรายวิชา ดังนั้นหนังสือแต่ละเล่มจะมีเซตของคำสำคัญที่เกี่ยวข้องกับหนังสือเช่นกัน




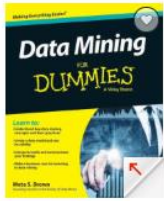



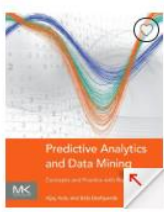
BARNES & NOBLE Books ▾ data mining 🔍 

< Favorite Paperbacks: Buy 2, Get the 3rd Free | 15% Off Your Order - Code: LABORDAY15 | Young Adult Thrillers: Buy One, Get One 50% Off >

Books NOOK Books NOOK Textbooks Newsstand Teens Kids Toys Games & Collectibles Gift Home & Office 

Books / Computers

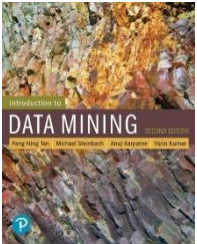
1 - 20 of 408 results for data mining Show: 20 ▾ Sort by: Top Matches ▾  

			
<p>Data Mining and Knowledg... by Evangelos Triantaphyllou ☆☆☆☆☆ Paperback \$233.39 \$289.00 See All Formats</p>	<p>Data Science for Business... by Foster Provost ★★★★★ Paperback \$32.80 \$49.99 See All Formats</p>	<p>Python Data Mining Quick ... by Nathan Greenelch ☆☆☆☆☆ Paperback \$29.99 See All Formats</p>	<p>Data Mining For Dummies by Meta S. Brown ☆☆☆☆☆ Paperback \$31.49 \$34.99 See All Formats</p>
			
<p>Mining the Social Web: Dat... by Matthew A. Russell ☆☆☆☆☆ Paperback \$44.99 \$49.99 See All Formats</p>	<p>Machine Learning for Begin... by Matt Henderson ★★★★☆ Paperback \$19.99 See All Formats</p>	<p>Data Science for Business... by Travis Goleman ☆☆☆☆☆ NOOK Book from \$2.99</p>	<p>Predictive Analytics and Da... by Vijay Kotu ☆☆☆☆☆ Paperback from \$32.58 \$59.95 See All Formats</p>

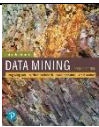
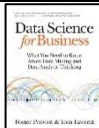
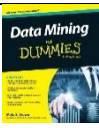

ภาพที่ 3-15 หน้าเว็บไซต์ Barnes & Noble จากการค้นหาหนังสือด้วยคำสำคัญคือ “data mining” โดยระบุหมวดหนังสือที่ทำการค้นหาเป็นหมวด “Computers”

เมื่อรวบรวมหนังสือทั้งหมดที่เกี่ยวข้องกับคำอธิบายรายวิชาแล้ว ต่อไปทำการคำนวณความครอบคลุมและความไม่ครอบคลุมระหว่างคำอธิบายรายวิชาและหนังสือทั้งหมดที่เกี่ยวข้องกับคำอธิบายรายวิชา โดยเริ่มต้นจากการเปรียบเทียบคำสำคัญระหว่างคำอธิบายรายวิชาและหนังสือที่เกี่ยวข้องกับคำอธิบายรายวิชาแสดงดังตารางที่ 3-3 แสดงตัวอย่างการเปรียบเทียบคำสำคัญระหว่างเซตของคำสำคัญของคำอธิบายรายวิชาและเซตของคำสำคัญของหนังสือชื่อ “Introduction to Data Mining” รหัสหนังสือ “9780133128901” ที่เกี่ยวข้องกับคำอธิบายรายวิชา จากตัวอย่างที่แสดงหนังสือทั้งหมดที่เกี่ยวข้องกับคำอธิบายรายวิชาที่จะถูกนำมาเปรียบเทียบเช่นเดียวกัน ข้อมูลที่ได้จากการเปรียบเทียบเซตของคำสำคัญจะนำมาคำนวณค่าความครอบคลุม โดยคำนวณได้จากสมการที่ 8 ที่ได้กล่าวไว้ข้างต้น แสดงตัวอย่างดังตารางที่ 3-4 อีกทั้งยังคำนวณค่าความไม่ครอบคลุมที่สามารถคำนวณได้จากสมการที่ 9 ที่ได้กล่าวไว้ข้างต้นเช่นกัน แสดงตัวอย่างดังตารางที่ 3-5 และจัดอันดับหนังสือที่มีค่าความครอบคลุมมากที่สุดไปย้งน้อยที่สุด

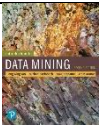

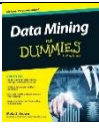

ตารางที่ 3-3 การเปรียบเทียบคำสำคัญระหว่างเซตของคำสำคัญของคำอธิบายรายวิชาของวิชา “Data Mining” และเซตของคำสำคัญของหนังสือชื่อ “Introduction to Data Mining” รหัสหนังสือ “9780133128901” ที่เกี่ยวข้องกับคำอธิบายรายวิชา

	data mining	fundamental data mining	data mining concepts	data types	data types for data mining	data mining techniques	pattern mining	association-rule mining	classification	clustering	outlier analysis	anomaly detection	data mining tools
Course description of Data Mining	1	1	1	1	1	1	1	1	1	1	1	1	1
9780133128901 	1	0	1	1	0	0	1	1	1	1	0	1	1

ตารางที่ 3-4 การคำนวณค่าความครอบคลุมระหว่างคำอธิบายรายวิชาของวิชา “Data Mining” และหนังสือทั้งหมดที่เกี่ยวข้องเกี่ยวกับคำอธิบายรายวิชา

Books related to course description of Data Mining	Coverage calculation $C(c_i, b_j) = \frac{ K_{c_i} \cap K_{b_j} }{ K_{c_i} }$	Coverage
9780133128901 	$C(c_{Data Mining}, b_1) = \frac{ 10 }{ 13 }$	0.769
9781449361327 	$C(c_{Data Mining}, b_2) = \frac{ 8 }{ 13 }$	0.615
9781118893173 	$C(c_{Data Mining}, b_3) = \frac{ 9 }{ 13 }$	0.692
9781948580496 	$C(c_{Data Mining}, b_4) = \frac{ 5 }{ 13 }$	0.385
⋮	⋮	⋮

ตารางที่ 3-5 การคำนวณค่าความไม่ครอบคลุมระหว่างคำอธิบายรายวิชาของวิชา “Data Mining” และหนังสือทั้งหมดที่เกี่ยวข้องกับคำอธิบายรายวิชา


Books related to course description of Data Mining	UnCoverage calculation $U(c_i, b_j) = 1 - C(c_i, b_j)$	UnCoverage
9780133128901 	$U(c_{Data Mining}, b_1) = 1 - 0.769$	0.231
9781449361327 	$U(c_{Data Mining}, b_1) = 1 - 0.615$	0.385
9781118893173 	$U(c_{Data Mining}, b_1) = 1 - 0.692$	0.308
9781948580496 	$U(c_{Data Mining}, b_1) = 1 - 0.385$	0.615
⋮	⋮	⋮

หลังการคำนวณค่าความครอบคลุมและค่าความไม่ครอบคลุมของหนังสือทั้งหมดที่เกี่ยวข้องกับคำอธิบายรายวิชา แล้วทำการจัดอันดับแล้ว จะได้รายการแนะนำหนังสือที่ถูกจัดอันดับค่าความครอบคลุมที่มีค่ามากที่สุดไปยังค่าความครอบคลุมที่มีค่าน้อยที่สุด โดยพิจารณาจำนวนรายการแนะนำจากค่าความครอบคลุมมากที่สุด N เล่ม ซึ่งงานวิจัยนี้ค่า N ที่น่าสนใจได้แก่ 10, 30, 50 และ 100 ตามลำดับ ตารางที่ 3-6 แสดงรายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความครอบคลุมและพิจารณาค่า N = 10 และตาราง ที่ 3-7 แสดงรายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าไม่ความครอบคลุมและพิจารณาค่า N = 10



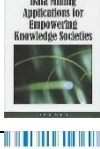


ตารางที่ 3-6 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความครอบคลุม

List of books recommendation (Coverage)		
Top-N	Book	Coverage
1	9780123814791 	0.923
2	9781439862094 	0.846
3	9781491985045 	0.846
4	9780128042915 	0.769
5	9780133128901 	0.769
6	9781118877432 	0.692
7	9781118893173 	0.692
8	9781118879368 	0.692
9	9781449361327 	0.615




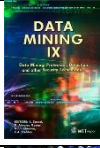

ตารางที่ 3-6 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความครอบคลุม (ต่อ)

List of books recommendation (Coverage)		
Top-N	Book	Coverage
10	 9781118618042	0.615

ตารางที่ 3-7 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความไม่ครอบคลุม

List of books recommendation (UnCoverage)		
Top-N	Book	UnCoverage
1	 9780817647988	0.077
2	 9789814271066	0.077
3	 9781599046570	0.077
4	 9781586038984	0.077
5	 9783642065767	0.077

ตารางที่ 3-7 รายการแนะนำหนังสือของรายวิชา “Data Mining” โดยจัดอันดับจากค่าความไม่ครอบคลุม (ต่อ)

List of books recommendation (UnCoverage)		
Top-N	Book	UnCoverage
6	9781591402220 	0.077
7	9781852338671 	0.077
8	9783642072376 	0.077
9	9781845641108 	0.077
10	9781558604896 	0.077

ตัวอย่างข้างต้นเป็นการสร้างรายการแนะนำหนังสือของวิชา “Data Mining” ทั้งนี้งานวิจัยนี้ได้ทำการรวบรวมคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ เพื่อใช้เป็นข้อมูลรับเข้าจำนวน 72 วิชา และข้อมูลหนังสือจำนวน 176,878 เล่ม เพื่อสร้างรายการแนะนำหนังสืออ่านเพิ่มเติมสำหรับนิสิต/นักศึกษาใช้ในการเลือกหนังสืออ่านเพิ่มเติมจากในชั้นเรียนได้ง่ายและสะดวกมากขึ้น เพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเอง

บทที่ 4

ผลการดำเนินงาน

ในบทนี้จะอธิบายถึงผลการดำเนินงาน เพื่อประเมินระบบแนะนำหนังสืออ่านเพิ่มเติมที่ได้แนะนำไว้ข้างต้น งานวิจัยนี้ได้ทำการรวบรวมคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ จากสาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา เพื่อใช้เป็นข้อมูลรับเข้าจำนวน 72 วิชา และสร้างรายการแนะนำหนังสืออ่านเพิ่มเติมที่เกี่ยวข้องกับคำอธิบายรายวิชา โดยพิจารณาการแนะนำหนังสือจำนวน N เล่ม (Top- N recommendation) โดยในงานวิจัยนี้พิจารณาค่า N ที่น่าสนใจที่ 10, 30, 50 และ 100 ตามลำดับ (ค่า N ที่น่าสนใจเหล่านี้มักถูกนำมาใช้ในระบบแนะนำ และเป็นค่า N ที่น่าสนใจในการพิจารณารูปแบบในด้านการทำเหมืองข้อมูล) จากวิธีการที่นำเสนอในการแนะนำหนังสือ จะเห็นได้ว่าการสกัดคำสำคัญเป็นวิธีการที่มีความสำคัญต่อประสิทธิภาพของระบบ ดังนั้นการประเมินระบบแนะนำหนังสืออ่านเพิ่มเติมจึงแบ่งออกเป็น 2 ส่วนได้แก่ 1) การวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ และ 2) การประเมินระบบแนะนำหนังสืออ่านเพิ่มเติม ดังต่อไปนี้

4.1 การวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ

จากหัวข้อที่ 3.3 การสกัดคำสำคัญคือวิธีการที่สำคัญของระบบแนะนำหนังสืออ่านเพิ่มเติม ที่สามารถส่งผลต่อประสิทธิภาพของระบบ ทั้งนี้เพื่อประเมินว่าคำสำคัญที่สกัดออกมาเป็นคำสำคัญที่ถูกต้องหรือไม่ งานวิจัยนี้ได้รับความช่วยเหลือจากผู้เชี่ยวชาญ (อาจารย์ผู้สอนรายวิชา) เพื่อบ่งบอกถึงคำสำคัญที่ถูกต้องและไม่ถูกต้อง ดังนั้นจึงทำการวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ โดยการวัดค่าความถูกต้อง (Precision), ค่าความแม่นยำ (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) จะถูกนำมาประยุกต์ใช้ในการวัดประสิทธิภาพของการสกัดคำสำคัญ สามารถคำนวณได้จากข้อมูลภายในตาราง Confusion matrix เป็นตารางแสดงข้อมูลระหว่างข้อมูลความเป็นจริงจริงและข้อมูลทำนาย โดยข้อมูลแกนนอนคือข้อมูลความเป็นจริง ส่วนข้อมูลแกนตั้งคือข้อมูลที่ทำนายแสดงดังตารางที่ 4-1 ดังต่อไปนี้

ตารางที่ 4-1 ตาราง Confusion matrix ระหว่างข้อมูลความเป็นจริงจริงและข้อมูลทำนาย

		System	
		Extracted	Not extracted
Actual	Keyword	TP	FN
	Not keyword	FP	TN

จากการใช้วิธีการสกัดคำสำคัญและมีผู้เชี่ยวชาญให้ความถูกต้องนั้นจะสามารถแทนข้อมูลใน Confusion matrix โดยแกนนอนคือข้อมูลคำสำคัญที่ถูกต้องและไม่ถูกต้อง ส่วนตั้งคือข้อมูลคำสำคัญที่ถูกสกัดและไม่ถูกสกัดออกมาจากคำอธิบายรายวิชา ดังนั้นค่า True Positive (TP), True Negative (TN), False Positive (FP) และ False Negative (FN) สามารถอธิบายได้ดังนี้

TP คือ คำสำคัญที่ถูกสกัดได้ และเป็นคำสำคัญที่ถูกต้อง

TN คือ คำสำคัญที่ไม่ถูกสกัด และเป็นคำสำคัญที่ไม่ถูกต้อง

FP คือ คำสำคัญที่ถูกสกัดได้ และเป็นคำสำคัญที่ไม่ถูกต้อง

FN คือ คำสำคัญที่ไม่ถูกสกัด และเป็นคำสำคัญที่ถูกต้อง

จากข้อมูลใน Confusion matrix สามารถคำนวณค่าความถูกต้อง, ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวม ได้จากสมการที่ 10, 11 และ 12 ตามลำดับดังนี้

$$Precision = \frac{TP}{TP + FP} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$F\text{-measure} = 2 \times \frac{precision \times recall}{precision + recall} \quad (12)$$

ตัวอย่างการคำนวณค่าความถูกต้อง, ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวมของรายวิชา “Data Mining” จากภาพที่ 3-14 แสดงคำสำคัญที่ถูกผู้เชี่ยวชาญระบุถึงความถูกต้อง สามารถนำข้อมูลมาสร้าง Confusion matrix ได้ตารางที่ 4-2

ตารางที่ 4-2 ตาราง Confusion matrix ระหว่างข้อมูลคำสำคัญที่ถูกต้องและคำสำคัญที่สกัดได้ของรายวิชา “Data Mining”

		System	
		Extracted	Not extracted
Actual	Keyword	13	0
	Not keyword	3	0

ดังนั้นจึงสามารถคำนวณค่าความถูกต้อง, ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวมของการสกัดคำสำคัญจากอธิบายรายวิชาได้ดังนี้

$$Precision = \frac{TP}{TP + FP} = \frac{13}{13 + 3} = 0.813 \quad \text{คิดเป็นเปอร์เซ็นต์คือ 81.3\%}$$

$$Recall = \frac{TP}{TP + FN} = \frac{13}{13 + 0} = 1 \quad \text{คิดเป็นเปอร์เซ็นต์คือ 100\%}$$

$$F\text{-measure} = 2 \times \frac{precision \times recall}{precision + recall} \quad \text{คิดเป็นเปอร์เซ็นต์คือ 89.7\%}$$

$$= 2 \times \frac{0.813 \times 1}{0.813 + 1} = 0.897$$

นอกจากนี้งานวิจัยนี้ยังทำการเปรียบเทียบวิธีการสกัดคำสำคัญกับขั้นตอนวิธี “Termine” (Frantzi et al., 2000) และขั้นตอนวิธี “RAKE” (Rose et al., 2010) ซึ่งเป็นวิธีการสกัดคำสำคัญโดยนำวิธีการทางภาษาศาสตร์ (Linguistic approach) มาประยุกต์ใช้และได้รับความนิยม

จากข้อมูลดังกล่าวข้างต้นสามารถทราบถึงจำนวนคำสำคัญที่ระบบสกัดออกมาได้ทั้งหมด, จำนวนคำสำคัญที่ระบบสกัดออกมาได้และมีความถูกต้อง, ค่าความถูกต้อง, ค่าความแม่นยำ และค่าประสิทธิภาพโดยรวม แสดงดังตารางที่ 4-3

ตารางที่ 4-3 เปรียบเทียบการวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ, ขั้นตอนวิธี “Termine” และขั้นตอนวิธี “RAKE”

Approach	Extracted keywords	Correct keywords	Precision	Recall	F-measure
SBS keyword extraction	1447	1253	91.3	95.3	93.1
Termine	1365	523	91.1	40.6	55.1
RAKE	1695	837	70.1	62.6	66.0

จากตารางที่ 4-3 จะเห็นได้ว่าขั้นตอนวิธี “Termine” มีค่าความถูกต้อง 91.1%, ค่าความแม่นยำ 40.6% แสดงให้เห็นว่าขั้นตอนวิธี “Termine” สามารถสกัดคำสำคัญได้ถูกต้องสูงแต่ด้วยค่าความแม่นยำที่ค่อนข้างน้อยแสดงให้เห็นว่าไม่สามารถที่จะสกัดคำสำคัญออกมาได้ครอบคลุมทั้งหมดของคำอธิบายรายวิชา ทำให้ค่าประสิทธิภาพโดยรวมมีค่า 55.1% ต่อมา ขั้นตอนวิธี “RAKE” มีค่าความถูกต้อง 70.1%, ค่าความแม่นยำ 62.6% แสดงให้เห็นว่าขั้นตอนวิธี “RAKE” สามารถสกัดคำสำคัญถูกต้องได้น้อยกว่าขั้นตอนวิธี “Termine” แต่จากค่าความแม่นยำแสดงให้เห็นว่าขั้นตอนวิธี “RAKE” สามารถสกัดคำสำคัญได้ครอบคลุมมากกว่าขั้นตอนวิธี “Termine” และด้วยค่าความแม่นยำที่สูงกว่านี้ ทำให้ค่าประสิทธิภาพโดยรวมมีค่า 66.0% มากกว่าขั้นตอนวิธี “Termine” และสุดท้ายวิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความถูกต้อง 91.3%, ค่าความแม่นยำ 95.3% แสดงให้เห็นว่าสามารถสกัดคำสำคัญได้ถูกต้องมากกว่าทั้ง 2 ขั้นตอนวิธีที่นำมาเปรียบเทียบ เนื่องด้วยคำสำคัญที่ใช้ในระบบสามารถสกัดคำสำคัญได้ครอบคลุมผลลัพธ์ของคำสำคัญของทั้ง 2 ขั้นตอนวิธีสกัดได้ทั้งหมด และจากค่าความแม่นยำสามารถแสดงได้ว่าวิธีการสกัดคำสำคัญที่ใช้ในระบบสกัดคำสำคัญออกมาได้ครอบคลุมคำอธิบายรายวิชาได้สูง เนื่องด้วยกฎทางภาษาศาสตร์ที่สร้างขึ้นมีการประกอบคำที่เกี่ยวข้องเข้าด้วยกันโดยพิจารณาจากคำบุพบทหรือคำสันธาน ทำให้สามารถสกัดคำสำคัญได้มากกว่าทั้ง 2 ขั้นตอนวิธีที่นำมาเปรียบเทียบ

4.2 การประเมินระบบแนะนำหนังสืออ่านเพิ่มเติม

การประเมินประสิทธิภาพผลลัพธ์ของระบบแนะนำหนังสืออ่านเพิ่มเติม จะพิจารณาจากค่าความครอบคลุมโดยเฉลี่ย (Average coverage) และค่าความครอบคลุมสะสม (Total coverage) (พิจารณา ค่าความไม่ครอบคลุมโดยเฉลี่ย (Average uncoverage) และค่าความไม่ครอบคลุมสะสม (Total uncoverage) เช่นกัน) ของหนังสือทั้งหมดในระบบแนะนำ

จากหัวข้อที่ 3.5 ได้อธิบายไว้ว่าจากการคำนวณค่าความครอบคลุม และค่าความไม่ครอบคลุม ระหว่างคำอธิบายรายวิชา c_i และรายการหนังสือที่เกี่ยวข้อง $B_{c_i} = \{b_j, \dots, b_k\}$ จะได้รายการค่าความครอบคลุม $C_{c_i} = \{C(c_i, b_j), \dots, C(c_i, b_k)\}$ โดยเรียงลำดับจากหนังสือที่มีความเกี่ยวข้องมากที่สุดไปยังหนังสือที่มีความเกี่ยวข้องน้อยที่สุด และรายการค่าความไม่ครอบคลุม $U_{c_i} = \{U(c_i, b_j), \dots, U(c_i, b_k)\}$ โดยเรียงลำดับจากหนังสือที่มีความไม่เกี่ยวข้องมากที่สุดไปยังหนังสือที่มีความไม่เกี่ยวข้องน้อยที่สุด ดังนั้นค่าความครอบคลุมโดยเฉลี่ยตามจำนวนหนังสือ N เล่มที่น่าสนใจสามารถคำนวณได้จากรายการค่าความครอบคลุม C_{c_i} ดังสมการที่ 13

$$averageCov_{c_i} = \frac{\sum_{j=0}^N C_{(c_i, b_j)}}{N} \quad (13)$$

กำหนดให้ N คือ จำนวนหนังสือ N เล่มที่น่าสนใจได้แก่ 10, 30, 50 หรือ 100

$C_{(c_i, b_j)}$ คือ ค่าความครอบคลุมระหว่างคำอธิบายรายวิชา c_i และหนังสือ b_j ที่เกี่ยวข้องกับคำอธิบายรายวิชา c_i สามารถคำนวณได้จากสมการที่ 8

และค่าความไม่ครอบคลุมโดยเฉลี่ยสามารถคำนวณได้จากรายการค่าความไม่ครอบคลุม U_{c_i} ดังสมการที่ 14

$$averageUnCov_{c_i} = \frac{\sum_{j=0}^N U_{(c_i, b_j)}}{N} \quad (14)$$

กำหนดให้ N คือ จำนวนหนังสือ N เล่มที่น่าสนใจได้แก่ 10, 30, 50 หรือ 100

$U_{(c_i, b_j)}$ คือ ค่าความครอบคลุมระหว่างคำอธิบายรายวิชา c_i และหนังสือ b_j ที่เกี่ยวข้องกับคำอธิบายรายวิชา c_i สามารถคำนวณได้จากสมการที่ 9

ในส่วนของคุณค่าความครอบคลุมสะสมสามารถคำนวณได้โดยพิจารณาจากรายการหนังสือที่เกี่ยวข้องกับคำอธิบายรายวิชา จากนั้นจะพิจารณาความเกี่ยวเนื่องกันระหว่างคำสำคัญที่มีอยู่ในหนังสือและคำสำคัญที่มีอยู่ในคำอธิบายรายวิชา และทำการนับจำนวนสะสมของคำสำคัญที่มีความเหมือนกันและยังไม่เคยปรากฏมาก่อน ดังนั้นค่าความครอบคลุมสะสมตามจำนวนหนังสือ N เล่มที่น่าสนใจสามารถคำนวณได้จากสมการที่ 15 ดังต่อไปนี้

$$totalCov_{c_i} = \frac{|(K_{c_i} \cap K_{b_j}) \cup (K_{c_i} \cap K_{b_{j+1}}) \cup \dots \cup (K_{c_i} \cap K_{b_N})|}{|K_{c_i}|} \quad (15)$$

กำหนดให้ N คือ จำนวนหนังสือ N เล่มที่น่าสนใจได้แก่ 10, 30, 50 หรือ 100

K_{c_i} คือ เซตของคำสำคัญของคำอธิบายรายวิชา c_i

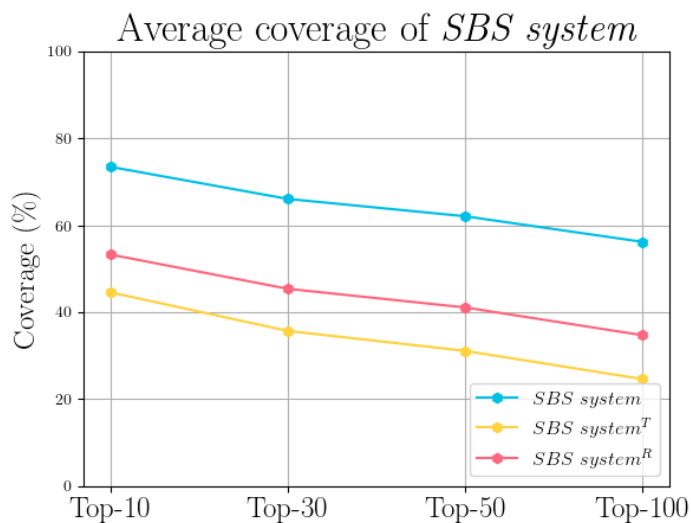
K_{b_j} คือ เซตของคำสำคัญของหนังสือ b_j ที่เกี่ยวเนื่องกับคำอธิบายรายวิชา c_i

และค่าความไม่ครอบคลุมสะสมสามารถคำนวณได้จากค่าตรงกันข้ามกับค่าความครอบคลุมสะสมดังสมการที่ 16

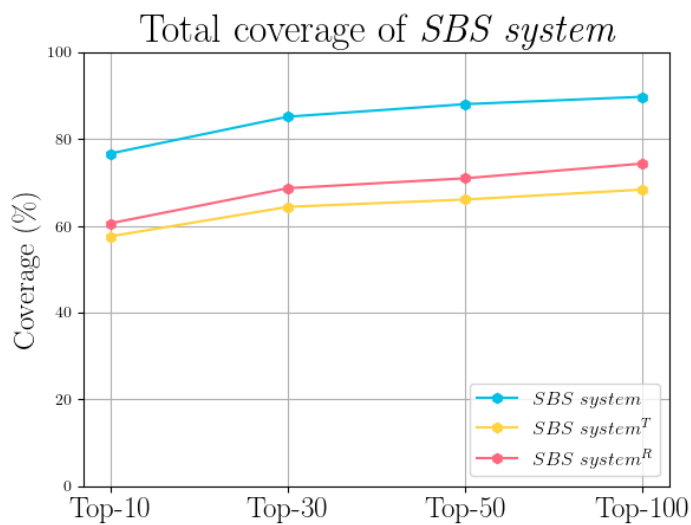
$$totalUnCov_{c_i} = 1 - totalCov_{c_i} \quad (16)$$

กำหนดให้ N คือ จำนวนหนังสือ N เล่มที่น่าสนใจได้แก่ 10, 30, 50 หรือ 100

นอกจากนี้ยังเปรียบเทียบผลลัพธ์ค่าความครอบคลุมโดยเฉลี่ย, ค่าความครอบคลุมสะสม, ค่าความไม่ครอบคลุมโดยเฉลี่ย และค่าความไม่ครอบคลุมสะสม ระหว่างระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบ (SBS system), ระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยขั้นตอนวิธี “Termine” (SBS system^T) และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยขั้นตอนวิธี “RAKE” (SBS system^R) ภาพที่ 4-1 แสดงการเปรียบเทียบค่าความครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญ และภาพที่ 4-2 แสดงการเปรียบเทียบค่าความครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญตามลำดับ



ภาพที่ 4-1 การเปรียบเทียบผลลัพธ์ค่าความครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญ

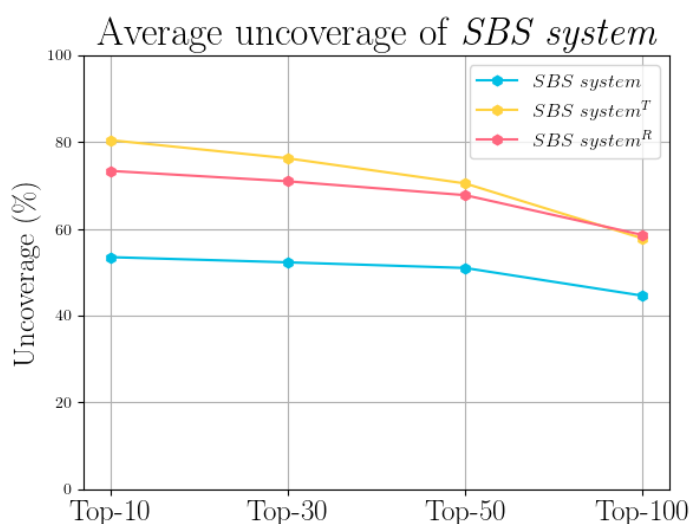


ภาพที่ 4-2 การเปรียบเทียบผลลัพธ์ค่าความครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีการสกัดคำสำคัญ

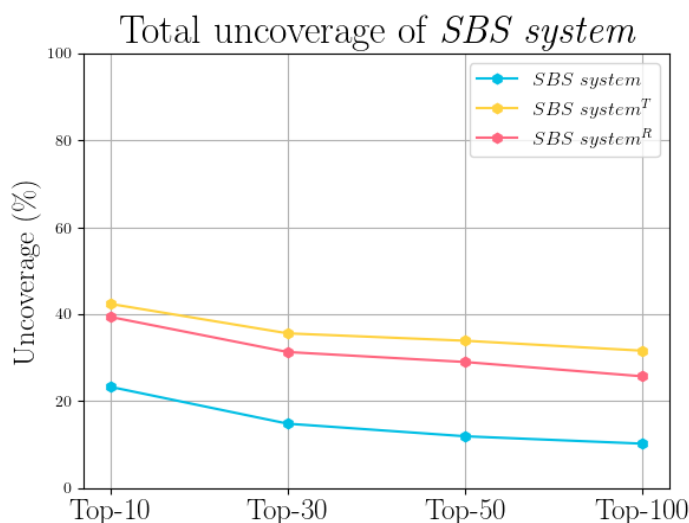
จากภาพที่ 4-1 จะสังเกตได้ว่าเมื่อจำนวนในการแนะนำหนังสือเพิ่มมากขึ้น ค่าความครอบคลุมโดยเฉลี่ยจะลดลง เห็นได้ชัดว่าเมื่อทำการแนะนำหนังสือในจำนวนที่มากขึ้นมีโอกาสที่จะได้หนังสือที่เกี่ยวข้องกับคำอธิบายรายวิชาน้อยลง (หนังสือที่แนะนำทั้งหมดมีการจัดอันดับที่เกี่ยวข้องกับคำอธิบายรายวิชาเรียงลำดับจากหนังสือที่มีความเกี่ยวข้องมากที่สุดไปยังหนังสือที่มีความเกี่ยวข้องน้อยที่สุด และพิจารณาการแนะนำหนังสือจากความเกี่ยวข้องมากที่สุด N เล่ม) ดังนั้นสามารถบอกได้ว่าหนังสือที่ถูกจัดอันดับอยู่ในอันดับต้น ๆ นั้นมีความเกี่ยวข้องกับคำอธิบายรายวิชามากที่สุด จะเห็นได้ว่าโดยค่าความครอบคลุมโดยเฉลี่ยที่ $N = 10$ จะมีค่าความครอบคลุมโดยเฉลี่ยมากที่สุดในทุก ๆ ระบบแนะนำหนังสืออ่านเพิ่มเติม บ่งบอกได้ว่านิสิต/นักศึกษาจะถูกนำหนังสือที่มีความเกี่ยวข้องเกี่ยวกับเนื้อหาในชั้นเรียนมากที่สุดก่อน ส่งเสริมให้การเลือกอ่านหนังสืออ่านเพิ่มเติมนั้นมีคุณภาพยิ่งขึ้น และยังสังเกตได้ว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความครอบคลุมโดยเฉลี่ยสูงกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” หมายความว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบสามารถสร้างรายการแนะนำหนังสือที่มีความเกี่ยวข้องเกี่ยวกับคำอธิบายรายวิชาได้มากกว่าทั้ง 2 ระบบแนะนำหนังสืออ่านเพิ่มเติมที่นำมาเปรียบเทียบ

ในทางตรงกันข้าม ภาพที่ 4-2 ค่าความครอบคลุมสะสมจะเพิ่มมากขึ้น เมื่อจำนวนในการแนะนำหนังสือเพิ่มมากขึ้น เนื่องจากจำนวนหนังสือที่แนะนำเพิ่มมากขึ้นก็จะมีค่าความเหมือนของคำสำคัญระหว่างคำสำคัญที่มีอยู่ในหนังสือและคำสำคัญที่มีอยู่ในคำอธิบายรายวิชาปรากฏมากขึ้นไปด้วย จะเห็นได้ว่าโดยค่าความครอบคลุมโดยเฉลี่ยที่ $N = 100$ จะมีค่าความครอบคลุมสะสมมากที่สุดในทุก ๆ ระบบแนะนำหนังสืออ่านเพิ่มเติม บ่งบอกได้ว่าหนังสือแต่ละเล่มจะมีความเกี่ยวข้องเกี่ยวกับคำอธิบายรายวิชาในเนื้อหาที่แตกต่างกันออกไป เมื่อเลือกอ่านหนังสือเพิ่มมากขึ้นจะสามารถเพิ่มพูนความรู้ความเข้าใจในเนื้อหาต่าง ๆ ที่เกี่ยวข้องกับคำอธิบายรายวิชาได้มากยิ่งขึ้นไปด้วย และสังเกตได้ว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความครอบคลุมสะสมสูงกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” หมายความว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบสามารถสร้างรายการแนะนำหนังสือได้ครอบคลุมเนื้อหาในคำอธิบายรายวิชาได้มากกว่าทั้ง 2 ระบบแนะนำหนังสืออ่านเพิ่มเติมที่นำมาเปรียบเทียบ

นอกจากนี้ยังพิจารณาค่าความไม่ครอบคลุมโดยเฉลี่ย (Average uncoverage) และค่าความไม่ครอบคลุมสะสม (Total uncoverage) เพื่อพิจารณาหนังสือที่มีค่าสำคัญที่เพิ่มมากขึ้นหรือแตกต่างออกไปจากค่าสำคัญในคำอธิบายรายวิชา โดยค่าความไม่ครอบคลุมโดยเฉลี่ย สามารถคำนวณในทางตรงกันข้ามกับค่าความครอบคลุมโดยเฉลี่ย และค่าความไม่ครอบคลุมสะสม สามารถคำนวณในทางตรงกันข้ามกับค่าความครอบคลุมสะสม เช่นเดียวกัน ภาพที่ 4-3 แสดงการเปรียบเทียบค่าความไม่ครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติม และภาพที่ 4-4 แสดงการเปรียบเทียบค่าความไม่ครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติม ตามลำดับ



ภาพที่ 4-3 การเปรียบเทียบผลลัพธ์ค่าความไม่ครอบคลุมโดยเฉลี่ยของระบบแนะนำหนังสืออ่านเพิ่มเติม ด้วยแต่ละวิธีสกัดความสำคัญ



ภาพที่ 4-4 การเปรียบเทียบผลลัพธ์ค่าความไม่ครอบคลุมสะสมของระบบแนะนำหนังสืออ่านเพิ่มเติมด้วยแต่ละวิธีสกัดความสำคัญ

ด้วยเหตุที่รายการหนังสือที่แนะนำทั้งหมดมีการจัดอันดับที่ไม่เกี่ยวเนื่องกับคำอธิบายรายวิชาเรียงลำดับจากหนังสือที่มีความไม่เกี่ยวเนื่องมากที่สุดไปยังหนังสือที่มีความไม่เกี่ยวเนื่องน้อยที่สุด และพิจารณาการแนะนำหนังสือจากความไม่เกี่ยวเนื่องมากที่สุด N เล่ม จากภาพที่ 4-3 จะสังเกตเห็นได้ว่าค่าความไม่ครอบคลุมโดยเฉลี่ยก็จะลดลง เมื่อจำนวนในการแนะนำหนังสือที่ไม่เกี่ยวเนื่องเพิ่มมากขึ้น แสดงให้เห็นว่าหนังสือที่มีความไม่เกี่ยวเนื่องกับคำอธิบายรายวิชามากที่สุดจะถูกจัดอันดับอยู่ต้น ๆ ของรายการแนะนำ ทำให้นิสิต/นักศึกษาสามารถเลือกอ่านหนังสือที่มีเนื้อหาเพิ่มเติมมากขึ้นจากในชั้นเรียน ได้เรียนรู้เนื้อหาใหม่ ๆ เพื่อเพิ่มพูนทักษะและองค์ความรู้ของตนเอง และจะเห็นได้ว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความไม่ครอบคลุมโดยเฉลี่ยน้อยกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” เนื่องจากวิธีการสกัดคำสำคัญที่ใช้ในระบบสามารถสกัดคำสำคัญได้ถูกต้องและครอบคลุมคำอธิบายรายวิชา จึงทำให้ค่าความไม่ครอบคลุมโดยเฉลี่ยน้อยกว่าทั้ง 2 ระบบแนะนำหนังสืออ่านเพิ่มเติมที่นำมาเปรียบเทียบ

นอกจากนี้ ในภาพที่ 4-4 จะสังเกตเห็นได้ว่าค่าความไม่ครอบคลุมสะสมจะลดลง เมื่อจำนวนการแนะนำหนังสือมากยิ่งขึ้น อันเนื่องมาจากค่าความไม่ครอบคลุมสะสมคำนวณได้จากค่าตรงกันข้ามกับความครอบคลุมสะสม ดังนั้นเมื่อจำนวนหนังสือที่แนะนำเพิ่มมากยิ่งขึ้นก็จะมีค่าความเหมือนของค่าสำคัญระหว่างค่าสำคัญที่มีอยู่ในหนังสือและค่าสำคัญที่มีอยู่ในคำอธิบายรายวิชาปรากฏมากขึ้นไปด้วย ทำให้ค่าความไม่เกี่ยวเนื่องระหว่างหนังสือกับคำอธิบายรายวิชานั้นลดลงไปด้วย และสังเกตได้ว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดค่าสำคัญที่ใช้ในระบบมีค่าความไม่ครอบคลุมสะสมน้อยกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดค่าสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดค่าสำคัญด้วยวิธีการ “RAKE” เพราะว่าวิธีการสกัดค่าสำคัญที่ใช้ในระบบสามารถสกัดค่าสำคัญออกมาได้ครอบคลุมเนื้อหาในคำอธิบายรายวิชา จึงทำให้ค่าความไม่ครอบคลุมสะสมน้อยกว่าทั้ง 2 ระบบแนะนำหนังสืออ่านเพิ่มเติมที่นำมาเปรียบเทียบ

บทที่ 5

สรุปและอภิปรายผล

งานวิจัยนี้นำเสนอระบบแนะนำหนังสืออ่านเพิ่มเติม (Supplementary book suggestion system) หรือเรียกว่า “SBS system” เพื่อส่งเสริมการอ่านและการเรียนรู้ด้วยตนเองภายใต้คำค้นของผู้ใช้ เนื่องจากการอ่านหนังสือเป็นปัจจัยสำคัญต่อการพัฒนาตนเองให้รอบรู้ รวมไปถึงการพัฒนากระบวนการทางความคิดได้ อีกทั้งในปัจจุบันมีระบบค้นหาหนังสือและระบบแนะนำหนังสือต่าง ๆ มากมาย เพื่อช่วยให้ผู้อ่านมีความสะดวกในการค้นหาและเลือกหนังสือที่ตนเองสนใจได้ดียิ่งขึ้น อาทิเช่น การค้นหาหนังสือจากความสอดคล้องของคำค้นหนึ่ง ๆ กับชื่อหนังสือหรือสารบัญหนังสือ การแนะนำหนังสือจากความนิยมของหนังสือ การแนะนำหนังสือจากข้อคิดเห็นที่ได้รับจากผู้อ่าน การแนะนำหนังสือจากพฤติกรรมการยืมหนังสือในห้องสมุด เป็นต้น แต่อย่างไรก็ตามแนวคิดทั้งหมดไม่ได้มุ่งเน้นในการแนะนำหนังสือจากความสอดคล้องของหนังสือที่เกี่ยวข้องกับการศึกษาในชั้นเรียน และไม่ได้มุ่งเน้นในการแนะนำหนังสือที่เกี่ยวข้องกับศาสตร์ที่มีคำศัพท์เฉพาะทางค่อนข้างมากกับคำค้น (คำค้นหนึ่ง ๆ สามารถมีได้หลายรูปแบบ อาทิเช่น คำหนึ่ง ๆ คำหลายคำ คำอธิบายรายวิชาที่ประกอบไปด้วยเนื้อหาต่างๆ ประมวลผลรายวิชา) ที่ซึ่งช่วยให้ผู้อ่านได้รับการแนะนำที่สอดคล้องกับเนื้อหาที่ตนเองสนใจมากยิ่งขึ้น

ดังนั้นงานวิจัยนี้จึงมุ่งเน้นในการแนะนำหนังสือโดยพิจารณาจากความสอดคล้องของหนังสือที่เพิ่มเติมจากการศึกษาในชั้นเรียน ที่ซึ่งเป็นหนังสือที่เกี่ยวข้องกับคำค้นที่เฉพาะเจาะจงเป็นหลัก โดยมุ่งเน้นที่จะทำการแนะนำหนังสืออ่านเพิ่มเติมจากการศึกษาในชั้นเรียน หรือ มหาวิทยาลัยผ่านการพิจารณาคำอธิบายรายวิชาและ/หรือประมวลรายวิชา โดยรูปแบบการแนะนำหนังสืออ่านเพิ่มเติมจะเป็นการสร้างรายการของหนังสือที่มีความเกี่ยวข้องกับคำค้นหรือกลุ่มของคำค้นที่ผู้อ่านกำหนด (คำค้นอาจเป็นคำหนึ่ง ๆ กลุ่มของคำ คำอธิบายรายวิชาและ/หรือประมวลรายวิชา) โดยหนังสือในรายการที่ถูกแนะนำจะถูกเรียงลำดับตามความเกี่ยวเนื่องของเนื้อหา (เรียงจากมากไปน้อยและแสดงถึงความเกี่ยวเนื่องของหนังสือกับคำค้น)

ระบบแนะนำหนังสืออ่านเพิ่มเติม ที่นำเสนอประกอบไปด้วย 5 ขั้นตอนการดำเนินงานหลัก 1) การรวบรวมคำศัพท์เฉพาะทางด้านวิทยาการคอมพิวเตอร์ 2) การรวบรวมข้อมูลคำอธิบายรายวิชาทางด้านวิทยาการคอมพิวเตอร์ 3) การสกัดคำสำคัญจากคำอธิบายรายวิชาและการสกัดคำสำคัญจากหนังสือ 4) การค้นหาหนังสือที่มีความเกี่ยวข้องกับคำสำคัญในคำอธิบายรายวิชา และ 5) การคำนวณความเกี่ยวข้องของหนังสือที่ครอบคลุมคำสำคัญในคำอธิบายรายวิชา การจัดอันดับหนังสือที่มีความครอบคลุมคำสำคัญมากไปยังหนังสือที่มีความครอบคลุมคำสำคัญน้อย และการสร้างรายการแนะนำหนังสืออ่านเพิ่มเติมที่เกี่ยวข้องกับคำอธิบายรายวิชา โดยพิจารณาการแนะนำหนังสือจำนวน N เล่ม โดยในงานวิจัยนี้พิจารณาค่า N ที่น่าสนใจที่ 10, 30, 50 และ 100 ตามลำดับ

ในการประเมินประสิทธิภาพของระบบแนะนำหนังสืออ่านเพิ่มเติมแบ่งออกเป็น 2 ส่วน ได้แก่

1) การวัดประสิทธิภาพของการสกัดคำสำคัญที่ใช้ในระบบ ที่ซึ่งจะเห็นได้ว่าการสกัดคำสำคัญเป็นวิธีการที่มีความสำคัญต่อประสิทธิภาพของระบบ ในการทดลองได้ทำการเปรียบเทียบวิธีการสกัดคำสำคัญกับขั้นตอนวิธี “Termine” และขั้นตอนวิธี “RAKE” ซึ่งเป็นวิธีการสกัดคำสำคัญโดยนำวิธีการทางภาษาศาสตร์มาประยุกต์ใช้และได้รับความนิยม จากนั้นทำการวัดประสิทธิภาพของการสกัดคำสำคัญระหว่างวิธีการสกัดคำสำคัญที่ใช้ในระบบ โดยเทียบกับวิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และวิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” ผ่านการการวัดค่าความถูกต้อง (Precision), ค่าความแม่นยำ (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) ตามลำดับ โดยจากผลลัพธ์ของการวัดประสิทธิภาพของการสกัดคำสำคัญจะสังเกตได้ว่าขั้นตอนวิธี “Termine” มีค่าความถูกต้อง 91.1%, ค่าความแม่นยำ 40.6% แสดงให้เห็นว่าขั้นตอนวิธี “Termine” สามารถสกัดคำสำคัญได้ถูกต้องสูงแต่ด้วยค่าความแม่นยำที่ค่อนข้างน้อยแสดงให้เห็นว่าไม่สามารถที่จะสกัดคำสำคัญออกมาได้ครอบคลุมทั้งหมดของคำอธิบายรายวิชา ทำให้ค่าประสิทธิภาพโดยรวมมีค่า 55.1% ต่อมา ขั้นตอนวิธี “RAKE” มีค่าความถูกต้อง 70.1%, ค่าความแม่นยำ 62.6% แสดงให้เห็นว่าขั้นตอนวิธี “RAKE” สามารถสกัดคำสำคัญถูกต้องได้น้อยกว่าขั้นตอนวิธี “Termine” แต่จากค่าความแม่นยำแสดงให้เห็นว่าขั้นตอนวิธี “RAKE” สามารถสกัดคำสำคัญได้ครอบคลุมมากกว่าขั้นตอนวิธี “Termine” และด้วยค่าความแม่นยำที่สูงกว่านี้ ทำให้ค่าประสิทธิภาพโดยรวมมีค่า 66.0% มากกว่าขั้นตอนวิธี “Termine” และสุดท้ายวิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความถูกต้อง 91.3%, ค่าความแม่นยำ 95.3% แสดงให้เห็นว่าสามารถสกัดคำสำคัญได้ถูกต้องมากกว่าทั้ง 2 ขั้นตอนวิธีที่นำมาเปรียบเทียบ เนื่องด้วยคำสำคัญที่ใช้ในระบบสามารถสกัดคำสำคัญ

ได้ครอบคลุมผลลัพธ์ของคำสำคัญของทั้ง 2 ขั้นตอนวิธีสกัดได้ทั้งหมด และจากค่าความแม่นยำสามารถแสดงได้ว่าวิธีการสกัดคำสำคัญที่ใช้ในระบบสกัดคำสำคัญออกมาได้ครอบคลุมคำอธิบายรายวิชาได้สูง เนื่องด้วยกฎทางภาษาศาสตร์ที่สร้างขึ้นมีการประกอบคำที่เกี่ยวข้องเข้าด้วยกันโดยพิจารณาจากคำบุพบทหรือคำสันธาน ทำให้สามารถสกัดคำสำคัญได้มากกว่าทั้ง 2 ขั้นตอนวิธีที่นำมาเปรียบเทียบ

2) การประเมินประสิทธิภาพของระบบแนะนำหนังสืออ่านเพิ่มเติม โดยพิจารณาจากค่าความครอบคลุมโดยเฉลี่ย (Average coverage) และค่าความครอบคลุมสะสม (Total coverage) (พิจารณาค่าความไม่ครอบคลุมโดยเฉลี่ย (Average uncoverage) และค่าความไม่ครอบคลุมสะสม (Total uncoverage)) ซึ่งทำการเปรียบเทียบผลลัพธ์ค่าความครอบคลุมโดยเฉลี่ย, ค่าความครอบคลุมสะสม, ค่าความไม่ครอบคลุมโดยเฉลี่ย และค่าความไม่ครอบคลุมสะสม ระหว่างระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบ (SBS system), ระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” (SBS system^T) และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” (SBS system^R) โดยผลลัพธ์ของการเปรียบเทียบแสดงให้เห็นว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญที่ใช้ในระบบมีค่าความครอบคลุมโดยเฉลี่ย และค่าความครอบคลุมสะสมสูงกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE” และมีค่าความไม่ครอบคลุมโดยเฉลี่ย และค่าความไม่ครอบคลุมสะสมน้อยกว่าระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “Termine” และระบบแนะนำหนังสืออ่านเพิ่มเติมที่ใช้วิธีการสกัดคำสำคัญด้วยวิธีการ “RAKE”

จากการประเมินระบบแนะนำหนังสืออ่านเพิ่มเติมจะเห็นได้ว่าวิธีการสกัดคำสำคัญที่ใช้ในระบบที่มีค่าความถูกต้อง (Precision), ค่าความแม่นยำ (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) สูง ทำให้ค่าความครอบคลุมของคำสำคัญระหว่างคำอธิบายรายวิชาและหนังสือที่เกี่ยวข้องมีค่าสูงขึ้นเช่นเดียวกัน ส่งผลให้รายการหนังสือที่ถูกแนะนำจากระบบแนะนำหนังสืออ่านเพิ่มเติมมีความเกี่ยวข้องกับคำอธิบายรายวิชามากยิ่งขึ้น นำไปสู่การแนะนำหนังสืออย่างมีประสิทธิภาพซึ่งจะช่วยให้ผู้อ่านสามารถตัดสินใจที่จะเลือกหนังสืออ่านเพิ่มเติมได้ง่ายและสะดวกยิ่งขึ้น

บรรณานุกรม

- Chandak, M., Girasem, S., & Mukhopadhyay, D. (2015). Introducing Hybrid Technique for Optimization of Book Recommender System. *Procedia Computer Science*, 45, 23 – 31.
- Chen, L.-C., Kuo, P.-J., & Liao, I-E. (2015). 2015. Ontology-based library recommender system using MapReduce. *Cluster Computing*, 18(1), 113 – 121.
- Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International journal on digital libraries*, 3(2), 115-130.
- Garrido, A. L., Ilarri, S. (2014). TMR: A Semantic recommender system using topic maps on the items descriptions. *Proceedings of the 11th European Conference of Web Semantic*, 213 – 217.
- Garrido, A. L., Pera, M. S., Ilarri, S. (2014). SOLE-R, a Semantic and Linguistic Approach for Book Recommendations, *Proceedings of IEEE 14th International Conference on Advanced Learning Technologies (ICALT)*, 524 – 528.
- Hulth, A. (2003, July). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the 2003 conference on Empirical methods in natural language processing* (pp. 216-223). Association for Computational Linguistics.
- Kim, J. K., Kim, H. K., Oh, H. Y., & Ryu, R. U. (2010). A group recommendation system for online communities. *International Journal of Information Management: The Journal for Information Professionals*, 30(3), 212 – 219.

- Amphawan, K., Lenca, P. (2015), Mining top-k frequent-regular closed patterns. *Expert Systems with Applications*, 42(21), 2015, 7882-7894.
- Linden, G., Smith, B., & York, J. (2003). Amazon.com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), 76-80.
- Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*.
- Ng, Y-K. (2016). Recommending Books for Children Based on the Collaborative and Content-Based Filtering Approaches. *Computational Science and Its Applications -- ICCSA 2016*, 9789, 302-317.
- Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1-20
- Saxena, S., Singh, B., & Khan, Z. (2014). Frequent pattern mining technique for improving book lending recommendation service. *Journal of Basic and Applied Engineering Research*, 1(10), 38-52.
- Teng, Y., Zhang, L., Tian, Y., & Li, X. (2015). A Novel FAHP Based Book Recommendation Method by Fusing Apriori Rule Mining. *Proceedings of the 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 237 – 243.
- Vaz, P. C., Matos, D. M., & Martins, B. (2012). Stylometric relevance-feedback towards a hybrid book recommendation algorithm. *Proceedings of the fifth ACM workshop on Research advances in large digital book repositories and complementary media*, 13 – 16.

ภาคผนวก



2018

ICAICTA

International Conference on Advanced Informatics:
Concepts, Theory and Applications

Conference : 14-17 August, 2018 | Krabi, Thailand

ICAICTA2018

@Krabi, Thailand

August 14 - 17, 2018

Call for Papers

ICAICTA 2018 - International Conference on Advanced Informatics: Concept Theory and Applications

Prospective authors are invited to submit full papers of no more than six (6) pages including results, according to template provided below. Papers will be subject to double blind peer review in which the reviewers do not know the author's identity. In order to make blind reviewing possible, authors must omit their names and affiliations from the paper.

Conference website : <http://services.informatics.buu.ac.th/icaicta2018/>

Scope of Conference

Computational Science and Engineering

- High Performance Computing
- Distributed Systems and Pervasive Computing
- Numerical Methods and Simulations
- Game Theory and Optimizations
- Cryptography and Data Hiding

Media Technology and Information Systems

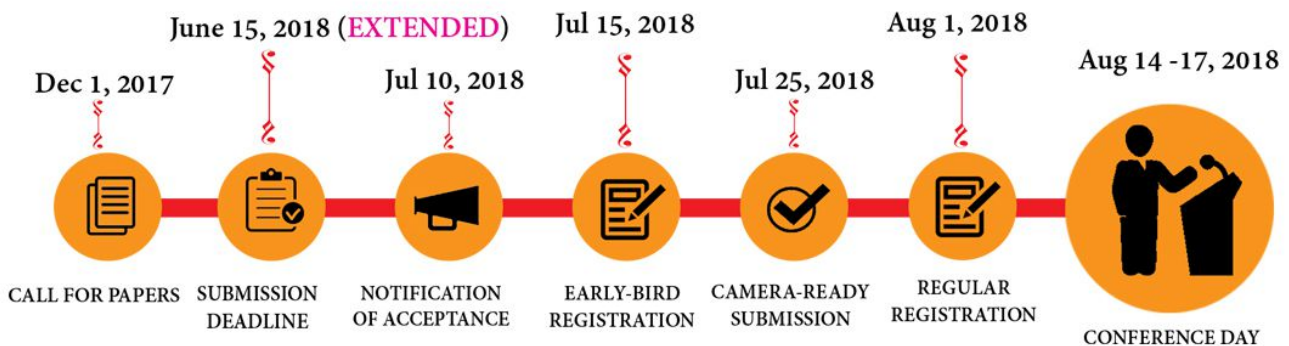
- Computer Vision, Graphics, and Image Processing
- Natural Language and Speech Processing
- Human Computer Interactions
- Multimedia Services and Applications
- Audit, Security and Governance
- Intelligent Robotics

Intelligent Systems and Data Sciences

- Information Retrieval and Data Mining
- Database and Programming
- Artificial Intelligence and Machine Learning
- Expert and Recommendation Systems
- Learning Systems and Knowledge Modeling

Computation for Science

- Computational Chemistry
- Computational Physics
- Computational Biology
- Computational Social Science



IEEE

IEEE Xplore®
Digital Library



Supplementary Book Suggestion for Computer Science Courses

Benchamawan Chaisoongnoen[§], Komate Amphawan[†]
Computational Innovation Laboratory,
Faculty of Informatics, Burapha University
Chonburi, 20131, Thailand
Email: benchamawan.gxii@gmail.com[§], komate@gmail.com[†]

Aekapop Bunpeng
Faculty of Informatics, Burapha University
Chonburi, 20131, Thailand
Email: aekapop@buu.ac.th

Abstract—On-line selling website is currently famous and popular. There are several websites selling products and/or services including on-line book-selling websites. At present, the current book-selling websites usually apply recommender systems to recommend a book or a set of books to customers. However, the recommender systems mostly focus on recommending books that users usually view or buy together and also on books having high review rates. This may cause failure to recommend books that cover most required contents, for example, books related to a course description of a course students have registered. To address on this issue, we here introduce an alternative recommender system called *Supplementary Books Suggestion system (SBS system)* to create a list of supplementary books related/relevance to a course description of a course in computer science domain by regarding relevance between a book and a course description. This can help students easily find supplementary books to read and also may help to encourage the students doing self-learning. Experiments on real course descriptions were conducted to investigate the effectiveness of the *SBS system* in the terms of precision, recall, F-measure and average (also total) coverage/uncoverage of contents between a list of supplementary books and a course description.

Keywords—Book recommender system, Keyword extraction, Supplementary books, Course description;

I. INTRODUCTION

Currently, it is the era of 4.0 based on the growth of the Internet and technology. People use more and more Internet in daily life in order to assist their living and to serve their needs such as finding goods/products, services, travel, places, friends and so on. Meanwhile, in the business aspect, it is the era that focuses on digital transformation and innovation. Many businesses start to improve their product or service by applying AI and/or IoT. Moreover, to improve marketing, understanding individual (personalize) needs is really important. Based on this, recommendation systems have become essential tools to assist users to find what they want and it is currently applied in a wide range of businesses such as music, movie, clothes, and so on.

In addition, recommendation systems are also widely applied in on-line book-selling industry. Amazon, O'Reilly, Goodreads, Barnes & Noble are well-known websites applying recommendation systems to suggest books to customers based

on relevance, clicks, buying behaviors, reviews and so on. By this, several techniques are proposed to improve recommendation performance, for example, (i) Book recommendation system based on combine features of content based filtering, collaborative filtering and association mining [1], (ii) Book recommendation system based on considering both table of contents (ToC) and association rule mining [2], (iii) Book Recommender System Using Fuzzy Linguistic Quantifiers [3] and so on.

Besides, there are several approaches to book recommendation using different criteria such as (i) Recommendation model for personalized book lists [4] aiming to recommend books to groups of people having common interests and then evaluate the accuracy satisfaction of the model by using questionnaires, (ii) Book recommendation system for digital library [5] trying to identify relationship between books that users interest through their loaning histories, (iii) Recommendation system using hybrid filtering (collaborative and content-based filtering) incorporate with best price identification [6], [7] proposing to recommend books that meet users need and to inform about on-line book-selling website with the lowest price, etc. However, previous book recommendation systems (mentioned above) face with several issues to provide the relevant book to user needs since most of them do not take into account book contents at a deep level and regard only the mere information of books provided on web pages along with metadata and rating. This may cause failure to recommend books that cover most required contents, for example, books related to a course description of a course students have registered.

To cope with this issue, we here introduce an alternative approach on book recommender system called *Supplementary Books Suggestion system* (also called *SBS system* for short) in order to create a list of supplementary books related/relevance to a course description of a course in computer science domain by regarding relevance between a book and a course description. This list of supplementary books can help students easily find supplementary books (or should read books) to read and also may help to encourage the students doing self-learning. Experiments on real course descriptions were conducted to investigate the effectiveness of the proposed *SBS system* in the terms of precision, recall, F-measure and average (also total) coverage/uncoverage of contents between a list of supplementary books relevant to a course description.

II. SUPPLEMENTARY BOOK SUGGESTION SYSTEM

In this section, we here describe details of our proposed supplementary book suggestion system (also called *SBS system* in short). The system aims to assist students to easily see a list of “should-read books” related to a computer science course they have registered. Moreover, this system can help to provide add-on information and ability to perform self-learning and also help to extend the body of knowledge of students. The *SBS system* consists of five computational steps (as shown in Fig. 1) : 1) Terminology gathering, 2) Course descriptions input, 3) Keyword extraction, 4) Relevance books searching, 5) Coverage calculation and ranking, respectively. Each of the above steps can be described in details as follows.

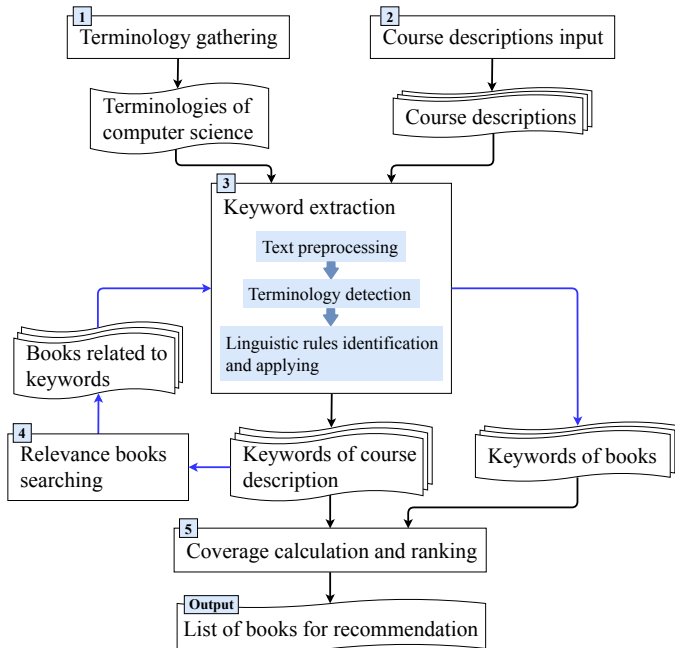


Fig. 1: Framework of the supplementary book suggestion system

A. Terminology gathering

In computer science domain, there is a large amount of reserved words and terminologies with a specific meaning, for example, “natural language processing”, “data mining”, “data structure”, “data science”, “big data” and so on. Most of these are well-known in the domain which can help to identify specific objects or actions. To study in this domain, students and/or anyone interested should have a background or know about meaning of reserved words and terminologies since course titles, course descriptions, book title and table of contents of books contain numerous of these terminologies with extension words, for instance, “data science concepts” composing from 1) “data science”—a terminology and 2) “concepts”—an extension word, respectively. In addition, identifying terminologies in the course titles, course descriptions, book title and table of contents of books can help to know about important topics to study in each course. Thus, to automatic extraction of the topics to study, terminologies in computer science domain should be prior gathered.

Based on our *SBS system*, terminologies from the lists of

computer science terms of *Oxford University*¹ and *Labautopedia*² are gathered into our corpus. Last, by regarding and removing redundant terminologies from both sites, our corpus contains 6,527 terminologies.

B. Course descriptions input

The second step of *SBS system* is to collect course descriptions in order to know main topics related to each course. To do that, course descriptions in the Computer Science curriculum of Faculty of Informatics, Burapha University [8] are considered and used as an example input. Then, the title and details of course description (in English) of a course are collected into database (as highlighted in blue in Fig. 2).

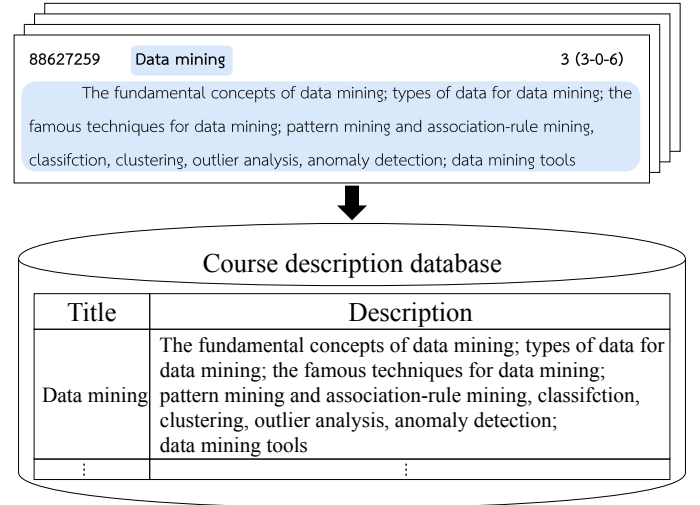


Fig. 2: Example of course descriptions input

C. Keywords extraction

Whenever a course description is input (as shown in Fig. 2), the process of keyword extraction is performed to determine important objects or actions hidden in the course description. This task consists of three consecutive steps as follows (see Fig. 3):

- 1) *Text preprocessing*—performs on (i) *Stop word removal*—removes stop words except for prepositions and conjunctions since we have to keep these, for example, the word “and” is a conjunction which can help to split the noun phrase “unstructured and structured data” to be “unstructured data” and “structured data”, (ii) *Word error correction*—corrects words from typos or input errors by spelling mistake correction (SMC) method [9], (iii) *Sentence tokenization*—splits sentences or phrases in the course description to be individual sentences, and (iv) *Part of speech tagging*—identifies type of words by using Stanford corpus [10] (Noted “JJ”, “JJR”, “JJS” refers to Adjective, “NN”, “NNS”, “NNP”, “NNPS” refers to Noun, “IN” refers to preposition and “CC” refers to conjunction, respectively).

¹<http://www.oxfordreference.com/view/10.1093/acref/9780199688975.001.0001/acref-9780199688975>

²http://www.labautopedia.org/mw/List_of_programming_and_computer_science_terms

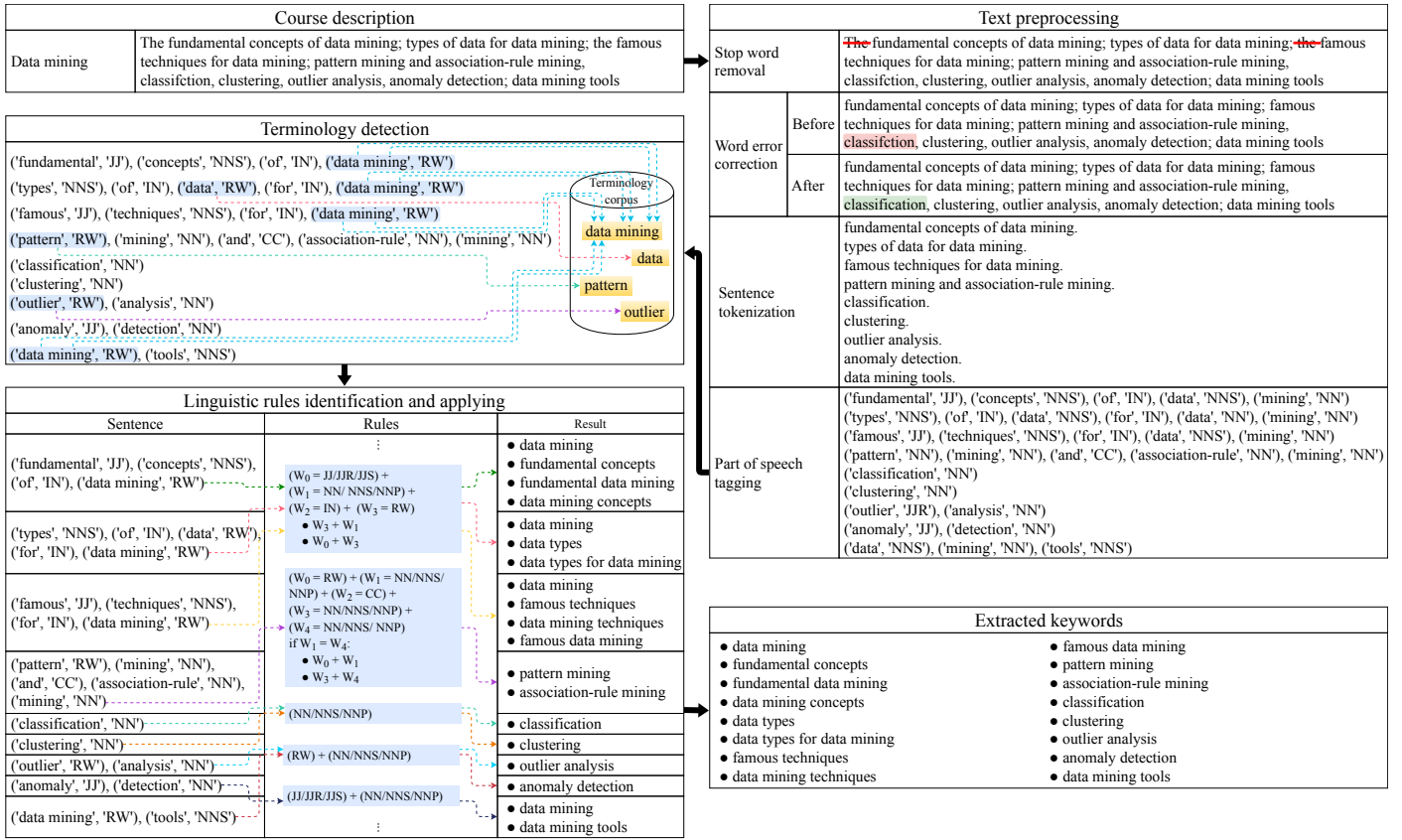


Fig. 3: Example of keyword extraction from an input course description

- 2) *Terminology detection*—finds terminologies in the course description by comparing n-grams of words in each sentence of the course description and terminologies gathered from the first step. For example, the first sentence of the course description contains (‘fundamental’, ‘JJ’), (‘concepts’, ‘NNS’), (‘of’, ‘IN’), (‘data’, ‘NNS’), (‘mining’, ‘NN’). From the sentence, the words ‘data’ and ‘mining’ are grouped together to be ‘data mining’ by the concept of 2-grams. Then, it is compared with terminologies contained in the terminology corpus and then identified as ‘reserved word’, *i.e.* ‘RW’ (see Fig. 3).
- 3) *Linguistic rules identification and applying*—creates linguistic rules by applying the concept of domain-specific term extraction of [11], [12] and then applies these rules to identify keywords hidden in each sentence of the course description. To create rules, there are two cases to be considered: (i) rules with terminology and (ii) rules without terminology, respectively. For the first type of rules, each terminology in the considered sentence is firstly regarded. It is thus merged with its surrounding noun and adjective and then identified as keywords. On the other hand, rules without terminology try to identify noun phrases being in the form of combination between nouns, adjective and/or preposition, respectively. For example, there is a created rule sequentially containing “adjective (JJ) + noun (NN or NNS or NNP) + preposition (IN) + terminology (RW)”. It can con-

verse the above sequence of words to be keywords of (i) “terminology (RW) + noun (NN or NNS or NNP)” and (ii) “adjective (JJ) + terminology (RW)”, respectively. For the first sentence from the previous step *i.e.* “fundamental (JJ) + concepts (NNS) + of (IN) + data mining (RW)”, it can be extracted to identify keywords by the above rule as (i) “data mining concepts” (RW + NNS) and (ii) “fundamental data mining” (JJ + RW), respectively. Noted that all of the rules we have created and used in the *SBS system* can be viewed and downloaded from <http://staff.informatics.buu.ac.th/~komate/SBSsystem>.

D. Relevance books searching

After extracting keywords $K_{c_i} = \{k_{c_i,1}, k_{c_i,2}, \dots, k_{c_i,n}\}$ from a course description c_i , the searching relevance books is performed. Each keyword $k_{c_i,x} \in K_{c_i}$ is thus considered and used to search for related books from <https://www.barnesandnoble.com> (*i.e.* on-line book-selling website focusing on academic books). By this, a list of books in which the important information each book *e.g.* title, descriptions, table of contents (TOC) and other essential information (see Fig. 4) is collected into our database and set to be relevant to the course description. For example, as in Fig. 4, the keyword “data mining” is considered and used for searching for related books. From the figure, the related book is “Introduction to Data mining” is collected with title = “Introduction to Data Mining”, author = “Ping-Ning Tan, Michael Steinbacj, Anuj Kapatne, Vipin Kumar”,

price = “108.50”, *isbn13* = “9780133128901”, *publisher* = “Pearson”, *publicationdate* = “01/18/2018” *edition* = “2”, *bookdescription* = “Introducing the fundamental concepts and algorithms of data mining. Introduction to Data Mining, 2nd Edition , gives a comprehensive overview of the background and general themes of data mining and is designed to be useful to students, instructors, researchers, and professionals.” etc. and *toc* = “Introduction. Data. Classification: Basic Concepts and Techniques. Classification: Alternatives Techniques. Association Analysis: Basic Concepts and Algorithms. Association Analysis: Advanced Concepts. Cluster Analysis: Basic Concepts and Algorithms. Cluster Analysis: Additional Issues and Algorithms. Anomaly Detection. Avoiding False Discoveries”, respectively.

From these information, *bookdescription* and *toc* are then considered and transferred to *Keyword extraction* step (Section II-C) in order to get keywords. Then, the keywords from *bookdescription* and *toc* are merged together and collected as the set of keywords of the related book. For example, the first sentence from *book-description* is “Introducing the fundamental concepts and algorithms of data mining”. It then transferred to *Keyword extraction* where we can get “fundamental data mining”, “data mining concepts”, and “data mining algorithm” as keywords for the book, respectively.

Fig. 4: Example of collecting of a book related to the keyword “Data mining” from the course description of Data mining course

E. Coverage calculation and ranking

The last step of the *SBS system* is the calculation of coverage (similarities) between the keywords extracted from a course description and the keywords from each of its related

books and then ranking related books based on their coverages. To calculate similarity between a course description c_i with keywords $K_{c_i} = \{k_{c_i,1}, k_{c_i,2}, \dots, k_{c_i,n}\}$ and a book b_j related to c_i with keywords $K_{b_j} = \{k_{b_j,1}, k_{b_j,2}, \dots, k_{b_j,m}\}$, the percentage of *coverage* and *uncoverage* [13] between the book b_j and the course description c_i is applied as follows:

$$C(c_i, b_j) = \frac{|K_{c_i} \cap K_{b_j}|}{|K_{c_i}|}$$

$$U(c_i, b_j) = 1 - C(c_i, b_j)$$

Therefore, as described in Section II-D, each course description c_i has a list of related books $B_{c_i} = \{b_j, \dots, b_k\}$ in which each book b_j associates with the set of keywords $K_{b_j} = \{k_{b_j,1}, k_{b_j,2}, \dots, k_{b_j,m}\}$. Therefore, by using the two formulas as above, there will be a list of coverage values $C_{c_i} = \{C(c_i, b_j), \dots, C(c_i, b_k)\}$ and a list of uncoverage values $U_{c_i} = \{U(c_i, b_j), \dots, U(c_i, b_k)\}$ expressing similarities and dissimilarities between the course description c_i and the list of related books B_{c_i} , respectively. Thus, the list of recommended books can be generated by ranking coverage values in descending order and filtered out only the top- N books with highest coverage (Noted that this paper specifies the value of N to be 10, 30, 50 and 100, respectively). On the other hand, the list of recommended booked can also be created in the same manner as above by using uncoverage value. Thus can help to gain a list of books having details (keywords) extending from details (keywords) in the course description.

III. EXPERIMENTAL STUDY

In this section, we here describe experimental studies conducted to evaluate our proposed *SBS system*. Seventy-two course descriptions from the Computer Science curriculum of Faculty of Informatics, Burapha University [8] are considered and collected. To create a list of supplementary books related to a course description, the number of recommended books is set to be 10, 30, 50 and 100, respectively. Noted that these numbers are usually used for top- N recommendations [14], [15] and mining top- k (N -most) interesting patterns [16], [17].

As detailed in Section II, we can see that keyword extraction is the major computation of the *SBS system* affecting to its effectiveness. Thus, we then also observe the performance of the keyword extraction used in the *SBS system* and make a comparative study with *Termine* [11] and *RAKE* [18] (two famous keyword extraction techniques). Three popular metrics, recall, precision and F-measure, are applied to observe the accuracy of the three keyword extraction techniques in which each metric can be defined as follows.

$$precision = \frac{\text{number of relevant items retrieved}}{\text{number of items retrieved}}$$

$$recall = \frac{\text{number of relevant items retrieved}}{\text{number of relevant items}}$$

$$F\text{-measure} = 2 \times \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

As shown in Fig. 5, the performance of our keyword extraction in comparison with *Termine* and *RAKE* are shown. Anyhow, to evaluate whether an extracted keyword is a correct keyword or not, we have labeled keywords on course descriptions based on the help from expertises (lecturers on the courses). Based on this, the number of extracted keywords, the number of correct extracted keywords, recall, precision, and F-measure of the three techniques can be calculated. From the figure, we can see that our technique outperforms both *Termine* and *RAKE* techniques in all metrics. Based on *Termine*, noun and preposition phrases are first considered to create a list of candidate keywords and then sent the list to expertises in order to identify keywords. However, *Termine* does not take into account domain terminologies and other linguistic rules (such as combination of words, combination of words with a terminology) which can help to improve its performance. Meanwhile, *RAKE* mainly focuses on considering “word co-occurrence” by applying the n-grams technique to identify keywords. It then regards on a frequency of occurrence of co-occurrence words, ranks co-occurrence words by descending order of their frequency and selects one-third highest frequency co-occurrence words to be candidate keywords. Last, the keywords are thus identified from the list of candidates by expertises. However, for our keyword extraction technique, it applies several linguistic rules including noun phrases, adjective phrases, preposition phrases, term phrases (*i.e.* terminology + noun, terminology + adj, and terminology + preposition) and coordination of phrases (*i.e.* generating a keyword based on combinations between nouns, adjective and/or preposition occurring surrounding of the conjunction “and” or “or”). These can help to increase the accuracy of keyword extraction. Hence, the precision of our technique is 91.3% and the recall is 95.3% causing the F-measure is quite high as 93.1%. Thanks to the terminologies prior collected in the database and the created rules that mostly cover all the case of keywords in the course descriptions.

Approach	Extracted keywords	Correct keywords	Precision	Recall	F-measure
Our keyword extraction	1447	1253	91.3	95.3	93.1
Termine	1365	523	91.1	40.6	55.1
RAKE	1695	837	70.1	62.6	66.0

Fig. 5: Evaluation of keyword extraction used in *SBS system*

In addition, to investigate the quality of recommendation results, the average and total coverage (also for uncoverage) of all supplementary books from the *SBS system* are calculated and scrutinized. The average coverage (also for uncoverage) is calculated by averaging coverage (uncoverage) values of all supplementary books related to course descriptions. Meanwhile, the total coverage (uncoverage) is computed by considering a list of supplementary books related to a course description. Then, all of matched keywords (unmatched keywords for uncoverage) between keywords of the supplementary books and keywords of the course description is accumulated. Last, the number of matched keywords (unmatched keywords) of all course description is counted and averaged. Furthermore, we thus also observe and compare the coverage and uncoverage

of our *SBS system* (in the terms of average and total) with that of the *SBS system* using *Termine* or *RAKE* in stead of using our keyword extraction technique (denoted these systems as *SBS system^T* and *SBS system^R*).

Figure 6 shows the average and total coverage of the three systems. From the figure, it can be observed that as the number of recommended books increases, the average coverage decreases. Obviously, with higher number of required recommended books, there are more chance to get less relevance books related to a course description (Noted that all of recommended books are ranked by descending order of their courage and then selected by the top-*N* rank). Conversely, the total coverage of the three systems increases as the number of recommended books increases. It is because higher number of recommended books leads to having more topics related to course descriptions. Last, it can be seen that our *SBS system* can give higher average and total coverage than the other two systems (Thanks to our keyword extraction technique).

The average and total uncoverage of the three systems are also observed and illustrated in Fig. 7 to see the expansion of topics extended from course descriptions. By this, a (another) list of books related to each course description having several additional contents is created. This list can help students to get/know more topics related to the course to learn about. The average and total uncoverages are calculated on the contrary of the average and total coverage. By this, all books in the list of recommended books related to a course description are ranked by descending order of uncoverage value (the coverage value is not used on this basis). From the figure, it can be seen that the average uncoverage of the three systems decreases as the number of required recommended books increases. As mentioned above, all of recommended books are ranked by descending order of uncoverage value. Thus, higher number of books returned causes lower value of average coverage. Moreover, the total uncoverage of the three systems also decreases as the number of required recommended books increases. Higher number of books suggested makes the systems take more and more books with low uncoverage to be results. Last, we can also observe that our *SBS system* gives lower average and total coverage than the other two systems. Based on *SBS system*, it usually extracts correct keywords and then retrieves relevance books based on the correct keywords. This leads the *SBS system* to generates low uncoverage lists of recommended books in comparison with the other two systems.

IV. CONCLUSION

In this paper, we have introduced a new book recommender system called *Supplementary Book Suggestion system (SBS system)*. The system can suggest a list of should read books related to a course description of a course in computer science domain. This can help students easily find supplementary books to read and also may help to encourage the students doing self-learning. Based on a course, its course description is regarded to extract important keywords in order to understand/know about objects and actions concerned with the course. These keywords are thus used to match between the course and its relevant books. Experiments on real courses were done to investigate precision, recall and F-measure of our keyword extraction technique in a comparison to *RAKE* and *Termine* since the performance of keyword extraction affect

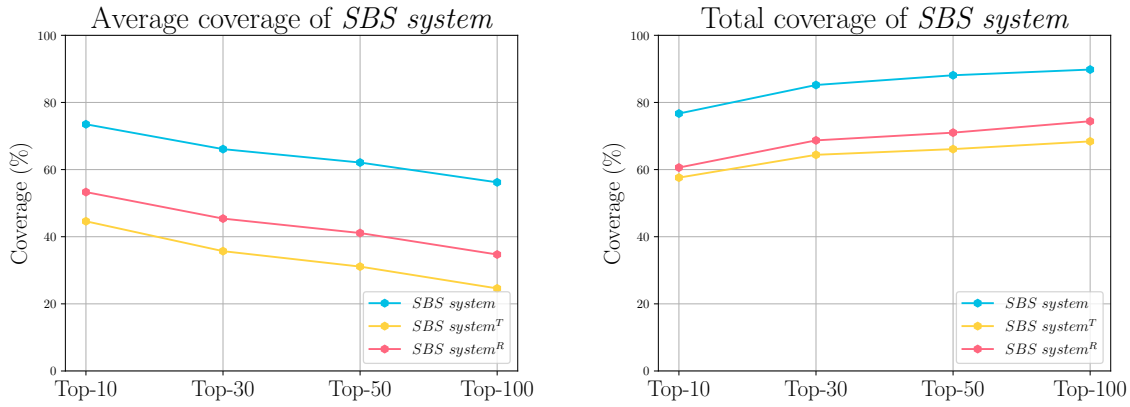


Fig. 6: Average and total coverage of *SBS system*, *SBS system*^T and *SBS system*^R

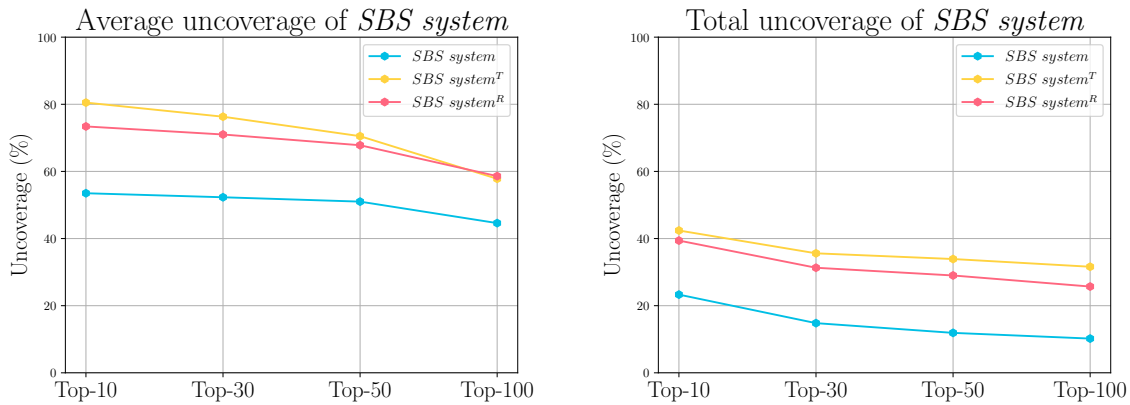


Fig. 7: Average and total uncoverage of *SBS system*, *SBS system*^T and *SBS system*^R

to performance of *SBS system*. In addition, the average and total coverages (also for uncoverage) of the list of books to the course description are also observed. From all of the experiment, it can be seen that our keyword extraction technique and *SBS system* can effectively suggest supplementary books related to a course.

ACKNOWLEDGEMENT

This work was financially supported by a research grant of Burapha University through the National Research Council of Thailand (Grant No. 16/2561). In addition, the authors would like to give a special thanks to all lecturers of Faculty of Informatics, Burapha University on help and supports.

REFERENCES

- [1] A. S. Tewari, A. Kumar, and A. G. Barman, "Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining," in *2014 IEEE International Advance Computing Conference (IACC)*, 2014, pp. 500–503.
- [2] Z. Ali, S. Khusro, and I. Ullah, "A hybrid book recommender system based on table of contents (toc) and association rule mining," in *Proceedings of the 10th International Conference on Informatics and Systems*, 2016, pp. 68–74.
- [3] S. S. Sohail, J. Siddiqui, and R. Ali, *Book Recommender System Using Fuzzy Linguistic Quantifiers*. Singapore: Springer Singapore, 2017, pp. 47–60.
- [4] S. Maneewongvatana and S. Maneewongvatana, "A recommendation model for personalized book lists," in *2010 10th International Symposium on Communications and Information Technologies*, 2010, pp. 389–394.
- [5] P. Jomsri, "Book recommendation system for digital library based on user profiles by using association rule," in *Fourth edition of the International Conference on the Innovative Computing Technology (INTECH 2014)*, 2014, pp. 130–134.
- [6] A. S. Tewari and K. Priyanka, "Book recommendation system based on collaborative filtering and association rule mining for college students," in *2014 International Conference on Contemporary Computing and Informatics (IC3I)*, 2014, pp. 135–138.
- [7] A. Goel, D. Khandelwal, J. Mundhra, and R. Tiwari, "Intelligent and integrated book recommendation and best price identifier system using machine learning," in *Intelligent Engineering Informatics*, 2018, pp. 397–412.
- [8] B. U. Faculty of Informatics, "Curriculum of computer science," <https://www.informatics.buu.ac.th/2016/wp-content/uploads/2014/11/4.6.2.37-BSc-CS-ProgramSpec-cs.pdf>, 2016.
- [9] S. Sharma and S. Gupta, "A correction model for real-word errors," *Procedia Computer Science*, vol. 70, pp. 99 – 106, 2015.
- [10] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, 2003, pp. 173–180.
- [11] K. Frantzi, S. Ananiadou, and H. Mima, "Automatic recognition of multi-word terms: the c-value/nc-value method," *International Journal on Digital Libraries*, vol. 3, no. 2, pp. 115–130, 2000.

- [12] K. Balachandran and S. Ranathunga, "Domain-specific term extraction for concept identification in ontology construction," in *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, 2016, pp. 34–41.
- [13] M. Ge, C. Delgado-Battenfeld, and D. Jannach, "Beyond accuracy: Evaluating recommender systems by coverage and serendipity," in *Proceedings of the Fourth ACM Conference on Recommender Systems*, 2010, pp. 257–260.
- [14] H. K. Kim, H. Y. Oh, J. C. Gu, and J. K. Kim, "Commenders: A recommendation procedure for online book communities," *Electronic Commerce Research and Applications*, vol. 10, no. 5, pp. 501–509, 2011.
- [15] X. Ning and G. Karypis, "Slim: Sparse linear methods for top-n recommender systems," in *2011 IEEE 11th International Conference on Data Mining*, 2011, pp. 497–506.
- [16] A. Wai-chee Fu, R. Wang-wai Kwong, and J. Tang, "Mining n-most interesting itemsets," in *Foundations of Intelligent Systems*, 2010, pp. 59–67.
- [17] K. Amphawan and P. Lenca, "Mining top-k frequent-regular closed patterns," *Expert Systems with Applications*, vol. 42, no. 21, pp. 7882 – 7894, 2015.
- [18] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic keyword extraction from individual documents," in *Text Mining: Applications and Theory*, 2010, pp. 1 – 20.