การแยกกลุ่มเอกสารภาษาไทยโดยใช้ความถี่เอกสาร

THAI DOCUMENT CLUSTERING USING DOCUMENT FREQUENCY

ไพฑูรย์ นุชแจ้ง

PAITOON NUCHJANG

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า

บัณฑิตวิทยาลัย

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ.2547

ISBN 974-9680-40-5

THAI DOCUMENT CLUSTERING USING DOCUMENT FREQUENCY

PAITOON NUCHJANG

A THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENT FOR THE DEGREE OF

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING

SCHOOL OF GRADUATE STUDIES

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2004

ISBN 974-9680-40-5

หัวข้อวิทยานิพนธ์	การแยกกลุ่มเอกสารภาษาไทยโดยใช้ความถี่เอกสาร
นักศึกษา	นายไพฑูรย์ นุชแจ้ง
รหัสประจำตัว	42061130
ปริญญา	วิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชา	วิศวกรรมไฟฟ้า
พ.ศ.	2547
อาจารย์ผู้ควบคุมวิทยานิพนธ์	รศ.ดร. บุญธีร์ เครือตราชู
อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม	รศ.ดร. ชม กิ้มปาน

บทคัดย่อ

วิทยานิพนธ์นี้เสนอการจัดกลุ่มเอกสารภาษาไทยออกเป็นกลุ่มๆ ด้วยระบบคอมพิวเตอร์ ซึ่งวิทยานิพนธ์ฉบับนี้จะแบ่งออกเป็น 2 ส่วน ในส่วนแรกจะเป็นการตัดคำภาษาไทย ส่วนที่ 2 จะเป็นวิธีการจัดกลุ่มเอกสารภาษาไทย โดยระบบจะต้องทำการตัดคำภาษาไทยออกจากประโยคก่อน และแก้ความคลุมเครือของการตัดคำภาษาไทยซึ่งจะใช้วิธี Bigram จากนั้นระบบจะนำเอาคำที่ได้ไปเทียบหากลุ่มเอกสาร เอกสารที่คล้ายกันจะอยู่ในกลุ่มเดียวกัน ในการพิจารณาจัดกลุ่มเอกสารภาษาไทย จะพิจารณาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน, จำนวนความถี่คำ,การเหมือนกันของคำระหว่างเอกสาร และจำนวนเอกสารหรือความถี่เอกสารที่มีคำเหมือนกันปรากฏอยู่ ทั้งหมดนี้ถูกนำมาวัดผลลัพธ์การจัดกลุ่มด้วยเครื่องมือวัดคือ ค่า Precision, Recall และ ค่า F-measure เปรียบเทียบกับการใช้มนุษย์จัดกลุ่มเอกสาร ในวิทยานิพนธ์นี้จะเน้นศึกษาดูการเหมือนกันของคำระหว่างเอกสาร รวมกับการหาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน และศึกษาจำนวนเอกสารที่มีคำเหมือนกันปรากฏอยู่เป็นหลักเปรียบเทียบกับวิธีอื่น ผลของการจัดกลุ่มที่ได้วิธีพิจารณาการเหมือนกันของคำระหว่างเอกสารร่วมกับการหาระยะห่างระหว่างกลุ่มคำที่เหมือนกัน เอกสารในกลุ่มจะมีเนื้อหาใกล้เคียงกันมากหรือน้อยตามค่าเทรดโฮลที่กำหนด ถ้ากำหนดค่าเทรดโฮลน้อยเอกสารจะมีเนื้อหาใกล้เคียงกันมาก แต่จะได้กลุ่มเอกสารจำนวนมากด้วย และระบบมักจะไม่หยุดทำงานเนื่องจากเอกสารมีการเปลี่ยนกลุ่มไปมา ในขณะที่การจัดกลุ่มแบบพิจารณาตามจำนวนเอกสารที่มีคำเหมือนกันปรากฏอยู่ ระบบจะหยุดทำงานได้และการจัดกลุ่มก็ดีกว่าทุกวิธีที่เปรียบเทียบ

Thesis Title	Thai Document Clustering using Document Frequency
Student	Mr. Paitoon Nuchjang
Student ID.	42061130
Degree	Master of Engineering
Programme	Electrical Engineering
Year	2004
Thesis Advisor	Assoc.Prof.Dr. Boontee Kruatrachue
Thesis Coordinate Advisor	Assoc.Prof.Dr. Chom Kimpan

ABSTRACT

This thesis presents a Thai document clustering by the computer. It is divided into two parts. First part shows a Thai word parsing and second part shows a clustering. The system has to parse Thai words from sentences and modify an ambiguity on Thai word parsing by Bigram before it brings the acquired keywords to find out groups and to gather in a cluster. That is the same document is in the same group. This research considers Thai document clustering from keywords distance, keywords frequency, similarity of keywords and quantity of documents or documents frequency which has the same keywords. All of these considerations are measured by Precision, Recall and F-measure in comparison with a clustering from human. This thesis emphasizes the study of similarity of keywords in company with keywords distance, and documents frequency which has the same keywords compared with the other methods. The results are that the similarity of keywords in company with keywords distance, the documents are comparable in contents more or less after assigned by threshold. If the threshold value is less, the contents will be very comparable and have many groups too. But sometimes the system does not stop because the documents are changed all the time. In the meanwhile, the consideration of documents frequency which has the same keywords can stop the system and the document clustering is the best.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้อย่างดี ด้วยคำแนะนำและคำปรึกษาเกี่ยวกับ การแยกกลุ่มเอกสารภาษาไทยโดยใช้ความถี่เอกสาร รวมทั้งได้ทดสอบการตรวจเทียบจาก รศ.ดร. บุญธีร์ เครือตราชู ซึ่งเป็นอาจารย์ผู้ควบคุมวิทยานิพนธ์ และ รศ.ดร. ชม กิ้มปาน อาจารย์ผู้ควบคุมวิทยานิพนธ์ร่วม ผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์จากท่านและขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบพระคุณบิดาและมารดา ซึ่งท่านทั้งสองเป็นผู้ให้การสนับสนุน ให้ความช่วยเหลือและเป็นกำลังใจให้กับผู้วิจัยอย่างยิ่ง จนสามารถทำให้งานวิจัยนี้ลุล่วงไปได้ด้วยดี ผู้วิจัยรู้สึกซาบซึ้งและขอกราบขอบพระคุณท่านทั้งสองเป็นอย่างสูง

ขอขอบคุณเพื่อนๆ นักศึกษาทุกคนที่ช่วยเหลือให้คำแนะนำต่างๆพร้อมทั้งช่วยตรวจเทียบและแก้ไขทฤษฎีและอื่นๆที่ผิดพลาดจนสำเร็จสมบูรณ์ยิ่งขึ้น และยังให้กำลังใจต่อผู้วิจัยอย่างใกล้ชิดตลอดมา

คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์ฉบับนี้ ผู้วิจัยขอมอบแด่ผู้มีพระคุณทุกท่าน

ไพฑูรย์ นุชแจ้ง

สารบัญ

หน้า

สารบัญตาราง

สารบัญรูป

สารบัญรูป(ต่อ)

สารบัญรูป(ต่อ)

บทที่1
บทนำ

1.1 ความเป็นมาของงานวิจัย

ในปัจจุบันนี้เทคโนโลยีทางด้านระบบไมโครคอมพิวเตอร์ ได้รับการพัฒนาให้มีประสิทธิภาพมากขึ้น ทำให้มีการใช้คอมพิวเตอร์อย่างกว้างขวางในงานที่แตกต่างกันออกไป โดยเฉพาะการพัฒนาซอฟต์แวร์ประเภทปัญญาประดิษฐ์ (Artificial Intelligence) ในด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing) หรือ NLP ก็เป็นศาสตร์สาขาหนึ่งของปัญญาประดิษฐ์ เนื่องจากมีการคาดหวังว่า จะสามารถนำไปประยุกต์ใช้งานด้านต่างๆได้มากมาย เช่น การจัดแยกกลุ่มเอกสาร (Document Clustering) หรือการที่จะพยายามให้เครื่องจักรหรือคอมพิวเตอร์สามารถจำลองการทำงานของมนุษย์ หรือเข้าใจในภาษามนุษย์ การจัดแยกเอกสารออกเป็นหมวดหมู่หรือกลุ่มก็เพื่อต้องการนำประโยชน์จากระบบคอมพิวเตอร์มาใช้ เพื่อลดการทำงานของมนุษย์ลง เนื่องจากว่าการจะแยกลุ่มเอกสารใดๆนั้นต้องเสียเวลาและบุคลากรจำนวนหนึ่ง ยิ่งถ้าเอกสารมีจำนวนมากจะยิ่งต้องใช้เวลาและบุคลากรมากขึ้นไปอีก การจัดแยกกลุ่มเอกสารก็มีการทำหลายวิธี แต่ละวิธีก็มีข้อดีข้อเสียแตกต่างกัน ยิ่งการแยกกลุ่มเอกสารต่างๆด้วยระบบภาษาไทย ซึ่งยังพัฒนาไปได้ไม่มากนัก ทั้งนี้เพราะภาษาของแต่ละประเทศต่างๆมีโครงสร้างและระบบการใช้ที่เป็นเอกลักษณ์เฉพาะตัว รวมทั้งข้อยกเว้นทางไวยากรณ์ที่เกิดจากความเคยชินในการใช้ภาษาของประชาชนในประเทศนั้นๆ หรืออิทธิพลจากประเทศที่อยู่ข้างเคียง ภาษาไทยก็ได้รับผลกระทบเหล่านี้เช่นกัน

ดังนั้นในการศึกษาโครงสร้างของภาษาไทย เพื่อจะนำมาประยุกต์ใช้ในการแยกกลุ่มเอกสารภาษาไทยให้ได้ถูกต้องนั้น ปัญหาส่วนหนึ่งจึงเกิดจากลักษณะของภาษาไทยนั่นเองคือ ข้อความภาษาไทยที่ประกอบขึ้นเป็นประโยคมีลักษณะเป็นการนำหน่วยคำมาเขียนติดกัน โดยไม่มีช่องว่างหรือเครื่องหมายระบุการจบของหน่วยคำ (Morpheme) บ้างข้อความจะมีการนำประโยคภาษาไทยหลายประโยคมาผสมกันเป็นข้อความเดียว และมีการปรับแต่งข้อความให้สวยงามกระทัดรัด ทำให้การแยกหน่วยคำและการแยกประโยคออกมาจากข้อความทำได้ยากขึ้น โดยบางประโยคก็มีลักษณะกำกวมแยกหน่วยคำได้ลำบาก เช่น ตานอนตากลม เป็นต้น ซึ่งต้องอาศัยบริบทที่อยู่โดยรอบจึงจะได้ความหมายที่แท้จริง จากปัญหากรณีนี้ มีผลทำให้การวิเคราะห์โครงสร้างไวยากรณ์ของประโยคภาษาไทยมีความยุ่งยากซับซ้อนมากขึ้นเช่นเดียวกัน เพราะต้องมีการแยกแยะหน่วยคำและประโยคออกมาให้ได้เสียก่อน ซึ่งต้องอาศัยวิธีการต่างๆรวมทั้งคำศัพท์ในพจนานุกรมภาษาไทย และไวยากรณ์ภาษาไทยเข้ามาช่วย เพื่อแก้ไขปัญหาทางด้านภาษาศาสตร์ที่ประสบอยู่ นอกจากปัญหาในเรื่องการแยกและวิเคราะห์หน่วยคำในภาษาไทยแล้ว ยังต้องหาวิธี

ในการจะแยกกลุ่มเอกสารออกมาจัดหมวดหมู่และกลุ่มได้อย่างไรถึงจะมีความถูกต้อง ของสมาชิกในกลุ่มเอกสารมากที่สุดด้วย ซึ่งเอกสารในกลุ่มเดียวกันควรจะมีเนื้อหาเหมือนกันหรือใกล้เคียงกัน โดยในวิทยานิพนธ์นี้ได้ทดลองหลายวิธีเปรียบเทียบกัน ทั้งพิจารณาระยะห่างระหว่างคำในเวคเตอร์ , การพิจารณาคำที่เหมือนกัน, การพิจารณาให้น้ำหนักคำตามความถี่คำ และการพิจารณาให้น้ำหนักคำตามจำนวนเอกสารที่ปรากฏคำนั้นๆ เป็นต้น ซึ่งอาศัยเครื่องมือ 3 อย่าง คือ ค่า Precision , Recall และ F- measure ในการวัดผลการจัดกลุ่มแต่ละวิธีว่าวิธีไหนให้ผลลัพธ์ที่ดีแตกต่างกันอย่างไร วิธีการจัดกลุ่มเอกสารที่ดีนั้นสามารถนำไปประยุกต์ใช้งานอย่างอื่นได้อีก นอกเหนือจากการจัดกลุ่มเอกสารภาษาไทยทั่วไปแล้ว ยังประยุกต์ใช้ในการสืบค้นเอกสารต่าง และการค้นหาฎีกาทางกฎหมายที่เคยมีการตัดสินคดีความไว้แล้วก็ได้ ซึ่งช่วยลดการทำงานของนักกฎหมายลง

1.2 จุดประสงค์และขอบเขตของงานวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาทดลองประยุกต์ระบบการวิเคราะห์ประโยคภาษาไทย และนำเอาหน่วยคำภาษาไทยที่ได้จากการตัดคำ มาใช้ในการแยกกลุ่มเอกสารภาษาไทยโดยการใช้ตัวแบ่งกลุ่มเอกสารแบบหลายๆวิธีเปรียบเทียบกัน ขอบเขตของงานวิจัยนี้เป็นการแยกกลุ่มเอกสารภาษาไทยจากบทคัดย่อวิทยานิพนธ์ในสาขาต่างๆออกเป็นกลุ่มๆ โดยมีรูปแบบประโยคที่ใช้ในการวิเคราะห์ต้องประกอบด้วยประโยคที่มีลักษณะดังต่อไปนี้

- เป็นประโยคภาษาไทยที่มีเครื่องหมาย และอักษรภาษาอื่นปนอยู่ก็ได้
- รูปแบบประโยคถูกต้องตามหลักภาษาไทย
- การแยกกลุ่มเอกสารต่างๆจะพิจารณาเฉพาะหน่วยคำภาษาไทยเท่านั้น

การพิจารณาแยกกลุ่มเอกสาร จะตรวจสอบความคล้ายของกลุ่มเอกสาร โดยในงานวิจัยนี้ได้นำเสนอกระบวนการแยกกลุ่มเอกสารภาษาไทยไว้ ดังนี้

- การวิเคราะห์หน่วยคำภาษาไทยและการแยกแยะหน่วยคำออกจากประโยค
- การวิเคราะห์รูปประโยค และหาประโยคที่ถูกต้องเพื่อนำเอาหน่วยคำสำคัญที่แยกได้มาใช้ในการแยกกลุ่มเอกสารภาษาไทย
- การหาวิธีการแยกเอกสาร โดยพิจารณาในมุมมองของการเหมือนกันของคำ,การพิจารณาระยะห่างของกลุ่มคำในเวคเตอร์, การให้น้ำหนักคำโดยดูตามความถี่คำที่เหมือนกันของเอกสาร และการพิจารณาให้น้ำหนักคำตามจำนวนเอกสารที่มีคำเหมือนกัน เพื่อจะหาวิธีที่ดีที่สุดเหมาะสมที่สุด ว่ามีวิธีไหนที่จะทำให้สมาชิกในกลุ่มเอกสารมีเนื้อหาใกล้เคียงกันมากที่สุด
- จะใช้ค่า Precision , Recall และ F-measure เป็นเครื่องมือวัดผลลัพธ์การจัดกลุ่มเอกสารที่ได้ว่าวิธีไหนให้ผลลัพธ์ที่ดีที่สุด โดยเปรียบเทียบกับการจัดกลุ่มด้วยมนุษย์

วิธีที่ดีที่สุด จะต้องได้ค่าทั้ง3เครื่องมือวัดสูงที่สุด ซึ่งนั้นหมายถึงว่ากลุ่มเอกสารนั้นๆ จะมีเนื้อหาของเอกสารภายในกลุ่มใกล้เคียงกันมากที่สุดด้วย

1.3 สมมติฐานของการศึกษา

การวิเคราะห์ประโยคภาษาไทยด้วยระบบคอมพิวเตอร์ ดังที่กล่าวมามีการพัฒนาและนำไปประยุกต์ใช้มากมาย ทั้งการแปลภาษา , การตรวจสอบเอกสาร , การแยกแยะข้อความเอกสารออกจากรูปภาพ เป็นต้น ในวิทยานิพนธ์นี้จึงมีแนวความคิดว่าน่าจะนำไปใช้ในการแยกกลุ่มเอกสารภาษาไทยออกเป็นกลุ่มๆหรือหมวดหมู่ได้บ้าง เพื่อเป็นแนวทางแก้ไขและช่วยประหยัดเวลาและทรัพยากรบุคคลในการจัดแยกเอกสารต่างๆ ทั้งนี้มีผู้วิจัยท่านอื่นที่ผ่านมาได้วางแนวทางไว้บ้างแล้วในการวิเคราะห์ระบบภาษาไทยและการการประยุกต์ใช้ในการแยกกลุ่มเอกสาร แต่ละวิธีก็มีข้อดีต่างกันแล้วแต่วิธีการที่ใช้ ดังนั้นในงานวิจัยนี้จึงได้พยายามทำการปรับปรุงข้อดีของการวิเคราะห์ประโยคภาษาไทยให้ดีขึ้นทั้งในแง่การแยกแยะหน่วยคำ , การป้อนฐานความรู้ทางภาษาไทย และการวิเคราะห์รูปประโยคภาษาไทย เพื่อมุ่งเน้นไปในการแยกกลุ่มเอกสารภาษาไทยเป็นสำคัญ และทดลองหาวิธีการแยกกลุ่มเอกสารในมุมมองอื่น โดยพิจารณาดูความคล้ายของเอกสารหลายๆแบบหลายๆวิธี โดยตั้งสมมติฐานว่าเอกสารที่อยู่ในกลุ่มเดียวกันเนื้อหาควรอยู่ใกล้กัน และเอกสารในกลุ่มเดียวกันน่าจะมีคำซ้ำๆกัน การใช้น้ำหนักคำโดยคำนึงถึงความถี่คำ,ความถี่เอกสาร,การเหมือนกันของคำ และ ระยะห่างกลุ่มคำ จะใช้ในการกำหนดกลุ่มเอกสารได้ว่าจะอยู่ในกลุ่มใด ซึ่งจะเอาทั้งหมดสมมติฐานนี้มาใช้ในการพิจารณา

1.4 ทฤษฎีและแนวคิดในการวิจัย

ในการแยกเอกสารประโยคภาษาไทยด้วยระบบคอมพิวเตอร์นั้น จะต้องวิเคราะห์ประโยคและทำการแยกแยะหน่วยคำออกมาให้ได้ก่อน ซึ่งมีงานวิจัยไว้หลายวิธีแต่ที่น่าสนใจและกล่าวถึงในวิทยานิพนธ์ฉบับนี้มี 2 วิธี คือ

- วิธี Fast word matching
- วิธี Complete word matching

วิธีทั้งสองอันต่างก็มีข้อดีกันคนละแบบ ในวิทยานิพนธ์ฉบับนี้ได้ลองใช้วิธีการตัดคำโดยเอาแบบอย่างคล้ายกับวิธีทั้งสองมาประยุกต์กับการเทียบคำในพจนานุกรม (Matching) และเพิ่มให้ระบบสามารถตัดคำและเครื่องหมายที่ไม่มีในพจนานุกรมภาษาไทยออกไปได้ด้วย ซึ่งจะเน้นพิจารณาเฉพาะหน่วยคำที่มีในพจนานุกรมภาษาไทยเท่านั้น เพื่อประโยชน์ในการนำเอาหน่วยคำเหล่านั้นมาพิจารณาหาหน่วยคำสำคัญในการแยกกลุ่มเอกสารภาษาไทยต่อไปเป็นหลัก วิธีการตัดหน่วยคำแบบนี้จะพิจารณาการแยกหน่วยคำจากทางซ้ายมือไปทางขวามือของประโยค ซึ่งขอเรียกว่าวิธี Left search matching ผลจากการแยกหน่วยคำแล้วของทั้ง 3 วิธีในบางครั้งยังได้ผลลัพธ์

มากกว่า 1 ผลลัพธ์ ดังนั้นจึงต้องใช้วิธีทางไวยากรณ์ภาษาไทยเข้ามาช่วยพิจารณาอีกครั้งหนึ่ง เพื่อจะได้ผลลัพธ์ประโยคที่ถูกต้องและสามารถแยกหน่วยคำแล้วเหลือเพียง 1 ผลลัพธ์เท่านั้น ซึ่งก็มีผู้วิจัยไว้อยู่หลายวิธีแต่ที่เกี่ยวข้องกับวิทยานิพนธ์นี้มีน่าสนใจ คือ วิธี Transition Network ซึ่งเป็นการพิจารณาลำดับของหน่วยคำว่าเรียงกันอย่างไร ตามหลักไวยากรณ์ภาษาไทย

ในบางครั้งการหาผลลัพธ์ของการตัดคำที่ได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ให้เหลือผลลัพธ์เดียวจะใช้ตรวจสอบจากกฎไวยากรณ์ภาษาไทยเพียงอย่างเดียว ผู้วิจัยจึงได้นำเสนอวิธีที่เรียกว่า Bigram โดยใช้กฎไวยากรณ์ภาษาไทยและการหาค่าของความน่าจะเป็นร่วมกัน เพื่อเป็นการช่วยเพิ่มค่าน้ำหนักในโครงสร้างไวยากรณ์เพิ่มขึ้นอีกอย่างหนึ่ง ในการพิจารณาหาผลลัพธ์ของการตัดคำภาษาไทยที่มีผลลัพธ์มากกว่า 1 ผลลัพธ์ให้เหลือผลลัพธ์เดียว เพื่อให้เหมาะสมกับการใช้ประยุกต์กับการแยกกลุ่มเอกสารภาษาไทยได้ดีขึ้น

เมื่อได้หน่วยคำภาษาไทยที่แยกออกจากประโยคภาษาไทยได้แล้ว ก็จะต้องนำมาหาวิธีแยกเอาหน่วยคำที่เป็นหน่วยคำสำคัญจริงๆของเอกสารมาพิจารณา เพื่อจะใช้ในการแยกกลุ่มเอกสารต่อไป ในการพิจารณาจะคำนึงถึงคำสำคัญในเอกสารเป็นหลักว่าเอกสารกลุ่มเดียวกันน่าจะมีคำที่เหมือนกันมากกว่ากลุ่มอื่น แต่การแบ่งกลุ่มเอกสารนั้นจะใช้คำเหมือนกันหรือความถี่คำมาแบ่งก็ไม่น่าจะได้ผลดีนัก น่าจะพิจารณาในส่วนของจำนวนเอกสารกับคำในเอกสารบ้าง รวมทั้งเอกสารที่อยู่ในกลุ่มเดียวกันก็น่าจะมีระยะห่างของกลุ่มคำสำคัญใกล้ๆกัน แต่ระยะเท่าไรจึงจะเหมาะสมทำให้ได้สมาชิกในกลุ่มดีที่สุด การพิจารณาทั้งหมดที่กล่าวมา จะใช้เป็นตัวแบ่งกลุ่มเอกสารในการควบคุมและปรับค่าน้ำหนักในระบบโครงข่ายประสาทให้ได้ดี และจะใช้ค่า Precision , Recall และ F- measure เป็นตัววัดผลการจัดกลุ่ม ว่าวิธีไหนให้ค่าตัววัดทั้ง3สูงสุดเมื่อเทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์

1.5 ขั้นตอนของการศึกษา

การแยกกลุ่มเอกสารภาษาไทยจำเป็นต้องมีการเตรียมข้อมูลก่อน เนื่องจากประโยคภาษาไทยไม่สามารถแยกคำเป็นคำๆออกมาพิจารณาได้ง่ายๆ และไม่สามารถบอกการจบประโยคได้ง่ายๆเช่นกัน เนื่องจากลักษณะของประโยคภาษาไทยจะเขียนหน่วยคำติดต่อกันไป ทำให้คอมพิวเตอร์ต้องมีระบบการตรวจสอบเพิ่มขึ้น ก่อนจะนำเอาหน่วยคำสำคัญที่ได้มาใช้แยกเอกสาร ในวิทยานิพนธ์นี้จะมีขั้นตอนการทำงานดังนี้

1 การแยกหน่วยคำภาษาไทยออกจากประโยค โดยพิจารณาเฉพาะหน่วยคำที่มีในพจนานุกรมภาษาไทยเท่านั้น ไม่พิจารณาเครื่องหมายหรือหน่วยคำในภาษาอื่นซึ่งจะใช้วิธี Left search matching

2 การวิเคราะห์ประโยคจากผลลัพธ์ของวิธี Left search matching ที่ได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ ให้เหลือผลลัพธ์เดียวโดยใช้กฎไวยากรณ์ภาษาไทยแบบพิจารณา Transition Network ที่เรียกว่า วิธี Viterbi Algorithm ร่วมกับวิธีการหาความน่าจะเป็นแบบวิธี N-gram

3. เมื่อได้ผลลัพธ์จากการแยกหน่วยคำภาษาไทยแล้ว ก็จะพิจารณาคำสำคัญเหล่านั้นว่าจะจัดกลุ่มอย่างไรในฐานข้อมูลได้บ้าง ต้องทำการกรองคำที่ไม่น่าจะเป็นตัวบ่งบอกเนื้อหาในเอกสารออกก่อนเพื่อระบบจะได้ทำงานได้ดีมีประสิทธิภาพมากขึ้น การพิจารณาคำสำคัญเหล่านั้นว่าเหมือนกับเอกสารใดกลุ่มใด เริ่มจากทดลองหาความเหมือนของคำดูก่อนว่ามีผลอย่างไร และถ้าทำร่วมกับการพิจารณาระยะห่างของสมาชิกคำในเอกสารในรูปเวกเตอร์จะต้องหาค่าเทรสโฮลที่เหมาะสมมากน้อยเท่าไรที่จะทำให้ได้กลุ่มเอกสารเหมาะสม นอกจากนี้ได้ทดลองพิจารณาเพิ่มในส่วนของการให้น้ำหนักคำตามค่าความถี่คำ ในการจัดกลุ่มเอกสาร เปรียบเทียบกับการให้น้ำหนักคำตามจำนวนเอกสาร ด้วยว่าผลการจัดกลุ่มจะเป็นเช่นไร

4. เมื่อได้ทำการจัดกลุ่มเอกสารตาม สมมติฐานวิธีต่างๆแล้ว ได้ทำการให้น้ำหนักคำในแต่ละกลุ่มเอกสารด้วย เพื่อจะดูผลของน้ำหนักคำแต่ละคำในกลุ่มเอกสารของแต่ละวิธีว่ามีความแตกต่างกันอย่างไร เพราะค่าน้ำหนักคำเหล่านี้จะแปรผันกับจำนวนสมาชิกในกลุ่มเอกสารที่จัดได้ ซึ่งถ้าจัดเอกสารได้ดีสามารถนำน้ำหนักคำ นี้มาช่วยในการตัดคำที่ไม่ใช่ตัวบ่งบอกเนื้อหาในเอกสารทิ้งได้ ซึ่งคือคำที่มีน้ำหนักน้อยๆ ส่วนคำที่มีน้ำหนักมากๆยังบอกได้อีกด้วยว่าคำไหนคือคำสำคัญในเอกสารในกลุ่มนั้นๆด้วย หรือนำไปใช้ในการให้น้ำหนักคำในการจัดเอกสารอื่นต่อไปได้

5. ทำการเปรียบเทียบดูผลการทดสอบว่า การจัดกลุ่มเอกสารวิธีใดได้ผลลัพธ์ที่ดีที่สุดเปรียบเทียบกับการจัดกลุ่มด้วยมนุษย์ โดยใช้ค่า Precision, Recall และ ค่า F-measure ในการวัดว่าวิธีการจัดกลุ่มแบบไหนดีที่สุด วิธีที่ดีที่สุดควรจะต้องได้ค่าทั้ง3 เครื่องมือวัดสูงสุด และให้ผลการจัดกลุ่มเป็นอย่างไรบ้างในแต่ละแบบ โดยค่าทั้ง 3 เครื่องมือวัดนั้น แต่ละค่าจะบอกคุณลักษณะผลการจัดกลุ่มที่ได้ว่าเป็นอย่างไร ดีหรือไม่ดี

บทที่ 2
การทำพจนานุกรมและการแยกหน่วยคำ

ในการแยกแยะหน่วยคำที่เขียนเรียงติดกันอยู่ในประโยคภาษาไทยด้วยระบบคอมพิวเตอร์ จะต้องมีการกำหนดกระบวนการ หรืออัลกอริทึมของการค้นหาหน่วยคำในประโยคดังกล่าว แล้วสร้างเป็นซอฟต์แวร์และซอฟต์แวร์ที่สร้างขึ้นบนพื้นฐานของอัลกอริทึม จะทำหน้าที่ในการค้นหาและตัดสินใจแยกแยะหน่วยคำออกจากประโยคเอง โดยอาศัยฐานข้อมูลคำศัพท์ในพจนานุกรมภาษาไทยเป็นหลัก ซึ่งต้องสร้างขึ้นเพื่อใช้กับซอฟต์แวร์ของการแยกหน่วยคำนั้นๆ

2.1 โครงสร้างของพจนานุกรมไทย

การจัดเก็บข้อมูลสำหรับพจนานุกรมด้วยคอมพิวเตอร์ ต้องคำนึงถึงจุดมุ่งหมายของการใช้งานเป็นหลักเพื่อจัดรูปแบบโครงสร้างของข้อมูลให้เหมาะสม โครงสร้างของข้อมูลต้องทำให้เรียกค้นข้อมูลได้รวดเร็วถูกต้องแม่นยำ สามารถปรับปรุงแก้ไขข้อมูลได้ง่ายขยายเพิ่มเติมคำใหม่ได้ โครงสร้างข้อมูลที่ใช้ในการจัดเก็บด้วยคอมพิวเตอร์มีหลายรูปแบบ ดูเพิ่มเติมได้ใน[2] ดังเช่น

2.1.1 โครงสร้างการจัดเก็บข้อมูลแบบลำดับดัชนี

เป็นโครงสร้างในยุคแรกๆของการจัดเก็บพจนานุกรมคำศัพท์ ที่เหมาะสมกับจำนวนข้อมูลที่ไม่มากนัก โครงสร้างแบบนี้ง่ายต่อการจัดเก็บเพิ่มเติมแก้ไข โครงสร้างแบบลำดับดัชนี มีการค้นหาแบบลำดับฉะนั้นการค้นหาคำศัพท์ที่ต้องการจึงต้องทำการเปรียบเทียบกับคำศัพท์ในพจนานุกรมทุกตัว จึงจะรู้ว่าคำศัพท์ที่ต้องการนั้นมีในพจนานุกรมหรือไม่ คล้ายกับการเรียงคำศัพท์ในหนังสือพจนานุกรม ซึ่งจะเห็นได้ว่า โครงสร้างแบบนี้ไม่เหมาะสมสำหรับการเก็บข้อมูลที่มีจำนวนมากและต้องการค้นหาข้อมูลอย่างรวดเร็ว

2.1.2 โครงสร้างการจัดเก็บข้อมูลแบบตารางดัชนีและไบนารีทรี

(Table Index and binary tree)

แบ่งออกเป็น 2 ส่วน คือ ส่วนของข้อมูลและส่วนของตาราง

ส่วนของข้อมูล : ในการเก็บข้อมูลจะคล้ายกับต้นไม้ โดยเปรียบเทียบข้อมูลไหนมีค่ามากกว่ากัน ซึ่งในที่นี้ใช้รหัสแอสกีของข้อมูลเป็นตัวเปรียบเทียบ

ส่วนของตาราง : เป็นส่วนที่ใช้เก็บตำแหน่งเริ่มต้นของข้อมูล โดยการพิจารณาจากตัวอักษร 2 ตัวแรกของข้อมูลเป็นหลัก เพราะแทนที่จะเริ่มต้นที่ด้านบนสุดของต้นไม้ ก็สามารถที่จะข้ามไปยังตำแหน่งเริ่มต้นของข้อมูลที่ขึ้นต้นด้วยอักษรชุดนั้นได้เลย

2.1.3 โครงสร้างการจัดเก็บแบบดัชนี 3 ชั้น และเรียงลำดับ (3 Index and Sequential)

เป็นโครงสร้างพจนานุกรมที่พัฒนาขึ้นเพื่อการวิเคราะห์โครงสร้างประโยคภาษาไทย และการแยกคำของประโยคภาษาไทยโดยวิธี Fast word matching มีหลักการพื้นฐานคล้ายแบบลำดับดัชนี แต่จะมีการเพิ่มดัชนีชั้นที่ 1 ชั้นที่ 2 และชั้นที่ 3 เพื่อการเข้าถึงข้อมูลได้อย่างรวดเร็วมีหลักการ คือ ใช้ตัวอักขระ 2 ตัวแรกของคำศัพท์ที่ต้องการจัดเก็บ เป็นดัชนีตัวที่1 และ 2 ตามลำดับ และนำเอาจำนวนตัวอักขระของคำศัพท์นั้นเป็นดัชนีตัวที่ 3 ซึ่งจะได้กลุ่มของคำศัพท์ออกมา ภายในกลุ่มของคำศัพท์นี้จะเรียงกันแบบลำดับ

2.1.4 การจัดเก็บฐานข้อมูลพจนานุกรมคำศัพท์แบบเชิงสัมพันธ์ (Relational Database)

การเก็บข้อมูลแบบนี้ เป็นโครงสร้างที่ช่วยงานได้หลายอย่างเช่น ลดการซ้ำซ้อนของข้อมูล ความสามารถในการใช้ข้อมูลร่วมกันได้ หรือมีระบบรักษาความปลอดภัยให้ใช้งาน เป็นต้น ลักษณะการเก็บข้อมูลจะแสดงรายละเอียดของข้อมูลและความสัมพันธ์ระหว่างข้อมูล อยู่ในรูปของตารางซึ่งในแต่ละ ตารางจะประกอบด้วย คอลัมน์(Column) โดยชื่อคอลัมน์จะต้องมีชื่อไม่ซ้ำกัน และสามารถแสดงความสัมพันธ์ระหว่างข้อมูลอยู่ในรูปของตารางได้ โดยไม่มีตัวชี้มาเกี่ยวข้องในการแสดงความสัมพันธ์นี้ แต่สามารถมีตัวชี้ (index) มาเกี่ยวข้องได้เพื่อประโยชน์ในการเพิ่มความเร็วในการจัดการข้อมูลเท่านั้น

การออกแบบพจนานุกรมที่ใช้ในวิทยานิพนธ์ฉบับนี้ จะใช้ร่วมกันระหว่างโครงสร้างการจัดเก็บพจนานุกรมแบบดัชนี 3 ชั้นเรียงลำดับ กับแบบการจัดเก็บฐานข้อมูลพจนานุกรมคำศัพท์แบบเชิงสัมพันธ์ โดยคำศัพท์ในพจนานุกรมไทยทั้งหมดจะถูกจัดเก็บ ในโปรแกรมของ Microsoft Access เพื่อใช้อ้างอิงเป็นฐานข้อมูลของคำศัพท์ โดยมีการเรียงลำดับตามค่าอักษรเหมือนในคำศัพท์พจนานุกรมไทย นอกจากนี้ในตารางฐานข้อมูลคำศัพท์ ยังเพิ่มช่องอักขระ 2 ตัวแรกของคำศัพท์ไว้ด้วย เพื่อประโยชน์ในการสืบค้นข้อมูลได้เร็วขึ้นในกระบวนการของการตัดคำ เพราะในกระบวนการแยกคำหรือตัดคำ ด้วยวิธี Left search matching ซึ่งในวิทยานิพนธ์นี้ใช้เป็นหลัก จะทำการหาอักขระ 2 ตัวแรกของคำศัพท์ก่อนพอพบอักขระ 2 ตัวแรกของคำศัพท์แล้ว จึงหาคำศัพท์ในกลุ่มอักขระ2 ตัวแรกเดียวกันนี้ว่า มีคำศัพท์ใดตรงกับประโยคที่ต้องการแยกหน่วยคำบ้าง ก็จะทำการแยกหน่วยคำนั้นๆออกมาจากประโยคได้ ซึ่งวิธีนี้จะทำให้ระบบมีประสิทธิภาพกว่าการที่ต้องหาคำศัพท์ทุกคำในพจนานุกรม นอกเหนือจากนี้ยังทำ Backtrackingไว้ด้วย และยังเพิ่มประเภทของคำศัพท์เข้าไป เช่น คำนาม , สรรพนาม , กริยา ,วิเศษณ์ , บุพบท , สันธาน และอุทาน เป็นต้น เพื่อใช้ในการวิเคราะห์ประโยคโดยอาศัยไวยากรณ์ภาษาไทยเข้ามารวมในการพิจารณาด้วย รูปตารางฐานข้อมูลคำศัพท์ได้แสดงไว้ในภาคผนวก ก.

2.2 การแยกหน่วยคำออกจากประโยค

ปัญหาของระบบประมวลผลภาษาไทย อยู่ที่ลักษณะพื้นฐานของภาษาไทยโดยเฉพาะส่วนของหน่วยคำ เพราะมีการนำหน่วยคำมาเขียนติดต่อกันเป็นประโยคโดยไม่มีเครื่องหมายหรือช่องว่างบอกการสิ้นสุดของหน่วยคำแต่อย่างใด จึงมีผลทำให้การวิเคราะห์โครงสร้างของประโยคมีความยุ่งยากซับซ้อนขึ้น ได้มีผู้วิจัยค้นคว้าอัลกอริทึมสำหรับการแยกแยะหน่วยคำออกจากประโยคไว้หลายวิธีขึ้นอยู่กับวัตถุประสงค์ที่ต้องการนำไปใช้งาน โดยมีพจนานุกรมคำศัพท์ภาษาไทยที่เก็บบันทึกไว้ในคอมพิวเตอร์เป็นฐานข้อมูลในการเปรียบเทียบคำ ดังตัวอย่าง เช่น [2]

2.2.1 วิธีเปรียบเทียบจากพจนานุกรม

เป็นวิธีตัดแยกหน่วยคำภาษาไทยออกจากประโยค โดยใช้หลักการเปรียบเทียบกับพจนานุกรมที่เก็บบันทึกไว้ เช่น ข้อความตัวอย่าง " คนไทย "

ขั้นตอนการทำงานมีดังนี้

- ข้อมูลตัวแรกคือ " ค " ระบบจะมีตัวชี้และชี้ไปที่พจนานุกรมหมวด ค
- ข้อมูลตัวต่อมาคือ " น " ตัวชี้จะชี้ ณ. ตำแหน่งคำว่า " คน " ในพจนานุกรม
- ตัวที่3 คือ " ไ " ก็จะพบว่า " คนไ " ไม่มีในพจนานุกรม ทำให้ทราบจุดสิ้นสุดของคำได้ทันที และตัวชี้จะชี้ไปที่ " ไ " เป็นการเริ่มต้นคำใหม่

วิธีการเปรียบเทียบพจนานุกรมแบบนี้สามารถแยกหน่วยคำได้ ง่ายต่อการนำไปใช้ แต่ยังมีปัญหาต้องแก้ เช่น คำที่มีความหมายกำกวม จึงมีการเพิ่มหลักการของ Backtracking เพื่อช่วยให้คำที่ถูกแยกแยะถูกต้องมากที่สุด ดังข้อความตัวอย่าง " หมอกลัว "

ขั้นตอนการทำงานแบบมี Backtracking

- เมื่อเปรียบเทียบมาถึงคำว่า " หมอ " ซึ่งตรงกับคำในพจนานุกรม ระบบก็จะทำเครื่องหมายระบุว่า สามารถแยกเป็นหน่วยคำได้
- เปรียบเทียบต่อก็จะได้คำว่า " หมอก " ซึ่งก็มีในพจนานุกรมเช่นกัน ระบบก็จะทำเครื่องหมายว่าสามารถแยกเป็นหน่วยคำได้และเป็นคำที่ยาวกว่า
- เปรียบเทียบต่อเป็นคำว่า " หมอกล " ซึ่งไม่มีในพจนานุกรม ระบบถือว่าคำที่แยกได้เบื้องต้นคือคำว่า " หมอก "
- เมื่อเปรียบเทียบคำที่เหลือ คือ " ลัว " จะไม่พบในพจนานุกรม ระบบจะย้อนหลังโดยถือเป็นคำแยกได้คำแรก คือ " หมอ " แล้วจึงเปรียบเทียบคำว่า " กลัว " ใหม่อีกครั้ง

วิธีการเช่นนี้ทำให้สามารถแยกคำที่กำกวมได้ แต่ก็มีปัญหาอีกหลายอย่างที่วิธีการเปรียบเทียบพจนานุกรมยังให้คำตอบที่ชัดเจนไม่ได้

2.2.2 วิธี Longest Matching

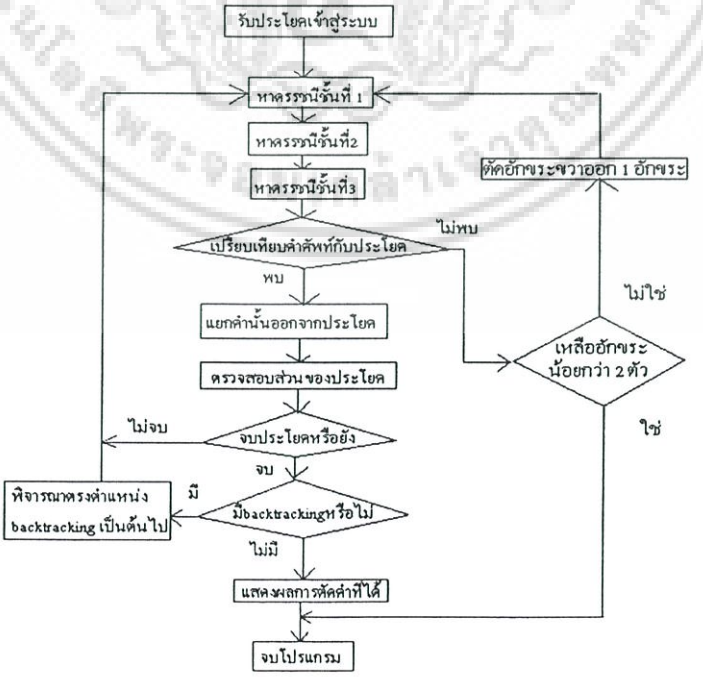
ซึ่งวิธีนี้เป็นการเลือกคำที่ยาวที่สุดไปเปรียบเทียบก่อน ถ้าไม่พบในพจนานุกรมระบบจะทำการลดความยาวของคำลงทีละอักขระ แล้วทำการเปรียบเทียบกับพจนานุกรมใหม่จนได้หน่วยคำออกมา เช่น ข้อความว่า " ความก้าวหน้าทางด้านวิทยาศาสตร์มีความสำคัญ " ผลของการแยกหน่วยคำทั้งหมดเป็นดังนี้

ตารางที่ 2.1 ผลของการแยกแยะหน่วยคำโดยวิธี Longest Matching

ส่วนของคำที่ยาวที่สุด	ส่วนที่เหลือ
ความก้าวหน้า	ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ
ทาง	ด้านวิทยาศาสตร์มีบทบาทสำคัญ
ด้าน	วิทยาศาสตร์มีบทบาทสำคัญ
วิทยาศาสตร์	มีบทบาทสำคัญ
มี	บทบาทสำคัญ
บทบาท	สำคัญ
สำคัญ	

2.2.3 วิธีการแยกหน่วยคำแบบ Fast word matching

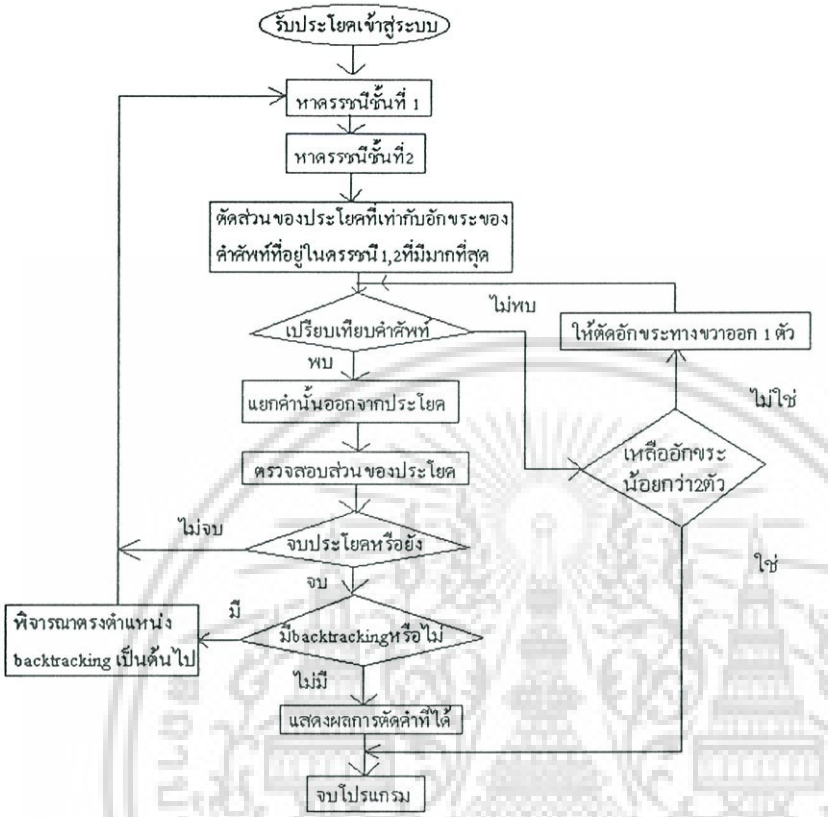
ซึ่งวิธีนี้ได้ถูกนำเสนอในงานวิจัยของคุณ สมศักดิ์ จันวัน [1]



รูปที่ 2.1 แผนผังการทำงานของระบบ Fast word matching

2.2.4 วิธีการแยกหน่วยคำแบบ Complete word matching

เป็นงานวิจัยคุณ สิงห์ ตรงงาม [2]



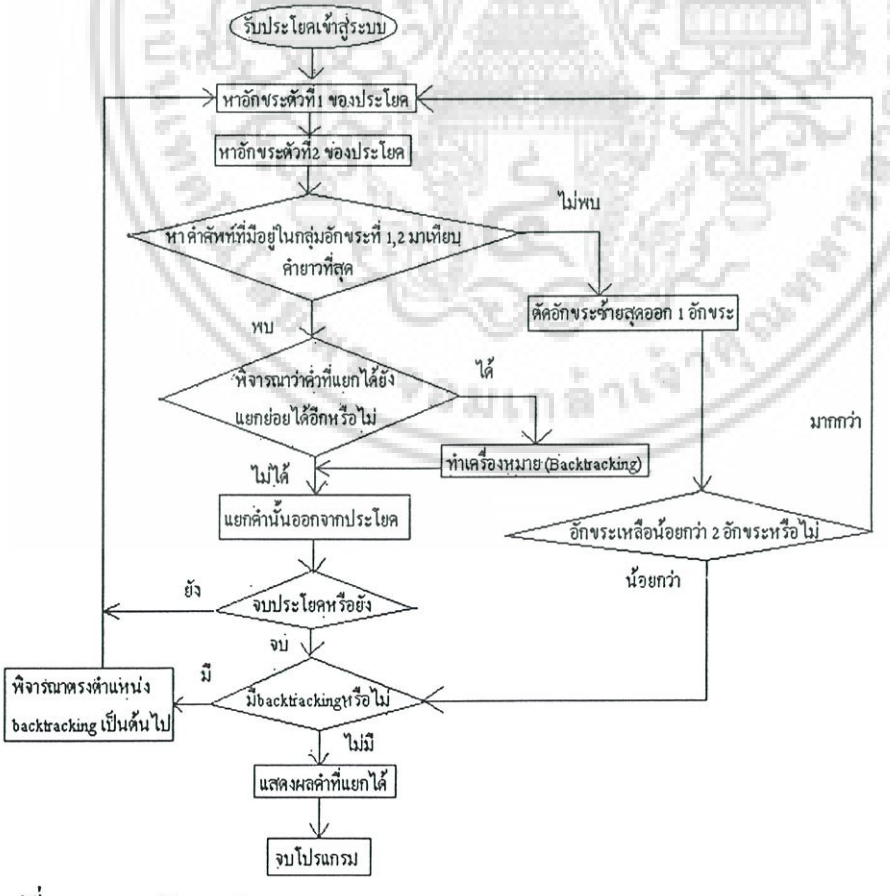
รูปที่ 2.2 แผนผังการทำงานของระบบ Complete word matching

2.2.5 การแยกหน่วยคำด้วยวิธี Left search matching

จากการแยกแยะหน่วยคำดังที่กล่าวมาแล้วทั้งหมด ต่างก็มีข้อดีข้อเสียของแต่ละวิธีต่างๆ กันไป แต่ทั้งหมดก็เป็นพื้นฐานหรือแนวทางในการพัฒนาวิธีอื่นๆต่อไปเพื่อให้มีประสิทธิภาพดีขึ้น ในวิทยานิพนธ์นี้ก็เช่นกันได้นำวิธี Left search matching มาใช้ในการแยกแยะหน่วยคำภาษาไทยออกจากประโยคในอีกวิธีหนึ่ง โดยประยุกต์เอาข้อดีของทั้ง Fast word matching และ Complete word matching มาใช้

วิธี Left search matching นี้จะใช้ค่าดัชนีของ อักขระตัวที่ 1 และ 2 ของคำที่อยู่ซ้ายมือสุดในประโยคที่ต้องการค้นหา มาเปรียบเทียบกับคำศัพท์ที่อยู่ในพจนานุกรมในกลุ่มอักขระที่ 1 และ 2 เดียวกัน ถ้าพบคำศัพท์ก็ให้พิจารณาคำศัพท์ ในกลุ่มเดียวกันนี้อีกว่า คำที่พบยังสามารถแยกเป็นคำอื่นได้อีกหรือไม่ ถ้าทำได้ก็ให้ทำเครื่องหมายไว้ว่าสามารถแยกคำออกไปได้อีกที่เรียกว่าการทำ (Backtracking) แล้วตัดเอาหน่วยคำที่ยาวที่สุดออกมาก่อน ในกรณีที่ไม่พบคำศัพท์จะตัดอักขระซ้ายมือสุดออก1อักขระ แล้วหาอักขระตัวที่ 1 และ 2 ของคำที่อยู่ซ้ายมือสุดในประโยคที่ต้องการค้นหาใหม่ ต่อจากนั้นก็หาคำที่เหลืออีกในประโยคจนแยกคำออกมาจากประโยคได้ทั้งหมด และถ้า

คำไหนแยกคำออกไปได้อีกก็ให้ทำเครื่องหมายไว้เช่นกันทุกหน่วยคำ ส่วนคำที่แยกเป็นคำอื่นอีกไม่ได้ก็แยกออกจากประโยคได้เลยโดยไม่ต้องมีเครื่องหมายใดๆ เมื่อแยกหน่วยคำได้จนหมดประโยคแล้วก็กลับมาพิจารณาคำที่มีเครื่องหมายอยู่อีกครั้ง โดยพิจารณาคำที่มีเครื่องหมายทางขวามือสุดของหน่วยคำที่แยกออกจากประโยคว่า มีการแบ่งเป็นคำอื่นอะไรได้อีกที่ไม่ซ้ำกับคำเดิมที่มีเครื่องหมาย แล้วทำการแยกหน่วยคำประโยคใหม่ ถ้าพบก็ให้หาคำย่อยอื่นลงไปอีกว่ามีหรือไม่ ถ้าไม่พบก็เอาเครื่องหมายออกแล้วแยกคำใหม่ออกมา แต่ถ้าพบก็แยกคำย่อยอันนั้นออกมาแล้วใส่เครื่องหมายอีกครั้ง แล้วหาหน่วยคำอื่นถัดไปของประโยคทำเช่นนี้จนแยกหน่วยคำได้จบประโยคอีกครั้ง จึงนำประโยคที่แยกหน่วยคำใหม่นี้เปรียบเทียบกับประโยคที่แยกหน่วยคำอันอื่นที่ผ่านมาแล้ว ว่าแยกหน่วยคำได้ซ้ำกันไหม ถ้าไม่ซ้ำกันก็แสดงผลการแยกหน่วยคำที่ได้ แต่ถ้าซ้ำกันก็ไม่แสดงผล ให้กลับไปหาคำที่มีเครื่องหมายคำต่อไป แล้วทำการแยกหน่วยคำต่อเช่นเดิมดังที่กล่าวมาทำจนแยกหน่วยคำในประโยคได้หมดแล้วและไม่มีเครื่องหมายที่คำใดๆอีกก็เป็นอันเสร็จกระบวนการแยกหน่วยคำแบบ Left search matching การตัดอักขระที่ไม่พบคำศัพท์ของวิธีนี้จะต่างจาก 2 วิธีแรกคือจะตัดอักขระทางซ้ายสุดแทนทางขวามือ เพื่อจะป้องกันคำที่เขียนผิด, คำภาษาอังกฤษและเครื่องหมายต่างๆที่อยู่ข้างหน้าประโยค ระบบการตัดคำจะได้ทำงานได้ไม่เกิด error เพราะหาคำศัพท์ไม่พบ



รูปที่ 2.3 แผนผังการทำงานของระบบ Left search matching

ขั้นตอนการทำงานแบบ Left search matching จากประโยคตัวอย่าง " หลานมารอกราบป้า "

1. หาอักขระ 2 ตัว แรกซ้ายมือ คือ หล
2. จะตัดคำออกมาได้คือ หลาน และ หลา
3. ทำเครื่องหมายคำว่า (/!หลาน*/) เพราะตัดคำได้มากกว่า 1คำ
4. แยกคำว่า (/!หลาน*/) ออกจากประโยค
5. จะได้ประโยคที่เหลือคือ " มารอกราบป้า "
6. หาอักขระ2ตัวแรก ซ้ายมือของคำถัดไปมาพิจารณาจะได้คำว่า " มา "
7. ตัดคำออกมาได้คือ " มาร " และ " มา "
8. คำที่แยกได้จากข้อ 1- 7 คือ " /!หลาน*/!มาร*/ "
9. จะได้ประโยคที่เหลือคือ " อกราบป้า "
10. หาอักขระ 2 ตัวแรก ซ้ายมือของคำถัดไปมาพิจารณาจะได้คำว่า " อก "
11. ตัดคำออกมาได้ คือ " อก "
12. คำที่แยกได้จาก 1- 11 คือ " /!หลาน*/!มาร*/อก/ "
13. จะได้ประโยคที่เหลือคือ " ราบป้า "
14. ทำดังข้อ 1- 13 ไปจนจบประโยค
15. เมื่อจบประโยคให้พิจารณาคำตรงตำแหน่งที่มีเครื่องหมาย(!*)ที่อยู่ขวามือสุดใหม่ โดยแทนคำที่ไม่ซ้ำกับคำที่มีเครื่องหมายแต่อยู่ในกลุ่มอักขระ2ตัวแรกเดียวกัน
16. แล้วแยกเป็นคำออกมาจากประโยค เหมือน 1- 14 ใหม่อีกครั้ง
17. ทำจนหน่วยคำที่แยกออกมาได้ไม่มีเครื่องหมายและแยกได้หมดทั้งประโยค

ดังนั้นจะได้ผลลัพธ์ประโยคการตัดคำทั้งหมด คือ

/ หลาน / มาร / อก / ราบ / ป้า /

/ หลาน / มา / รอก / ราบ / ป้า /

*** / หลาน / มา / รอ / กราบ / ป้า / *** (ประโยคที่ถูกต้อง)

สรุปผลจากการประมวลผลการแยกแยะหน่วยคำ เมื่อเทียบกับวิธี Fast word matching และ Complete word matching แล้ววิธี Left search matching จะมีประสิทธิภาพที่ดีสำหรับการนำมาใช้แยกกลุ่มเอกสารภาษาไทยเพราะ เป็นการเทียบหน่วยคำโดยตรงกับพจนานุกรมไม่ได้เทียบกับวลีที่ยาวหรือประโยคทั้งประโยค และการตัดอักขระทางซ้ายมือสุดที่ละ 1 อักขระยังช่วยในการแก้ไขกรณีหาคำศัพท์หรือเครื่องหมายที่ไม่พบในพจนานุกรมได้ ส่วนด้านการแยกหน่วยคำจะมีความถูกต้องได้ใกล้เคียงกัน ผลการเปรียบเทียบวิธีทั้ง 3 ดูได้จากภาคผนวก ก. แต่ถึงอย่างไรก็ตามผลการแยกหน่วยคำของทั้ง 3 วิธี บางครั้งก็ได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ จึงต้องทำการหาวิธีที่จะทำให้แยกหน่วยคำได้เพียงผลลัพธ์เดียวเท่านั้น

บทที่ 3
การวิเคราะห์โครงสร้างประโยคภาษาไทย

ในการวิเคราะห์โครงสร้างประโยคภาษาไทยด้วยระบบคอมพิวเตอร์ของงานวิจัยนี้ ในขั้นตอนแรกเราจะต้องทำการประมวลผลเพื่อแยกแยะหน่วยคำ (Word Parsing)จากประโยคออกเป็นหน่วยคำ(Morpheme) ที่มีความหมายตามพจนานุกรมเสียก่อน ถ้าได้ผลลัพธ์ของการแยกหน่วยคำมากกว่าหนึ่งผลลัพธ์แล้ว จะต้องทำให้เหลือผลลัพธ์เดียวก่อนจะไปทำการประยุกต์ใช้งานอื่นได้ การจะหาผลลัพธ์ของการตัดคำให้ได้ผลลัพธ์เดียวนั้น จะต้องใช้หลักไวยากรณ์ภาษาไทยมาช่วยวิเคราะห์การแยกหน่วยคำ

การประยุกต์ประโยคภาษาไทยเพื่อการใช้งานอื่นๆนั้น จนถึงการใช้แยกกลุ่มเอกสารด้วยนั้น ขั้นแรกจะต้องมีการแยกหน่วยคำในประโยคก่อน เพราะเนื่องจากประโยคภาษาไทยมีลักษณะการเขียนคำที่เรียงติดกันและไม่ทราบจุดสิ้นสุดของประโยคได้ง่ายนักเหมือนอย่างภาษาอังกฤษที่แยกหน่วยคำชัดเจน นอกจากนี้ภาษาและสำนวนของภาษาไทยก็มีความซับซ้อนมาก ดังนั้นการแยกหน่วยคำนั้นจะต้องแยกเฉพาะคำที่เป็นคำศัพท์ในภาษาไทย นั่นคือเป็นคำที่มีความหมายและบรรจุอยู่ในพจนานุกรมไทย นอกจากนั้นการแยกหน่วยคำด้วยระบบซอฟต์แวร์คอมพิวเตอร์จะต้องได้ผลลัพธ์รวดเร็วและถูกต้อง ซึ่งในวิทยานิพนธ์นี้ได้เสนอผลงานวิจัยเกี่ยวกับระบบแยกแยะหน่วยคำไทยด้วยวิธี Left search matching ดังที่กล่าวมาแล้ว ในบทที่ 2

แต่เมื่อทำการแยกแยะหน่วยคำภาษาไทยจากประโยคต้นแบบ (Source Language) นั้นในบางครั้งสำหรับบางประโยค อาจจะให้ผลลัพธ์ของการแยกแยะหน่วยคำได้หลายแบบ ซึ่งหน่วยคำในแต่ละแบบที่แยกได้อาจไม่ถูกต้องในด้านความหมายและโครงสร้างของประโยค ฉะนั้นจึงจำเป็นต้องมีกระบวนการตัดสินผลของการแยกหน่วยคำที่ถูกต้องของความสัมพันธ์ของหน่วยคำที่แยกได้ กระบวนการนี้เป็นศาสตร์หนึ่งของการวิเคราะห์โครงสร้างภาษาไทยนั่นเอง

ในการวิเคราะห์โครงสร้างของประโยคภาษาไทยนั้น สิ่งที่ต้องใช้เป็นฐานความรู้คือ หลักไวยากรณ์ของการใช้หน่วยคำในภาษาไทยเกี่ยวกับการจัดแบ่งประเภทของกลุ่มคำความสัมพันธ์ทางโครงสร้างของคำ ตลอดจนการเรียงลำดับของคำในประโยคนั้น สิ่งเหล่านี้เป็นความรู้พื้นฐานทางด้านภาษาศาสตร์

3.1 ชนิดของคำในประโยคภาษาไทย

ประโยค ถือเป็นถ้อยคำที่มีความครบบริบูรณ์ ประโยคหนึ่งๆแบ่งออกเป็น 2 ภาค คือภาคประธานและภาคแสดง ประโยคจะประกอบด้วยวลีหลายๆวลีมารวมกัน เช่น นามวลี ,กริยาวลี , บุพบทวลี เป็นต้น

วลี คือ กลุ่มคำที่มีคำติดต่อกันตั้งแต่ 2 คำขึ้นไป ซึ่งมีความหมายติดต่อกันเป็นเรื่องเดียว แต่เป็นเพียงส่วนหนึ่งของประโยค และไม่มีเนื้อความครบถ้วนเป็นประโยค

คำ คือ อักษรที่ประสมกันแล้วมีความหมาย ซึ่งแบ่งออกเป็น 8 ชนิดได้แก่

1). คำนาม (Noun)

2). คำสรรพนาม (Pronoun)

3). คำกริยา (Verb)

4). คำวิเศษณ์ (Adverb)

5). คำคุณศัพท์ (Adjective)

6). คำบุพบท (Preposition)

7). คำสันธาน (Conjunction)

8). คำอุทาน (Interjection)

คำวิเศษณ์และคำคุณศัพท์ มักจะมีความหมายใกล้เคียงกันคือ คำที่ประกอบคำอื่นให้มีความหมายต่างออกไป นอกจากนี้ในพจนานุกรมคำศัพท์ภาษาไทยก็มักจะรวมคำวิเศษณ์กับคำคุณศัพท์เป็นคำเดียวกัน ไม่แยกชนิดเป็นประเภทออกมาชัดเจน ในวิทยานิพนธ์ฉบับนี้จึงขออธิบายรวมกันไป หรือถือว่าเป็นคำเดียวกัน แล้วแต่ว่ามันจะใช้ไปขยายคำใดถ้าขยายพวกคำกริยาก็มักจะเป็นคำวิเศษณ์ ถ้าขยายคำนามหรือคำอื่นๆก็ถือว่าเป็น คำคุณศัพท์ รายละเอียดเพิ่มเติมของคำทั้ง 8 ประเภท หาได้จากเอกสารอ้างอิง [1]

ประโยคภาษาไทยมีการกำหนดตำแหน่งของคำที่ปรากฏในประโยคพื้นฐาน (Simple Sentence) ทั่วๆ ไป ซึ่งมักประกอบด้วยส่วนของนามวลี (Noun Phrase : NP) กับกริยาวลี (Verb Phrase : VP) ซึ่งบางส่วนอาจถูกละไว้ เช่น วิเศษณ์วลี (Adverbial Phrase : ADVP)

ในประโยคธรรมดาทั่วๆ ไปอาจจะประกอบด้วยนามวลีเพียงอย่างเดียว หรือกริยาวลีเพียงอย่างเดียวก็ได้ ซึ่งเราก็จะเรียกเป็นประโยคเหมือนกัน แต่โดยทั่วไปแล้วประโยคภาษาไทยจะมีโครงสร้างพื้นฐานที่ประกอบเรียงตามลำดับได้ดังนี้

Sentence ----------→ (ADVP) NP (ADVP) + VP (ADVP)

ซึ่งแบ่งย่อยๆได้ 3 ประเภทดังที่กล่าวมา คือ

- นามวลี
- กริยาวลี
- วิเศษณ์วลี

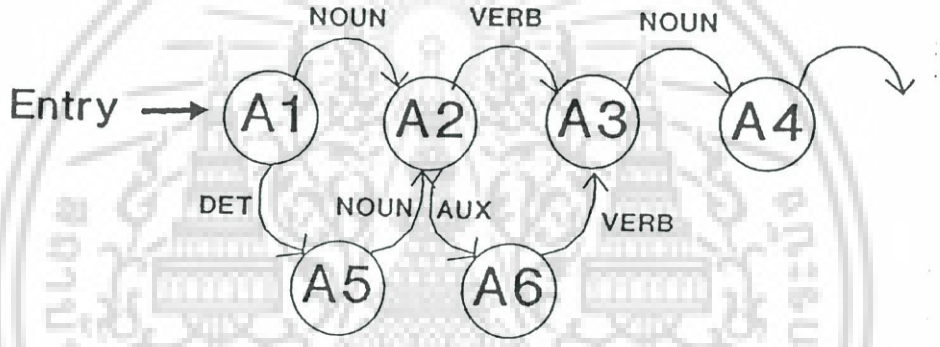
การประมวลผลโครงสร้างประโยคของภาษาไทยหรือการวิเคราะห์ภาษาไทย ทางด้านการประมวลผลภาษาธรรมชาตินั้นมีหลายวิธี ในวิทยานิพนธ์ฉบับนี้ได้นำเสนอวิธีการที่เรียกว่า Bigram วิธีนี้เป็นการใช้ทฤษฎีกราฟที่เกี่ยวข้องกับโนด(node) และ อาร์ค(Arc) คล้ายของวิธี ATN และ M-ATN ที่เป็นแบบระบบโครงข่าย ดูเพิ่มเติมได้ในเอกสารที่ [1] โดยนำคุณสมบัติทางโครงสร้าง

วลีมาช่วยร่วมกับทฤษฎีคณิตศาสตร์ของความน่าจะเป็น ซึ่งนำมาช่วยเพิ่มน้ำหนักโนดต่างๆในวิธี ATN และ M-ATN เพื่อประโยชน์ในการแก้ความคลุมเครือของการตัดคำแบบ Left search matching และนำเอาคำที่แยกได้ไปใช้ในการจัดหมวดหมู่เอกสารต่อไป

3.2 การทำงานของระบบประมวลผลแบบโครงข่าย หรือ แบบกราฟ

3.2.1 หลักการพื้นฐานของระบบ Transition Network : TN

ระบบ TN จะประกอบด้วยกลุ่มของโนดที่ถูกเชื่อมด้วยกลุ่มของอาร์ค ซึ่งเชื่อมต่อตามเงื่อนไขความสัมพันธ์ทางไวยากรณ์ เช่น คำนาม , คุณศัพท์ ,กริยา,บุพบท เป็นต้น โดยจะมีการจัดเรียงอาร์คตามหลักไวยากรณ์ภาษา ซึ่งโครงสร้างภาษามีดังรูปเป็นต้น



รูปที่ 3.1 ตัวอย่างความสัมพันธ์ทางโครงสร้างบางส่วนของระบบ TN

3.2.2 หลักการทำงานพื้นฐานของระบบ Augmented Transition Network : ATN

ATN มีพื้นฐานมาจาก TN โดยจะแสดงโครงสร้างประโยคของภาษาใดๆในรูปของโครงข่ายได้โดยการจัดกลุ่มของอาร์ค และโนดมาประกอบกันเป็นรูปของโครงข่าย โดยที่โนดจะแสดงหน่วยคำและอาร์คจะแสดงความสัมพันธ์ระหว่างหน่วยคำที่เชื่อมติดกัน ความสัมพันธ์ระหว่างหน่วยคำจะเป็นไปตามไวยากรณ์ทางด้านชนิดของคำเมื่อปรากฏในประโยค ข้อมูลทางภาษานี้จะเป็นตัวกำหนดการเชื่อมโยงโนดด้วยอาร์ค และการกำหนดความสัมพันธ์หรือเงื่อนไขของอาร์คนั้นๆด้วย

3.2.3 หลักการทำงานพื้นฐานของระบบ (Modified Augmented Transition Network): M-ATN

ระบบ M-ATN ของงานวิจัยนี้ได้พัฒนาขึ้นมาจากระบบ ATN โดยอาศัยโนดและอาร์ค โดยที่โนดแต่ละโนดจะต้องไม่มีชื่อซ้ำกันและจะถูกเชื่อมต่อกันด้วยอาร์ค หรือกลุ่มของอาร์คกลายเป็นโครงข่าย โดยที่อาร์คแต่ละอาร์คจะมีเงื่อนไขกำหนด

จากวิธีที่กล่าวมาทั้งหมดของระบบโครงข่าย แบบโนด และ กราฟ ก็ยังมีข้อที่ยังต้องแก้ไขอยู่บ้าง เพราะไวยากรณ์ภาษาไทยมีความสลับซับซ้อนมาก การจะแก้ไขเพิ่มเติมเปลี่ยนแปลงโนดและอาร์คค่อนข้างยุ่งยากทำได้ลำบาก ในวิทยานิพนธ์นี้จึงมีแนวคิดที่จะใช้วิธี Bigram ร่วมกับคณิตศาสตร์ความน่าจะเป็นมาใช้ในการเพิ่มน้ำหนักโนดและอาร์ค เพราะมีแนวคิดที่ว่าถ้าใส่ลำดับประเภทของหน่วยคำที่ถูกต้องทางไวยากรณ์ให้กับฐานข้อมูลไวยากรณ์อย่างถูกต้องแล้ว การพิจารณาประโยคที่ผ่านการตัดคำแล้วก็น่าจะมีความถูกต้องทางหลักภาษาไทยด้วย และถ้าจะมีการแก้ไขเพิ่มเติมหรือเปลี่ยนแปลงฐานข้อมูลก็น่าจะทำได้ง่ายกว่า ฉะนั้นในวิทยานิพนธ์ฉบับนี้จึงได้นำหลักการวิธี Bigram มาใช้ร่วมกันกับคณิตศาสตร์ความน่าจะเป็น เพื่อตรวจสอบประโยคผลลัพธ์ของการตัดคำกับไวยากรณ์ภาษาไทยว่าจะมีผลเช่นไร เพื่อจะนำประโยคผลลัพธ์ที่ได้นั้นไปใช้ในการแยกกลุ่มเอกสารภาษาไทยต่อไป

3.3 ฐานความรู้ที่ใช้ร่วมกับ วิธี Bigram

การทำขบวนการ Bigram เป็นขบวนการที่ต้องอาศัยหลักไวยากรณ์ภาษาไทยเก็บเป็นฐานความรู้เช่นกันเหมือนในระบบ ATN และ M-ATN ซึ่งจะอยู่ในรูปของการทำงานของโนด และอาร์ค ร่วมกับคณิตศาสตร์ความน่าจะเป็นในการพิจารณาผลลัพธ์ของการตัดคำแบบ Left search matching ที่ได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ให้เหลือผลลัพธ์เดียว ที่ถูกต้องตามหลักไวยากรณ์ภาษาไทย ฉะนั้นฐานความรู้ในหลักภาษาไทยจึงเป็นเรื่องสำคัญเพื่อใช้อ้างอิงพิจารณา ในการวิเคราะห์ประโยคภาษาไทยด้วยวิธี ATN และ M-ATN จะเก็บรวบรวมฐานข้อมูลซึ่งทำการเปลี่ยนแปลงแก้ไขเพิ่มเติมได้ยาก เพราะเก็บในตัวของ Source Program และนอกจากนี้รูปแบบของไวยากรณ์และหลักภาษาไทยยังมีความยากซับซ้อนมาก จึงยากที่จะจัดเก็บได้หมดทุกกรณี แต่ในวิธีการเก็บฐานความรู้ในแบบที่ป้อนรูปประโยคที่มีลำดับทางประเภทของคำที่เรียงกันอย่างถูกต้องทางหลักภาษา แล้วใช้วิธีจัดเก็บแบบคู่ลำดับก่อนหลังที่มีคณิตศาสตร์ความน่าจะเป็นเข้ามาช่วยจะทำให้ได้ผลที่ดีกว่า ซึ่งถ้าสามารถใส่ลำดับประเภทของคำตามหลักไวยากรณ์ได้หมดทุกรูปแบบประโยคผลลัพธ์ที่พิจารณาจากการตัดคำแบบ Left search matching ก็น่าจะถูกต้องตามหลักภาษาด้วย แต่ก็เป็นการทำที่ค่อนข้างยาก ถึงอย่างไรก็ยังเป็นวิธีที่ดีเพราะเราสามารถเพิ่มเติมเปลี่ยนแปลงแก้ไขลำดับประเภทของหน่วยคำได้ง่ายกว่า ใช้การจัดเก็บแบบวิธีของ ATN และ M-ATN

ซึ่งวิธีการจัดเก็บแบบคู่ลำดับประเภทของหน่วนคำก่อนหลังนี้และมีคณิตศาสตร์ความน่าจะเป็นร่วมด้วย จะนำมาใช้เป็นฐานข้อมูลการวิเคราะห์ประโยคแบบวิธี Bigram ประเภทของคำ เช่น

- คำนาม แทนด้วย N
- คำสรรพนาม แทนด้วย Pron
- คำกริยา แทนด้วย V
- คำวิเศษณ์ แทนด้วย Adv
- คำคุณศัพท์ แทนด้วย Adj
- คำบุพบท แทนด้วย Prep
- คำสันธาน แทนด้วย Conj
- คำอุทาน แทนด้วย Interj

การสร้างฐานความรู้ที่ใช้กับวิธี Bigram จะเป็นการเก็บข้อมูลจากรูปประโยคภาษาไทย ที่ผ่านการพิจารณาลำดับประเภทของหน่วยคำที่ถูกต้องตามหลักภาษาไทยแล้วเข้ามาเก็บไว้ โดยพิจารณาลำดับหน่วยคำของประโยคที่ป้อนเข้ามาเก็บในฐานข้อมูลทีละ 2 คำ จากซ้ายมือไปทางขวามือของประโยคฐานข้อมูลแล้วใช้สูตรสมการของความน่าจะเป็นเก็บค่าประเภทของหน่วยคำที่ป้อนเข้ามาเก็บไว้ในฐานข้อมูล สำหรับสูตรความน่าจะเป็นที่ใช้ในกรณีของการเก็บข้อมูลแบบ Bigram [8] คือ

$$Prob(X \mid Y) \cong \frac{Count\ (Y\ at\ position\ i-1\ and\ X\ at\ i)}{Count\ (Y\ at\ position\ i-1)} \qquad (3.1)$$

Prob(x|y) = ความน่าจะเป็นที่ y ตามด้วย x

Count (y at position i-1 and x at I) = จำนวนครั้งที่พบ y แล้วตามด้วย x

Count (y at position i-1) = จำนวนครั้งที่พบ y

การป้อนประโยคที่ถูกต้องและเก็บผลแบบคู่ลำดับก่อนหลัง ร่วมกับสมการความน่าจะเป็น (3.1)ไว้ในฐานข้อมูล ซึ่งกระบวนการทั้งหมดนี้จะทำงานด้วยโปรแกรมเรียนรู้และจดจำ ต่อจากนั้นจะนำผลของตารางฐานข้อมูลนี้ มาช่วยในการพิจารณาประโยคผลลัพธ์ที่ผ่านการแยกหน่วยคำด้วยวิธี Left search matching แล้ว และได้ผลลัพธ์มากกว่า 1 ผลลัพธ์มาพิจารณาแบบ Bigram (ตารางที่ใช้ช่วยพิจารณาวิธีของ Bigram ดูได้จากตารางที่ 3.1) ซึ่งจะทำให้ได้ผลลัพธ์ที่ถูกต้องตามหลักภาษาไทย ตัวอย่างประโยคฐานความรู้ที่ถูกต้องทางหลักภาษาไทยที่ป้อนเก็บไว้ในฐานข้อมูล ซึ่งจะใช้ในการหา ผลลัพธ์ของ Unknown ที่ได้จากการตัดคำแบบ Left search matching เช่น

- เปิด / ไฟ / ที่ / หน้าบ้าน / ด้วย = V / N / Prep / N / Prep
- เขา / ดี / มี / วาจา / ไพเราะ = Pron / Adj / V / N / Adj
- ปากกา / ของ / ฉัน / อยู่ / ที่ / เขา = N / Prep / Pron / V / Prep / Pron
- ต้นไม้ / โค่น / เพราะ / โดน / คน / ตัด = N / V / Conj / N / N / V
- ไฟฟ้า / หน้าบ้าน / และ / หลังบ้าน / เปิด = N / N / Conj / N / V
- นก / บิน / และ / คาบ / อาหาร = N / V / Conj / V / N
- คน / ที่ / เป็น / ครู / ต้อง / มี / ความอดทน = N / Pron / V / N / Adj / V / N

ขั้นตอนของการทำงานโปรแกรมเรียนรู้และจดจำ จากประโยคตัวอย่าง " เปิด / ไฟ / ที่ / หน้าบ้าน / ด้วย " มีดังนี้

1. เพิ่ม @ ไปที่หน้าประโยคที่ถูกแยกหน่วยคำไว้แล้วจากวิธี Left search matching และกำหนดประเภทของคำตามลำดับที่ถูกต้องตามหลักภาษาไทยแล้ว ดังนี้
   " เปิด / ไฟ / ที่ / หน้าบ้าน / ด้วย " = " @ / V / N / Prep / N / Prep "
2. พิจารณาคำจากทางซ้ายมือไปทางขวามือ โดยพิจารณาทีละ 2 คำ ตามลำดับดังนี้
   @ ⟶ V, V ⟶ N,N ⟶ Prep , Prep ⟶ N,N ⟶ Prep
3. เก็บค่าความน่าจะเป็น(Prob) ทีละคู่ ตามสมการ (3.1) แล้วนำไปเก็บในตารางฐานข้อมูล(ตารางที่3.1) เช่น

$$Prob(V \mid @) = \frac{Count(@V)}{Count\ @} = \frac{Count\_ii\ \dot{\pi}}{Count\_i\ \dot{\pi}} \frac{@\dot{V}}{@\dot{}} = \frac{3}{50} = 0.06$$

$$Prob(N \mid V) = \frac{Count(VN)}{Count\ V} = \frac{Count\_ii\ \dot{\pi}}{Count\_i\ \dot{\pi}} \frac{\dot{V}N}{\dot{V}} = \frac{27}{62} = 0.44$$

$$Prob(Prep \mid N) = \frac{Count(NPrep)}{CountN} = \frac{Count\_ii\ \dot{\pi}}{Count\_i\ \dot{\pi}} \frac{N\dot{}Prep}{N\dot{}} = \frac{9}{55} = 0.16$$

$$Prob(N \mid Prep) = \frac{Count(PrepN)}{Count\ Prep} = \frac{Count\_ii\ \dot{\pi}}{Count\_i\ \dot{\pi}} \frac{Prep\dot{N}o}{Prep\dot{}\ 1} = \frac{14}{22} = 0.63$$

$$Prob(Prep \mid N) = \frac{Count(NPrep)}{CountN} = \frac{Count\_ii\ \dot{\pi}}{Count\_i\ \dot{\pi}} \frac{N\dot{}Prep}{N\dot{}} = \frac{9}{55} = 0.16$$

ตารางที่ 3.1 ฐานข้อมูลที่ใช้ประกอบการวิเคราะห์ประโยคภาษาไทยแบบวิธี Bigram

category	count_i	pair	count_ii	estimate	totalword
@	50	@Adj	1	0.02	
@	50	@Interj	1	0.02	
@	50	@N	20	0.40	265
@	50	@Pron	25	0.50	
@	50	@V	3	0.06	
Adj	16	AdjAdj	2	0.13	
Adj	16	AdjConj	1	0.06	
Adj	16	AdjN	1	0.06	
Adj	16	AdjPrep	1	0.06	
Adj	16	AdjV	11	0.69	
Adv	5	AdvAdj	1	0.20	
Adv	5	AdvConj	1	0.20	
Adv	5	AdvInterj	1	0.20	
Adv	5	AdvPrep	1	0.20	
Adv	5	AdvPron	1	0.20	
Conj	15	ConjAdj	1	0.07	
Conj	15	ConjN	9	0.60	
Conj	15	ConjPron	2	0.13	
Conj	15	ConjV	3	0.20	
Interj	1	InterjPron	1	1.00	
N	55	NAdj	6	0.11	
N	55	NAdv	2	0.04	
N	55	NConj	7	0.13	
N	55	NN	6	0.11	
N	55	NPrep	9	0.16	
N	55	NPron	3	0.05	
N	55	NV	22	0.40	
Prep	22	PrepInterj	1	0.05	
Prep	22	PrepN	14	0.64	
Prep	22	PrepPron	6	0.27	
Prep	22	PrepV	1	0.05	
Pron	39	PronAdj	8	0.21	
Pron	39	PronAdv	2	0.05	

ตารางที่ 3.1 (ต่อ)

Pron	39	PronPrep	1	0.03
Pron	39	PronV	28	0.72
V	62	VAdj	7	0.11
V	62	VAdv	2	0.03
V	62	VConj	6	0.10
V	62	VN	27	0.44
V	62	VPrep	11	0.18
V	62	VPron	5	0.08
V	62	VV	4	0.06

category = ประเภทของหน่วยคำที่เก็บในฐานข้อมูล เช่น คำนาม(N) ,คำสรรพนาม(Pron) , คำกริยา(V), คำวิเศษณ์ (Adv) , คำคุณศัพท์ (Adj), คำบุพบท(Prep),คำสันธาน(Conj) และคำอุทาน(Interj)

count_i = จำนวนครั้งที่พบประเภทของหน่วยคำนั้นๆที่ถูกป้อนเข้ามาเก็บในฐานข้อมูล

pair = คู่ลำดับประเภทของหน่วยคำก่อนหลังเรียงจากซ้ายไปขวา

count_ii = จำนวนครั้งที่พบ pair ที่ป้อนมาเก็บในฐานข้อมูล

$$\text{estimate} = \frac{count\_ii\ (YX)}{count\_i\ (Y)} = \text{Prop}(X|Y)$$

3.4 การวิเคราะห์โครงสร้างประโยคภาษาไทยร่วมกับคณิตศาสตร์ความน่าจะเป็น

การวิเคราะห์ประโยคภาษาไทย ที่ได้จากการแยกหน่วยคำของประโยคกำกวมหรือประโยคที่แยกหน่วยคำได้มากกว่า 1 ผลลัพธ์ ในรูปของโนด และอาร์ค จะเป็นมีหลายวิธี ดังที่กล่าวมาแล้วเช่น วิธี ATN กับวิธี M-ATN เป็นต้น ในงานวิจัยนี้ได้ปรับปรุงใช้วิธี Bigram ขึ้นมาช่วยวิเคราะห์โครงสร้างภาษาไทยแทน ซึ่งวิธีที่คล้ายกับวิธี Bigram ก็มีหลายวิธี เช่น

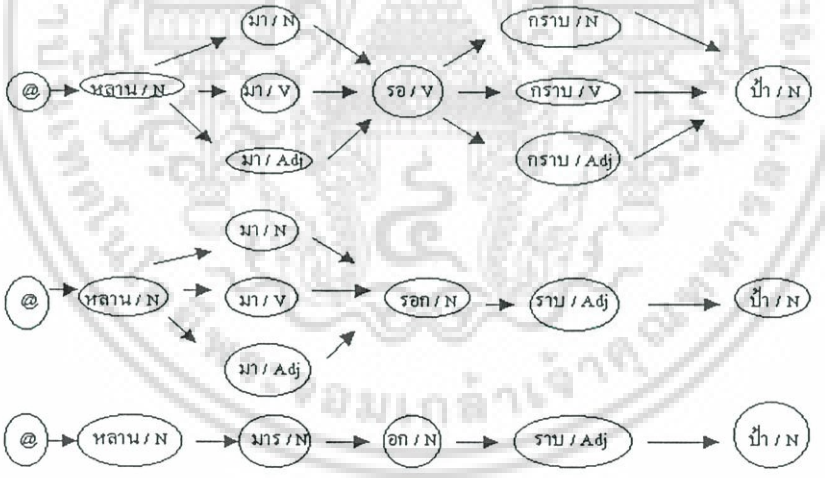
- วิธี N - gram models
- วิธี Conditional Probability (CP)
- วิธี Bayes ' s Theorem

3.4.1 วิธี N - gram models

เป็นวิธีการวิเคราะห์ประโยคภาษาไทยที่วิธี Bigram เป็นส่วนหนึ่งของวิธีนี้ โดยที่วิธี N - gram models จะพิจารณาหน่วยคำตามลำดับค่าของ N จากซ้ายไปขวาของประโยค เช่น N = Bi ก็เป็นการพิจารณาหน่วยคำ 2 หน่วยคำ , N = Tri ก็เป็นการพิจารณาหน่วยคำ 3 หน่วยคำ, N = Four ก็เป็นการพิจารณาหน่วยคำ 4 หน่วยคำ เป็นต้น

3.4.1.1 วิธี Bigram

วิธี Bigram นี้เป็นวิธีที่ใช้วิเคราะห์แยกแยะผลลัพธ์ที่ได้จากการตัดคำแบบ Left search matching ซึ่งเกิดจากประโยคต้นแบบที่กำกวมแล้วได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ จะต้องลดให้เหลือเพียงผลลัพธ์เดียว และถูกต้องตามหลักภาษาไทยด้วย โดยอาศัยตารางฐานข้อมูลหลักภาษาไทยจากที่กล่าวมาช่วยในการตรวจสอบวิเคราะห์ผล หลักการของวิธี Bigram นี้ถูกอ้างอิงในหนังสือ Natural Language Understanding [8] โดยวิธีนี้จะพิจารณาคำทีละ 2 คำจากซ้ายมือไปขวามือของประโยค โดยดูลำดับประเภทของหน่วยคำเทียบกับในตารางฐานข้อมูลที่ก็ใช้วิธีBigram เก็บข้อมูลร่วมกับคณิตศาสตร์ความน่าจะเป็นเช่นกัน ว่าผลลัพธ์ของประโยคกำกวมที่ผ่านการทำ Left search matching แล้ว ประโยคไหนจะถูกต้องทางหลักไวยากรณ์มากที่สุด โดยพิจารณาค่า Prob (estimate) ในตารางว่ามีค่าผลคูณของคู่ลำดับทั้งประโยค มีค่าความน่าจะเป็น(Prob) มากที่สุด ประโยคนั้นจะเป็นประโยคที่น่าจะถูกที่สุด ซึ่งถ้าพิจารณาก็เหมือนกับการหา node และ Arc ในระบบโครงข่าย Augmented Transition network [ATN] หรือ Modified Augmented Transition network [M-ATN] เพียงแต่มีค่าน้ำหนักความน่าจะเป็นมาควบคุมการกำหนดเงื่อนไขของ Arc อีกทีหนึ่ง ซึ่งขบวนการนี้จะหาเส้นทางที่ทำให้โครงข่ายใดมีค่าผลคูณของความน่าจะเป็นหรือ Prob มากที่สุดก็จะเป็นเส้นทางที่ถูกเลือกว่าเป็นประโยคที่ถูกต้อง วิธีการนี้เรียกว่า Viterbi Algorithm



รูปที่ 3.2 แผนผังการทำงานของ Viterbi Algorithm

ขั้นตอนการทำ Bigram มีขั้นตอนดังต่อไปนี้

1. นำประโยคผลลัพธ์ที่ได้จากการแยกแยะหน่วยคำ ด้วยวิธีLeft search matching แล้ว มาทำการวิเคราะห์หาลำดับประเภทของคำทีละ 2 คำ จากซ้ายไปขวาของประโยคซึ่งเทียบกับค่าความน่าจะเป็น(Prob) หรือ estimate ในตารางฐานข้อมูลหรือ Knowledge base ตารางที่ 3.1
2. นำค่า Prob ของคู่ลำดับทีละคู่มาคูณกัน จนหมดทั้งประโยคผลลัพธ์ที่ 1

3. แล้วทำการหาค่า Prob ของคู่ลำดับทีละคู่มาคูณกันอีกจนหมดประโยคผลลัพธ์ที่2 แล้วมาเปรียบเทียบกันว่าประโยคผลลัพธ์ที่ 1 กับประโยคผลลัพธ์ที่ 2 ใครมีค่าความน่าจะเป็นมากกว่ากัน ก็เก็บประโยคผลลัพธ์ที่มากกว่าไว้
4. ทำการหาค่า Prob ของคู่ลำดับทีละคู่มาคูณกันอีกจนหมดประโยคผลลัพธ์ถัดๆไป แล้วเปรียบเทียบผลดูว่าประโยคใดมีผลลัพธ์ของการคูณค่า Prob สูงสุด ประโยคนั้นจะเป็นประโยคที่น่าจะถูกต้องทางหลักภาษาไทยมากที่สุด แล้วแสดงผลประโยคนั้นออกมา

ตัวอย่างประโยคที่ใช้วิธี Bigram ในการประมวลผล เช่น

" หลานมารอกราบป้า " เมื่อผ่านวิธี Left search matching จะได้ประโยคผลลัพธ์ดังนี้

หลาน / มาร / อก / ราบ / ป้า

หลาน / มา / รอก / ราบ / ป้า

* **หลาน / มา / รอ / กราบ / ป้า** * (ซึ่งเป็นประโยคที่ถูกต้อง)

ฉะนั้นจำเป็นต้องทำการตัดสินว่าประโยคใดเป็นประโยคที่ถูกต้องที่สุดออกมาเพียงประโยคเดียว โดยจะต้องใช้วิธีทางไวยากรณ์ภาษาไทยเข้ามาช่วยแยกแยะและจะต้องอาศัยวิธี Bigram เข้าช่วยทำร่วมด้วย

ขั้นตอนการประมวลผลด้วยวิธี Bigram จากประโยคกำกวมตัวอย่าง (อาศัยแผนภาพ Viterbi Algorithm และตารางฐานข้อมูล 3.1)

หลาน / มา / รอ / กราบ / ป้า = @ / N / N / V / N / N

Prob(@NNVNN) = Prob(N|@) * Prob(N|N) * Prob(V|N) * Prob(N|V) * Prob(N|N)

= 0.40 * 0.11 * 0.40 * 0.44 * 0.11

= $8.51 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / V / V / V / N

Prob(@NVVVN) = Prob(N|@) * Prob(V|N) * Prob(V|V) * Prob(V|V) * Prob(N|V)

= 0.40 * 0.40 * 0.06 * 0.06 * 0.44

= $2.53 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / Adj / V / Adj / N

Prob(@NAdjVAdjN) = Prob(N|@) * Prob(Adj|N) * Prob(V|Adj) * Prob(Adj|V) * Prob(N|Adj)

= 0.40 * 0.11 * 0.69 * 0.11 * 0.06

= $2.00 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / N / V / V / N

Prob(@NNVVN) = Prob(N|@) * Prob(N|N) * Prob(V|N) * Prob(V|V) * Prob(N|V)

= 0.40 * 0.11 * 0.40 * 0.06 * 0.44

= $4.64 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / N / V / Adj / N

Prob(@NNVAdjN) = Prob(N|@) * Prob(N|N) * Prob(V|N) * Prob(Adj|V) * Prob(N|Adj)

= 0.40 * 0.11 * 0.40 * 0.11 * 0.06 = $1.16 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / V / V / N / N

Prob(@NVVNN) = Prob(N|@) * Prob(V|N) * Prob(V|V) * Prob(N|V) * Prob(N|N)

= 0.40 * 0.40 * 0.06 * 0.44 * 0.11

= $4.64 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / V / V / Adj / N

Prob(@NVVAdjN) = Prob(N|@) * Prob(V|N) * Prob(V|V) * Prob(Adj|V) * Prob(N|Adj)

= 0.40 * 0.40 * 0.06 * 0.11 * 0.06

= $6.33 * 10^{-5}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / Adj / V / V / N

Prob(@NAdjVVN) = Prob(N|@) * Prob(Adj|N) * Prob(V|Adj) * Prob(V|V) * Prob(N|V)

= 0.40 * 0.11 * 0.69 * 0.06 * 0.44

= $8.01 * 10^{-4}$

หลาน / มา / รอ / กราบ / ป้า = @ / N / Adj / V / N / N

Prob(@NAdjVNN) = Prob(N|@) * Prob(Adj|N) * Prob(V|Adj) * Prob(N|V) * Prob(N|N)

= 0.40 * 0.11 * 0.69 * 0.06 * 0.11

= $2.00 * 10^{-4}$

หลาน / มา / รอก / ราบ / ป้า = @ / N / N / N / Adj / N

Prob(@NNNAdjN) = Prob(N|@) * Prob(N|N) * Prob(N|N) * Prob(Adj|N) * Prob(N|Adj)

= 0.40 * 0.11 * 0.11 * 0.11 * 0.06

= $3.19 * 10^{-5}$

หลาน / มา / รอก / ราบ / ป้า = @ / N / V / N / Adj / N

Prob(@NVNAdjN) = Prob(N|@) * Prob(V|N) * Prob(N|V) * Prob(Adj|N) * Prob(N|Adj)

= 0.40 * 0.40 * 0.44 * 0.11 * 0.06

= $4.64 * 10^{-4}$

หลาน / มา / รอก / ราบ / ป้า = @ / N / Adj / N / Adj / N

Prob(@NAdjNAdjN) = Prob(N|@) * Prob(Adj|N) * Prob(N|Adj) * Prob(Adj|N) * Prob(N|Adj)

= 0.40 * 0.11 * 0.06 * 0.11 * 0.06

= $1.91 * 10^{-5}$

หลาน / มาร / อก / ราบ / ป้า = @ / N / N / N / Adj / N

Prob(@NNNAdjN) = Prob(N|@) * Prob(N|N) * Prob(N|N) * Prob(Adj|N) * Prob(N|Adj)

= 0.40 * 0.11 * 0.11 * 0.11 * 0.06

= $3.19 * 10^{-5}$

ผลจากการคำนวณวิธี Bigram แบบ Viterbi Algorithm จะได้ประโยคผลลัพธ์จากการตัดคำแบบ Left search matching ของประโยคกำกวม ที่มีค่าผลคูณความน่าจะเป็นของประโยคผลลัพธ์สูงสุด คือ

หลาน / มา / รอ / กราบ / ป้า = @ / N / N / V / N / N

Prob(@NNVNN) = Prob(N|@) * Prob(N|N) * Prob(V|N) * Prob(N|V) * Prob(N|N)

= 0.40 * 0.11 * 0.40 * 0.44 * 0.11

= $8.51 * 10^{-4}$

ซึ่งจะถือว่าประโยคผลลัพธ์นี้น่าจะเป็นประโยคที่ถูกต้อง ในกรณีที่ค่า Prob(?|?) ไม่มีในฐานข้อมูลจะให้ Prob(?|?) = 0.0001 เพื่อไม่ให้มีค่าเป็น 0 มิฉะนั้นจะทำให้ค่าผิดพลาด

3.4.1.2 วิธี Trigram

วิธี Trigram นี้เป็นวิธีที่ใช้วิเคราะห์แยกแยะผลลัพธ์ที่ได้จากการตัดคำแบบ Left search matching ซึ่งเกิดจากประโยคต้นแบบที่กำกวมแล้วได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ จะต้องลดให้เหลือเพียงผลลัพธ์เดียว และถูกต้องตามหลักภาษาไทยด้วย โดยอาศัยตารางฐานข้อมูลหลักภาษาไทยมาช่วยในการตรวจสอบวิเคราะห์ผลเช่นเดียวกับวิธีของ Bigram โดยวิธีนี้จะพิจารณาคำทีละ3 คำจากซ้ายมือไปขวามือของประโยค โดยดูลำดับประเภทของหน่วยคำเทียบกับในตารางฐานข้อมูลที่ใช้วิธี Trigram เก็บข้อมูลร่วมกับคณิตศาสตร์ความน่าจะเป็นเช่นกัน ว่าผลลัพธ์ของประโยคกำกวมที่ผ่านการทำ Left search matching แล้ว ประโยคไหนจะถูกต้องทางหลักไวยากรณ์มากที่สุด โดยพิจารณาค่า Prob (estimate) ในตารางว่ามีค่าผลคูณของคู่ลำดับทั้งประโยค มีค่าความน่าจะเป็น(Prob)มากที่สุด ประโยคนั้นจะเป็นประโยคที่น่าจะถูกที่สุดในทางหลักภาษาไทย

ในวิทยานิพนธ์ฉบับนี้ได้ทำการเปรียบเทียบวิธี Bigram กับ Trigram ไว้ด้วยเพื่อดูผลลัพธ์ของการวิเคราะห์ประโยคว่ามีข้อดีข้อเสียต่างกันอย่างไร โดยการเปรียบเทียบวิธีทั้ง 2 จะช่วยบอกให้ทราบได้ว่าในกรณีที่ N - gram สูงๆขึ้นไปจะดีกว่าค่า N - gram ต่ำๆหรือไม่ เพื่อจะใช้ประโยชน์ในการที่จะนำไปช่วยวิเคราะห์โครงสร้างภาษาไทยที่ได้จากการตัดหน่วยคำประโยคกำกวมของวิธี Left search matching ต่อไปได้ดีที่สุด ประโยคที่ป้อนเก็บในฐานข้อมูลแบบ Trigram ใช้ประโยคเดียวกันกับ Bigram เพียงแต่พิจารณาหน่วยคำทีละ 3 หน่วยคำ โดยใช้สูตรความน่าจะเป็นของ Trigram ร่วมด้วยเท่านั้นเพื่อจะใช้เปรียบเทียบกันได้ ซึ่งได้ยกตัวอย่างประโยค

ฐานข้อมูลดังที่ผ่านมาแล้วในหัวข้อฐานความรู้ที่ใช้ร่วมกับวิธี Bigram ส่วนสมการที่ใช้ในตารางฐานข้อมูลแบบ Trigram คือ

$$\Pr ob\,(Z \mid XY) = \frac{Count\ (XY\ at\ position\ i-1\ and\ Z\ at\ i)}{Count\ (XY\ at\ position\ i-1)} \qquad (3.2)$$

Prob (Z| XY) = ความน่าจะเป็นที่พบ X ,Y แล้วตามด้วย Z

Count (XY at position i-1 and Z at i) = จำนวนที่พบ X,Y แล้วตามด้วย Z

Count (XY at position i-1) = จำนวนที่พบ XY

ตัวอย่างประโยคกำกวมที่ใช้ทดสอบระหว่างวิธี Bigram กับ Trigram โดยก่อนทำการเปรียบเทียบวิธี ทั้ง 2 ต้องผ่านกระบวนการวิธี แยกหน่วยคำมาก่อนแล้ว และได้ผลลัพธ์การแยกหน่วยคำมากกว่า 1 ผลลัพธ์ เช่น

- หลานมารอกราบป้า
- เขามีเงินมากกว่าฉัน
- มารอกราบปู่ที่หน้าบ้าน
- ตากราบพระและขนธูปเทียน
- นกปากกลมกินกบและปลาในบ่อน้ำ
- หนังสือดีเรื่องปลาตากลมน่าอ่านมาก

ตารางที่ 3.2 ฐานข้อมูลที่ใช้ประกอบการวิเคราะห์ประโยคภาษาไทยแบบวิธี Trigram

category	count_ii	pair	count_iii	estimate
@Adj	1	@AdjV	1	1.00
@Interj	1	@InterjPron	1	1.00
@N	20	@NAdj	1	0.05
@N	20	@NConj	1	0.05
@N	20	@NN	2	0.10
@N	20	@NPrep	3	0.15
@N	20	@NPron	2	0.10
@N	20	@NV	11	0.55
@Pron	25	@PronAdj	5	0.20
@Pron	25	@PronV	20	0.80
@V	3	@VN	1	0.33
@V	3	@VPron	1	0.33
@V	3	@VV	1	0.33
AdjN	1	AdjNV	1	1.00
AdjV	9	AdjVAdj	2	0.22
AdjV	9	AdjVN	4	0.44
NAdj	3	NAdjAdj	1	0.33
NAdj	3	NAdjV	2	0.67
NN	4	NNV	2	0.50
NV	16	NVAdj	2	0.13
NV	16	NVN	5	0.31
NV	16	NVV	2	0.13
VAdj	5	VAdjN	1	0.20
VAdj	5	VAdjV	2	0.40
VN	18	VNN	2	0.11
VN	18	VNV	1	0.06
VN	18	VNAdj	3	0.16

category = ประเภทของลำดับหน่วยคำที่เก็บในฐานข้อมูล เช่น คำนาม(N) ,คำสรรพนาม(Pron) , คำกริยา(V), คำวิเศษณ์ (Adv) , คำคุณศัพท์ (Adj), คำบุพบท(Prep) , คำสันธาน(Conj) และ คำอุทาน(Interj)

count_i = จำนวนครั้งที่พบประเภทของลำดับหน่วยคำนั้นๆที่ถูกป้อนเข้ามาเก็บในฐานข้อมูล

pair = คู่ลำดับประเภทของหน่วยคำก่อนหลังเรียงจากซ้ายไปขวา แบบทีละ 3 หน่วยคำ (Trigram)

count_ii = จำนวนครั้งที่พบ pair ที่ป้อนมาเก็บในฐานข้อมูล

$$\text{estimate} = \frac{count\_ii\ XYZ}{count\_i\ XY} = \text{Prob}(Z|XY)$$

ตัวอย่างการคำนวณและพิจารณาแบบ Trigram จากประโยคการตัดคำแบบ Left search matching ที่เป็น Unknown เช่น

หลาน / มา / รอ / กราบ / ป้า = N / Adj / V / N / N

Prob(NAdjVNN) = Prob(Adj|@N) * Prob(V|NAdj) * Prob(N|AdjV) * Prob(N|VN)

= 0.05 * 0.67 * 0.44 * 0.11

= $1.62 * 10^{-3}$

หลาน / มา / รอก / ราบ / ป้า = N / V / N / Adj / N

Prob(NVNAdjN) = Prob(V|@N) * Prob(N|NV) * Prob(Adj|VN) * Prob(N|NAdj)

= 0.55 * 0.31 * 0.17 * 0.33

= $9.56 * 10^{-3}$

หลาน / มาร / อก / ราบ / ป้า = N / N / N / Adj / N

Prob(NNNAdjN) = Prob(N|@N) * Prob(N|NN) * Prob(Adj|NN) * Prob(N|NAdj)

= 0.10 * 0.0001 * 0.0001 * 0.33

= $3.30 * 10^{-10}$

ฉะนั้นวิธี Trigram จะให้ผลลัพธ์ประโยคที่มีผลคูณของความน่าจะเป็นสูงสุด คือ

หลาน / มา / รอก / ราบ / ป้า = N / V / N / Adj / N

Prob(NVNAdjN) = Prob(V|@N) * Prob(N|NV) * Prob(Adj|VN) * Prob(N|NAdj)

= 0.55 * 0.31 * 0.17 * 0.33

= $9.56 * 10^{-3}$

ซึ่งจะถือว่าประโยคผลลัพธ์นี้เป็นประโยคที่น่าจะถูกต้องเมื่อใช้วิธี Trigram ในกรณีที่ค่า Prob(?|??) ไม่มีในฐานข้อมูลจะให้ Prob(?|??) = 0.0001 เพื่อไม่ให้มีค่าเป็น 0 มิฉะนั้นจะทำให้ค่าผิดพลาด

ตารางที่ 3.3 ผลการเปรียบเทียบวิธี Bigram กับ Trigram

จำนวนประโยคที่เก็บในฐานข้อมูล	%ที่Bigramทำได้ถูกต้อง	% ที่ Trigram ทำได้ถูกต้อง
5	80	65
10	90	75
20	95	80
50	100	85

ผลการเปรียบเทียบวิธี Bigram กับวิธี Trigram วิธี Bigram จะให้ผลลัพธ์ที่ดีกว่าถูกต้องมากกว่าวิธี Trigram ส่วนเวลาในการประมวลผลจะใกล้เคียงกันมาก สาเหตุที่วิธี Bigramให้ผลลัพธ์ที่ดีกว่าเพราะ การตรวจสอบลำดับคำทีละ 2 คำ ในประโยคฐานข้อมูล แล้วนำมาเก็บในรูปความน่าจะ

เป็น จะมีความยืดหยุ่นกว่าการตรวจสอบลำดับคำทีละ 3 คำ แล้วเก็บในรูปความน่าจะเป็น ซึ่งจะเป็นการเจาะจงค่าประเภทของหน่วยคำจนเกินไปถึง 3 ตำแหน่งหน่วยคำ เมื่อนำมาตรวจสอบกับประโยคที่เป็น Unknown การตรวจสอบทีละ 2 คำ จะไม่เจาะจงหน่วยคำจนเกินไปทำให้ได้ค่าของประโยคผลลัพธ์ Unknown มีค่าที่ดีกว่าแบบตรวจทีละ 3 คำ การเจาะจงประเภทของหน่วยคำจนเกินไปทำให้ได้ผลลัพธ์ของประโยคผิดพลาดได้

3.4.2 วิธี Conditional Probability (CP)

เป็นวิธีการหาลำดับของเหตุการณ์ความน่าจะเป็นว่า เมื่อเกิดเหตุการณ์หนึ่งแล้วจะเกิดอีกเหตุการณ์หนึ่งมีค่าเป็นเท่าไร โดยเหตุการณ์ทั้งสองเป็นอิสระต่อกัน สมการของ Conditional Probability (CP) มีสมการดังนี้

$$Prob(A \mid B) = \frac{Prob(A \& B)}{Prob(B)} \qquad (3.3)$$

Prob(A|B) = ความน่าจะเป็นที่มีเหตุการณ์ A แล้วเกิดเหตุการณ์ B

Prob(A&B) = ความน่าจะเป็นที่มีเหตุการณ์ A และเกิดเหตุการณ์ B พร้อมกัน

Prob(B) = ความน่าจะเป็นที่เกิดเหตุการณ์ B

วิธี CP เมื่อนำมาเปรียบเทียบกับวิธี Bigram จะเป็นการเปรียบเทียบประเภทของหน่วยคำทีละ 2 หน่วยคำ เช่นเดียวกัน และเป็นการเปรียบเทียบจากซ้ายไปขวามือของประโยค Unknown จะต่างกันเพียงแต่สูตรสมการเท่านั้น ในส่วนฐานข้อมูลของวิธี CP จะใช้ตารางเดียวกันกับตารางวิธี Bigram เพราะเป็นการพิจารณาหน่วยคำทีละ 2 หน่วยคำเหมือนกัน ตัวอย่างประโยคที่ใช้ทดสอบระหว่างวิธี CP กับวิธี Bigram จะใช้ประโยคเดียวกันกับวิธี Bigram กับ Trigram ดังที่กล่าวมา

เมื่อนำทั้ง 3 วิธี คือ Bigram ,Trigram และ Condition Probability (CP) มาเปรียบเทียบกัน แล้วพบว่า วิธี Bigram จะให้ผลลัพธ์ที่ถูกต้องมากกว่า และยิ่งจำนวนประโยคฐานข้อมูลมีมากขึ้นค่าความถูกต้องก็จะเพิ่มขึ้น ในขณะที่เวลาในการประมวลผลของทั้ง 3 วิธีใกล้เคียงกัน ทั้ง 3 วิธีนี้จะใช้ในการช่วยวิเคราะห์ประโยคกำกวมที่เกิดจากการตัดหน่วยคำแบบ Left search matching มาแล้วและได้ผลลัพธ์มากกว่า 1 ผลลัพธ์ ซึ่งจะต้องแยกประโยคกำกวมออกเพื่อให้เหลือผลลัพธ์เดียว โดยใช้หลักไวยากรณ์ภาษาไทยในรูปแบบของ โนดและ อาร์ค ร่วมกับความน่าจะเป็นมาตรวจสอบ โดยอาศัยวิธีทั้ง 3 ข้างต้นช่วยจึงจะทำให้ได้ประโยคที่ถูกต้องทางหลักภาษาไทยออกมาได้ แต่ถึงอย่างไรก็ดีวิธีการ Bigram , Trigram และ Condition Probability (CP) จะมีประสิทธิภาพได้ดีก็ต้องอาศัยฐานข้อมูลของประโยคไวยากรณ์ที่ดีมีประสิทธิภาพด้วยจึงจะได้ผลลัพธ์ประโยคที่ถูกต้อง

สมบูรณ์ แต่จากการเปรียบเทียบผลลัพธ์ที่ได้วิธี Bigram จะดีที่สุด รองลงมาจะเป็นวิธีของ CP และ Trigram ตามลำดับ เมื่อเพิ่มประโยคในฐานข้อมูลมากขึ้นวิธี Trigram จะยิ่งมีค่าความถูกต้องสูงขึ้น ขณะที่วิธีของ CP มีค่าความถูกต้องคงที่ไม่เปลี่ยนแปลงไปตามค่าการเพิ่มขึ้นของฐานข้อมูลมากนัก แต่ในด้านผลลัพธ์ในการตัดสินประโยคกำกวมจากการตัดคำด้วยวิธี Left search matching มาแล้วนั้นจะให้เปอร์เซ็นต์ความถูกต้องค่อนข้างสูง แต่ก็ยังไม่ดีเท่าวิธี Bigram อาจจะเนื่องมาจากค่าตัวแปรที่ใช้ในวิธี CP จะต้องมีความเป็นอิสระต่อกันดังที่กล่าวมาแล้ว ซึ่งตามทฤษฎีของความเป็นอิสระต่อกันจะต้องได้ว่า Prob(A&B) = Prob(A) * Prob(B) แต่จากตารางที่ 3.1 ไม่ได้เป็นเช่นนั้น เช่น

$$Prob(Adj\&N) \neq Prob(Adj) * Prob(N)$$

$$1 \neq 16 * 55$$

ซึ่งแสดงว่าหน่วยคำที่เกิดขึ้นไม่ได้เป็นอิสระต่อกันจริง ทั้งนี้เพราะบางครั้งหน่วยคำในภาษาไทยจะไม่มีความเป็นอิสระกันก็ได้ เช่น Adv มักตามหลังคำกริยาที่มันขยาย หรือ Adj ก็จะต้องตามหลังคำนาม คำสรรพนาม ที่มันขยายอยู่ เป็นต้น ดังที่กล่าวมาแล้ว ส่วนวิธี Bigram ยิ่งเพิ่มประโยคฐานข้อมูลมากขึ้นค่าความถูกต้องก็ยิ่งมากขึ้นตามไปด้วย ฉะนั้นในวิทยานิพนธ์นี้จึงใช้วิธี Bigram มาทำการวิเคราะห์ประโยคกำกวมที่เกิดจากการแยกหน่วยคำแบบ Left search matching มาใช้ ผลการเปรียบเทียบโปรแกรมของทั้ง 3 วิธี มีอยู่ในภาคผนวก ก.

3.4.3 วิธี Bayes ' s Theorem

ข้อที่น่าสังเกตวิธี Condition Probability (CP) จะให้ผลลัพธ์ของการพิจารณาประโยคที่ถูกต้องจากการตัดคำได้เปอร์เซ็นต์ที่สูงใกล้เคียงกับวิธี Bigram น่าจะมาจากวิธีการของ Bayes ' s Theorem เนื่องจากสูตร สมการของ Bayes 's Theorem เมื่อพิจารณาแล้วเปรียบเทียบกับตารางที่ 3.1 ก็จะเป็น วิธีเดียวกับ วิธี Bigram นั้นเอง เพราะสมการของ Bayes ต้องอาศัย สมการ Condition Probability (CP) ดังสมการที่ (3.3)

ส่วนสมการของ Bayes ' s Theorem มีสูตรสมการดังนี้

$$Prob(A \mid B) = \frac{Prob(B \mid A) * Prob(A)}{Prob(B)} \qquad (3.4)$$

Prob(A|B) = ความน่าจะเป็นที่มีเหตุการณ์ A แล้วเกิดเหตุการณ์ B

Prob(B|A) = ความน่าจะเป็นที่มีเหตุการณ์ B แล้วเกิดเหตุการณ์ A

Prob(A) = ความน่าจะเป็นที่เกิดเหตุการณ์ A

Prob(B) = ความน่าจะเป็นที่เกิดเหตุการณ์ B

ซึ่งถ้าเปลี่ยนรูปสมการ Bayes สมการที่ (3.4) ใหม่โดยอาศัยสมการ CP จะได้

$$Prob(A \mid B) = \frac{Prob(B \,\&\, A)}{Prob(A)} * \frac{Prob(A)}{Prob(B)} = \frac{Prob(B \,\&\, A)}{Prob(B)} \qquad (3.5)$$

จากสมการที่ (3.5) และดูเทียบในตารางที่ 3.1 กับสมการของ Bigram สมการที่ (3.1) ก็จะคือ วิธีของ Bigram นั้นเอง ดังนั้นวิธีของ Bayes น่าจะให้ผลลัพธ์ของการวิเคราะห์ประโยคภาษาไทยได้เช่นกันกับวิธีของ Bigram และจากสมการของ Bayes นี้เองจึงน่าจะทำให้ ค่าของการวิเคราะห์ประโยคภาษาไทยแบบวิธี Condition Probability (CP) ได้ผลลัพธ์ที่มีเปอร์เซ็นต์ความถูกต้องของการพิจารณาการตัดคำค่อนข้างสูงใกล้เคียงกับวิธี Bigram มากนั้นเอง แต่ที่ยังไม่ถูกต้องดีที่สุดก็อาจจะเป็นเพราะ ค่าตัวแปรของทั้งสมการ Condition Probability (CP) และ ของ Bayes นั้นจะต้องเป็นค่าตัวแปรที่มีความเป็นอิสระนั้นเองดังที่กล่าวมาข้างต้น

บทที่ 4
แนวทางการจัดกลุ่มเอกสารภาษาไทย

การจะประยุกต์เอาประโยคภาษาไทย เพื่อใช้ในการแยกกลุ่มเอกสารภาษาไทยนั้นมีปัญหาและอุปสรรคอยู่หลายอย่างเช่นกัน เนื่องจากในตัวประโยคโครงสร้างของภาษาไทยเอง และรูปแบบทางไวยากรณ์ของภาษาไทยด้วย ดังนั้นจึงต้องมีการเตรียมข้อมูลและการหาขบวนการในการคัดกรองคำสำคัญ (Keywords) เพื่อจะนำมากำหนดในการแยกกลุ่มเอกสาร ซึ่งเมื่อแยกหน่วยคำจากประโยคภาษาไทยได้แล้ว ต้องมากำจัดคำที่ไม่น่าจะเป็นคำสำคัญทิ้งก่อน เช่น คำเชื่อมต่างๆ ร่วมทั้งคำอื่นๆด้วย (และ,หรือ,แล้ว,กัน,.... เป็นต้น) และเมื่อจัดกลุ่มเอกสารได้แล้วต้องหาเครื่องมือมาวัดและตรวจสอบผลลัพธ์การจัดกลุ่มเอกสารแต่ละวิธีด้วยว่าวิธีไหน จะให้ผลการจัดกลุ่มดีที่สุด ในที่นี้จะใช้ค่า Precision, Recall และค่า F- measure เป็นตัววัด เทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ ซึ่งการจัดกลุ่มเอกสารที่ดีเนื้อหาเอกสารในกลุ่มเดียวกันควรจะมีเนื้อหาใกล้เคียงกันมากกว่าเอกสารต่างกลุ่มกัน โดยในวิทยานิพนธ์นี้จะแยกกลุ่มเอกสารภาษาไทยจากบทคัดย่อวิทยานิพนธ์ปริญญาโทจากหลายสาขา ตัวอย่างบทคัดย่อวิทยานิพนธ์ดูได้จากภาคผนวก ข

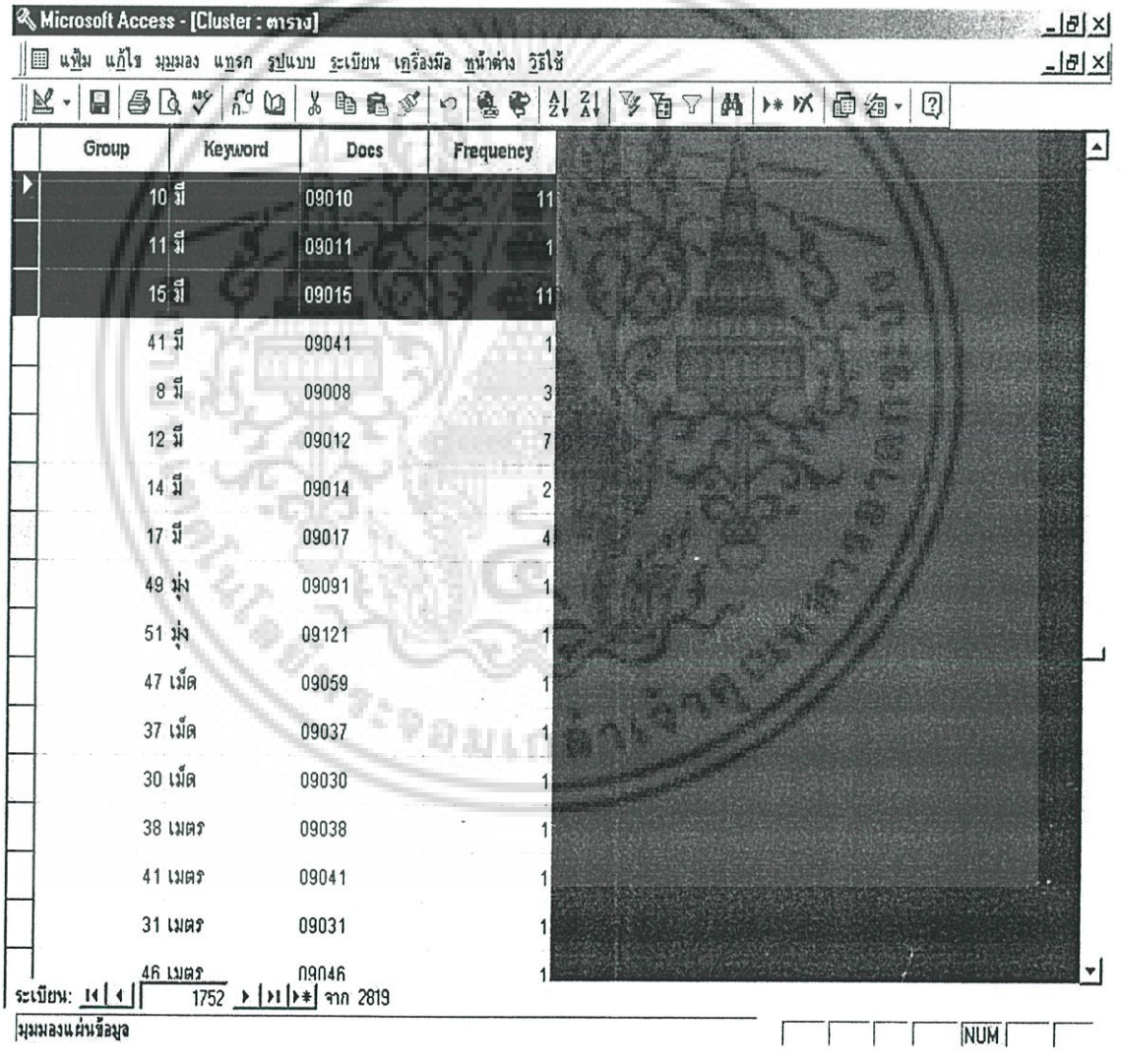
4.1 การกำจัดคำที่ไม่ใช่คำสำคัญในเอกสาร

การจัดกลุ่มเอกสารด้วยระบบคอมพิวเตอร์ จำเป็นต้องแยกคำเป็นคำๆออกจากประโยคก่อนเพื่อให้ระบบทำการวิเคราะห์ประโยคและเอกสารได้ เนื่องจากระบบคอมพิวเตอร์ไม่สามารถแยกประโยคภาษาไทยออกเป็นคำๆได้เอง จึงต้องมีการเตรียมขั้นตอนในการตัดคำภาษาไทยออกจากประโยคภาษาไทยดังกล่าวมาแล้วในตอนต้นก่อน ในการจัดกลุ่มเอกสารภาษาไทยเมื่อแยกคำออกมาได้ด้วยวิธี Bigram ดังที่กล่าวมาแล้ว จะได้คำจำนวนหนึ่งที่ไม่น่าจะใช่คำสำคัญ(Keywords) ของเอกสาร เช่น พวกคำเชื่อมต่างๆ(และ,หรือ,แล้ว,ถ้า,..... เป็นต้น) ซึ่งจะต้องเอาออกก่อนนำไปใช้วิเคราะห์แยกกลุ่มเอกสารเพื่อระบบการจัดกลุ่มเอกสารจะได้มีประสิทธิภาพมากขึ้นไม่เกิดการผิดพลาด คำที่รบกวนระบบการจัดกลุ่มเอกสารนี้จึงต้องถูกกำจัดออกก่อนที่จะนำเอาไปวิเคราะห์หากลุ่มเอกสาร โดยคำต่างๆที่แยกด้วยวิธี Bigramจากประโยคภาษาไทยแล้ว ทั้งหมดจะถูกจัดเก็บไว้ในตารางฐานข้อมูลของโปรแกรม Microsoft Access ซึ่งจะนำเอาฐานข้อมูลนี้ไปใช้ในการจัดกลุ่มเอกสารต่อไปตามวิธีการจัดกลุ่มเอกสารแบบต่างๆดังจะกล่าวต่อไป เพื่อจะหาวิธีที่ดีที่สุดในการจัดกลุ่มเอกสาร เปรียบเทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ โดยใช้เครื่องมือวัด Precision , Recall และ F-measure ดังนั้นการกำจัดคำที่รบกวนระบบด้วยโปรแกรม Microsoft access นั้นจึงทำได้ง่ายมาก เพียงแต่ทำการเรียงลำดับคำจากมากไปหาน้อย เราก็สามารถกำจัดคำรบกวนระบบการจัดกลุ่มเอกสารได้ โดยทำการ Delete คำเหล่านั้นออกจากฐานข้อมูลก็เป็นอัน

เสร็จเรียบร้อย คำที่เหลือในฐานข้อมูลก็จะเป็นคำสำคัญ(Keywords)จริงๆของระบบ ซึ่งจะนำไปใช้จัดกลุ่มเอกสารได้ ตัวอย่างคำที่รบกวนระบบที่ต้องกำจัดทิ้ง เช่น

ก็ , กว่า , กัน , กับ , การ , การศึกษา , ขนาด , ของ , เข้า , ควร , ความ , ค่า , คือ , จะ , จาก , ชนิด , ใช้ , ซิ , ซึ่ง , ด้วย , ดี , เดียว , โดย , ใด , ได้ , ต่อ , ต้อง , ตัว , ต่าง , ตาม , แต่ , ถึง , ทั้ง , ทาง , ทำ , ที่ , ทุก , นอก , น้อย , นัก , นั้น , นำ , นี้ , เนื่อง , ใน , บาง , บ้าง , แบบ , เป็น , ไป , พบ , เพียง , เพื่อ , มา , มาก , มี , ไม่ , ยัง , รวม , ร่วม , ละ , แล้ว , และ , ว่า , ไว้ , สอง , สาม , สี่ , ส่วน , สำคัญ , สำหรับ , สามารถ , สุด , หนึ่ง , หรือ , หา , ให้ , อย่าง , อย่างไร , อยู่ , อื่น , อีก

ตัวอย่างคำรบกวนระบบการจัดกลุ่มเอกสารดังกล่าวข้างบน ที่ต้องตัดทิ้งในฐานข้อมูล Microsoft Access แสดงไว้ในรูปที่ 4.1 (ส่วนที่เป็นแถบดำ คำว่า "มี")

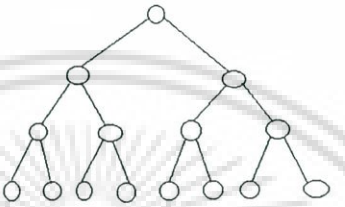


รูปที่ 4.1 แสดงตัวอย่างคำรบกวนระบบการจัดกลุ่มเอกสารที่ต้องกำจัดออกในฐานข้อมูล

4.2 วิธีการจัดกลุ่มเอกสารโดยทั่วไป

การจัดกลุ่มเอกสารโดยทั่วไปแบ่งได้เป็น 2 วิธีใหญ่ดังนี้

1. การจัดกลุ่มแบบแผนภาพต้นไม้(Tree Diagram) หรือแบบ Hierarchical Agglomerative Clustering มีวิธีการจัดกลุ่ม คือ เริ่มจากพิจารณาแต่ละเอกสารว่ามีคู่คล้ายกับเอกสารใดมากที่สุดแล้วจึงทำการรวมเข้ากับเอกสารนั้น ต่อจากนั้นจึงพิจารณาแต่ละคู่เอกสารที่รวมกันแล้วอีกว่า จะมีคู่คล้ายกับคู่เอกสารกลุ่มใดอีกมากที่สุดจึงรวมเข้าไปอีก ทำเช่นนี้ต่อไปจนเอกสารไม่เปลี่ยนแปลงอีกแล้วจึงหยุด



รูปที่ 4.2 แผนภาพต้นไม้ (Tree Diagram)

2. การจัดกลุ่มเอกสารแบบ Iterative Clustering เป็นวิธีการที่กล่าวถึง การจัดกลุ่มเอกสารที่เริ่มต้นมีจำนวนกลุ่มแน่นอน แล้วทำการดึงเอกสารสมาชิกในกลุ่มไปตรวจสอบกับกลุ่มอื่นรวมทั้งกลุ่มเดิมของตนเอง ว่าจะคล้ายกับกลุ่มไหนมากที่สุดก็จะเข้าไปรวมกับกลุ่มนั้น ความจริงวิธีนี้ก็คือต้องการตรวจสอบเพียงว่าเอกสารที่จัดกลุ่มแล้ว เป็นสมาชิกจริงในกลุ่มหรือไม่ รูปแบบโดยทั่วไปของวิธี Iterative Clustering มีดังนี้

2.1 สมมติให้มีจำนวนกลุ่มเท่ากับ K กลุ่ม เริ่มต้น

2.2 ให้เอาเอกสารแต่ละเอกสารออกมา หาค่าความคล้ายในแต่ละกลุ่มทุกกลุ่ม

2.3 ทำการรวมเอกสารนั้นกับกลุ่มที่คำนวณ ได้ค่าความคล้ายมากที่สุด

2.4 กลับไปทำในข้อ 2.2 ซ้ำอีก จนได้กลุ่มเอกสารที่ไม่เปลี่ยนกลุ่มอีกแล้วทุกเอกสาร

วิธีการจัดกลุ่มแบบ Iterative Clustering นี้ ผลการจัดกลุ่มจะได้จำนวนกลุ่มเท่าเดิม โดยปกติกลุ่มเอกสารจะไม่มีการลดจำนวนกลุ่มเหมือนแบบ hierarchical clustering แต่บางครั้งอาจมีการลดจำนวนกลุ่มได้เนื่องจาก สมาชิกในกลุ่มย้ายกลุ่มออกไปทั้งหมด สมาชิกกลุ่มอาจเปลี่ยนกลุ่มได้เมื่อเทียบกับก่อนเริ่มต้นทำ Iterative Clustering โดยระบบจะต้องทำจนกว่าเอกสารจะไม่เปลี่ยนกลุ่มอีกแล้วจึงจะหยุดได้ ซึ่งจะได้เอกสารที่เป็นสมาชิกจริงๆในกลุ่มนั้นๆ

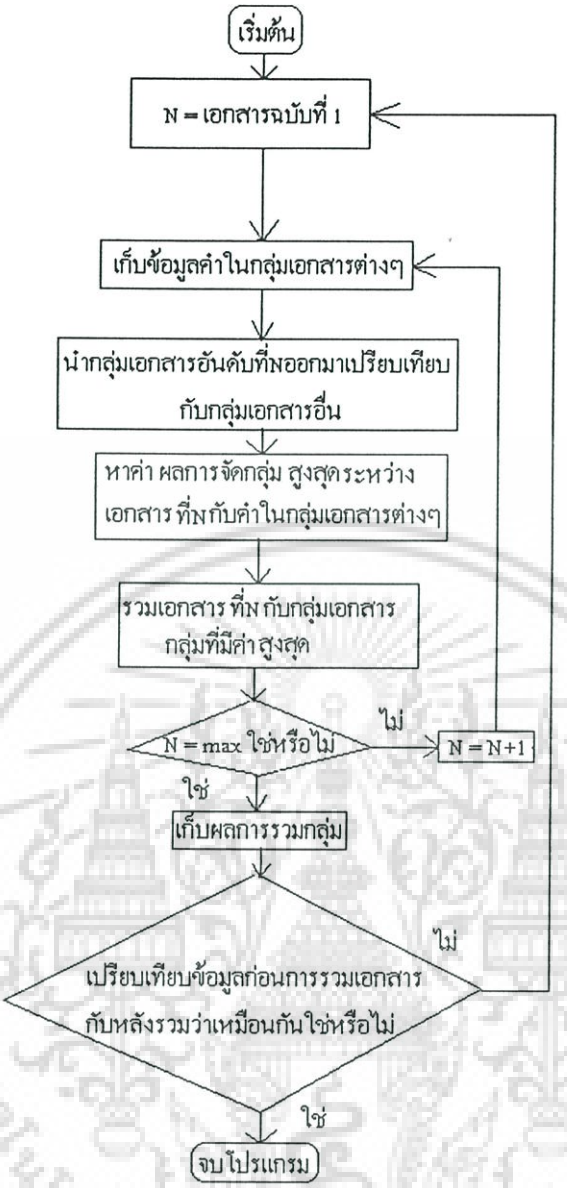
จากการกล่าวถึงวิธีการจัดกลุ่มดังกล่าว ในงานวิทยานิพนธ์นี้จึงใช้เป็นแนวทางในการจัดกลุ่มเอกสารภาษาไทยต่อไป ฉะนั้นเมื่อเราทำการตัดคำที่รบกวนระบบออกไปแล้ว เราจะได้ส่วนของคำสำคัญ(Keywords) เท่านั้นในฐานข้อมูลที่จะใช้ในการจัดกลุ่มเอกสารภาษาไทย โดยที่เราจะนำไปพิจารณาหาวิธีการจัดกลุ่มเอกสารจากคำสำคัญเหล่านี้ วิธีการจัดกลุ่มเอกสารภาษาไทยที่

ทดลองวิจัยในวิทยานิพนธ์นี้ จะใช้อัลกอริทึ่มในการจัดกลุ่มแบบเดียวกันคือแบบ Iterative Clustering จะต่างกันตรงใช้สูตรในการจัดกลุ่ม และวิธีการพิจารณาคำสำคัญในการจัดกลุ่มเอกสารภาษาไทย อัลกอริทึ่มการจัดกลุ่มเอกสารภาษาไทยจะเป็นดังนี้

1. ทำเอกสารแต่ละเอกสารเป็น 1 เอกสาร และ 1 กลุ่ม (1 Doc 1 Group) ซึ่งแต่ละเอกสารประกอบด้วยคำสำคัญต่างๆ

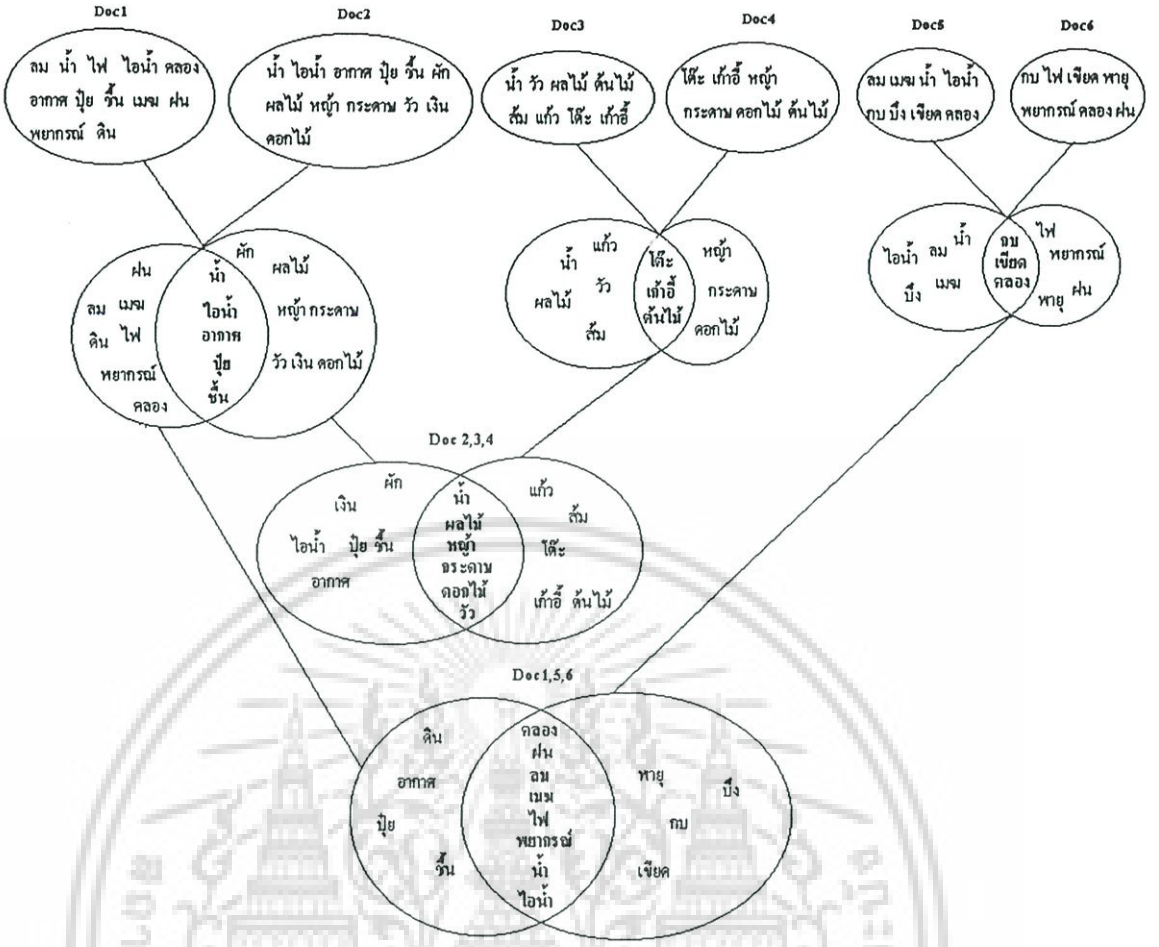
2. ทำการจัดกลุ่มเอกสารโดยเอาเอกสาร อันที่1 กลุ่มที่1 ไปหากลุ่มที่มันจะเข้าไปรวมได้ตามวิธีและสูตรของการจัดกลุ่ม ต่อจากนั้นก็นำเอาเอกสารที่2 กลุ่มที่2 ไปหากลุ่มที่มันจะเข้าไปรวมได้ตามวิธีและสูตรของการจัดกลุ่มเช่นกัน ทำเช่นนี้ไปเรื่อยๆจนหมดทุกเอกสาร ในกรณีสมมติเอกสารที่1กลุ่ม1 รวมกับเอกสารที่2 กลุ่ม 2 ก่อน เมื่อจะพิจารณาเอกสารที่2 ว่าจะรวมกับกลุ่มใดให้เอาเอกสารที่ 2 ออกมาจากกลุ่มที่รวมกับเอกสารที่ 1 ออกก่อน ต่อจากนั้นจึงนำไปหาว่าจะร่วมกับกลุ่มไหนตามวิธีและสูตรการจัดกลุ่ม โดยต้องเทียบเอกสารที่ 2 กับทุกกลุ่มจนกว่าจะหากลุ่มที่เหมาะสมแล้วรวมกลุ่มเข้าไปได้ ซึ่งเอกสารแต่ละเอกสารจะร่วมกันเองในรอบแรกระดับหนึ่ง

3. เมื่อครบรอบแรกในทุกเอกสารแล้ว จะทำการรวมในระดับถัดไปอีกรอบโดยทำเหมือนในข้อ 2 คือ จะเอาเอกสารแต่ละอันออกจากกลุ่มก่อน แล้วทำการหาตามวิธีและสูตรการจัดกลุ่มแต่ละวิธี โดยนำเอกสารที่ออกมาไปเทียบหากลุ่มทุกกลุ่มที่เหมาะสมว่าจะรวมกับกลุ่มไหนได้อีก ก็จะเข้าไปรวมกับกลุ่มเอกสารนั้นๆต่อไป ทำเช่นนี้จนครบทุกเอกสาร ก็จะได้อีกรอบ ทำเช่นนี้ไปจนกลุ่มเอกสารไม่เปลี่ยนแล้วจึงหยุด ในระยะแรกเอกสารจะรวมกันมีลักษณะคล้ายโครงสร้างต้นไม้ จากนั้นระบบก็จะเป็นแบบ Iterative clustering ต่อจากนั้นทำการเก็บผลการรวมกลุ่มเอกสารเอาไว้ทุกรอบ เพื่อจะนำมาพิจารณาหาวิธีการจัดกลุ่มที่ดีที่สุดต่อไป แผนภาพ อัลกอริทึ่มการจัดกลุ่มเอกสารแสดงในภาพข้างล่าง



รูปที่ 4.3 แผนภาพอัลกอริทึมในการจัดกลุ่มเอกสารภาษาไทย

สาเหตุที่ต้องจัดกลุ่มเอกสารแบบ Iterative Clustering โดยดึงเอกสารทีละกลุ่มไปหาค่าความคล้าย ในกลุ่มเอกสารต่างๆแต่ละรอบ โดยไม่เอากลุ่มที่รวมกันแล้วไปหากลุ่มอื่นเพื่อจะรวมกับกลุ่มอื่นอีกแบบวิธี Hierarchical Agglomerative Clustering เพราะ ป้องกันการรวมเอกสารบางเอกสารผิดกลุ่มได้ เนื่องจากเมื่อเอกสารถูกรวมกันในรอบแรกแล้ว พอรอบต่อไปเอกสารบางเอกสารในสมาชิกกลุ่มนั้นอาจไม่ถูกต้องนักที่จะถูกรวมไปกับกลุ่มอื่นด้วย สาเหตุเพราะคำสำคัญในเอกสารไม่เกี่ยวข้องอะไรมากนักกับกลุ่มใหม่ที่ไปรวม ดังตัวอย่างข้างล่างนี้



รูปที่4.4 แผนภาพแสดงการรวมกลุ่มเอกสารแบบพิจารณาวิธี Hierarchical Agglomerative Clustering

จากรูปที่4.4 กำหนดให้เริ่มต้น เอกสาร1เอกสาร เท่ากับ 1กลุ่ม (1Doc เท่ากับ 1 Group) เมื่อพิจารณาเอกสารฉบับที่ 1 หรือ กลุ่มที่1 จะมีคำสำคัญที่เหมือนกับเอกสารฉบับที่ 2 หรือกลุ่มที่ 2 มากที่สุดคือ จำนวน 5 คำ (**น้ำ,ไอน้ำ,อากาศ,ปุ๋ย,ชื้น**) จึงรวมเอกสารฉบับที่ 1 กับ 2 เป็นกลุ่มเดียวกัน กลายเป็นกลุ่มใหม่คือ กลุ่ม(1,2) ส่วนเอกสารฉบับที่ 3 หรือกลุ่มที่3 ก็มีคำสำคัญที่เหมือนกับเอกสารฉบับที่ 4 หรือกลุ่มที่4 มากที่สุดคือ จำนวน3คำ (**โต๊ะ,เก้าอี้,ต้นไม้**) จึงรวมเอกสารฉบับที่ 3 กับ 4 เป็นกลุ่มเดียวกัน กลายเป็นกลุ่มใหม่คือ กลุ่ม(3,4) เอกสารฉบับที่5 และ 6 ก็เช่นกันรวมเป็นกลุ่มใหม่คือ กลุ่มที่(5,6) โดยมีจำนวนคำที่เหมือนกัน 3 คำ (**กบ,เขียด,คลอง**) แต่เมื่อทำการรวมเอกสารกันในระดับต่อไปอีก วิธี Hierarchical Agglomerative Clustering จะเอาเอกสารกลุ่ม (1,2) ไปเทียบกับกลุ่ม(3,4) และกลุ่ม(5,6) เพื่อหาคำที่เหมือนกับกลุ่ม (1,2)มากที่สุดจะได้รวมกลุ่มต่อไปเป็นกลุ่ม(1,2,3,4) หรือ กลุ่ม (1,2,5,6) แต่ถ้าพิจารณาสมาชิกในคำสำคัญของกลุ่ม(1,2) แล้วจะพบว่า ถ้ากลุ่ม(1,2) รวมกับกลุ่ม (3,4) เป็นกลุ่ม (1,2,3,4) แล้ว เอกสารฉบับที่ 1 ไม่น่าจะถูกเข้าไป

รวมด้วยเลยเพราะ มีคำสำคัญเหมือนกับในกลุ่ม(3,4) อยู่น้อย ในขณะที่เอกสารฉบับที่ 2 นั้นสมควรจะเข้าไปรวมเพียงฉบับเดียวในกลุ่มที่(3,4)มากกว่า เพราะมีคำสำคัญเหมือนกันอยู่มาก ที่จริงเอกสารฉบับที่ 1 ควรจะไปรวมกับกลุ่ม(5,6) มากกว่าเพราะ มีคำสำคัญเหมือนกันมากกว่า หรือถ้ากลุ่ม(1,2) ไปรวมกับกลุ่ม(5,6) เอกสารฉบับที่ 2 ก็ไม่น่าจะเข้าไปรวมด้วยเพราะ คำสำคัญไม่เกี่ยวข้องกับกลุ่ม(5,6) มากเหมือนกับเอกสารฉบับที่1 จากสาเหตุนี้จึงสมควรจะทำการจัดกลุ่มเอกสารโดยพิจารณาการรวมกลุ่มเอกสารแบบพิจารณาทีละเอกสารในทุกรอบและทุกระดับของการรวมกลุ่มเอกสาร เพื่อป้องกันการผิดพลาดที่เกิดจากการสะสมคำสำคัญดังกล่าวมาข้างต้น เพราะถึงอย่างไรถ้าเอกสารมีคำสำคัญมากที่สุดในกลุ่มใดเอกสารก็จะเข้าไปรวมกับกลุ่มนั้นเองโดยอัตโนมัติอยู่แล้ว เนื่องจากเมื่อเอาเอกสารนั้นออกมาพิจารณากลุ่มก็ต้องทำการตรวจสอบทุกกลุ่มอยู่แล้ว ซึ่งถ้าเอกสารยังมีคำสำคัญเหมือนในกลุ่มเดิมมากที่สุดก็จะเข้าไปอยู่อย่างเดิม แต่ถ้าไม่ใช่ก็จะย้ายไปกลุ่มอื่นที่เหมาะสมกว่า ซึ่งก็จะเกิดการรวมกลุ่มสมาชิกเพิ่มขึ้นอย่างถูกวิธี และเมื่อเอกสารรวมกลุ่มกันจนถึงระดับหนึ่งแล้ว เอกสารก็จะไม่ย้ายกลุ่มอีกต่อไป หรือเหลือกลุ่มเอกสารเพียงกลุ่มเดียวก็ได้ ซึ่งทำให้เราแน่ใจได้ว่าแต่ละรอบหรือระดับการรวมกลุ่มเอกสารน่าจะมีความถูกต้องมากที่สุดของสมาชิกในกลุ่ม ต่อจากนั้นเราก็สามารถจะเลือกระดับใดก็ได้ไปพิจารณาผลลัพธ์การจัดกลุ่มเอกสารของแต่ละวิธี ว่าวิธีไหนจะให้ผลการจัดกลุ่มเอกสารที่ดีกว่ากัน ที่ทำให้สมาชิกในกลุ่มมีเนื้อหาใกล้เคียงกันมากที่สุด

การจัดกลุ่มเอกสารได้มีแนวทางการวิจัยไว้บางส่วน ซึ่งต่างก็มีหัวข้อน่าสนใจและเป็นแนวทางที่ดีให้น่าศึกษาค้นคว้าต่อมากพอสมควร ดังตัวอย่างต่อไปนี้

4.2.1 วิธีการจัดกลุ่มเอกสารแบบหาความเหมือนกันของคำเพียงอย่างเดียว (Similarity Keywords)

การจัดกลุ่มเอกสารภาษาไทยด้วยวิธีนี้ จะพิจารณาดูคำสำคัญในแต่ละเอกสารเทียบกันดังอัลกอริทึมที่กล่าวมาแล้วข้างต้น ซึ่งจะดูว่ามีคำสำคัญของเอกสารที่นำมาเทียบกับกลุ่มที่จะเข้าไปรวมว่ามีความเหมือนกันมากน้อยเพียงไร โดยไม่สนใจความถี่คำสำคัญและจำนวนคำสำคัญเลย สนใจเพียงว่ามีคำเหมือนกันมากที่สุดกับกลุ่มใด ก็จะเข้าไปรวมกับกลุ่มเอกสารนั้น

วิธีการนี้จะตรวจดูว่ามีคำสำคัญระหว่างเอกสารที่นำมาเปรียบเทียบ (Unknown) กับกลุ่มคำสำคัญ(Cluster)ในฐานข้อมูลว่าจะอยู่ในกลุ่มใด โดยมีสูตรการพิจารณาดังนี้

$$Similarity = \frac{\text{int } er \text{ sec}}{I_a + I_b - \text{int } er \text{ sec}} \qquad (4.1)$$

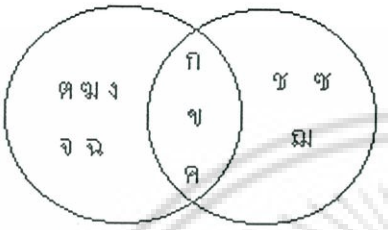
Similarity = ค่าความคล้ายคลึงของเอกสาร

intersec = จำนวนสมาชิกที่เหมือนกันในเอกสาร a และ b

I_a = จำนวนสมาชิกของเอกสาร a ทั้งหมด

I_b = จำนวนสมาชิกของเอกสาร b ทั้งหมด

ค่าของความคล้าย(Similarity) จะมีค่าสูงสุดคือ 1 และค่าต่ำสุดคือ 0 โดยค่า 1 จะบอกว่าเอกสารที่นำมาเปรียบเทียบ (Unknown) จะมีความคล้ายคลึงกับกลุ่มเอกสารนั้นๆมากที่สุด ส่วนค่า 0 จะหมายถึง เอกสารที่นำมาเปรียบเทียบกับกลุ่มเอกสารนั้นๆมีค่าความคล้ายต่ำที่สุด ค่าความคล้ายนี้ถ้ามองในรูปของเซตจะเป็นดังภาพข้างล่างนี้



รูปที่ 4.5 ภาพเปรียบเทียบความคล้ายของเอกสาร 2 เอกสาร

ตัวอย่างการคำนวณค่าความคล้าย (Similarity) (S) จากรูปที่ 4.5 โดยแทนด้วยสมการที่ 4.1 มี ดังนี้

$$S = \frac{3}{8+6-3} = \frac{3}{11} = 0.2727$$

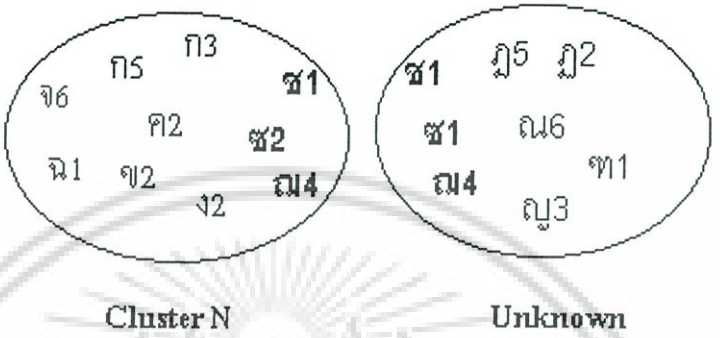
เอกสารทั้งสองในรูปภาพที่ 4.5 มีค่าความคล้ายกันอยู่เท่ากับ 0.2727

ผลของการจัดกลุ่มเอกสารด้วยวิธีหาความเหมือนกันของคำเพียงอย่างเดียว (Similarity Keywords) ดูได้จากหัวข้อผลการทดลอง

4.2.2 การพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency)(KF)

จากแนวทางและวิธีการจัดกลุ่มเอกสารด้วยวิธีพิจารณาการเหมือนกันของคำเพียงอย่างเดียว (Similarity Keywords) ยังน่าจะมีแนวทางอื่นที่ดีกว่าหรือไม่ในการพิจารณาการแบ่งกลุ่มเอกสารภาษาไทย ผู้วิจัยจึงได้ลองวิธีของการให้ค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Keywords Frequency) ซึ่งเป็นงานทดลองวิจัยของ Sahami Mehran [10] เป็นการจัดกลุ่มเอกสารอีกวิธีหนึ่งที่พิจารณาคำสำคัญ โดยพิจารณาในรูปของความน่าจะเป็นของความถี่คำในเอกสารที่ intersec กันเป็นหลักว่าจะเข้าไปรวมกันได้หรือไม่ ซึ่งเอกสาร unknown จะรวมกับกลุ่มที่มีความถี่คำสำคัญที่เหมือนกันมากที่สุด ในการศึกษาวิธี Keywords Frequency นี้ก็เพื่อจะเป็นแนวทางต่อไปในการจะปรับปรุงวิธีการจัดกลุ่มเอกสารภาษาไทยให้ดีขึ้น โดยการดูมุม

มองของคำสำคัญในลักษณะต่างๆที่น่าจะเกี่ยวเนื่องกันในการจัดกลุ่มเอกสาร ทั้งในแง่ของการเหมือนกันของคำเพียงอย่างเดียว เปรียบเทียบกับถ้าพิจารณาความถี่คำหรือจำนวนคำสำคัญด้วย โดยดูว่าผลจะแตกต่างและดีกว่ากันอย่างไรบาง ลักษณะการเปรียบเทียบเอกสารว่าจะรวมกลุ่มกันได้หรือไม่ ในวิธีการพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency)(KF) จะคล้ายกับรูปที่ 4.8



รูปที่ 4.6 ภาพแสดงการเปรียบเทียบการจัดกลุ่มแบบ PDO

จากรูปที่ 4.6 โดยแทนพยัญชนะไทยแทนคำ 1 คำในกลุ่ม และตัวเลขข้างพยัญชนะไทยคือ ความถี่ของจำนวนคำนั้นในกลุ่ม เช่น ณ6 หมายถึง คำว่า "ณ" มีความถี่ของคำอยู่ในกลุ่ม unknown เท่ากับ 6 คำ อักษรตัวทึบ(ช,ซ,ฌ) แสดงว่าคำนั้นเหมือนกันหรือ (intersec) กันกับกลุ่มเปรียบเทียบ สมการวิธีการจัดกลุ่มเอกสารแบบ การพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency) (KF) จะเป็นดังสมการที่ 4.2 และ 4.3

$$\sum_{w \in d_i \cap d_j} P(Y_i = w \mid d_i) \bullet P(Y_j = w \mid d_j) \qquad (4.2)$$

$$P(Y = w \mid d) = \frac{\xi(w,d)}{\sum_{w \in d} \xi(w,d)} \qquad (4.3)$$

P(Y = w|d) = โอกาสที่จะพบ w ในเอกสาร d

di คือ เอกสารกลุ่มที่ i

dj คือ เอกสารกลุ่มที่ j

$\xi(w,d)$ = จำนวนที่พบคำ w ในเอกสาร d

$\sum_{w \in d} \xi(w,d)$ = ผลรวมของความถี่ของคำที่ปรากฏในเอกสาร d

ตัวอย่างการคำนวณแบบการพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency)(KF) จากรูปที่ 4.6 โดยกำหนดให้ i = (Cluster N) , j = Unknown

$$PDO = \sum_{w \in di \cap dj} P(Yi = w|di) * P(Yj = w|dj)$$

$$= ((1/28) * (1/23)) + ((2/28) * (1/23)) + ((4/28) * (4/23))$$

$$= (0.036 * 0.043) + (0.071 * 0.043) + (0.143 * 0.174)$$

$$= 0.079 + 0.114 + 0.317$$

$$= 0.51$$

เพราะฉะนั้นค่าการคำนวณที่ได้แบบการพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency)(KF) จะเท่ากับ 0.51 ซึ่งถ้าเอา unknown ตรวจดูทุกกลุ่มแล้ว ตามสมการที่ 4.2 และตัวอย่างดังกล่าว ค่าที่คำนวณได้นี้กลุ่มไหนที่มีค่า PDO มากที่สุด ก็จะรวมเอกสาร unknown เข้ากับกลุ่มที่มีค่าPDO สูงสุดนั้น แล้วก็คำนวณตามสูตรนี้และอัลกอริที่มตามที่กล่าวมาแล้วทุกเอกสารและทุกกลุ่มเอกสาร จนเอกสารไม่มีการเปลี่ยนกลุ่มแล้วจึงหยุด แล้วนำผลลัพธ์ที่ได้มาวัดด้วยค่า Precision ,Recall และ F-measure เปรียบเทียบกับการจัดกลุ่มด้วยมนุษย์ และเปรียบเทียบกับวิธีการจัดกลุ่มอันอื่นๆต่อไป ว่าการจัดกลุ่มเอกสารได้ดีเพียงไร มีความน่าเชื่อถือมากหรือน้อย มีเปอร์เซ็นต์การจัดเอกสารได้ถูกกลุ่มกี่เปอร์เซ็นต์ และเนื้อหาเอกสารในกลุ่มใกล้เคียงกันมากน้อยอย่างไร

ผลการจัดกลุ่มเอกสารภาษาไทยด้วยวิธีพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกันระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency) (KF) ดูได้จากหัวข้อผลการทดลอง

บทที่ 5
การจัดกลุ่มเอกสารภาษาไทย

จากแนวทางการจัดกลุ่มเอกสารดังที่กล่าวมาแล้วในบทที่4 ซึ่งเป็นแนวทางในการวิจัยต่อเพื่อให้ได้การจัดกลุ่มเอกสารภาษาไทยที่ดีขึ้นนั้น ผลลัพธ์ของการจัดกลุ่มที่ได้ในบทที่ 4 ยังไม่ดีนักควรจะมีการปรับปรุงบ้างในบางอย่าง ถึงแม้จะรู้แนวทางในการจัดกลุ่มบางว่าควรต้องพิจารณาในส่วนของคำสำคัญแน่นอน แต่จะมองในมุมเพียงว่ามีคำที่เหมือนกันเพียงอย่างเดียวก็ไม่ดีนัก หรือจะมองเพียงแต่จำนวนคำที่เหมือนกันว่ามีจำนวนมากๆแล้วก็เข้าไปรวมกลุ่มเอกสารด้วยก็ไม่น่าจะถูกอีก เพราะคำบางคำมีจำนวนมากก็อาจไม่ใช่คำสำคัญ(Keywords)ก็ได้ ดังเช่น คำเชื่อมต่างๆ(และ,หรือ,กับ,แล้ว ฯลฯ) หรือคำที่มีอยู่ทั่วไป เช่น ทำ,ต่อไป,นี้,มี,เป็นต้นฯลฯ ดังที่กล่าวมาแล้วในส่วนที่ต้องกำจัดคำทิ้งไปที่จะมารบกวนระบบ

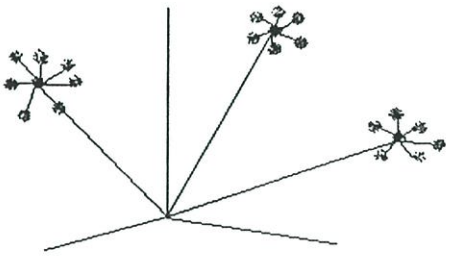
ฉะนั้นผู้วิจัยได้ทดลองหาแนวทางใหม่ดู โดยพิจารณาในมุมมองดังจะกล่าวถึงต่อไปซึ่งจะได้นำผลลัพธ์มาเปรียบเทียบกันในทุกวิธีเพื่อจะดูว่าวิธีไหนให้ผลลัพธ์ที่ดีที่สุด เมื่อใช้เกณฑ์การวัดผลด้วยเครื่องมือวัด Precision, Recall และ ค่า F-measure เมื่อเทียบกับการจัดกลุ่มด้วยมนุษย์เป็นตัวมาตราฐาน

5.1 การหาค่าระยะห่างของคำสำคัญระหว่างกลุ่มเอกสาร (Distance)

การพิจารณาระยะห่างของความถี่กลุ่มคำสำคัญ เป็นการพิจารณาค่าของ Normalize keywords ซึ่งเป็นค่าของความถี่คำสำคัญของคำนั้นๆที่พบบ่อยเท่าไรเมื่อเทียบกับคำสำคัญทั้งหมดของเอกสาร

$$Normalize\ \ keyword\ = \frac{Frequency\ \ keyword}{Total\ \ Frequency\ \ keyword} \qquad (5.1)$$

ค่า Normalize keyword นี้จะนำมาใช้ในการพิจารณา โดยตั้งสมมติฐานว่าเอกสารที่อยู่ในกลุ่มเดียวกันเนื้อหาในเอกสารหรือคำสำคัญต่างๆของเอกสาร น่าจะมีค่าระยะห่างของความถี่กลุ่มคำสำคัญอยู่ใกล้เคียงกัน เมื่อมองในรูปเวกเตอร์ 3 มิติ จะพบว่ากลุ่มต่างๆกระจายอยู่ในเวกเตอร์ โดยกลุ่มต่างๆเหล่านั้นจะมีจุดต่างๆแทนคำในเอกสารเป็นสมาชิกในกลุ่ม ซึ่งกลุ่มจุดเหล่านั้นจะแทนเอกสารในกลุ่มนั้นเอง



รูปที่ 5.1 กลุ่มเอกสารต่างๆในมุมมองของเวกเตอร์ 3 มิติ

โดยค่าระยะห่างของสมาชิกในกลุ่มจะเทียบจากสูตร

$$Distance = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + ...} \qquad (5.2)$$

Distance = ระยะห่างของ Normalize keyword ของเอกสารที่ 1 กับ เอกสารที่ 2

a_1 = Normalize keyword ของคำว่า a ในเอกสารที่ 1 (ฐานข้อมูล)

a_2 = Normalize keyword ของคำว่า a ในเอกสารที่ 2 (unknown)

b_1 = Normalize keyword ของคำว่า b ในเอกสารที่ 1 (ฐานข้อมูล)

b_2 = Normalize keyword ของคำว่า b ในเอกสารที่ 2 (unknown)

c_1 = Normalize keyword ของคำว่า c ในเอกสารที่ 1 (ฐานข้อมูล)

c_2 = Normalize keyword ของคำว่า c ในเอกสารที่ 2 (unknown)

ขั้นตอนการจัดกลุ่มแบบการหาความคล้ายของเอกสารที่พิจารณาคำที่เหมือนกัน ร่วมกับการหาค่าระยะห่างของคำสำคัญระหว่างกลุ่มเอกสารมีดังนี้ เมื่อได้หน่วยคำสำคัญที่แยกออกจากประโยคภาษาไทยแล้ว เราต้องทำการสอนระบบเพื่อให้ระบบทำการแยกกลุ่มเอกสารออกมาให้ได้ โดยเริ่มต้นในฐานข้อมูลจะมี 1Doc 1Group ก่อน จะต้องทำการสุ่มเอกสาร(unknown)เรียงลำดับตามหมายเลขเอกสาร ซึ่งเป็นตัวแทนของเอกสารแต่ละกลุ่มออกมาก่อน แล้วหาค่า Normalize keyword ของเอกสาร ต่อจากนั้นก็เริ่มทำการจัดกลุ่ม ซึ่งจะเอาเอกสาร Unknown มาเปรียบเทียบดูว่ามีความคล้ายกับกลุ่มเอกสารใดบ้างในฐานข้อมูลโดยใช้สมการ Similarity สมการที่ 4.1 หาดูว่าค่า Similarity ของเอกสาร Unknown เทียบกับเอกสารในฐานข้อมูลกลุ่มใดมีค่ามากที่สุด เอกสาร Unknown นี้ก็น่าจะอยู่ในกลุ่มเดียวกันกับเอกสารในฐานข้อมูลที่มีค่า Similarity มากที่สุด แต่ถ้าค่า Similarity มีค่าเป็น 0 ซึ่งหมายถึง ไม่มีเอกสารใดเลยในฐานข้อมูลคล้ายกับเอกสาร Unknown ก็จะเอาเอกสาร Unknown นี้มาสร้างเป็นกลุ่มเอกสารใหม่ในฐานข้อมูล กรณีที่เอกสาร Unknown คล้ายกับเอกสารในกลุ่มฐานข้อมูลกลุ่มใดมากที่สุด ต่อจากนั้นจะเอากลุ่มนั้นมาพิจารณาทดสอบหาระยะห่างของความถี่กลุ่มคำสำคัญ (Distance)ว่ามีค่าน้อยกว่าหรือเท่ากับค่า Distance ที่กำหนดหรือไม่ จากสมการที่ 5.2 โดยถ้าค่าน้อยกว่าหรือเท่ากับค่า Distance ที่กำหนดจะทำการรวม

เอกสาร Unknown นั้นกับกลุ่มเอกสารในฐานข้อมูลเป็นกลุ่มเดียวกัน แต่ถ้าDistance เกินกว่าค่าที่กำหนดจะเอาเอกสาร Unknown นั้นไปสร้างเป็นกลุ่มเอกสารใหม่ในฐานข้อมูล จะทำการสอน (Training)เช่นนี้จนกว่าจะไม่เกิดการเปลี่ยนแปลงกลุ่มเอกสารในฐานข้อมูลแล้ว จึงเก็บผลของค่า Distance ,สมาชิกในกลุ่มเอกสารแต่ละกลุ่ม และค่า Normalize Keyword เอาไว้ ต่อจากนั้นจึงทำการเปลี่ยนระยะหา Distance ค่าใหม่ แล้วทำการสอนใหม่อีก โดยเริ่ม 1Doc I Group แต่เปลี่ยนค่า Distance จนไม่เกิดกลุ่มเอกสารใหม่เช่นกัน พร้อมทั้งเก็บค่าต่างๆไว้ด้วยเช่นกัน แล้วจึงนำมาวิเคราะห์ผลลัพธ์ที่ได้ว่า จะกำหนดใช้ค่า Distance เท่าไรจึงเหมาะสมที่สุด

ค่า Distance ของเอกสารเริ่มต้นกลุ่มในฐานข้อมูล จะคิดเทียบกับจุดกำเนิด (Origin) ก่อนตามสมการที่ 5.3 ส่วนเอกสาร Unknown ที่จะนำมาเทียบในกลุ่มอันต่อๆไป เมื่อเกิดการรวมเอกสารบ้างแล้วในกลุ่ม ในสมการที่5.2 จะเทียบจากค่า Normalize keyword เฉลี่ยของแต่ละคำสำคัญในกลุ่มนั้นๆ ดังสมการที่ 5.4 (ในกลุ่มมีคำสำคัญคำเดียวกันมากกว่า 1 คำและหลายความถี่ เนื่องจากมีเอกสารหลายฉบับในกลุ่ม จึงคิดค่า Normalize เฉลี่ย)

$$Distance\ (D_o) = \sqrt{(a_1 - 0)^2 + (b_1 - 0)^2 + (c_1 - 0)^2 + ...} \qquad (5.3)$$

Distance(D_0) = ระยะห่างของ Normalize keyword ของเอกสารเริ่มต้นในฐานข้อมูลเทียบกับ Origin

a_1 = Normalize keyword ของคำว่า a ในเอกสารเริ่มต้นในฐานข้อมูล

b_1 = Normalize keyword ของคำว่า b ในเอกสารเริ่มต้นในฐานข้อมูล

c_1 = Normalize keyword ของคำว่า c ในเอกสารเริ่มต้นในฐานข้อมูล

$$Distance = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + ...} \qquad (5.4)$$

Distance = ระยะห่างของ Normalize keyword ของเอกสารที่ 1(ฐานข้อมูล) กับ เอกสารที่ 2 (Unknown)

a_1 = Normalize keyword ของคำว่า a เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

a_2 = Normalize keyword ของคำว่า a ในเอกสาร Unknown

b_1 = Normalize keyword ของคำว่า b เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

b_2 = Normalize keyword ของคำว่า b ในเอกสาร Unknown

c_1 = Normalize keyword ของคำว่า c เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดใน

ฐานข้อมูล

c_2 = Normalize keyword ของคำว่า c ในเอกสาร Unknown

เมื่อได้ค่า Distance มาแล้วก็จะพิจารณาค่า Distance นั้นว่าจะรวมเป็นกลุ่มเดียวกันกับกลุ่มเอกสารในฐานข้อมูลหรือไม่ ตามค่า Distance (FD) ที่ลองกำหนดดู ตามสูตรสมการที่ 5.5 ถ้าค่า Distance น้อยกว่าหรือเท่ากับค่า RD จะรวมเอกสาร Unknown เข้าเป็นกลุ่มเดียวกันกับกลุ่มในฐานข้อมูลที่เปรียบเทียบในกลุ่มนั้น แต่ถ้ามากกว่าจะเอาเอกสาร Unknown สร้างเป็นกลุ่มใหม่ในฐานข้อมูล ทำการสอนเช่นนี้ไปจนไม่เกิดกลุ่มใหม่ในแต่ละ Distance (FD) แล้วนำผลที่ได้มาพิจารณาหาค่า Distance(FD) ที่เหมาะสมต่อไป

$$RD = \frac{FD \times D_0}{100} \qquad (5.5)$$

RD = ค่าระยะห่าง Distance ที่ต้องการ

FD = ค่าระยะห่าง Distance ที่กำหนด (40,60,90,100,300)

D_0 = ค่าระยะห่าง Distance ของกลุ่มคำสำคัญในฐานข้อมูลเริ่มต้น สมการที่ 4.4

ในการคำนวณหาค่า Distance นั้นจากสมการ 5.2 และ 5.4 ในวิทยานิพนธ์นี้ ได้ทดลองใช้คำที่เป็นค่าตัวแปร normalize keywords ในสมการ โดยเอาเฉพาะคำที่ intersec กันมาแทนในสมการแล้วเก็บผลแต่ละ Distance เปรียบเทียบกับการแทน normalize keywords ในสมการทุกคำทั้งของ unknown และของกลุ่มเปรียบเทียบที่มีค่า Similarity สูงสุด (สมการที่ 4.1) ทั้งคำ intersec และไม่ intersec ดังแสดงในสมการข้างล่างนี้

สมการหาค่า Distance ระหว่าง unknown กับ กลุ่มที่มีค่าความคล้ายหรือค่าการเหมือนกันของคำสูงสุด ที่เอาเฉพาะคำสำคัญที่ intersec กันมาใช้คำนวณ จากสมการที่ 5.4 จะได้

$$Distance = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + ...}$$

Distance = ระยะห่างของ Normalize keyword ของเอกสารที่ 1(ฐานข้อมูล) กับ เอกสารที่ 2 (Unknown)

a_1 = Normalize keyword ของคำว่า a เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

a_2 = Normalize keyword ของคำว่า a ในเอกสาร Unknown

b_1 = Normalize keyword ของคำว่า b เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

b_2 = Normalize keyword ของคำว่า b ในเอกสาร Unknown

c_1 = Normalize keyword ของคำว่า c เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

c_2 = Normalize keyword ของคำว่า c ในเอกสาร Unknown

สมการหาค่า Distance ระหว่าง unknown กับ กลุ่มที่มีค่าความคล้ายหรือค่าการเหมือนกันของคำสูงสุดที่เอาคำสำคัญทั้งหมดทั้งที่ intersec และไม่ intersec กันมาใช้คำนวณ จะได้สมการดังนี้

$$Distance = \sqrt{\begin{aligned}&(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + \ldots..\\ &+ (d_1 - 0)^2 + (e_1 - 0)^2 + \ldots. + (0 - f_2)^2 + (0 - g_2)^2 \ldots..\end{aligned}} \quad (5.6)$$

Distance = ระยะห่างของ Normalize keyword ของเอกสารที่ 1(ฐานข้อมูล) กับ เอกสารที่ 2 (Unknown)

a_1 = Normalize keyword ของคำว่า a เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

a_2 = Normalize keyword ของคำว่า a ในเอกสาร Unknown

b_1 = Normalize keyword ของคำว่า b เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

b_2 = Normalize keyword ของคำว่า b ในเอกสาร Unknown

c_1 = Normalize keyword ของคำว่า c เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

c_2 = Normalize keyword ของคำว่า c ในเอกสาร Unknown

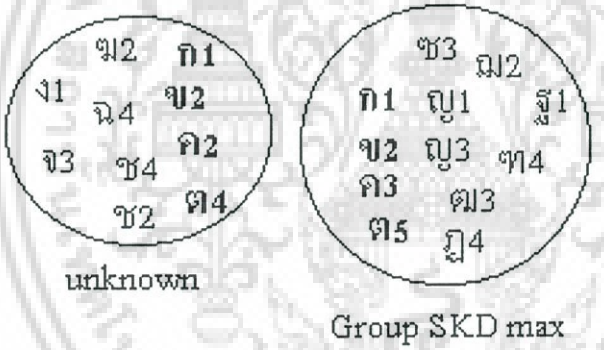
d_1 = Normalize keyword ของคำว่า d เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

e_1 = Normalize keyword ของคำว่า e เฉลี่ยในเอกสารกลุ่มที่มีค่าSimilarityมากสุดในฐานข้อมูล

f_2 = Normalize keyword ของคำว่า f ในเอกสาร Unknown

g_2 = Normalize keyword ของคำว่า g ในเอกสาร Unknown

นอกจากการหาค่าระยะห่างของคำสำคัญระหว่างกลุ่มเอกสาร (Distance) นี้แล้ว ยังได้ทดลองหาค่าความคล้ายของเอกสารแบบพิจารณาคำที่เหมือนกัน (Similarity Keywords) ร่วมด้วยไปพร้อมกันที่เรียกว่า วิธี Similarity Keywords Distance (SKD) โดยเริ่มแรกระบบจะทำการหาค่า Similarity Keywords ก่อนว่าเอกสารUnknown ที่จะนำมารวมเข้ากลุ่มมีค่า Similarity Keywords มากที่สุดในกลุ่มใด ต่อจากนั้นจึงทำการหาค่าระยะห่างของคำสำคัญระหว่างกลุ่มเอกสาร Distance ว่าเอกสาร Unknown นั้นเปรียบเทียบกับกลุ่มที่ มีค่า Similarity Keywords มากที่สุด อยู่ในระยะ Threshold ที่น้อยกว่าที่กำหนดหรือไม่ ถ้าน้อยกว่าก็ จะรวมเอกสาร Unknown นั้นเข้าไว้ในกลุ่ม แต่ถ้าค่า Distance มากกว่าค่า Threshold เอกสารนั้นก็จะไปสร้างเป็นกลุ่มเอกสารใหม่ จากวิธีนี้จะใช้ทดสอบดูผลลัพธ์การจัดกลุ่มเอกสารเปรียบเทียบกับ การหาความคล้ายของคำสำคัญระหว่างกลุ่มเอกสารแบบพิจารณาน้ำหนักตามความถี่คำ ว่าจะให้ผลลัพธ์อย่างไร ซึ่งดูผลการทดสอบได้ที่หัวข้อผลการทดลองข้างล่าง



รูปที่ 5.2 แสดงการพิจารณารวมกลุ่มเอกสารแบบ SKD

ตัวอย่างการคำนวณแบบ SKD พิจารณาจากรูปที่ 5.2 โดยแทนพยัญชนะไทยแทนคำ 1 คำในกลุ่ม และตัวเลขข้างพยัญชนะไทยคือ ความถี่ของจำนวนคำนั้นในกลุ่ม เช่น จ3 หมายถึง คำว่า "จ" มีความถี่ของคำอยู่ในกลุ่ม unknown เท่ากับ 3 คำ หรือพบ 3 ครั้งในเอกสารกลุ่ม unknown สมมติให้unknown หาค่า Similarityกลุ่มทุกกลุ่มในฐานข้อมูลแล้วได้กลุ่มที่มีค่าสูงสุดคือ Group SKD max จะคำนวณการหาตัวอย่างแบบ SKD ได้ดังนี้ แทนค่า Similarity ในสมการที่ 4.1

Similarity = intersec / (Ia+Ib - intersec) : Ia = unknown ; Ib = Group SKD max

= 4 / (9+11- 4)

= 0.25

หาค่า Distance ดูเฉพาะคำ intersec ของกลุ่ม Group SKD max ที่เทียบกับจุดกำเนิดก่อน Distance (D_0) โดยแทนค่าในสมการที่ 5.3

$$Distance = \sqrt{(a_1 - 0)^2 + (b_1 - 0)^2 + (c_1 - 0)^2 + ...}$$

Distance (D_0) = $((1/32)^2 + (2/32)^2 + (3/32)^2 + (5/32)^2)^{1/2}$

= $3.8 * 10^{-2}$

สมมติพิจารณาค่าหา Distance ดูเฉพาะคำ intersec ที่ threshold 60% โดยพิจารณาคำสำคัญระหว่าง unknown กับ Group SKD max แบบ intersec อย่างเดียว แทนค่าในสมการที่ 5.2 จะได้

$$Distance = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + ...}$$

Distance = $(((1/32)-(1/25))^2 + ((2/32)-(2/25))^2 + ((3/32)-(2/25))^2 + ((5/32)-(4/25))^2)^{1/2}$

= $1.285 * 10^{-7}$

พิจารณาค่าหา Distance ดูเฉพาะคำ intersec ที่ threshold 60 ดูว่าจะรวมเอกสาร unknown กับกลุ่ม Group SKD max หรือไม่ แทนค่าในสมการที่ 4.6

$$RD = \frac{FD \times D_0}{100}$$

RD = $(60 * (3.8*10^{-2})) /100$

= $2.28 *10^{-2}$

เพราะฉะนั้นจากค่าที่คำนวณได้ RD มีค่ามากกว่า Distance ดูเฉพาะคำ intersec ที่ threshold 60% ($1.285 * 10^{-7}$) จึงรวมเอกสาร unknown กับกลุ่ม Group SKD max นี้เข้าไว้ด้วยกัน

ตัวอย่างการคำนวณหาค่า Distance 60% เช่นกัน ที่พิจารณาคำทุกคำทั้ง intersec และไม่ intersec ของทั้ง unknown และ Group SKD max แทนค่าคำสำคัญทั้ง2 ในสมการที่5.6

$$Distance = \sqrt{\begin{matrix}(a_1 - a_2)^2 + (b_1 - b_2)^2 + (c_1 - c_2)^2 + ... \\ + (d_1 - 0)^2 + (e_1 - 0)^2 + ... + (0 - f_2)^2 + (0 - g_2)^2 .....\end{matrix}}$$

Distance = $(((1/32)-(1/25))^2 + ((2/32)-(2/25))^2 + ((3/32)-(2/25))^2 + ((5/32)-(4/25))^2 +$

$(3/32)^2 + (2/32)^2 + (1/32)^2 + (((1+3)/2)/32)^2 + (4/32)^2 + (3/32)^2 + (4/32)^2 +$

$(2/25)^2 + (1/25)^2 + (4/25)^2 + (3/25)^2 + (((4+2)/2)/25)^2)^{1/2}$

Distance = $1.22 * 10^{-1}$

เพราะฉะนั้นจากค่าที่คำนวณได้ RD ($2.28 * 10^{-2}$) มีค่าน้อยกว่า Distanceที่พิจารณาคำทุกคำทั้ง intersec และไม่ intersec ของทั้ง unknown และ Group SKD max ที่ threshold 60% ($1.22 * 10^{-1}$) จึงไม่รวมเอกสาร unknown กับกลุ่ม Group SKD max นี้เข้าไว้ด้วยกัน

5.2 การพิจารณาน้ำหนักของคำที่เหมือนกันตามจำนวนเอกสารในกลุ่ม (Document Frequency) (DF)

จากหัวข้อที่ 5.1 วิธีการจัดเอกสารแบบพิจารณาความคล้ายกันของคำสำคัญร่วมกับการหาระยะห่างระหว่างคำสำคัญ (Similarity Keywords Distance)(SKD) ซึ่งเป็นการพิจารณาดูเน้นไปที่คำสำคัญ (Keywords) เพียงอย่างเดียว ทำให้ได้ผลที่ไม่ดีนัก และต้องทำการทดสอบหาค่า ระยะห่างของกลุ่มคำสำคัญ Threshold หลายๆค่าจนกว่าจะได้ค่าที่เหมาะสมที่สุด ซึ่งถ้ากำหนดค่า Threshold น้อยๆ จะทำให้ได้เนื้อหาในกลุ่มเอกสารมีความใกล้เคียงกันสูงก็จริง แต่จะได้จำนวนกลุ่มเอกสารมากตามไปด้วย อันเนื่องจากเอกสารไม่สามารถเข้ากลุ่มใดได้ง่ายๆ จึงต้องออกมาสร้างเป็นกลุ่มใหม่เฉพาะ ทำให้ในแต่ละกลุ่มเอกสารมีจำนวนสมาชิกในกลุ่มน้อย แต่จะได้เนื้อหาเอกสารใกล้เคียงกันสูง โดยในทางตรงกันข้ามถ้ากำหนดค่า Threshold มากๆ จะทำให้ได้เนื้อหาเอกสารในกลุ่มไม่ใกล้เคียงกัน แต่จะได้จำนวนกลุ่มเอกสารน้อย เนื่องจากกลุ่มเอกสารจะมีจำนวนสมาชิกมากนั้นเอง และถ้ากำหนดค่า Threshold กลางๆ ระบบการจัดกลุ่มเอกสารก็จะไม่หยุดทำงานได้ง่ายๆ เนื่องจากเอกสารจะมีการเปลี่ยนกลุ่มไปมาตลอดเวลา ฉะนั้นจึงต้องหยุดระบบเองและนำมาหาค่า Threshold ที่เหมาะสมในการจัดกลุ่มเอกสารเอง จากเหตุผลดังกล่าวมานี้จึงได้ลองหาแนวทางใหม่ที่พิจารณาคำที่เหมือนกันตามจำนวนเอกสารดูบาง (Document Frequency)

แนวความคิดวิธีนี้คือ กลุ่มเอกสารที่กล่าวถึงเรื่องอะไรเอกสารสมาชิกในกลุ่มนั้น ก็น่าจะมีคำเหมือนๆกันปรากฏอยู่มากกว่าต่างกลุ่มกัน และคำที่ต่างกันในเอกสารกลุ่มเดียวกัน ควรที่จะถูกลดค่าน้ำหนักลงเพื่อไม่ให้เกิดค่าผิดพลาดจากคำที่ไม่ใช่คำสำคัญ (keywords) จริงๆ หรือจะกล่าวอีกในหนึ่งในแนวความคิดของ วิธีการพิจารณาคำสำคัญที่เหมือนกันตามจำนวนเอกสาร (Document Frequency) คือ เอกสารในกลุ่มเดียวกันน่าจะมีคำที่เหมือนกัน ฉะนั้นจากแนวคิดนี้จึงได้ถูกทำการทดลองวิจัยผลการจัดกลุ่มเอกสารภาษาไทยดู

การจัดกลุ่มเอกสารภาษาไทยด้วยวิธีนี้ (Document Frequency) จะพิจารณาเฉพาะจำนวนเอกสารในกลุ่ม (Cluster) กับคำสำคัญระหว่างเอกสาร Unknown ที่กำลังเปรียบเทียบเท่านั้น ไม่คำนึงถึงความถี่ของคำสำคัญเหมือนอย่างแบบ PDO หรือ Keyword Frequency และไม่เหมือนกับการหาความคล้ายกันของคำเพียงอย่างเดียว (Similarity Keywords) ดังที่กล่าวมาแล้วในแบบการหาความเหมือนกันของคำ (SKD) เพราะแบบ SKD จะนับจำนวนคำที่มีการ intersec ว่ามีกี่คำ แต่วิธี (DF) นี้จะพิจารณาดูว่ามีคำที่เหมือนกันในเอกสาร Unknown กับจำนวนเอกสารในกลุ่ม (Cluster) ที่กำลังเปรียบเทียบกันมากน้อยกี่เอกสาร โดยสนใจเฉพาะว่ามีคำสำคัญใน Unknown เหมือนกับคำในเอกสารในกลุ่ม (Cluster) กี่เอกสารก็นำจำนวนเอกสารในกลุ่ม Cluster มาบวกกัน ส่วนคำที่ไม่เหมือนกันก็เอา1คำของคำนั้นในเอกสาร unknown หารด้วยผลรวมจำนวนคำแต่ละคำในเอกสาร unknown เอามาลบกัน โดยไม่สนใจความถี่คำสำคัญในเอกสาร Unknown

เช่นกัน รูปแบบสมการการจัดกลุ่มเอกสารภาษาไทย โดยพิจารณาคำสำคัญที่เหมือนกันตามจำนวนเอกสารในกลุ่ม (Document Frequency) (DF) เป็นดังสมการที่ 5.7

$$DF = \sum_{w \in d_i \cap C_j} \left(\frac{DocF(w, C_j)}{\sum_{W \in C_j} DocF(w, C_j)} \right) - \sum_{w \in d_i \& w \notin C_j} \frac{1}{W\#(d_i)} \qquad (5.7)$$

$DocF(w, C_j)$ = จำนวนเอกสารในกลุ่มที่ j ที่พบคำ w

$W\#(d_i)$ = จำนวนคำที่พบในเอกสาร d_i

รูปที่ 5.3 ภาพตัวอย่างการจัดกลุ่มแบบ DF

การคำนวณแบบพิจารณาคำสำคัญที่เหมือนกันตามจำนวนเอกสารในกลุ่ม (Document Frequency) (DF) พิจารณาจากรูปที่ 5.3 สมมติให้กรอบสี่เหลี่ยมใหญ่ คือ กลุ่มเอกสาร (Cluster) ที่ประกอบด้วยเอกสารจำนวน 3 เอกสาร และแต่ละเอกสารประกอบด้วยคำ 3 คำ ที่จะมาเทียบหาว่าวงรีที่อยู่นอกกรอบสี่เหลี่ยมใหญ่ ซึ่งเป็นเอกสาร Unknown ประกอบด้วยคำ 3 คำ (ข1,จ4,ง5) จะรวมกลุ่มกับกรอบสี่เหลี่ยมใหญ่ Cluster ได้หรือไม่ โดยใช้สมการที่ 5.7 ทำการคำนวณจัดกลุ่มเอกสาร ระหว่างเอกสาร unknown กับกลุ่ม Cluster ซึ่งค่าของ Document Frequency (DF) จะมีค่าตั้งแต่ –1 ถึง 1 โดยที่ค่า –1 จะหมายถึงเอกสาร Unknown ไม่คล้ายกับกลุ่ม Cluster ใดเลย แต่ถ้าเป็น 1 หมายถึงเอกสาร Unknown เหมือนกับกลุ่มเอกสาร (Cluster) นั้นมากจนเป็นกลุ่มเอกสารเดียวกัน ตัวอย่างการคำนวณหาค่า Document Frequency (DF) จากรูปที่ 5.3 แทนด้วยสมการที่ 5.7 จะได้ค่าการคำนวณดังด้านล่าง วิธีนี้จะไม่พิจารณาความถี่คำ จะพิจารณาเฉพาะจำนวนเอกสารในกลุ่ม Cluster ที่มีคำในกลุ่มเอกสารที่เหมือนกับ Unknown และต่างกับ Unknown เท่านั้น ในตัวอย่างนี้ (รูปที่5.3) ส่วนที่ intersec กันมี 2 คำคือ “ข” ในCluster มี

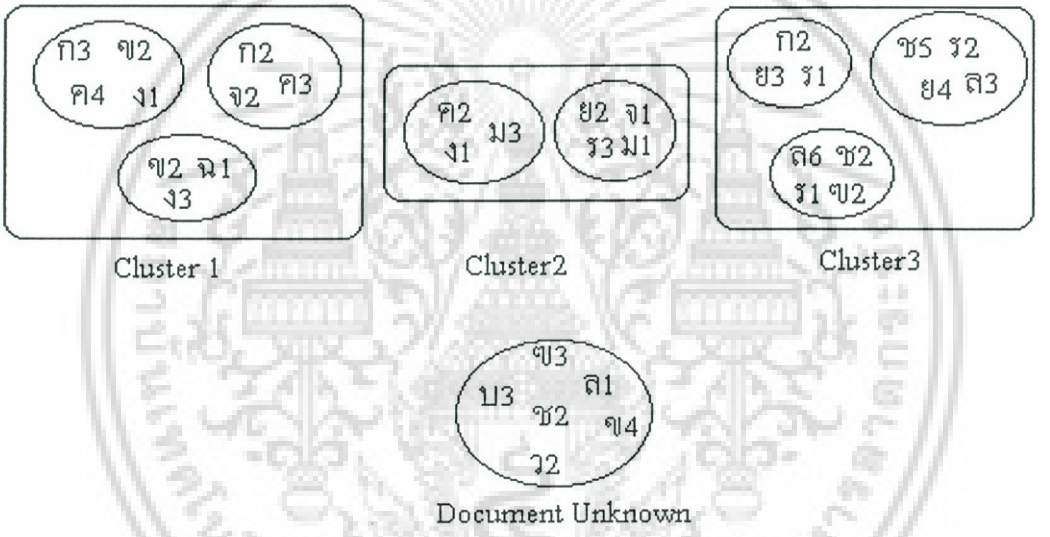
อยู่ 3 เอกสาร กับ "ง" ใน Cluster มี 1 เอกสาร ส่วนที่ไม่ intersecกันมี1คำ "จ" กลุ่มเอกสาร (Cluster) ในกรอบใหญ่มีทั้งหมด 9 คำ และเอกสาร(Unknown)ในวงรีมีทั้งหมด 3คำ

$$DF = (ข + ง) - จ$$

$$= [(3/9) + (1/9)] - (1/3)$$

$$= 0.11$$

ตัวอย่างการคำนวณวิธี การพิจารณาคำสำคัญที่เหมือนกันตามจำนวนเอกสารในกลุ่ม (Document Frequency) (DF) แล้วทำการเข้าไปรวมกลุ่มกับกลุ่มที่มีค่า Document Frequency (DF) มากที่สุด แสดงผลด้านข้างล่าง



รูปที่ 5.4 ภาพการพิจารณาการจัดกลุ่มเอกสารแบบ Document Frequency

จากรูปที่ 5.4 จะหาว่า Document Unknown จะเข้าไปรวมกลุ่มกับ Cluster ใดได้บางมีวิธีพิจารณาดังนี้ นำ Document Unknown ไปตรวจสอบกับ Cluster ทุก Cluster ก่อน ว่าจะได้ค่า Document Frequency (DF) มากที่สุดหรือไม่ตามสมการที่ 5.7 ดังนี้

พิจารณา Document Unknown เปรียบเทียบกับ Cluster 1

$$DF = \sum_{w \in d_i \cap C_j} \left(\frac{DocF(w, C_j)}{\sum_{w \in C_j} DocF(w, C_j)} \right) - \sum_{w \in d_i \& w \notin C_j} \frac{1}{W\#(d_i)}$$

DF = (ข) – (ฃ+ล+ช+ว+บ)

= (2 / 10) – (5)(1/6)

= - 0.63

พิจารณา Document Unknown เปรียบเทียบกับ Cluster 2 โดยใช้สมการที่ 5.7 หาผลลัพธ์ Document Frequency (DF) จะได้ค่าดังนี้

DF = – (ฃ+ล+ช+ว+บ+ข)

= – (6)(1/6)

= -1

พิจารณา Document Unknown เปรียบเทียบกับ Cluster 3 โดยใช้สมการที่ 5.7 หาผลลัพธ์ Document Frequency (DF) จะได้ค่าดังนี้

DF = (ฃ+ล+ช) – (ว+บ+ข)

= ((1/11)+ (2 /11) + (2 /11)) – (3)(1/6)

= (0.09 + 0.18 + 0.18) – 0.5

= - 0.05

จากผลการคำนวณหาค่า Document Frequency (DF) ของ เอกสาร Document Unknown เปรียบเทียบกับ Cluster 1 , Cluster2 และ Cluster3 นั้น จะได้ผลลัพธ์ของค่า Document Frequency (DF) ที่ Document Unknown เปรียบเทียบกับ Cluster 3 มีค่าสูงที่สุดคือ –0.05 ฉะนั้น เอกสาร Document Unknown จะเข้าไปรวมกับ Cluster3

ข้อดีของวิธีการพิจารณาคำที่เหมือนกันตามจำนวนเอกสาร (Document Frequency) เมื่อเทียบกับวิธีการหาความคล้ายของคำสำคัญร่วมกับการหาระยะห่างของคำสำคัญ คือ

1. ไม่ต้องทำการหาค่า Threshold ในการจัดกลุ่มเอกสาร
2. การพิจารณาคำสำคัญตามจำนวนเอกสารจะเป็นการ ป้องกันการผิดพลาดเนื่องจากการให้น้ำหนักคำบางคำที่ไม่น่าจะใช่คำสำคัญมากเกินไปได้
3. ค่าของ Threshold อาจมีการเปลี่ยนแปลงได้เมื่อตัวอย่างเอกสารที่นำมาจัดกลุ่มเปลี่ยนไป ทำให้ต้องหาค่า Threshold ที่เหมาะสมใหม่
4. การจัดกลุ่มเอกสารแบบการหาค่าความคล้ายของคำสำคัญร่วมกับการหาระยะห่างของคำสำคัญ ระบบมักจะไม่หยุดทำงาน เนื่องจากเอกสารจะมีการเปลี่ยนกลุ่มไปมาตลอดเวลา แต่วิธีการพิจารณาคำสำคัญตามจำนวนเอกสาร Document Frequency (DF) ระบบจะหยุดทำงานได้เนื่องจาก พิจารณาค่า DF ที่มากที่สุดในกลุ่มจึงจะเข้าไปรวมกลุ่ม

จากเหตุผลดังกล่าว วิธีการจัดเอกสารแบบพิจารณาคำสำคัญตามจำนวนเอกสาร Document Frequency (DF) จึงน่าจะได้ผลลัพธ์ที่ดี และจากการทดลองเปรียบเทียบกับทุกวิธี (การจัดเอกสารแบบพิจารณาความคล้ายของคำสำคัญเพียงอย่างเดียว , การจัดเอกสารแบบพิจารณาความถี่คำสำคัญ และการจัดเอกสารโดยพิจารณาความคล้ายของคำสำคัญร่วมกับการหาระยะห่างของคำสำคัญ) จะให้ผลลัพธ์ของการจัดกลุ่มเอกสารดีที่สุด โดยใช้ค่า Precision , Recall และค่า F-measure เป็นตัววัดผลเปรียบเทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ ผลลัพธ์ของการจัดกลุ่มเอกสารแบบพิจารณาคำสำคัญตามจำนวนเอกสาร แสดงไว้ที่หัวข้อผลการทดลองและสรุป

บทที่ 6
เครื่องมือวัดการจัดกลุ่มเอกสารและการให้น้ำหนักคำสำคัญ

6.1 การวัดผลการจัดกลุ่มเอกสารและการให้น้ำหนักคำสำคัญ

6.1.1 การวัดผลการจัดกลุ่มเอกสาร

เมื่อจัดกลุ่มเอกสารด้วยวิธีต่างๆดังกล่าวมาแล้ว จำเป็นต้องหาวิธีวัดว่าวิธีการจัดกลุ่มเอกสารวิธีไหนให้ผลลัพธ์การจัดกลุ่มได้ดีที่สุด ซึ่งเอกสารในกลุ่มเดียวกันควรจะมีเนื้อหาในกลุ่มคล้ายกันมากกว่าเอกสารต่างกลุ่มกัน โดยวิทยานิพนธ์ฉบับนี้ได้ใช้เครื่องมือวัดผลการจัดกลุ่มเอกสารภาษาไทย 3 เครื่องมือ คือ Precision, Recall และ F-measure เปรียบเทียบกับการจัดกลุ่มด้วยมนุษย์ สมการของค่าเครื่องมือวัดทั้ง 3 แสดงอยู่ด้านล่าง

$$\Pr ecision(P) = n_{ij} / n_j \tag{6.1}$$

$$\text{Re } call\ (R) = n_{ij} / n_i \tag{6.2}$$

n_{ij} = The number of members of class i in cluster j

n_j = The number of members of cluster j

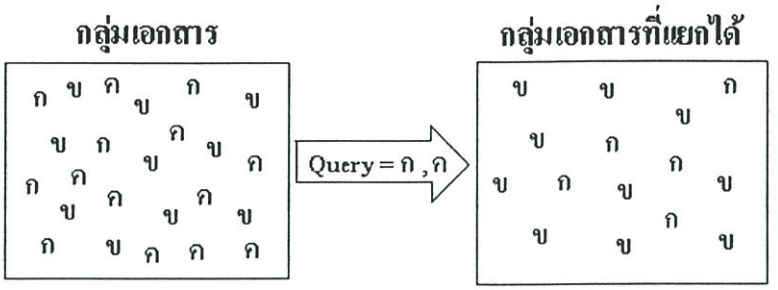
n_i = The number of members of class i

$$F(i,j) = \frac{(2 * \text{Re } call(i,j) * \Pr ecision(i,j))}{\Pr ecision(i,j) + \text{Re } call(i,j)} \tag{6.3}$$

F = F- measure

i = class i

j = cluster j



$$\text{Precision (P)} = \frac{ก}{ก+ข} = \frac{5}{15}$$

$$\text{Recall (R)} = \frac{ก}{ก+ค} = \frac{5}{14}$$

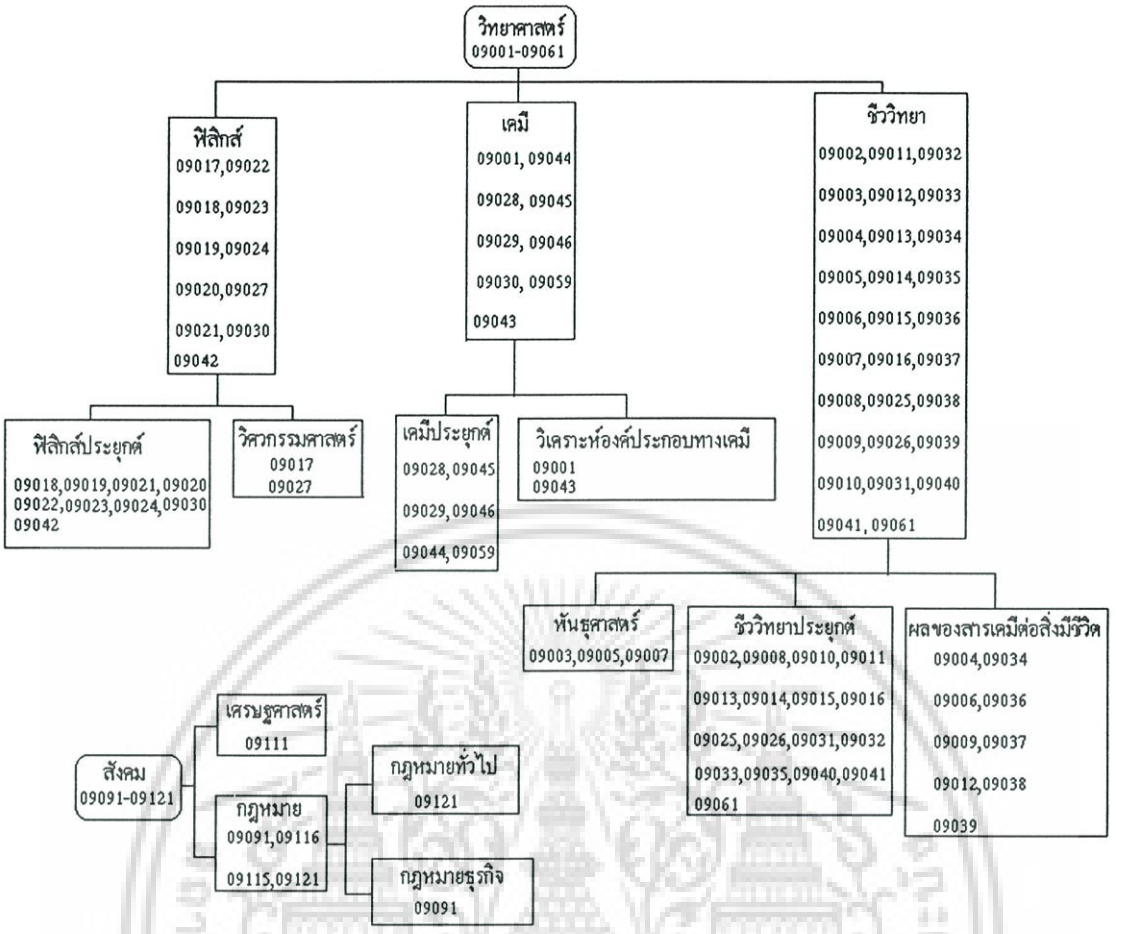
รูปที่ 6.1 ภาพอธิบายการหาค่า Precision และค่า Recall

ในรูปที่ 6.1 เป็นการหาค่า Precision และค่า Recall เปรียบเทียบให้ดูเข้าใจง่ายขึ้น ซึ่งค่าทั้ง2 จะใช้หาค่า F-measure ต่อไปได้อีก

โดยค่า **Precision** จะบอกว่าการจัดกลุ่มเอกสารนั้นจัดได้สมาชิกในกลุ่มถูกต้องไม่ผิดกลุ่มมากน้อยเท่าไร เมื่อเทียบกับการจัดกลุ่มด้วยมนุษย์

ค่า **Recall** จะบอกว่า ผลการจัดกลุ่มด้วยคอมพิวเตอร์นั้น จะเอาจำนวนหรือปริมาณสมาชิกที่ถูกต้องมามากหรือน้อยเท่าไร ในกลุ่มที่จัดโดยมนุษย์

ค่า **F-measure** จะบอกผลของค่าระหว่าง Precision และ ค่า Recall ในการจัดกลุ่มเอกสารนั้นๆว่าแตกต่างกันมากน้อยเท่าไร

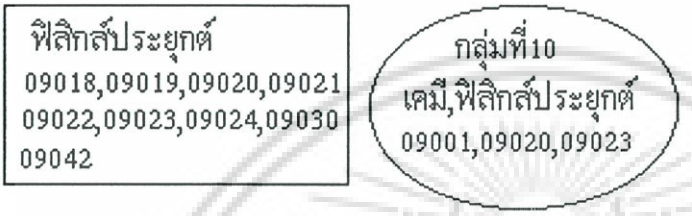


รูปที่ 6.2 ตัวอย่างการจัดกลุ่มเอกสารภาษาไทยด้วยมนุษย์

รูปที่ 6.3 ตัวอย่างผลลัพธ์การจัดกลุ่มเอกสารภาษาไทยด้วยระบบคอมพิวเตอร์

ตัวอย่างการคำนวณหาค่า Precision , Recall และ F-measure ของวิธีจัดกลุ่มเอกสารภาษาไทยด้วยระบบคอมพิวเตอร์ดังที่กล่าวมาแล้ว เทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ จากรูปที่ 6.3 เปรียบเทียบกับรูปที่ 6.2 โดยที่ตัวเลข 09xxx ใช้แทนชื่อสมาชิกเอกสารในกลุ่มนั้นๆ และใช้สมการที่ 6.1 ,6.2 และ 6.3 คำนวณตามลำดับจะได้ดังนี้

พิจารณาวัดผลการจัดกลุ่มเอกสารด้วยระบบคอมพิวเตอร์กลุ่มที่ 10 เทียบกับกลุ่ม ฟิสิกส์ประยุกต์ที่จัดโดยมนุษย์ เพราะ จัดกลุ่มโดยคอมพิวเตอร์มีเอกสารส่วนใหญ่ อยู่ในกลุ่มฟิสิกส์ประยุกต์



รูปที่ 6.4 ตัวอย่าง1เปรียบเทียบการจัดกลุ่มเอกสารด้วยมนุษย์กับคอมพิวเตอร์

ตัวอย่างการวัดผลการจัดกลุ่มเอกสารจากรูปข้างบนรูปที่ 6.4 มีเอกสารในกลุ่มฟิสิกส์ประยุกต์ คือ09020 กับ 09023 เป็นหลัก ฉะนั้นการวัดผลการจัดกลุ่มที่คำนวณได้มีดังนี้

Precision กลุ่มที่ 10 = 2 / 3 = 0.66

Recall กลุ่มที่ 10 = 2 / 9 = 0.22

F-measure กลุ่มที่ 10 = (2 * 0.22 * 0.66) / (0.22 + 0.66) = 0.33

รูปที่ 6.5 ตัวอย่าง2เปรียบเทียบการจัดกลุ่มเอกสารด้วยมนุษย์กับคอมพิวเตอร์

จากรูปด้านบนพิจารณาวัดผลการจัดกลุ่มเอกสารด้วยระบบคอมพิวเตอร์กลุ่มที่ 24 เทียบกับกลุ่ม เคมีประยุกต์ ที่จัดโดยมนุษย์ เพราะ จัดกลุ่มโดยคอมพิวเตอร์มีเอกสารส่วนใหญ่ อยู่ในกลุ่มเคมีประยุกต์ ฉะนั้นการวัดผลการจัดกลุ่มที่คำนวณได้มีดังนี้

Precision กลุ่มที่ 24 = 2 / 2 = 1

Recall กลุ่มที่ 24 = 2 / 6 = 0.33

F-measure กลุ่มที่ 24 = (2 * 0.33 * 1) / (1 + 0.33) = 0.50

ในการจัดกลุ่มเอกสารด้วยคอมพิวเตอร์ ตามวิธีการจัดกลุ่มต่างๆบางครั้งอาจมีเอกสารบ้างฉบับปนเข้ามาในกลุ่ม ทั้งที่ไม่ควรจะรวมเป็นสมาชิกในกลุ่มด้วย ฉะนั้นจึงจำเป็นต้องใช้เครื่องมือวัดผลการจัดกลุ่มเอกสารทั้ง 3 เครื่องมือ Precision , Recall และ F-measure ช่วยในการตัดสินผลการจัดกลุ่มว่าวิธีไหนจะดีที่สุด โดยการคำนวณผลการจัดกลุ่มจากเครื่องมือวัดทั้ง 3 อัน จะพิจารณาผลการจัดกลุ่มด้วยคอมพิวเตอร์ว่า สมาชิกกลุ่มส่วนใหญ่อยู่ในกลุ่มใดเมื่อเทียบกับการจัดกลุ่มด้วยมนุษย์ แล้วนำไปคำนวณตามสมการดังกล่าวมาแล้วข้างต้น ผลลัพธ์ของวิธีใดที่ให้ค่า Precision , Recall และค่า F-measure มากที่สุด จะเป็นวิธีการจัดกลุ่มที่ดีที่สุด จากตัวอย่างการคำนวณข้างบน กลุ่มที่24 เคมีประยุกต์ที่จัดด้วยคอมพิวเตอร์ จะจัดกลุ่มได้ดีกว่ากลุ่มที่ 10 ฟิสิกส์ประยุกต์

6.1.2 การให้น้ำหนักคำสำคัญ

เมื่อเราได้ผลการจัดกลุ่มที่ดีที่สุดแล้ว เราจะทราบว่าภายในกลุ่มต่างๆที่มีสมาชิกเอกสารอยู่นั้นมีเอกสารใดอยู่บ้าง และแต่ละกลุ่มประกอบด้วยคำสำคัญอะไรบ้างจำนวนเท่าไร จากนั้นเราก็สามารถจะมาคำนวณหาค่าการให้น้ำหนักคำเหล่านั้นได้ ซึ่งจะช่วยให้เรารู้ได้ว่าคำสำคัญเหล่านั้น สำคัญจริงมากน้อยเท่าไรในกลุ่มนั้น คำบางคำมีความสำคัญมากอาจปรากฏในกลุ่มเอกสารน้อย ในขณะที่คำที่ไม่ใช่คำสำคัญหลักในการบ่งบอกกลุ่มที่แน่ชัดอาจปรากฏบ่อยได้ การให้น้ำหนักคำจะช่วยเราได้ เหมือนกับเปรียบเทียบว่าการทำเรื่องย่อความนั้นเอง ที่ต้องเอาแต่คำหรือข้อความที่กระชับและได้เนื้อหาใจความ นอกจากนี้ยังช่วยเราในการคำนวณการจัดกลุ่มเอกสาร unknown อันอื่นต่อไปได้อีกด้วย โดยเอาน้ำหนักคำเหล่านี้มาช่วยในการคำนวณได้ หรือจะตัดคำสำคัญที่มีค่าน้ำหนักน้อยออกไปจากฐานข้อมูลก็ได้ ซึ่งคำนั้นอาจเป็นคำรบกวนระบบที่เราไม่ได้ตัดทิ้งไปก็เป็นไปได้

ในการให้น้ำหนักคำสำคัญนั้น จะพิจารณาในรูปของความน่าจะเป็น(Probability) โดยอาศัยสมการดังนี้

$$P(Y = w \mid d) = \frac{\xi(w,d)}{\sum_{w \in d} \xi(w,d)} \qquad (6.4)$$

$P(Y = w \mid d)$ = โอกาสที่จะพบคำ w ในเอกสาร d

$\xi(w,d)$ = จำนวนครั้งที่พบคำว่า w ในเอกสารd

$\sum_{w \in d} \xi(w,d)$ = ผลรวมความถี่ของคำที่ปรากฏในเอกสารd

$$P(Y = w \mid M) = \frac{\sum_{doc \in D} \xi(w,d)}{\sum_{doc \in D} \sum_{w \in doc} \xi(w,d)} \quad (6.5)$$

$P(Y = w \mid M)$ = ความน่าจะเป็นของเหตุการณ์ที่พบคำ w ในกลุ่มเอกสาร Cluster

$\sum_{doc \in D} \xi(w,d)$ = ผลรวมจำนวนครั้งที่พบคำwในเอกสาร d ที่เป็นสมาชิกในกลุ่มD

$\sum_{doc \in D} \sum_{w \in doc} \xi(w,d)$ = ผลรวมจำนวนครั้งที่พบคำ w ในเอกสารd และทั้งหมดในกลุ่มเอกสารD

รูปที่ 6.6 ตัวอย่างการจัดกลุ่มเอกสารที่ 1กลุ่ม (1Cluster)

สมการที่ 6.4 เป็นสมการที่พิจารณาเฉพาะคำที่ปรากฏในเอกสารเท่านั้น ส่วนสมการที่ 6.5 เป็นคำที่พบในกลุ่มเอกสารทั้งกลุ่ม ซึ่งเป็นการเฉลี่ยค่าให้เหมาะสมของเอกสารทั้งหมดในกลุ่ม จากรูปที่ 6.6 เป็นตัวอย่างการจัดกลุ่มเอกสารที่ประกอบด้วย 1กลุ่ม (1Cluster) ที่ประกอบด้วยสมาชิกเอกสาร 3 ฉบับ แต่ละฉบับมีคำสำคัญอยู่ 3 คำ โดยเลขข้างหลังคำ(พยัญชนะ) จะแทนจำนวนความถี่คำที่พบในเอกสารแต่ละฉบับ ตัวอย่างการคำนวณการให้น้ำหนักคำ ตามสมการที่ 6.4 และ 6.5 แทนค่าในสมการที่ 6.4 พิจารณาเอกสารฉบับ (ก2 ข3 ค1) จะได้

$$P(Y = w \mid d) = \frac{\xi(w,d)}{\sum_{w \in d} \xi(w,d)}$$

$$= (2/6), (3/6), (1/6)$$

$$= 0.33, 0.5, 0.16$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ก2 ข3 ค1) จะได้ค่าน้ำหนักของคำว่า "ก,ข,ค" เท่ากับ 0.33 , 0.5 , 0.16 ตามลำดับ

แทนค่าในสมการที่ 6.4 พิจารณาเอกสารฉบับ (ข5 ง9 ฆ7) จะได้

$$P(Y = w \mid d) = \frac{\xi(w,d)}{\sum_{w \in d} \xi(w,d)}$$

$$= (5 / 21), (9 / 21), (7 / 21)$$

$$= 0.238, 0.428, 0.333$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ข5 ง9 ฆ7) จะได้ค่าน้ำหนักของคำว่า "ข,ง,ฆ" เท่ากับ 0.238 , 0.428 , 0.333 ตามลำดับ

แทนค่าในสมการที่ 6.4 พิจารณาเอกสารฉบับ (ข6 ค3 ฆ4) จะได้

$$P(Y = w \mid d) = \frac{\xi(w,d)}{\sum_{w \in d} \xi(w,d)}$$

$$= (6 / 13), (3 / 13), (4 / 13)$$

$$= 0.461, 0.230, 0.307$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ข6 ค3 ฆ4) จะได้ค่าน้ำหนักของคำว่า "ข,ค,ฆ" เท่ากับ 0.461 , 0.230 , 0.307 ตามลำดับ

แทนค่าในสมการที่ 6.5 พิจารณาเอกสาร(ก2 ข3 ค1) เทียบกับกลุ่ม Clusterจะได้

$$P(Y = w \mid M) = \frac{\sum_{doc \in D} \xi(w,d)}{\sum_{doc \in D} \sum_{w \in doc} \xi(w,d)}$$

$$= (2 / 40), ((3 + 5 + 6) / 40), ((1 + 3) / 40)$$

$$= 0.05, 0.35, 0.1$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ก2 ข3 ค1) จะได้ค่าน้ำหนักของคำว่า "ก,ข,ค" เท่ากับ 0.05 , 0.35 , 0.1 ตามลำดับ

แทนค่าในสมการที่ 6.5 พิจารณาเอกสาร(ข5 ง9 ฆ7) เทียบกับกลุ่ม Cluster จะได้

$$P(Y = w \mid M) = \frac{\sum_{doc \in D} \xi(w,d)}{\sum_{doc \in D} \sum_{w \in doc} \xi(w,d)}$$

$$= ((3 + 5 + 6) / 40), (9 / 40), ((7 + 4) / 40)$$

$$= 0.35, 0.225, 0.275$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ข5 ง9 ฆ7) จะได้ค่าน้ำหนักของคำว่า "ข,ง,ฆ" เท่ากับ 0.35 , 0.225 , 0.275 ตามลำดับ

แทนค่าในสมการที่ 6.5 พิจารณาเอกสาร(ข6 ค3 ฆ4) เทียบกับกลุ่ม Clusterจะได้

$$P(Y = w \mid M) = \frac{\sum_{doc \in D} \xi(w,d)}{\sum_{doc \in D} \sum_{w \in doc} \xi(w,d)}$$

$$= ((3 + 5 + 6)/40), ((1 + 3)/40), ((4 + 7)/40)$$

$$= 0.35 , 0.1 , 0.275$$

เพราะฉะนั้น ค่าน้ำหนักในเอกสารฉบับ (ข6 ค3 ฆ4) จะได้ค่าน้ำหนักของคำว่า "ข,ค,ฆ" เท่ากับ 0.35 , 0.1 , 0.275 ตามลำดับ

รูปแบบการทำน้ำหนักคำบางครั้งจะทำในรูป ค่าเฉลี่ยเลขคณิต (arithmetic mean)(**AM**) ที่มีสมการดังนี้

$$P_{AM}(Y = w \mid d, M) = \frac{P(Y = w \mid d)}{2} + \frac{P(Y = w \mid M)}{2} \qquad (6.6)$$

หรืออาจทำการให้น้ำหนักคำ ในรูปแบบของค่าเฉลี่ยเรขาคณิต(geometric mean)(**GM**) ดังสมการ

$$P_{GM}(Y = w \mid d, M) = P(Y = w \mid d)^{1/2} \times P(Y = w \mid M)^{1/2} \qquad (6.7)$$

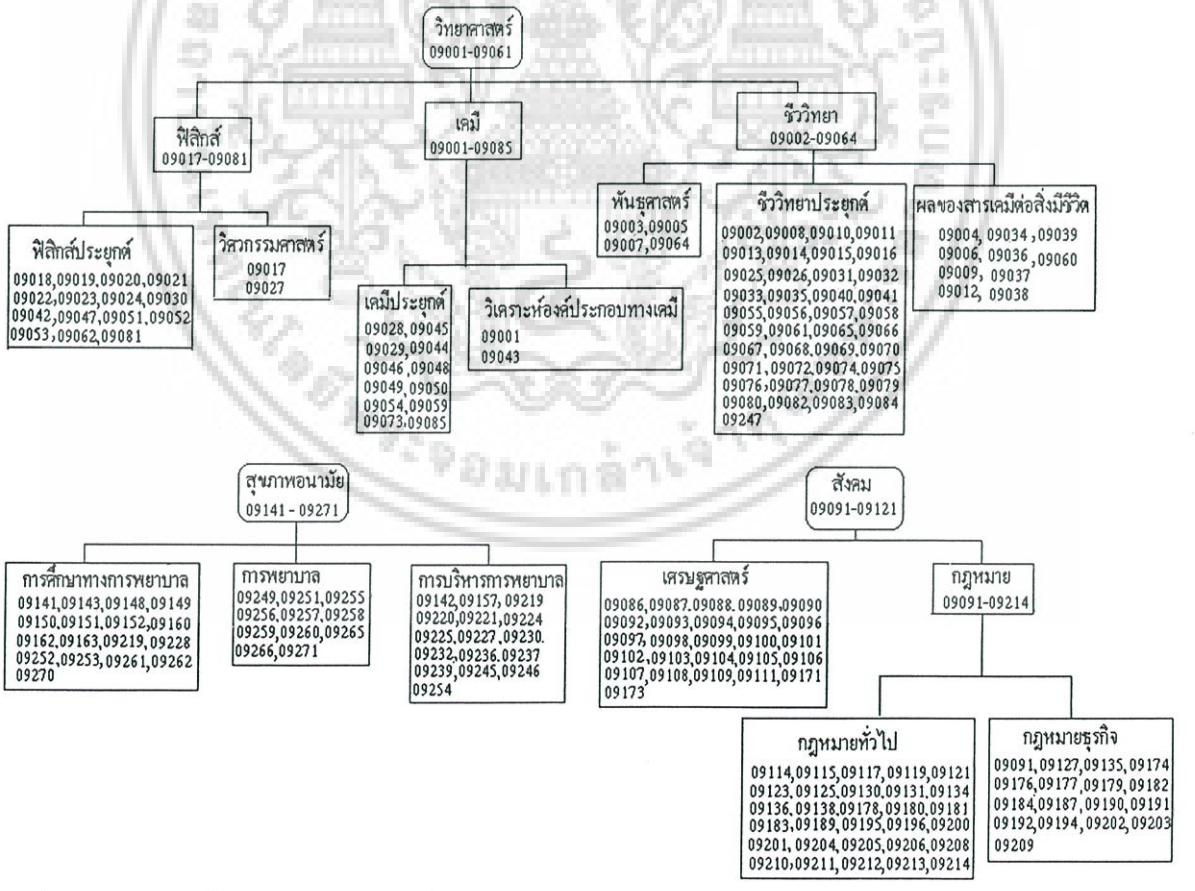
ในสมการที่ 6.7 การประมาณค่าของการให้น้ำหนักคำแบบ geometric mean ไม่ได้เป็นแบบ true Probability distribution จึงต้องปรับเปลี่ยนสมการแบบ geometric mean บ้าง ให้เป็นรูปของ Normal factor ซึ่งจะเรียกว่าสมการใหม่นี้ว่า normalized geometric mean (**NGM**) ดังนี้

$$P_{NGM}(Y = w \mid d, M) = \frac{P(Y = w \mid d)^{1/2} \times P(Y = w \mid M)^{1/2}}{\sum_{w \in W} P(Y = w \mid d)^{1/2} \times P(Y = w \mid M)^{1/2}} \qquad (6.8)$$

การจะให้น้ำหนักคำได้จะต้องผ่านการจัดกลุ่มเอกสารเรียบร้อยก่อน ถึงจะคำนวณหาค่าน้ำหนักคำตามสมการดังกล่าวมาแล้วได้ เพราะน้ำหนักคำบางคำในเอกสารอันเดียวกัน อาจมีค่าแตกต่างกันได้ถ้าใช้วิธีการจัดกลุ่มเอกสารต่างกัน เนื่องจากผลลัพธ์การจัดกลุ่มเอกสารต่างกันจึงมีสมาชิกในกลุ่มเอกสารต่างกันด้วย ทำให้จำนวนคำสำคัญแตกต่างกันน้ำหนักคำจึงต่างกัน ดังนั้นน้ำหนักคำใดจะถูกต้องหรือไม่ก็ขึ้นกับวิธีการจัดกลุ่มเอกสารด้วย น้ำหนักคำที่ดีจะสามารถนำไปประยุกต์ใช้ เป็นตัวช่วยในการจัดกลุ่มเอกสารฉบับอื่นๆได้ถูกกลุ่มต่อไป

บทที่ 7
ผลการทดลองและสรุปการจัดกลุ่มเอกสารภาษาไทย

การจัดกลุ่มเอกสารที่ดีควรจะต้องได้จำนวนกลุ่มที่เหมาะสม ไม่มากไปหรือน้อยไป เพราะ ถ้าจำนวนกลุ่มมากไปแสดงว่าสมาชิกเอกสารมีจำนวนน้อย เอกสารที่ควรจะรวมกันได้ก็ไม่รวมกัน และถ้าจำนวนกลุ่มน้อยเกินไปสมาชิกเอกสารในกลุ่มจะมีมาก ซึ่งอาจจะเป็นการเอาเอกสารที่ไม่ควรรวมอยู่ในกลุ่มเข้ามารวมด้วย หรือเป็นกลุ่มที่มีขนาดเนื้อหากว้างมาก ฉะนั้น สมาชิกเอกสารในแต่ละกลุ่มที่ได้ ควรจะต้องมีเนื้อหาใกล้เคียงกันมากกว่าเอกสารต่างกลุ่มกัน และมีขนาดกลุ่มที่เหมาะสม โดยในงานวิจัยนี้จะใช้เครื่องมือวัดดังที่กล่าวมาแล้ว เป็นตัววัดวิธีการจัดกลุ่มเอกสารว่าดีหรือไม่อย่างไร งานวิจัยนี้ได้เอาจำนวนกลุ่มที่ใกล้เคียงกันมาวัดผลเปรียบเทียบกัน วิธีที่ดีที่สุดจะต้องได้ค่าPrecision ,Recall และค่า F-measure สูงที่สุดเมื่อเทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ ตัวอย่างเอกสารที่จัดกลุ่มได้แสดงในภาคผนวก ข ส่วนผลการวัดการจัดกลุ่มเอกสารจากเครื่องมือวัดแสดงในตารางข้างล่าง



รูปที่ 7.1 เอกสารที่จัดกลุ่มด้วยมนุษย์ที่ใช้เป็นมาตราฐานวัดผลการจัดกลุ่ม

แทนชื่อกลุ่มเอกสารที่จัดโดยมนุษย์ด้วยรหัสตัวเลข และละ09,090,0900จากรหัสเอกสาร 09XXX เพื่อให้รูปแบบของตารางข้อมูลดูกระทัดรัด

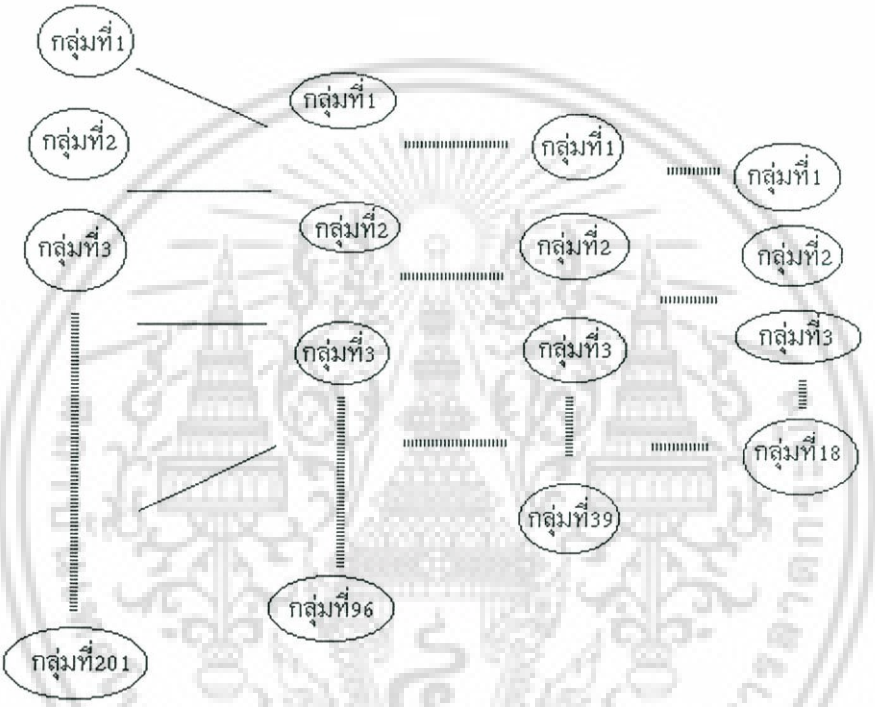
ตารางที่ 7.1 การแทนชื่อกลุ่มเอกสารที่จัดโดยมนุษย์ด้วยรหัสตัวเลข

<table>
<tr><td colspan="8">ชื่อกลุ่มเอกสาร
รหัส</td></tr>
<tr><td colspan="8">วิทยาศาสตร์
1</td></tr>
<tr><td colspan="2">ฟิสิกส์
11</td><td colspan="2">เคมี
12</td><td colspan="4">ชีววิทยา
13</td></tr>
<tr><td>ฟิสิกส์ประยุกต์
111</td><td>วิศวกรรมศาสตร์
112</td><td>เคมีประยุกต์
121</td><td>วิเคราะห์องค์ประกอบทางเคมี
122</td><td>พันธุศาสตร์
131</td><td>ชีววิทยาประยุกต์
132</td><td colspan="2">ผลของสารเคมีต่อสิ่งมีชีวิต
133</td></tr>
<tr><td colspan="8">สุขภาพอนามัย
2</td></tr>
<tr><td colspan="2">การศึกษาทางการพยาบาล
21</td><td colspan="2">การพยาบาล
22</td><td colspan="4">การบริหารการพยาบาล
23</td></tr>
<tr><td colspan="8">สังคม
3</td></tr>
<tr><td colspan="3">เศรษฐศาสตร์
31</td><td colspan="5">กฎหมาย
32</td></tr>
<tr><td colspan="3"></td><td colspan="2">กฎหมายทั่วไป
321</td><td colspan="3">กฎหมายธุรกิจ
322</td></tr>
</table>

7.1 ผลลัพธ์วิธีการจัดกลุ่มเอกสารแบบพิจารณาคำที่เหมือนกัน ร่วมกับการหาค่าระยะห่างของคำระหว่างกลุ่มเอกสาร (Similarity Keyword and Distance)(SKD)

7.1.1 ผลลัพธ์การหาความเหมือนของคำสำคัญเพียงอย่างเดียว

ผลการทดลองจัดกลุ่มเอกสารด้วยเครื่องคอมพิวเตอร์ แบบวิธีพิจารณาคำสำคัญที่ไม่พิจารณาน้ำหนักคำตามความถี่คำและน้ำหนักคำตามจำนวนเอกสาร ดูแต่คำ intersec กันเพียงอย่างเดียว ไม่รวมการหาค่า Distance ด้วย ได้ผลการรวมกลุ่มแบบ Iterative Clustering ดังภาพข้างล่างนี้



รูปที่ 7.2 ผลการรวมกลุ่มเอกสารภาษาไทยด้วยวิธี SKD ดูเฉพาะคำ intersec ไม่รวมการหาค่า Distance

ตารางที่ 7.2 ผลการจัดกลุ่มวิธีSKDดูเฉพาะคำ intersec ไม่รวมการหาค่า Distance เทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	Precision	Recall	F-measure
1	132	8,10,11,13,14,15,25,31,33	1	0.21	0.36
2	131	3	1	0.25	0.4
3	132	64, 66,76	0.6	0.05	0.1
4	133,132	4,6,9,71,75,79,82	1	0.13	0.23

ตารางที่ 7.2 (ต่อ)

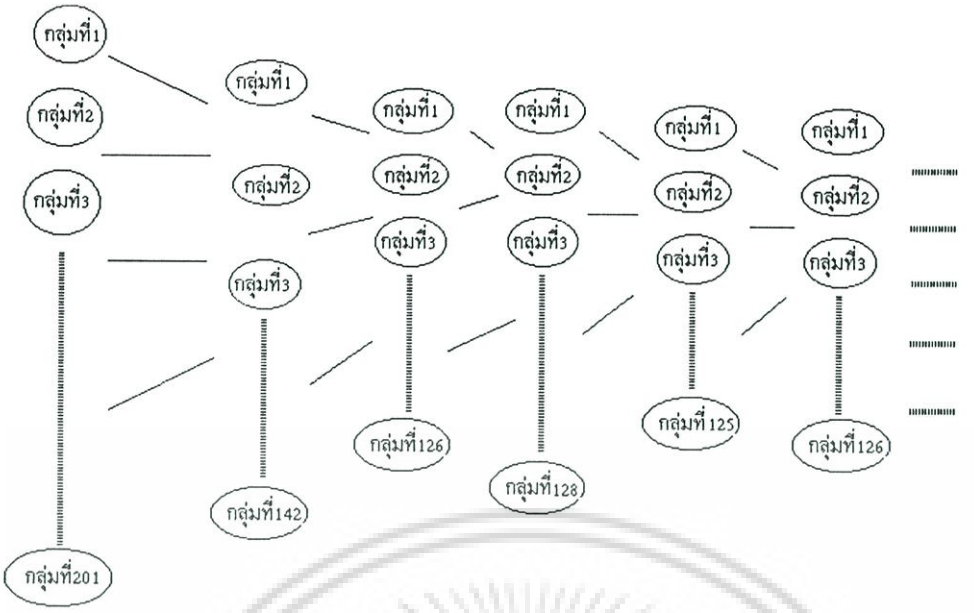
5	111	30,47	1	0.13	0.23
6	11	27,50,53,80	0.5	0.11	0.18
7	112,132	17,67	0.5	0.26	0.34
8	121	28,29,44,46,48,49,51,52	0.75	0.46	0.57
9	111,121	42,45,60,62,85	0.4	0.14	0.2
10	133	36,37,38,39,40,68	0.66	0.4	0.5
11	132	43,55,56,57,58,65,69,72,73	0.77	0.17	0.26
12	133,132	12 ,41	0.5	0.06	0.11
13	132	32	1	0.02	0.04
14	111	22	1	0.06	0.12
15	132	16,59,61,70,74,77,78,81,83 ,84	0.9	0.21	0.39
16	131	5	1	0.25	0.4
17	132	2, 34,35	0.6	0.04	0.75
18	131, 132	7,26	0.5	0.14	0.21
19	122 ,111	1,23	0.5	0.28	0.36
20	111	18	1	0.06	0.13
21	121	54	1	0.08	0.14
22	111	24	1	0.06	0.13
23	31	98,99,101,104,171,173,184 ,202	0.75	0.23	0.35
24	21,23	143,148,149,150,151,152, 157,163,219,224,225,228, 230,232,239,245,255	0.94	0.48	0.64
25	31	97	1	0.03	0.07
26	31	93,95,106,174,191	0.6	0.12	0.19
27	321	125,130,131,181,183,187, 192,213,214	0.77	0.24	0.37
28	32	177,182,190,206,208,209, 210,211,212	1	0.19	0.32

ตารางที่ 7.2 (ต่อ)

29	31	89	1	0.03	0.07
30	321	114,119,123,136,195	1	0.17	0.29
31	321	115,117,127,134,176,179, 180,189	0.75	0.2	0.32
32	32	135,138,194,200,201,203, 204,205	1	0.17	0.29
33	31	88,90,100,102,103,105,107 ,108,109	1	0.35	0.51
34	321	91,121,196,226	0.5	0.06	0.12
35	23	94,96,220,221,254,256	0.5	0.19	0.27
36	21,22	86,160,227,249,251,252, 253,258,259,261,262,265, 271	0.84	0.39	0.53
37	31	87,92,111	1	0.11	0.2
38	2	20,141,142,162,236,237, 246,257,260,270	0.9	0.2	0.33
39	111	19,21,47	1	0.2	0.33
Precision เฉลี่ย	0.81				
Recall เฉลี่ย	0.17				
F-measure เฉลี่ย	0.29				

7.1.2 ผลลัพธ์การหาความเหมือนกันของคำสำคัญร่วมกับการหาค่าระยะห่างของคำสำคัญระหว่างกลุ่มเอกสาร (Similarity Keyword and Distance)(SKD)

พิจารณาวิธี SKD เอาเฉพาะคำที่ intersec มาหาระยะ (Distance) 60 จะได้ผลการจัดกลุ่มเอกสารดังใน รูปที่ 7.3



รูปที่ 7.3 ผลการรวมกลุ่มเอกสารภาษาไทยวิธี SKD ที่ค่าDistance 60เฉพาะคำ intersec

ตารางที่ 7.3 ผลการจัดกลุ่มวิธี SKD ที่Distance 60 เฉพาะคำ intersec เทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	Precision	Recall	F-measure
1	131	3	1	0.25	0.4
2	133,121	6,44	0.5	0.09	0.16
3	132	8,11	1	0.05	0.09
4	111	19	1	0.07	0.13
5	111	18,22	1	0.13	0.23
6	133	12	1	0.1	0.18
7	133	36	1	0.1	0.18
8	132	2,31	1	0.05	0.09
9	121	28	1	0.08	0.15
10	111	42	1	0.07	0.12
11	112	27	1	0.5	0.67
12	132	58	1	0.02	0.04
13	133, 132	4,66	0.5	0.06	0.1
14	122	1	1	0.5	0.67

ตารางที่ 7.3 (ต่อ)

15	131	5	1	0.25	0.4
16	112,132	17,61	0.5	0.26	0.35
17	111	24	1	0.07	0.13
18	111	30	1	0.07	0.13
19	31	101	1	0.04	0.07
20	31	97	1	0.04	0.07
21	21	141	1	0.06	0.11
22	111	20	1	0.07	0.13
23	31	100,109	1	0.08	0.14
24	31	87	1	0.04	0.07
25	321	130	1	0.03	0.07
26	321	115	1	0.03	0.07
27	321	123	1	0.03	0.07
28	322	174	1	0.06	0.11
29	321	134	1	0.03	0.07
30	321	114,196	1	0.07	0.13
31	322	182,194	1	0.11	0.21
32	321	211	1	0.03	0.07
33	322	135	1	0.06	0.11
34	133,132,121	9,16,50	0.3	0.07	0.11
35	23	221	1	0.06	0.11
36	22	258	1	0.09	0.17
37	22	249,251	1	0.18	0.3
38	31	96	1	0.04	0.07
39	131	7	1	0.25	0.4
40	132	10,14	1	0.05	0.09
41	132	13	1	0.02	0.05
42	132	15,25,26	1	0.07	0.14
43	111	21	1	0.07	0.13
44	111	23	1	0.07	0.13

ตารางที่ 7.3 (ต่อ)

45	121	29,48	1	0.17	0.29
46	132	32,35	1	0.05	0.09
47	132	33	1	0.02	0.05
48	133	34	1	0.1	0.18
49	133,132	37,74	0.5	0.06	0.11
50	133	38,39	1	0.2	0.33
51	132	40	1	0.02	0.05
52	132	41	1	0.02	0.05
53	122	43	1	0.5	0.67
54	121	45	1	0.08	0.15
55	121	46	1	0.08	0.15
56	111	47,62	1	0.13	0.24
57	111	49,52,53	0.67	0.13	0.22
58	111,132	51,80,81,83	0.5	0.09	0.16
59	121	54	1	0.08	0.15
60	132	55,68	1	0.05	0.09
61	132	56,59,77	1	0.07	0.14
62	132	57,72,78	1	0.07	0.14
63	133	60	1	0.1	0.18
64	132	64,71,82	0.67	0.05	0.09
65	132	65,67,70,75	1	0.1	0.18
66	132	69,247	1	0.05	0.09
67	121	73	1	0.08	0.15
68	132	76	1	0.02	0.05
69	132	79	1	0.02	0.05
70	132	84	1	0.02	0.05
71	121	85	1	0.08	0.15
72	31	86	1	0.04	0.07
73	31	88	1	0.04	0.07
74	31,22	89,266	0.5	0.07	0.11

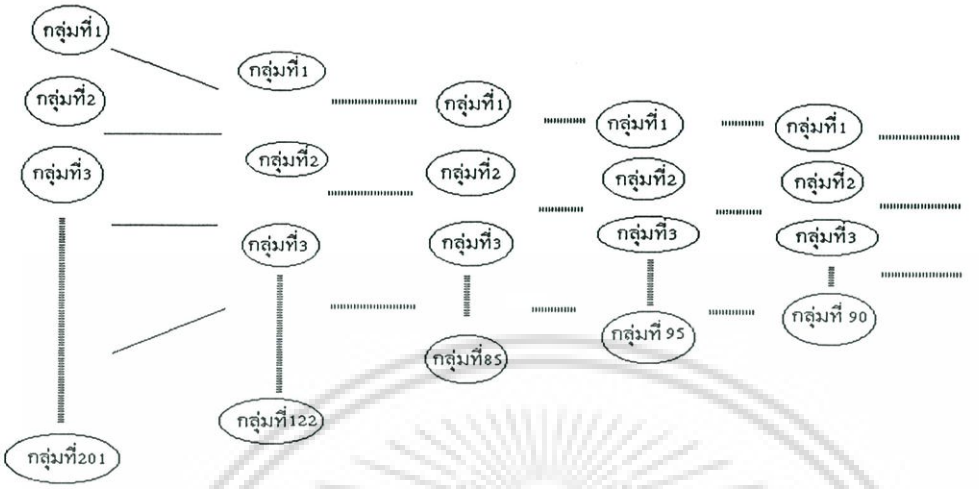
ตารางที่ 7.3 (ต่อ)

75	31	90	1	0.04	0.07
76	32	91,121	1	0.02	0.04
77	31	92,94	1	0.08	0.14
78	31	93	1	0.04	0.07
79	31	95	1	0.04	0.07
80	31	98	1	0.04	0.07
81	31	99	1	0.04	0.07
82	31	102	1	0.04	0.07
83	31	103,108,173	1	0.11	0.2
84	31	104	1	0.04	0.07
85	31	105	1	0.04	0.07
86	31	106	1	0.04	0.07
87	31	107	1	0.04	0.07
88	31	111	1	0.04	0.07
89	321	117,119,131	1	0.1	0.18
90	321	125,138	1	0.07	0.13
91	322	127	1	0.06	0.11
92	321	136	1	0.03	0.07
93	23	142,224,227,256, 270	0.6	0.19	0.29
94	21	143,150,157	0.67	0.12	0.2
95	21	148	1	0.06	0.11
96	21	149,160,252	1	0.18	0.31
97	23	151,219,237,260	0.5	0.13	0.2
98	21	152,163,228	1	0.18	0.3
99	21	162	1	0.06	0.11
100	31	171	1	0.04	0.07
101	322	176,202	1	0.12	0.21
102	321	177,192,209	1	0.18	0.3
103	321	178	1	0.03	0.06

ตารางที่ 7.3 (ต่อ)

104	322	179	1	0.06	0.11
105	32	180,184,187,208	1	0.09	0.16
106	321	181,183,190,201	0.75	0.1	0.18
107	321	189,195	1	0.07	0.13
108	322	191,203	1	0.12	0.2
109	321	200,205,206,213	1	0.13	0.24
110	321	204	1	0.03	0.06
111	321	210	1	0.03	0.06
112	321	212	1	0.03	0.06
113	321	214	1	0.03	0.06
114	23	220	1	0.06	0.12
115	23	225,232,245	1	0.19	0.37
116	23	230,236,259	0.66	0.13	0.21
117	23	239,254	1	0.13	0.22
118	23	246	1	0.06	0.12
119	21	253	1	0.06	0.12
120	22	255	1	0.09	0.16
121	22	257	1	0.09	0.16
122	21	261	1	0.06	0.12
123	21	262	1	0.06	0.12
124	22	265	1	0.09	0.16
125	22	271	1	0.09	0.16
Precision เฉลี่ย	0.94				
Recall เฉลี่ย	0.09				
F-measure เฉลี่ย	0.15				

◆ พิจารณาวิธี SKD มาหาระยะ (Distance) 90 เอาเฉพาะคำที่ intersec จะได้ผลการจัดกลุ่มเอกสารดังในรูปที่ 7.4



รูปที่ 7.4 ผลการรวมกลุ่มเอกสารภาษาไทยวิธีSKDหาค่า Distance 90 เฉพาะคำintersec

ตารางที่ 7.4 ผลการจัดกลุ่มวิธีSKDหาค่า Distance 90 เฉพาะคำ intersec เทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	ค่า Precision	ค่า Recall	ค่า F-measure
1	131	5	1	0.25	0.4
2	121	46	1	0.08	0.15
3	122,132	1,11	0.5	0.26	0.34
4	31	97	1	0.04	0.07
5	131	3	1	0.25	0.4
6	132	2	1	0.02	0.05
7	11	19,27	1	0.12	0.21
8	121	50	1	0.08	0.15
9	132	15	1	0.02	0.05
10	111	42	1	0.07	0.13
11	133,112	6,17	0.5	0.3	0.38
12	133	4,12,36	1	0.3	0.46
13	132	58	1	0.02	0.05

ตารางที่ 7.4 (ต่อ)

14	132	56,59	1	0.05	0.09
15	132	25,61,69	1	0.07	0.13
16	132	8,13,26	1	0.07	0.13
17	31	92	1	0.04	0.08
18	31	87,94	1	0.08	0.14
19	31,322	89,91	0.5	0.05	0.09
20	31	90,93,99,101, 105	1	0.19	0.32
21	321	138	1	0.03	0.06
22	22	258	1	0.09	0.16
23	21	141	1	0.06	0.11
24	322	174	1	0.06	0.11
25	321	115	1	0.03	0.06
26	321	123,189	1	0.07	0.12
27	322	125,182,194	0.67	0.11	0.2
28	31	100	1	0.04	0.07
29	322	203	1	0.06	0.11
30	321	134,181	1	0.07	0.13
31	322	180,184,187	0.67	0.11	0.3
32	321	135,183,190, 205,206,211, 212	0.7	0.17	0.27
33	31	96	1	0.04	0.07
34	21	149	1	0.06	0.11
35	23	224,227,255	0.67	0.13	0.2
36	23	152,160,163, 228	1	0.24	0.38
37	22	249,251	1	0.18	0.3
38	131	7	1	0.25	0.4

ตารางที่ 7.4 (ต่อ)

39	132,121	9,16,28,85, 247	0.4	0.1	0.16
40	132	10,14	1	0.05	0.09
41	111	18,21,24	1	0.2	0.3
42	111	20,23	1	0.13	0.24
43	111	22	1	0.07	0.13
44	121	29,44,45,48	1	0.33	0.5
45	111	30,47	1	0.13	0.24
46	132	31,40	1	0.05	0.09
47	132	32,33,35,41	1	0.1	0.18
48	132	34,39,66,68, 71	0.6	0.07	0.13
49	133	37,38,74	0.67	0.2	0.39
50	132	43,57,65,70, 77	0.8	0.1	0.18
51	111	49,52,53	0.67	0.13	0.22
52	111	51	1	0.07	0.13
53	121	54	1	0.08	0.15
54	132	55,60,67,79, 82	0.8	0.1	0.17
55	111	62	1	0.07	0.13
56	131, 132	64,76	0.5	0.14	0.22
57	132,121	72,73	0.5	0.05	0.1
58	132	75	1	0.024	0.05
59	132	78,80,81,83, 84	0.8	0.01	0.02
60	31	86,98,271	0.67	0.08	0.14
61	31	88	1	0.04	0.07
62	31	95,103,104, 171	1	0.15	0.27

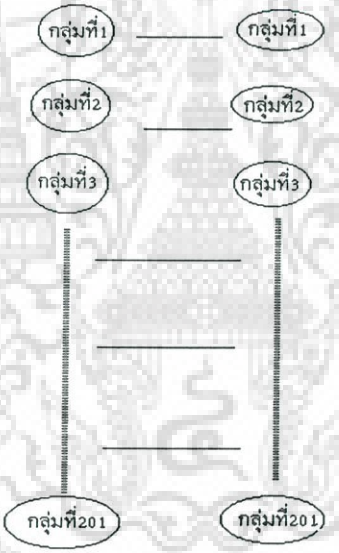
ตารางที่ 7.4 (ต่อ)

63	31	102,107,108, 111	1	0.15	0.27
64	31	106	1	0.04	0.07
65	31	109	1	0.04	0.07
66	32	114,127,179, 200,201	1	0.1	0.18
67	321	117,119,131	1	0.1	0.18
68	321	121	1	0.03	0.06
69	321	130,178,196, 208	1	0.13	0.24
70	321	136	1	0.03	0.06
71	23	142,219,221, 230,236,259	0.83	0.31	0.52
72	21	143,148,150, 151,157,162, 239	0.71	0.3	0.38
73	31	173	1	0.038	0.07
74	322	176,202	1	0.12	0.21
75	322	177,191,204, 209	0.75	0.18	0.29
76	321	195,210	1	0.07	0.13
77	321	213	1	0.03	0.06
78	321	214	1	0.3	0.06
79	23	220	1	0.06	0.12
80	23	225,232,245, 270	0.75	0.19	0.3
81	23	237,246,254	1	0.19	0.32
82	22	252,256,260, 261,266	0.6	0.27	0.38
83	21	253	1	0.06	0.11

ตารางที่ 7.4 (ต่อ)

84	22	257	1	0.09	0.17
85	21,22	262,265	0.5	0.07	0.13
Precision เฉลี่ย	0.90				
Recall เฉลี่ย	0.11				
F-measure เฉลี่ย	0.19				

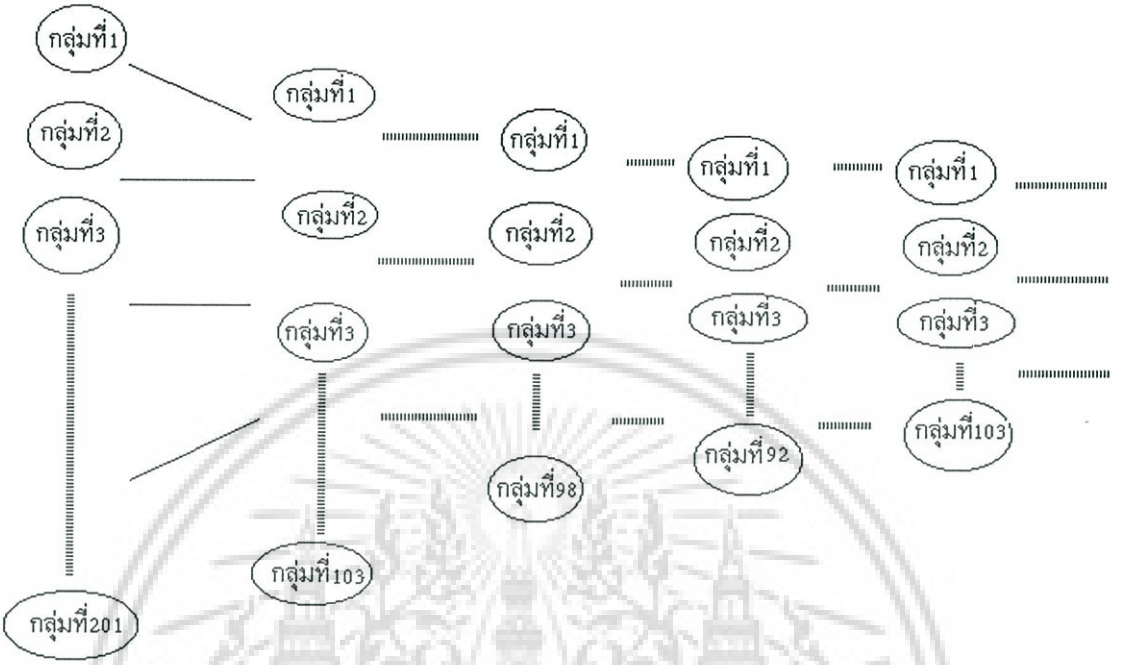
♦ พิจารณาวิธี SKD มาหาระยะ (Distance) 40 เอาทุกคำทั้งเอกสารUnknown และกับกลุ่มเปรียบเทียบ จะได้ผลการจัดกลุ่มเอกสารดังในรูปที่ 7.5



รูปที่ 7.5 ผลการรวมกลุ่มเอกสารภาษาไทยวิธีSKDหาค่า Distance 40 พิจารณาทุกคำทั้งเอกสาร Unknown และกับกลุ่มเปรียบเทียบ

จากรูปที่ 7.5 ไม่เกิดการรวมกลุ่มเอกสาร เพราะค่า Distance 40 ที่เอาทุกคำสำคัญมาคำนวณ (ทั้งคำ intersec และไม่ intersec) มีค่าน้อยเกินไปจึงไม่เกิดการรวมเอกสารเพราะ ไม่มีเอกสารกลุ่มใดคล้ายกันเลยใน Distance 40 เอกสารทุกกลุ่มต่างเป็นกลุ่มเฉพาะของตนเอง ทำให้ได้ค่า Precision 100 % แต่ค่า Recall เฉลี่ยจะต่ำที่สุด เพราะได้กลุ่มที่มีขนาดกลุ่มเล็กเกินไปมีสมาชิกน้อยมากในกลุ่ม

◆ พิจารณาวิธี SKD มาหาระยะ (Distance) 100 เอาทุกคำทั้งเอกสารUnknown และกับกลุ่มเปรียบเทียบ จะได้ผลการจัดกลุ่มเอกสารดังในรูปที่ 7.6



รูปที่ 7.6 ผลการรวมกลุ่มเอกสารภาษาไทยวิธีSKDหาค่า Distance 100 พิจารณาทุกคำทั้งเอกสาร Unknown และกลุ่มเปรียบเทียบ

ตารางที่ 7.5 ผลการจัดกลุ่มวิธีSKDหาค่า Distance 100พิจารณาทุกคำทั้งเอกสารUnknown และกับกลุ่มเปรียบเทียบและเทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	Precision	Recall	F-measure
1	132	31	1	0.02	0.05
2	121	50	1	0.08	0.15
3	132	58	1	0.02	0.05
4	131	3	1	0.25	0.4
5	133	6	1	0.1	0.18
6	132	8	1	0.02	0.05
7	132	2	1	0.02	0.05
8	111	18,22	1	0.13	0.24

ตารางที่ 7.5 (ต่อ)

9	111	20	1	0.07	0.13
10	132	32	1	0.02	0.05
11	133	4,36,37,38	1	0.4	0.57
12	111	30	1	0.07	0.13
13	11	19,27	1	0.12	0.21
14	132	10,14	1	0.05	0.09
15	111	42	1	0.07	0.13
16	132	56	1	0.02	0.05
17	112	17	1	0.5	0.67
18	132	55,60,67,75,82	0.8	0.1	0.17
19	111	24	1	0.07	0.13
20	31	93,97	1	0.08	0.15
21	31	87,89	1	0.08	0.15
22	31	90,103	1	0.08	0.15
23	321	115	1	0.03	0.06
24	322	135	1	0.06	0.11
25	31	100,101,104,202	0.75	0.16	0.26
26	31,322	107,174	0.5	0.05	0.09
27	32	91,121	1	0.04	0.08
28	321	134,136	1	0.07	0.13
29	321	125,177,196,204	0.75	0.1	0.18
30	322	182,184,194,200	0.75	0.18	0.3
31	23	224	1	0.06	0.12
32	23	220	1	0.06	0.12
33	21	141,143	1	0.12	0.21
34	23	237,239,254,255,257	0.6	0.19	0.29
35	22	249,251,253	0.67	0.18	0.28
36	22	258,265	1	0.18	0.3
37	122	1	1	0.5	0.67
38	131	5,64,69,266	0.5	0.5	0.5

ตารางที่ 7.5 (ต่อ)

39	131	7	1	0.25	0.4
40	133,132	9,85,247	0.33	0.06	0.04
41	132	11,13	1	0.05	0.09
42	133	12	1	0.1	0.18
43	132	15	1	0.02	0.04
44	132	16,74	1	0.05	0.09
45	111	21	1	0.07	0.13
46	111	23	1	0.07	0.13
47	132	26	1	0.02	0.05
48	121	28,29,44,45,48,49,52	0.86	0.5	0.95
49	132	33	1	0.02	0.05
50	133	34,39,71	0.67	0.2	0.15
51	132	35	1	0.02	0.05
52	132	40,41	1	0.05	0.09
53	132	43,57,59,65,72,77,78	0.86	0.15	0.25
54	121	46	1	0.08	0.15
55	111	47,62	1	0.12	0.21
56	111	51,53,80,81,84	0.6	0.2	0.3
57	121	54	1	0.08	0.15
58	132	61,83	1	0.05	0.09
59	132	66,68	1	0.05	0.09
60	132	70	1	0.02	0.05
61	121	73	1	0.08	0.15
62	132	76	1	0.02	0.05
63	132	79	1	0.02	0.05
64	31	86,98,271	0.67	0.08	0.14
65	31	88	1	0.04	0.07
66	31	92,94	1	0.08	0.15
67	31	95,106	1	0.08	0.15
68	31	96	1	0.04	0.07

ตารางที่ 7.5 (ต่อ)

69	31	99	1	0.04	0.07
70	31	102	1	0.04	0.07
71	31	105,107,108,173	1	0.15	0.27
72	31	109	1	0.04	0.07
73	31	111	1	0.04	0.07
74	322	114,176,187	0.67	0.12	0.2
75	321	117,119,123,131	1	0.13	0.24
76	321	130,178,214	1	0.1	0.18
77	321	138	1	0.03	0.06
78	21,22	142,150,151,157,162, 230,236,259,270	0.44	0.24	0.33
79	23	148,219,221,225,232, 245,246	0.86	0.38	0.52
80	21	149,152,160,163,228, 252,260	0.71	0.29	0.41
81	31	171	1	0.04	0.07
82	322	179	1	0.06	0.11
83	321	180,208,211,212,213	1	0.17	0.29
84	321	181,183,190,201,205	0.8	0.13	0.23
85	321	189,195	1	0.07	0.13
86	322	191	1	0.06	0.11
87	322	192	1	0.06	0.11
88	322	203	1	0.06	0.11
89	321	206	1	0.03	0.06
90	322	209	1	0.06	0.11
91	321	210	1	0.03	0.06
92	21	227,256,261,262	0.5	0.11	0.18
Precision เฉลี่ย	0.93				
Recall เฉลี่ย	0.11				

ตารางที่ 7.6 (ต่อ)

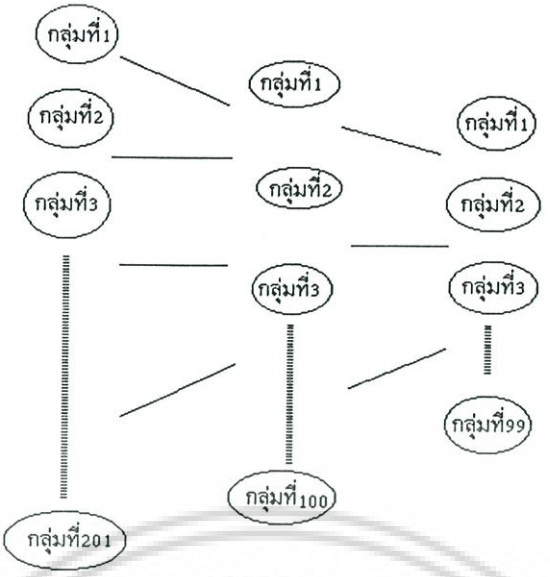
7	112,131	17,67	0.5	0.38	0.43
8	121	28,29,44,46,48,49,51,52	0.75	0.5	0.63
9	111, 121	42,45,60,62,85	0.4	0.15	0.22
10	133	36,37,38,39,40,68	0.67	0.4	0.5
11	132	43,55,56,57,58,65,69,72, 73	0.78	0.17	0.28
12	133,132	12,41	0.5	0.06	0.11
13	132	32	1	0.02	0.04
14	111	22	1	0.07	0.13
15	132	16,59,61,70,74,77,78,81, 83,84	0.9	0.22	0.35
16	131	5	1	0.25	0.4
17	132	2,34,35	0.67	0.05	0.09
18	131,132	7,26	0.5	0.14	0.22
19	122,111	1,23	0.5	0.29	0.37
20	111	18	1	0.07	0.13
21	121	54	1	0.08	0.15
22	111	24	1	0.07	0.13
23	31	98,99,101,171,173,84	0.83	0.19	0.31
24	21	143,148,149,150,152,160, 163,228,230,236,237,239, 257,259,260	0.53	0.47	0.5
25	31	97	1	0.04	0.08
26	31	93,95	1	0.08	0.15
27	321	125,130,180,181,183,190, 195,201,213	0.89	0.27	0.41
28	321	136,182,204,205,208,211, 212	0.86	0.2	0.32
29	322	89,107,127,174,176,179, 187,196,202,266	0.6	0.35	0.44

ตารางที่ 7.6 (ต่อ)

30	321	114,119,123,131,191,200	0.83	0.17	0.28
31	321	115,117,134,178,189,192, 210,214	0.75	0.2	0.32
32	32	135,138	1	0.04	0.08
33	31	88,90,100,102,103,105, 106,108,109,	1	0.35	0.52
34	322	91,121,177,194,203,206, 209	0.71	0.29	0.41
35	23	94,96,220,221,225,232, 245,254,256	0.67	0.38	0.48
36	22	86,224,227,249,251,252, 253,255,261,265,271	0.5	0.45	0.47
37	31	87,92,111	1	0.12	0.21
38	21	20,141,142,151,157,162, 219,246,258,262,270	0.45	0.29	0.35
39	111	19,21,247	0.67	0.13	0.22
Precision เฉลี่ย	0.77				
Recall เฉลี่ย	0.20				
F-measure เฉลี่ย	0.29				

7.2 ผลลัพธ์การพิจารณาค่าน้ำหนักคำตามความถี่คำที่เหมือนกัน ระหว่างกลุ่มเอกสาร (Probabilistic Document Overlap)(PDO) หรือ (Keywords Frequency)(KF)

วิธีพิจารณาให้น้ำหนักคำตามความถี่คำสำคัญ ซึ่งจะดูว่าเอกสาร unknown ใกล้เคียงกับกลุ่มใดมากที่สุดตามสูตรการหาค่า PDO ก็จะเข้าไปรวมกับกลุ่มนั้น ได้ผลการทดลองดังนี้



รูปที่ 7.8 ผลการรวมกลุ่มเอกสารวิธี PDO

ตารางที่ 7.7 ผลการจัดกลุ่มวิธี PDO เทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	Precision	Recall	F-measure
1	132	2,4,5	1	0.05	0.1
2	131	3,6,7	0.67	0.5	0.57
3	132,133	8,9	0.5	0.06	0.11
4	132	10,11	1	0.05	0.1
5	132,133	12,13	0.5	0.06	0.11
6	132	14,15	1	0.05	0.1
7	132,112	16,17	0.5	0.26	0.34
8	111	18,19	1	0.13	0.23
9	111	20,23	1	0.13	0.23
10	111,132	24,32	0.5	0.05	0.1
11	132,133	33,34	0.5	0.06	0.11
12	133	36,37	1	0.2	0.33
13	133	38,39	1	0.2	0.33
14	132	40,41	1	0.05	0.1
15	132,322	61,91	0.5	0.04	0.07

ตารางที่ 7.7 (ต่อ)

16	31,321	111,121	0.5	0.04	0.07
17	111	21,22	1	0.13	0.23
18	132	25,26	1	0.05	0.1
19	112,121	27,28	0.5	0.29	0.37
20	121,111	29,30	0.5	0.08	0.06
21	132	31,35	1	0.05	0.1
22	111,122	42,43	0.50	0.29	0.37
23	121	44,45	1	0.08	0.15
24	121,111	46,47	0.5	0.08	0.06
25	121	48,49	1	0.08	0.15
26	121,111	50,51	0.5	0.08	0.06
27	111	52,53	1	0.13	0.23
28	121,132	54,55	0.5	0.05	0.1
29	132	56,57	1	0.05	0.1
30	132	58,59	1	0.05	0.1
31	133,111	60,62	0.5	0.09	0.15
32	131,132	64,65	0.5	0.14	0.22
33	132	66,67	1	0.05	0.1
34	132	68,69	1	0.05	0.1
35	132	70,71	1	0.05	0.1
36	121	1,72,73	0.67	0.13	0.22
37	132	74,75	1	0.05	0.1
38	132	76,77	1	0.05	0.1
39	132	78,79	1	0.05	0.1
40	132,111	80,81	0.5	0.05	0.1
41	132	82,83	1	0.05	0.1
42	132,121	84,85	0.5	0.05	0.1
43	31	86,87	1	0.04	0.08
44	31	88,89	1	0.04	0.08

ตารางที่ 7.7 (ต่อ)

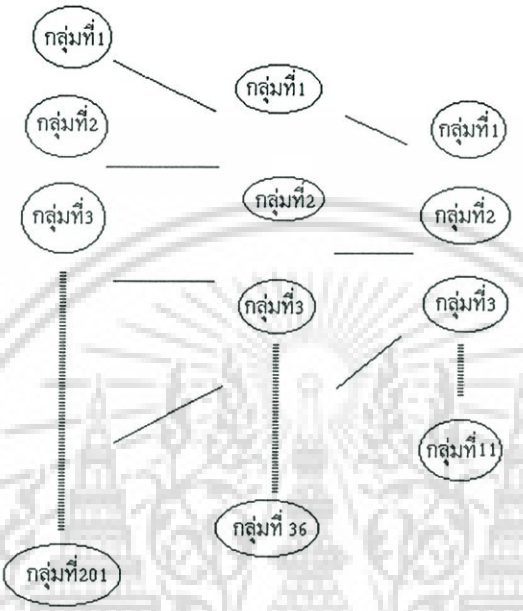
45	31	90,92	1	0.04	0.08
46	31	93,94	1	0.04	0.08
47	31	95,96	1	0.04	0.08
48	31	97,98	1	0.04	0.08
49	31	99,100	1	0.04	0.08
50	31	101,102	1	0.04	0.08
51	31	103,104	1	0.04	0.08
52	31	105,106	1	0.04	0.08
53	31	107,108	1	0.04	0.08
54	31,321	109,114	0.5	0.04	0.07
55	321	115,117	1	0.07	0.13
56	321	119,123	1	0.07	0.13
57	32	125,127	1	0.04	0.08
58	321	130,131	1	0.07	0.13
59	32	134,135	1	0.04	0.08
60	321	136,138	1	0.07	0.13
61	21,23	141,142	0.5	0.06	0.11
62	21	143,148	1	0.12	0.21
63	21	149,150	1	0.12	0.21
64	21	151,152	1	0.12	0.21
65	23,21	157,160	0.5	0.06	0.11
66	21	162,163	1	0.12	0.21
67	31	171,173	1	0.08	0.15
68	322	174,176	1	0.12	0.21
69	32	177,178	1	0.04	0.08
70	32	179,180	1	0.04	0.08
71	32	181,182	1	0.04	0.08
72	32	183,184	1	0.04	0.08
73	32	187,189	1	0.04	0.08
74	322	190,191	1	0.12	0.21

ตารางที่ 7.7 (ต่อ)

75	322	192,194	1	0.12	0.21
76	321	195,196	1	0.07	0.13
77	321	200,201	1	0.07	0.13
78	322	202,203	1	0.12	0.21
79	321	204,205	1	0.07	0.13
80	321	206,208	1	0.07	0.13
81	32	209,210	1	0.04	0.08
82	321	211,212	1	0.07	0.13
83	321	213,214	1	0.07	0.13
84	23	219,220	1	0.13	0.23
85	23	221,224	1	0.13	0.23
86	23	225,227	1	0.13	0.23
87	21,23	228,230	0.5	0.06	0.11
88	23	232,236	1	0.13	0.23
89	23	237,239	1	0.13	0.23
90	23	245,246	1	0.13	0.23
91	132,22	247,249	0.5	0.06	0.11
92	22,21	251,252	0.5	0.8	0.06
93	21,23	253,254	0.5	0.06	0.11
94	22	255,256	1	0.18	0.31
95	22	257,258	1	0.18	0.31
96	22	259,260	1	0.18	0.31
97	21	261,262	1	0.12	0.21
98	22	265,266	1	0.18	0.31
99	21,22	270,271	0.5	0.08	0.14
Precision เฉลี่ย	0.87				
Recall เฉลี่ย	0.10				
F-measure เฉลี่ย	0.15				

7.3 ผลลัพธ์การพิจารณาน้ำหนักของคำที่เหมือนกันตามจำนวนเอกสารในกลุ่ม (Document Frequency) (DF)

หรือแบบวิธีพิจารณาจำนวนเอกสารกับคำสำคัญ ซึ่งจะดูว่าเอกสาร unknown ใกล้เคียงกับกลุ่มใดมากที่สุดตามสูตรการหาค่า DF ก็จะเข้าไปรวมกับกลุ่มนั้น ได้ผลการทดลองดังนี้



รูปที่ 7.9 ผลการรวมกลุ่มเอกสารวิธี DF

ตารางที่ 7.8 ผลการจัดกลุ่มวิธี DF เทียบกับการจัดกลุ่มด้วยมนุษย์

ลำดับกลุ่มที่จัดด้วยคอมพิวเตอร์	กลุ่มที่จัดโดยมนุษย์	สมาชิกในกลุ่มที่จัดโดยคอมพิวเตอร์	Precision	Recall	F-measure
1	131	3,7	1	0.5	0.66
2	133	6,9	1	0.2	0.33
3	132	8,11,13	1	0.073	0.136
4	132	10,14	1	0.048	0.093
5	132	2	1	0.02	0.047
6	111	20,23	1	0.13	0.24
7	111	18,21,22,24	1	0.2	0.44
8	132	32,33	1	0.048	0.093
9	133	36,37	1	0.2	0.33
10	133	34,38,39	1	0.3	0.46

ตารางที่ 7.8 (ต่อ)

11	321	31,35,40,41,42,46,47,50,51 ,52,53,54,55,56,57,59,60, 61,62,64,65,66,67,68,69, 70,71,72,73,74,75,76,77, 78,79,80,81,82,83,84,85, 86,87,88,89,90,91,92,93, 94,95,96,97,98,99,100,101, 102,103,104,105,106,107, 108,109,111,114,115,117, 119,121,123,125,127,130, 131,134,135,136,138,171, 173,174,176,177,179,180, 181,182,183,184,187,189, 190,191,192,194,195,196, 200,201,202,203,204,205, 206,208,209,210,211,212, 213,214,220,227,239,246, 249,251,253,255,256,257, 258,259,260,261,262,265, 266,270,271	0.26	1	0.41
12	132	15	1	0.024	0.05
13	132	26	1	0.024	0.05
14	121	28	1	0.08	0.142
15	132	16	1	0.024	0.05
16	112	17	1	0.5	0.66
17	121	44,45	1	0.15	0.26
18	111	30	1	0.066	0.13
19	121	48,49	1	0.15	0.27
20	112	27	1	0.5	0.66
21	122	43	1	0.25	0.4
22	131	5	1	0.25	0.4

ตารางที่ 7.8 (ต่อ)

23	133	12	1	0.1	0.18
24	133	4	1	0.1	0.18
25	132	58	1	0.024	0.047
26	122	1	1	0.5	0.66
27	132	25	1	0.02	0.047
28	21	141,150,151,162	1	0.23	0.38
29	21	149,160,228	1	0.18	0.3
30	111	19	1	0.07	0.13
31	23	142,224	1	0.06	0.12
32	23	221,225,232,245	1	0.25	0.4
33	23	157,230,236,237	1	0.25	0.4
34	21	143,148	1	0.11	0.21
35	21	219	1	0.06	0.11
36	21	152,163,252	1	0.18	0.3
Precision เฉลี่ย	0.98				
Recall เฉลี่ย	0.20				
F-measure เฉลี่ย	0.28				

◆ สรุปผลเฉลี่ยเครื่องมือวัดทุกวิธีการจัดกลุ่มเอกสารภาษาไทยทั้งหมด เทียบกับการจัดกลุ่มด้วยมนุษย์ เป็นดังตารางข้างล่าง

ตารางที่ 7.9 แสดงผลเฉลี่ยเครื่องมือวัดวิธีการจัดกลุ่มเอกสารภาษาไทยทั้งหมด

วิธีการจัดกลุ่มเอกสาร	จำนวนกลุ่ม	Precision เฉลี่ย	Recall เฉลี่ย	F-measure เฉลี่ย
SKD ไม่รวมค่าDistance	39	0.81	0.17	0.29
SKD รวม Distance 60 (intersec)	125	0.94	0.09	0.15
SKD รวม Distance 90 (intersec)	85	0.90	0.11	0.19
SKD รวม Distance 100 ทั้ง intersec และ ไม่ intersec	92	0.93	0.11	0.18
SKD รวม Distance 300 ทั้ง intersec และ ไม่ intersec	39	0.77	0.2	0.29
PDO	99	0.87	0.10	0.15
DF	36	0.98	0.20	0.28

7.4 สรุปผลการทดลองวิจัย

1. เครื่องมือวัดผลการจัดกลุ่มเอกสารภาษาไทย เมื่อเปรียบเทียบกับการจัดกลุ่มเอกสารด้วยมนุษย์ มีความหมายดังนี้

- ค่า Precision จะบ่งบอกว่าวิธีการจัดกลุ่มเอกสารด้วยคอมพิวเตอร์สามารถจัดกลุ่มเอกสารได้ถูกต้องตรงกับการจัดกลุ่มด้วยมนุษย์มากน้อยเท่าไร โดยมีค่าตั้งแต่ 1ถึง 0 ซึ่งค่า เท่ากับ 1 หมายถึง คอมพิวเตอร์จัดกลุ่มเอกสารได้ถูกต้องมีสมาชิกในกลุ่ม เหมือนกับที่จัดโดยมนุษย์ ส่วนค่าเท่ากับ 0 หมายถึง คอมพิวเตอร์จัดกลุ่มเอกสารไม่มีสมาชิกในกลุ่มเหมือนที่จัดโดยมนุษย์เลย หรือ อาจกล่าวได้ว่า ค่า Precision นี้ จะพิจารณาคุณภาพของสมาชิกที่ถูกจัดกลุ่มด้วยคอมพิวเตอร์ว่ามีเนื้อหาใกล้เคียงกันหรือไม่

- ค่า Recall จะบอกว่า กลุ่มเอกสารที่จัดด้วยคอมพิวเตอร์ มีสมาชิกเหมือนในกลุ่มที่จัดโดยมนุษย์จำนวนมากน้อยเท่าไร ซึ่งมีค่าตั้งแต่ 1 ถึง 0 โดยที่ 1 หมายถึง กลุ่มเอกสารที่จัดด้วยคอมพิวเตอร์มีปริมาณสมาชิกทุกตัวเหมือนในกลุ่มเอกสารที่จัดด้วยมนุษย์ ส่วนค่า 0 หมายถึง ไม่มีปริมาณสมาชิกในกลุ่มที่จัดด้วยคอมพิวเตอร์ เหมือนในกลุ่มที่จัดด้วยมนุษย์เลย แม้แต่ตัวเดียว หรือ กล่าวได้ว่าค่า Recall นี้จะบ่งบอกปริมาณของสมาชิกที่ถูกจัดกลุ่มด้วยคอมพิวเตอร์ว่าถูกต้องมากเท่าไร

- ค่า F- measure จะบอกถึงค่าความแตกต่างระหว่างค่า Precision กับค่า Recall ว่าแตกต่างกันมากน้อยเท่าไร ซึ่งมีค่าอยู่ระหว่าง 1 ถึง 0 เช่นกัน โดยที่ 1หมายถึง ค่า Precision เท่ากับค่า Recall นั้นคือ การจัดกลุ่มด้วยคอมพิวเตอร์ สมาชิกในกลุ่มมีจำนวนสมาชิก

และเหมือนกับในกลุ่มที่จัดด้วยมนุษย์ทุกอย่าง ส่วนค่า 0 คือ ค่า Precision และค่า Recall เท่ากับ 0 ทั้งคู่ นั้นคือ เอกสารที่จัดด้วยคอมพิวเตอร์สมาชิกในกลุ่มไม่เหมือนกับ กลุ่มที่จัดโดยมนุษย์เลย

2. วิธีพิจารณาค่าความคล้ายแบบไม่คิดน้ำหนักและความถี่ (SKD) พิจารณาแต่คำที่เหมือนกันอย่างเดียว ไม่ดูค่า Distance ร่วมด้วย จะจัดกลุ่มที่ทำให้สมาชิกในกลุ่มผิดกลุ่มมาก นั้นหมายถึงการจัดกลุ่มเอกสารจะใช้ วิธีดูแต่คำที่เหมือนกันอย่างเดียว โดยไม่สนใจอย่างอื่นเลยแล้วทำการรวมกลุ่มกันนั้นไม่เหมาะสม

3. วิธี SKD ที่ใช้ร่วมกับการหาค่า Distance ผลการจัดกลุ่มเกิด การ Error เนื่องจากการสะสม keywords น้อย (Error เนื่องจากการสะสม Keywords หมายถึง เมื่อเอกสารมีการรวมกลุ่มจะมี keywords รวมกันเพิ่มขึ้น เมื่อนำเอกสารอื่นมาเปรียบเทียบกับกลุ่มนี้ ก็มีโอกาสจะเข้ารวมกับกลุ่มนี้มากกว่ากลุ่มอื่นได้) การจัดกลุ่มวิธีนี้จะสามารถควบคุมสมาชิกในกลุ่มได้ว่า ต้องการให้สมาชิกในกลุ่มมีเนื้อหาใกล้เคียงกันมากหรือน้อยเท่าไรก็ได้ ซึ่งถ้าสมาชิกมีเนื้อหาในกลุ่มใกล้กันมาก จะได้จำนวนกลุ่มมาก แต่ถ้าสมาชิกในกลุ่มมีเนื้อหาไม่ใกล้เคียงกัน จะมีจำนวนกลุ่มน้อยเนื่องจากในกลุ่มแต่ละกลุ่มมีเอกสารเป็นสมาชิกอยู่จำนวนมากนั้นเอง วิธีนี้การควบคุมสมาชิกในกลุ่มและจำนวนกลุ่มเอกสารไม่ดีนัก ต้องทำการปรับลดค่า Distance หลายๆค่าจนกว่าจะได้กลุ่มที่เหมาะสม

4. วิธี SKD ที่ใช้ร่วมกับการหาค่า Distance ก็มีปัญหาในการจัดกลุ่มด้วยคอมพิวเตอร์เช่นกัน นั้นคือ บางเอกสารสามารถเข้าไปรวมกลุ่มได้มากกว่า 1 กลุ่ม ที่ Distance(Threshold) ที่เรากำหนดทำให้ระบบการจัดกลุ่มเอกสารไม่หยุดทำงานเนื่องจากเอกสาร มีการเปลี่ยนกลุ่มไปมาตลอดเวลา การจะหาค่า Distance (Threshold) ที่เหมาะสมได้ต้องขึ้นกับตัวอย่างเอกสารที่นำมาจัดกลุ่ม และต้องทดสอบค้นหาเลือกเอาเฉพาะค่า Distance ที่เหมาะสมเอาเอง

5. วิธี PDO จะเป็นการพิจารณาความถี่ของคำที่เหมือนกันระหว่างเอกสาร Unknown กับกลุ่มเปรียบเทียบ ซึ่งจะให้ผลลัพธ์ไม่ดีนักเนื่องมาจากคำสำคัญ(Keywords) ในเอกสารหรือคำซึ่งเป็นตัวบ่งบอกเนื้อหาในเอกสาร อาจมีจำนวนน้อยกว่าคำอื่นที่ไม่ใช่ Keywords ก็ได้ เช่น คำเชื่อมต่างๆ (และ,แต่,ว่า,หรือ ฯลฯ) , ดู , ปฏิบัติ , เป็นต้น, ทำ, ไป ฯลฯ ทำให้การรวมกลุ่มเอกสารผิดพลาด

6. การจัดกลุ่มเอกสารภาษาไทยที่ใช้วิธี การหาความคล้ายของคำสำคัญระหว่างกลุ่มเอกสารแบบพิจารณาน้ำหนักคำที่เหมือนกันตามจำนวนเอกสาร (Document Frequency) (DF) จะให้ผลการจัดกลุ่มที่มีค่าจากเครื่องมือวัด คือ Precision , Recall และ F-measure มีค่าดีที่สุด และได้ขนาดจำนวนกลุ่มเอกสารรวมทั้งสมาชิกในกลุ่มเอกสารที่เหมาะสมกว่าวิธีอื่น แต่มีข้อเสียบางคือ ในบางกลุ่มขณะที่ทำการจัดเอกสารด้วยคอมพิวเตอร์ จะเกิดการ Error เนื่องจากการสะสม keywords ได้ทำให้เกิดการดึงเอาเอกสารต่างๆมารวมในกลุ่มมากเกินไป ดูในตารางที่ 7.8 กลุ่มที่ 11 กลุ่มกฎหมายทั่วไป แต่ก็ถือว่าเป็นวิธีการจัดกลุ่มเอกสารภาษาไทยที่ดี เพราะระบบจะหยุดทำงานได้เนื่องจากเอกสารไม่มีการเปลี่ยนกลุ่มไปมา ไม่เหมือนกับวิธี SKD และไม่ต้องหาค่า Threshold ด้วย นอกจากนี้เนื้อหาเอกสารและปริมาณเอกสารในกลุ่มก็ดีกว่าวิธีอื่นๆ สิ่งที่ควรปรับปรุงคือเรื่อง

Error เนื่องจากการสะสมของ Keywords ในบางกลุ่ม ซึ่งระบบอาจต้องมีการตรวจสอบซ้ำกับกลุ่มเอกสารที่ใช้อ้างอิง (การจัดกลุ่มด้วยมนุษย์)

7. จากการทดลองวิจัยนี้ ค่า Recall และค่า F-measure น้อย น่าจะมาจากสาเหตุ การจัดกลุ่มด้วยมนุษย์ที่นำมาเปรียบเทียบ มีขนาดกลุ่มใหญ่เกินไป ถ้าแยกเป็นกลุ่มย่อยลงไปได้อีก จะทำให้ค่าทั้ง 2 มีมากขึ้นได้

8. หลังการจัดกลุ่มด้วยคอมพิวเตอร์แล้ว มาทำการให้น้ำหนักคำสำคัญในเอกสาร ของแต่ละกลุ่มเอกสารที่จัดกลุ่มด้วยคอมพิวเตอร์จะแตกต่างกันบ้างทั้งที่เป็นเอกสารฉบับเดียวกัน และคำเดียวกัน ทั้งนี้เพราะขึ้นกับสมาชิกในกลุ่มเอกสารนั้นๆว่ามีคำสำคัญมากน้อยอย่างไร อันเป็นผลจากวิธีการจัดกลุ่มเอกสารนั้นเอง ตัวอย่างผลการให้น้ำหนักคำดูได้จากภาคผนวก ข นอกจากนี้ผลการให้น้ำหนักคำสามารถนำไปประยุกต์ใช้ในการจัดกลุ่มเอกสาร unknown ต่อไปได้ หรือจะตัดเอาคำที่มีน้ำหนักน้อยทิ้งไปก็ได้เพราะคำนั้นอาจเป็นตัวรบกวนระบบการจัดเอกสารในคราวต่อไปได้

เอกสารอ้างอิง

[1] สมศักดิ์ จันวัน. " ระบบวิเคราะห์โครงสร้างภาษาไทยด้วยคอมพิวเตอร์ ." วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2534.

[2] สิงห์ ตรงงาม. " ระบบการวิเคราะห์ประโยคภาษาไทยที่มีการละประธานที่ซ้ำกันในประโยค." วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมไฟฟ้า บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2540.

[3] นรเศรษฐ์ จันทสูตร. " เท็กโพรเซสซิ่งอะแคปทีพเรโซแน้นเทียรีนิวรอลเน็ตเวิร์ค. " สอบ หัวข้อเค้าโครงวิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 2545.

[4] อัญชลี วานิชทวีวัฒน์. " การจดจำอักขระภาษาไทยตัวพิมพ์โดยโครงข่ายประสาทเทียมแบบจำลองเซลท์ออร์แกนไนซิ่งแมบซ์ ." วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ บัณฑิตวิทยาลัย, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง. 254

[5] นววรรณ พันธุเมธา และ วินัย ภู่ระหงษ์. " เอกสารสอนชุดวิชาภาษาไทย 3." ครั้งที่ 8. กรุงเทพฯ : ชวนพิมพ์. 2542.

[6] ไพฑูรย์ นุชแจ้ง. "การหาผลลัพธ์การตัดคำภาษาไทยแบบ Left search matching ด้วยวิธี N-Gram." วิศวสารลาดกระบัง. ปีที่18, ฉบับที่3, กันยายน 2544. หน้า 55-60.

[7] ไพฑูรย์ นุชแจ้ง. "การแก้ความคลุมเครือการตัดคำภาษาไทย ด้วยวิธี Bigram เทียบกับวิธี CP." วิศวกรรมสาร มก. ปีที่15, ฉบับที่45, ธันวาคม 2544. หน้า 26-34.

[8] Allen Jam. "Natural Language Understanding." 2nd Ed. Redwood city : The Benjamin/ Cummings Publishing Company, Inc. 1995.

[9] Maureen Caudill and Charles Butler. "Understanding neural networks:Computer Explorations." Massachusetts: The MIT press, 1993.

[10] Sahami Mehran, "A Probabilistic Approach to Full-text Document Clustering", Computer Science Department, Stanford University, pp1-17,1999.

[11] Steinbach Michael, "A Comparison of Document Clustering techniques", Department of Computer Science and Engineering, University of Minnesota, pp1-5, 1998.

ภาคผนวก ก.

ตัวอย่างตารางข้อมูลและผลการเปรียบเทียบภาษาไทย

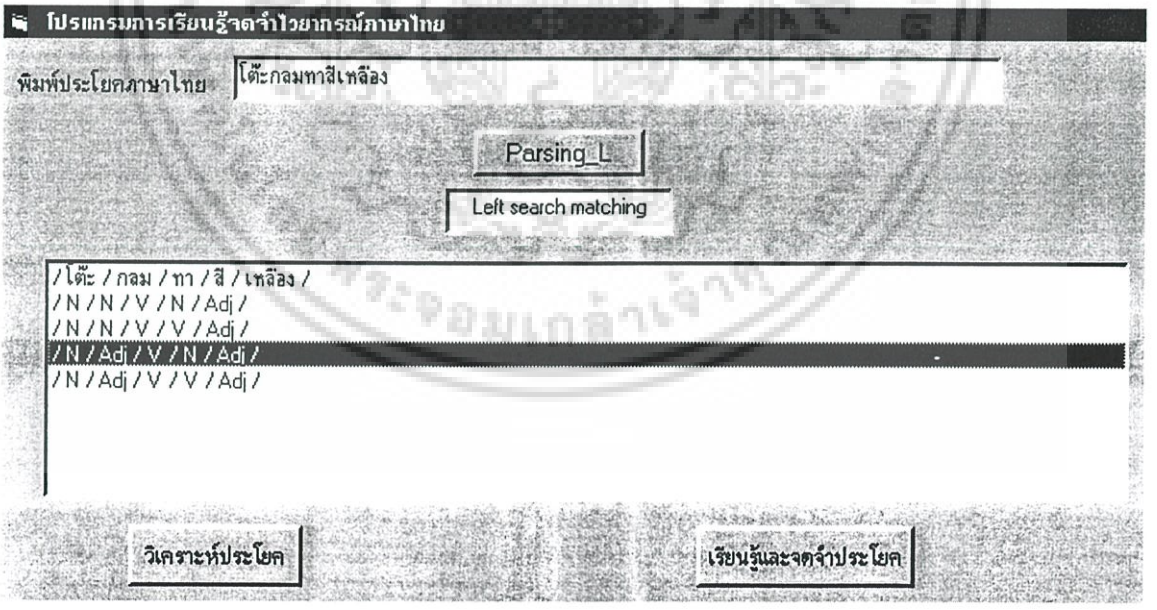
ตารางที่ ก.1 การเก็บฐานข้อมูลคำศัพท์เพื่อใช้ตัดคำภาษาไทย

Microsoft Access - [TableWord : ตาราง]

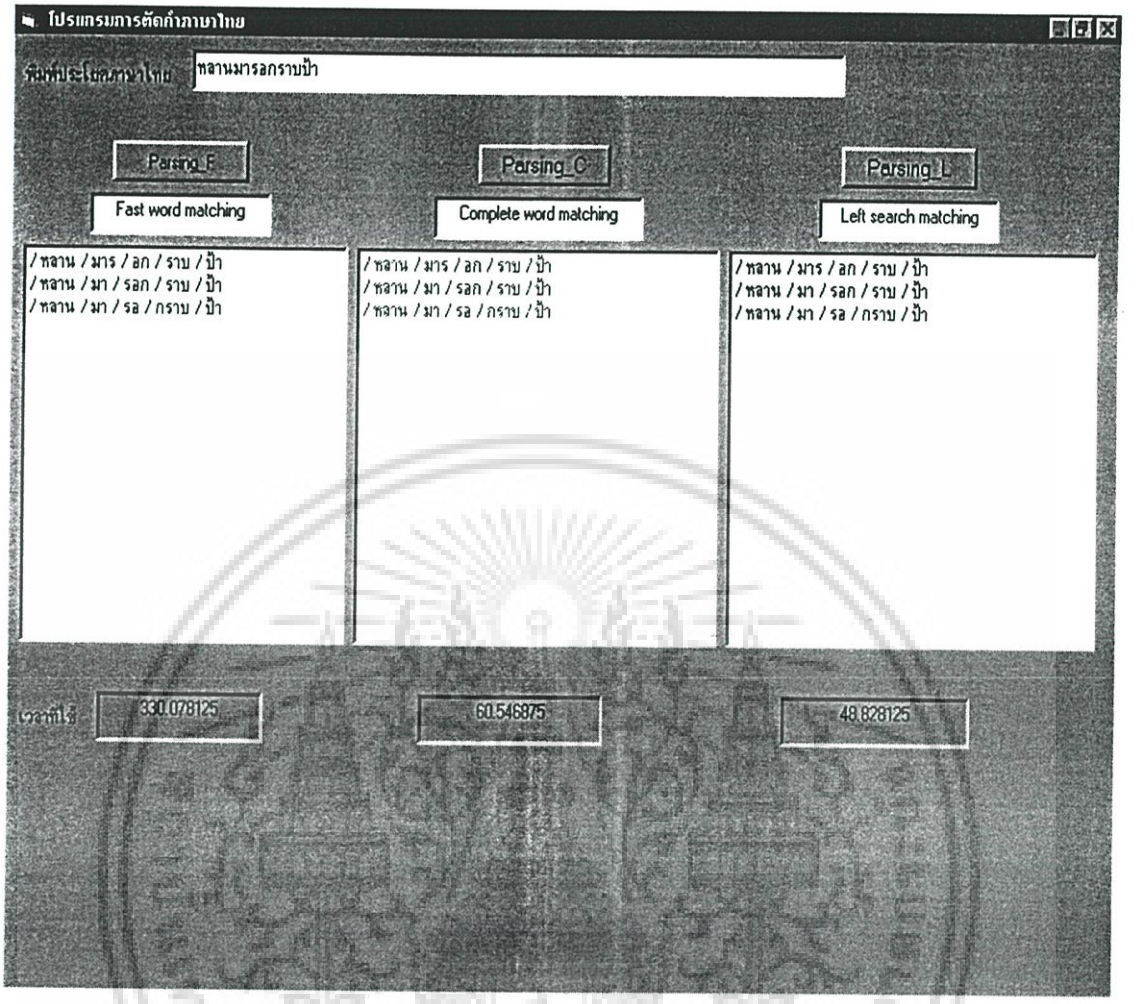
Group	Word	Leng	SameMeaning	Noun	Verb	Pronoun	Preposition	Adjective	Adverb	Conjunction
เท	เทียน	5		☑	☐	☐	☐	☐	☐	☐
เท	เที่ยว	6		☑	☑	☐	☐	☐	☐	☐
เท	เท่านั้น	8		☐	☐	☐	☐	☑	☐	☐
โท	โทรทัศน์	8	ทีวี	☑	☐	☐	☐	☐	☐	☐
เธ	เธอ	3		☐	☐	☑	☐	☐	☐	☐
แบ	แบบ	3		☑	☐	☐	☐	☐	☐	☐
แผ	แผน	3		☑	☐	☐	☐	☐	☐	☐
เพ	เพราะ	5		☐	☐	☐	☐	☐	☑	☑
เพ	เพียง	5		☐	☐	☐	☐	☑	☐	☐
เพ	เพาะปลูก	8		☐	☑	☐	☐	☐	☐	☐
เพ	เพื่อน	6		☑	☐	☐	☐	☐	☐	☐
ไพ	ไพเราะ	6		☐	☐	☐	☐	☐	☑	☐
แฟ	แฟ็กซ์	6		☑	☐	☐	☐	☐	☐	☐
ไฟ	ไฟ	2	ไฟฟ้า	☑	☐	☐	☐	☐	☐	☐
ไฟ	ไฟฟ้า	5	ไฟ	☑	☐	☐	☐	☐	☐	☐
มื	มือ	3		☑	☐	☐	☐	☐	☐	☐
ยื	ยืน	3		☐	☑	☐	☐	☐	☐	☐

ระเบียน: 1 จาก 353 NUM

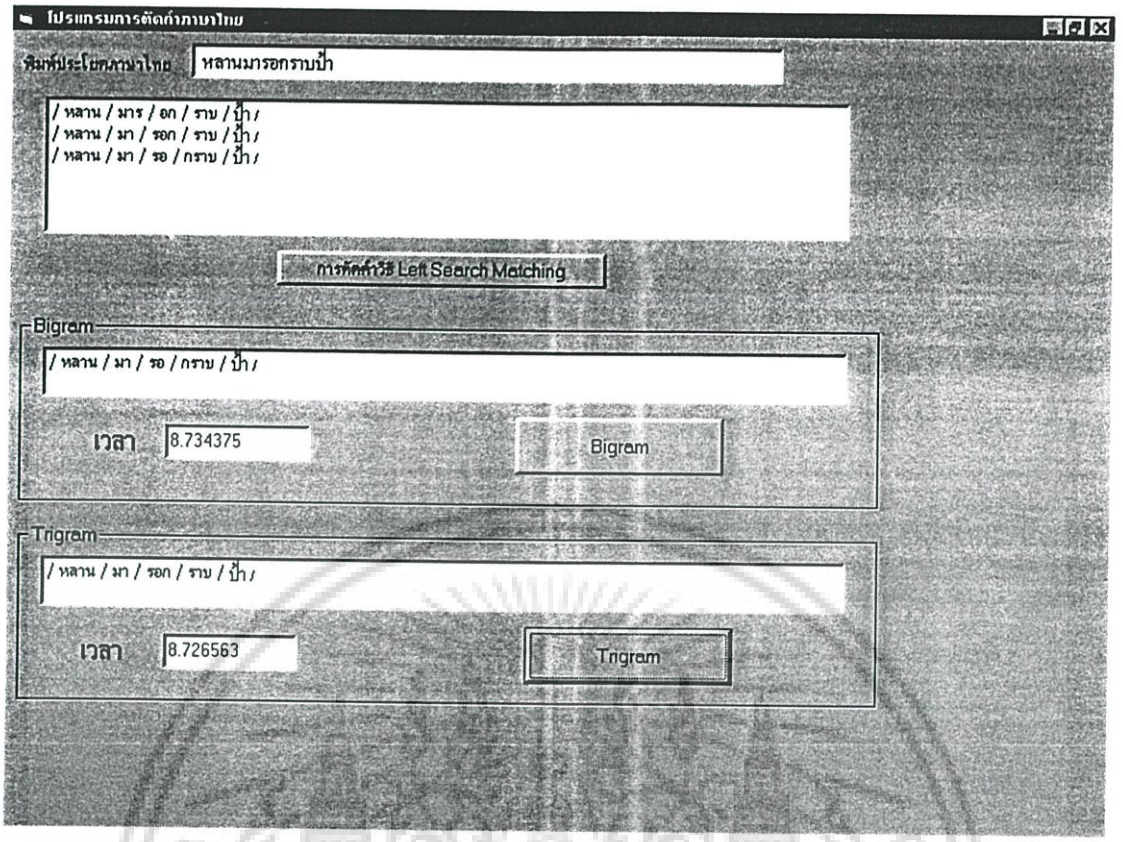
กลุ่มคำ



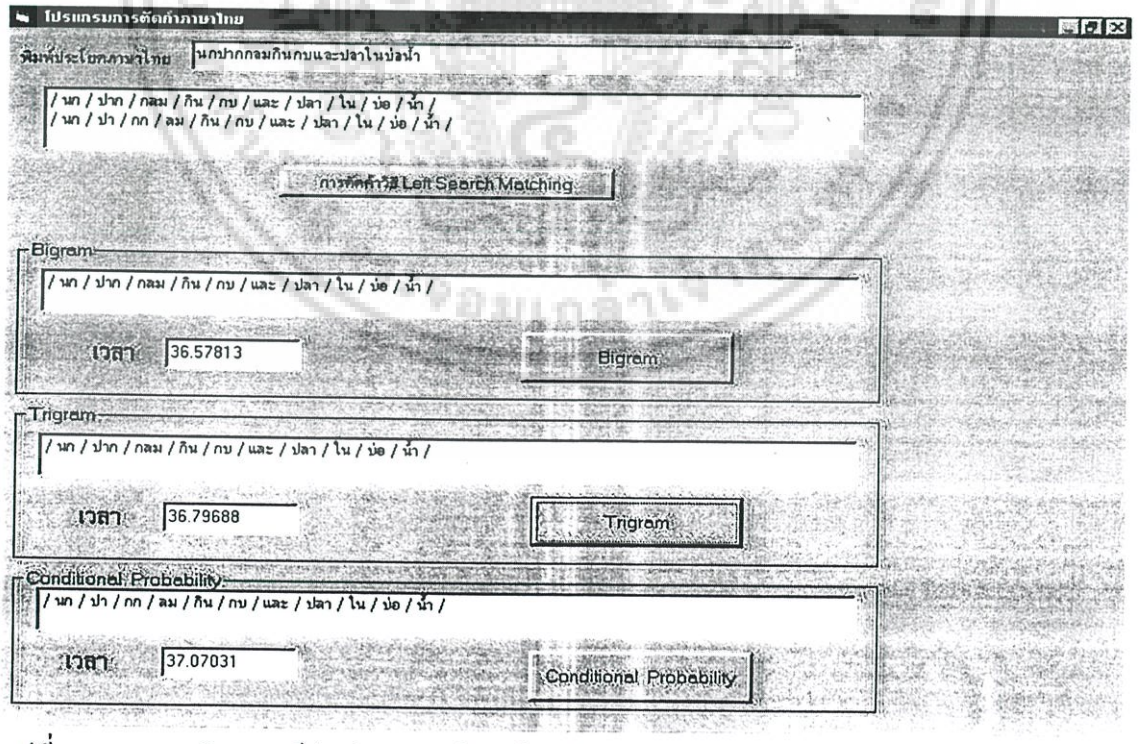
รูปที่ ก.1 โปรแกรมเรียนรู้และจดจำไวยากรณ์ภาษาไทย



รูปที่ ก.2 การเปรียบเทียบวิธีแยกหน่วยคำแบบ Fast word matching , Complete word matching และ Left search matching



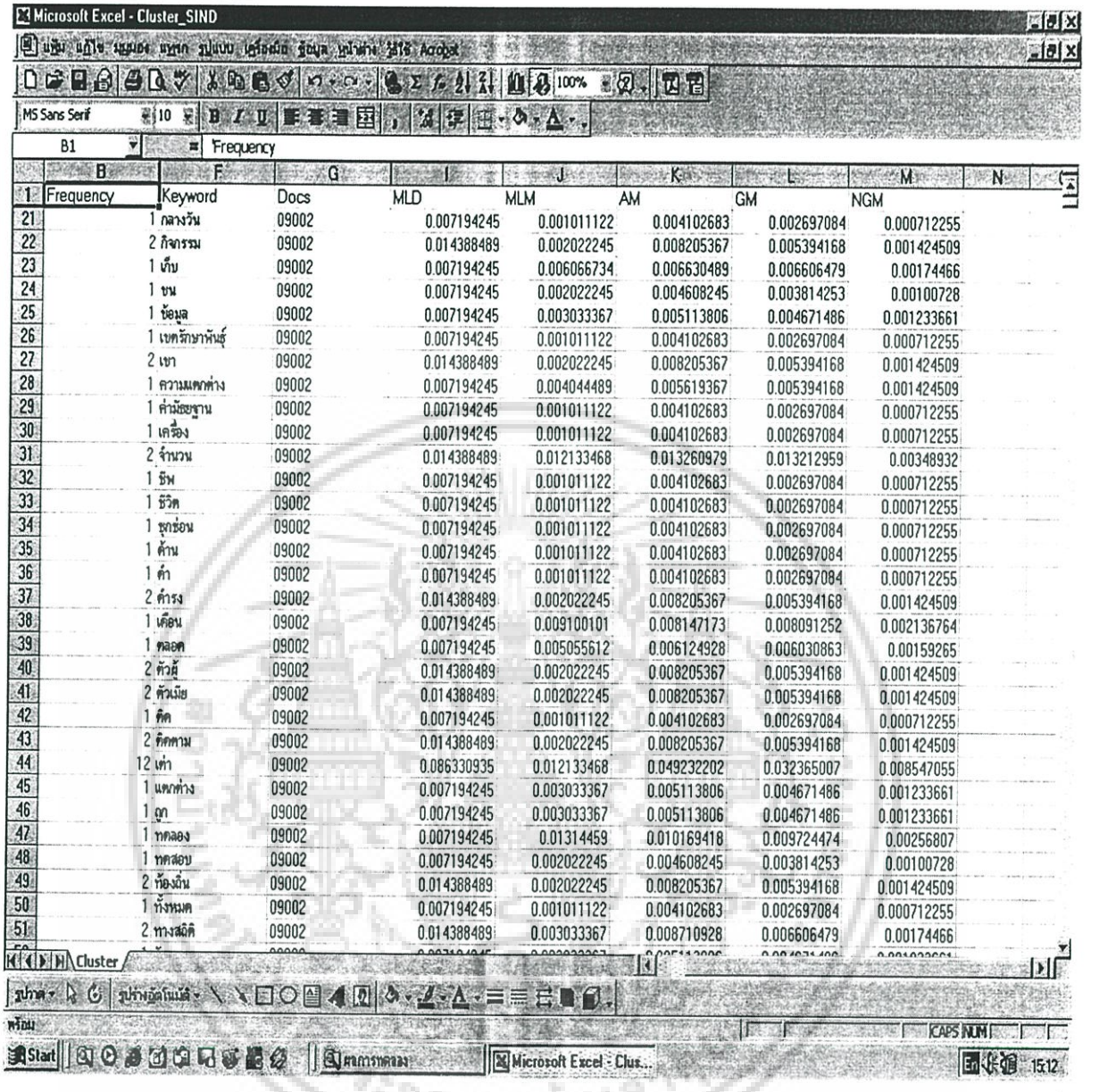
รูปที่ ก.3 ผลการวิเคราะห์ประโยคภาษาไทยด้วยวิธี Bigram กับวิธี Trigram ด้วยคอมพิวเตอร์



รูปที่ ก.4 ผลการวิเคราะห์ประโยคภาษาไทยด้วยวิธี Bigram ,Trigram และ Condition Probability (CP) ด้วยคอมพิวเตอร์

ภาคผนวก ข.

ตัวอย่างข้อมูลการจัดกลุ่มเอกสารภาษาไทย



Microsoft Excel - Cluster_SIND

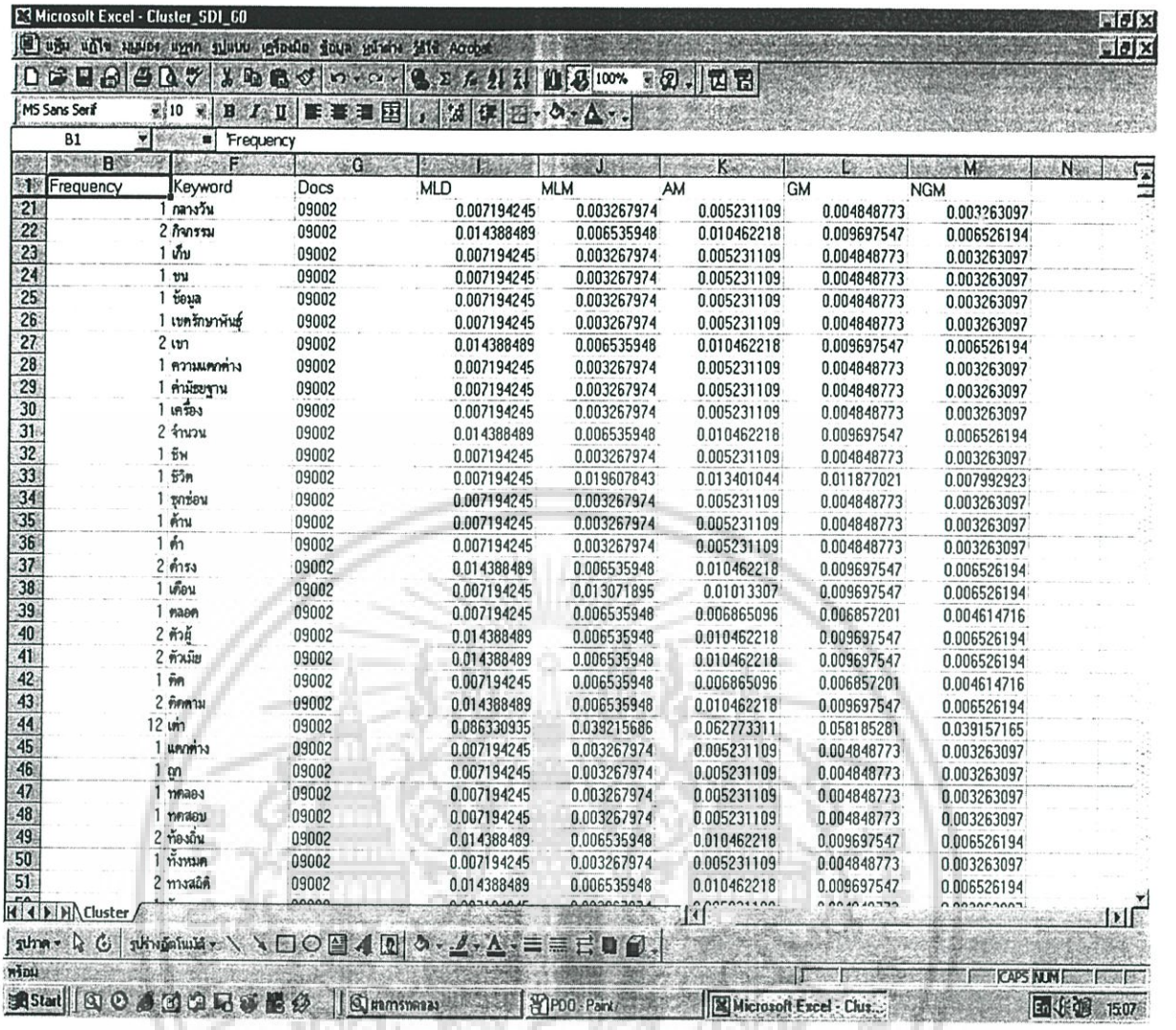
B1 = Frequency

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21	1	กลางวัน	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
22	2	กิจกรรม	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
23	1	เก็บ	09002	0.007194245	0.006066734	0.006630489	0.006606479	0.00174466	
24	1	ขน	09002	0.007194245	0.002022245	0.004608245	0.003814253	0.00100728	
25	1	ข้อมูล	09002	0.007194245	0.003033367	0.005113806	0.004671486	0.001233661	
26	1	เขตรักษาพันธุ์	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
27	2	เขา	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
28	1	ความแตกต่าง	09002	0.007194245	0.004044489	0.005619367	0.005394168	0.001424509	
29	1	คำนิยฐาน	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
30	1	เครื่อง	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
31	2	จำนวน	09002	0.014388489	0.012133468	0.013260979	0.013212959	0.00348932	
32	1	ชิพ	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
33	1	ชีวิต	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
34	1	ซุกซ่อน	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
35	1	ด้าน	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
36	1	ดำ	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
37	2	ดำรง	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
38	1	เดือน	09002	0.007194245	0.009100101	0.008147173	0.008091252	0.002136764	
39	1	ตลอด	09002	0.007194245	0.005055612	0.006124928	0.006030863	0.00159265	
40	2	ตัวผู้	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
41	2	ตัวเมีย	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
42	1	ติด	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
43	2	ติดตาม	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
44	12	เต่า	09002	0.086330935	0.012133468	0.049232202	0.032365007	0.008547055	
45	1	แตกต่าง	09002	0.007194245	0.003033367	0.005113806	0.004671486	0.001233661	
46	1	ถูก	09002	0.007194245	0.003033367	0.005113806	0.004671486	0.001233661	
47	1	ทดลอง	09002	0.007194245	0.01314459	0.010169418	0.009724474	0.00256807	
48	1	ทดสอบ	09002	0.007194245	0.002022245	0.004608245	0.003814253	0.00100728	
49	2	ท้องถิ่น	09002	0.014388489	0.002022245	0.008205367	0.005394168	0.001424509	
50	1	ทั้งหมด	09002	0.007194245	0.001011122	0.004102683	0.002697084	0.000712255	
51	2	ทางสถิติ	09002	0.014388489	0.003033367	0.008710928	0.006606479	0.00174466	

Cluster

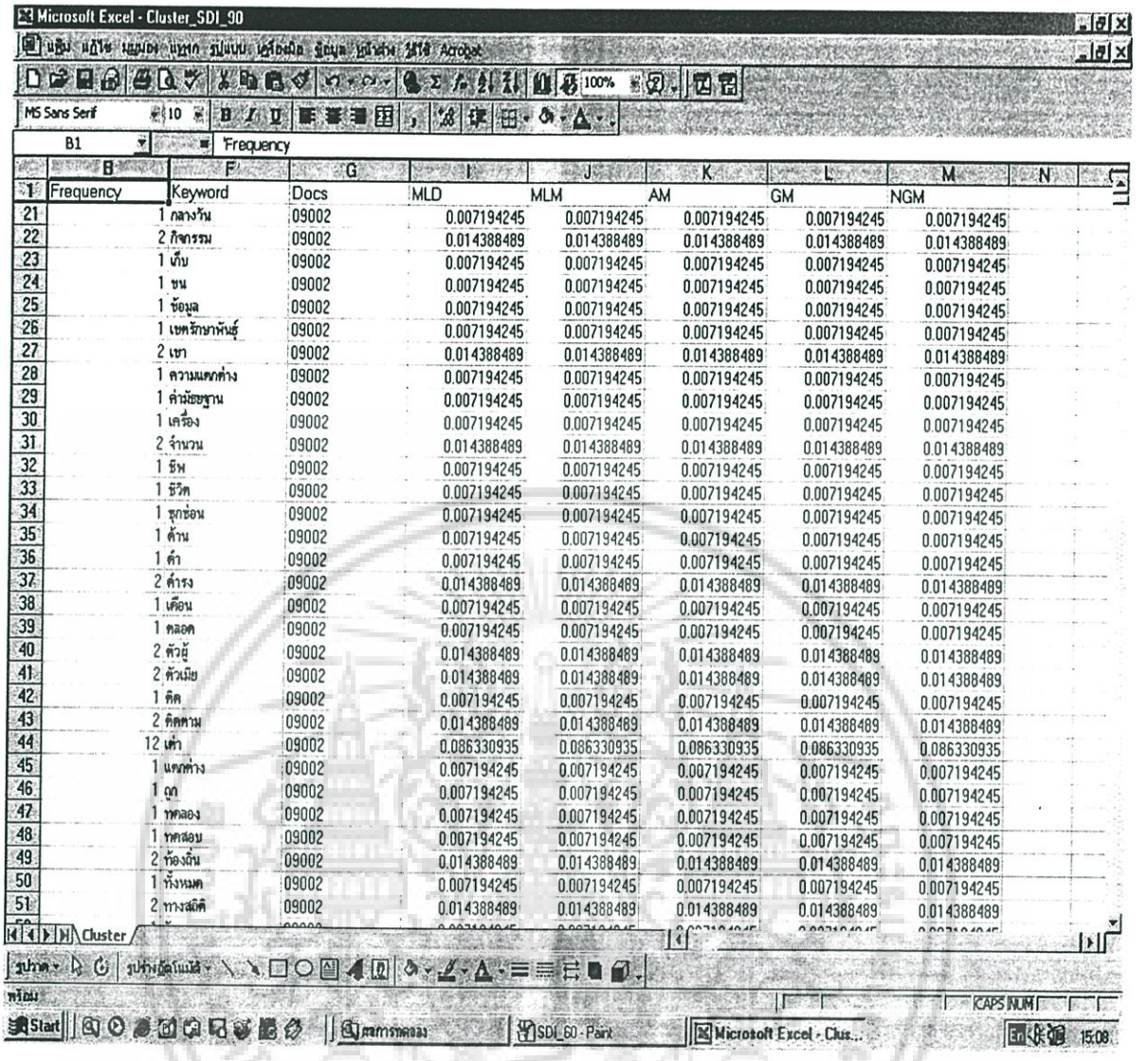
Start | ผลการค้นหา | Microsoft Excel - Clus... | 15:12

รูปที่ ข.6 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำเพียง อย่างเดียว



	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21	1	กลางวัน	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
22	2	กิจกรรม	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
23	1	เก็บ	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
24	1	ขน	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
25	1	ข้อมูล	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
26	1	เขตรักษาพันธุ์	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
27	2	เขา	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
28	1	ความแตกต่าง	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
29	1	ค่ามัธยฐาน	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
30	1	เครื่อง	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
31	2	จำนวน	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
32	1	ชีพ	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
33	1	ชีวิต	09002	0.007194245	0.019607843	0.013401044	0.011877021	0.007992923	
34	1	ซุกซ่อน	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
35	1	ด้าน	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
36	1	ดำ	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
37	2	คำรง	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
38	1	เดือน	09002	0.007194245	0.013071895	0.01013307	0.009697547	0.006526194	
39	1	ตลอด	09002	0.007194245	0.006535948	0.006865096	0.006857201	0.004614716	
40	2	ตัวผู้	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
41	2	ตัวเมีย	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
42	1	ติด	09002	0.007194245	0.006535948	0.006865096	0.006857201	0.004614716	
43	2	ติดตาม	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
44	12	เต่า	09002	0.086330935	0.039215686	0.062773311	0.058185281	0.039157165	
45	1	แตกต่าง	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
46	1	ถูก	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
47	1	ทดลอง	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
48	1	ทดสอบ	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
49	2	ท้องถิ่น	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	
50	1	ทั้งหมด	09002	0.007194245	0.003267974	0.005231109	0.004848773	0.003263097	
51	2	ทางสถิติ	09002	0.014388489	0.006535948	0.010462218	0.009697547	0.006526194	

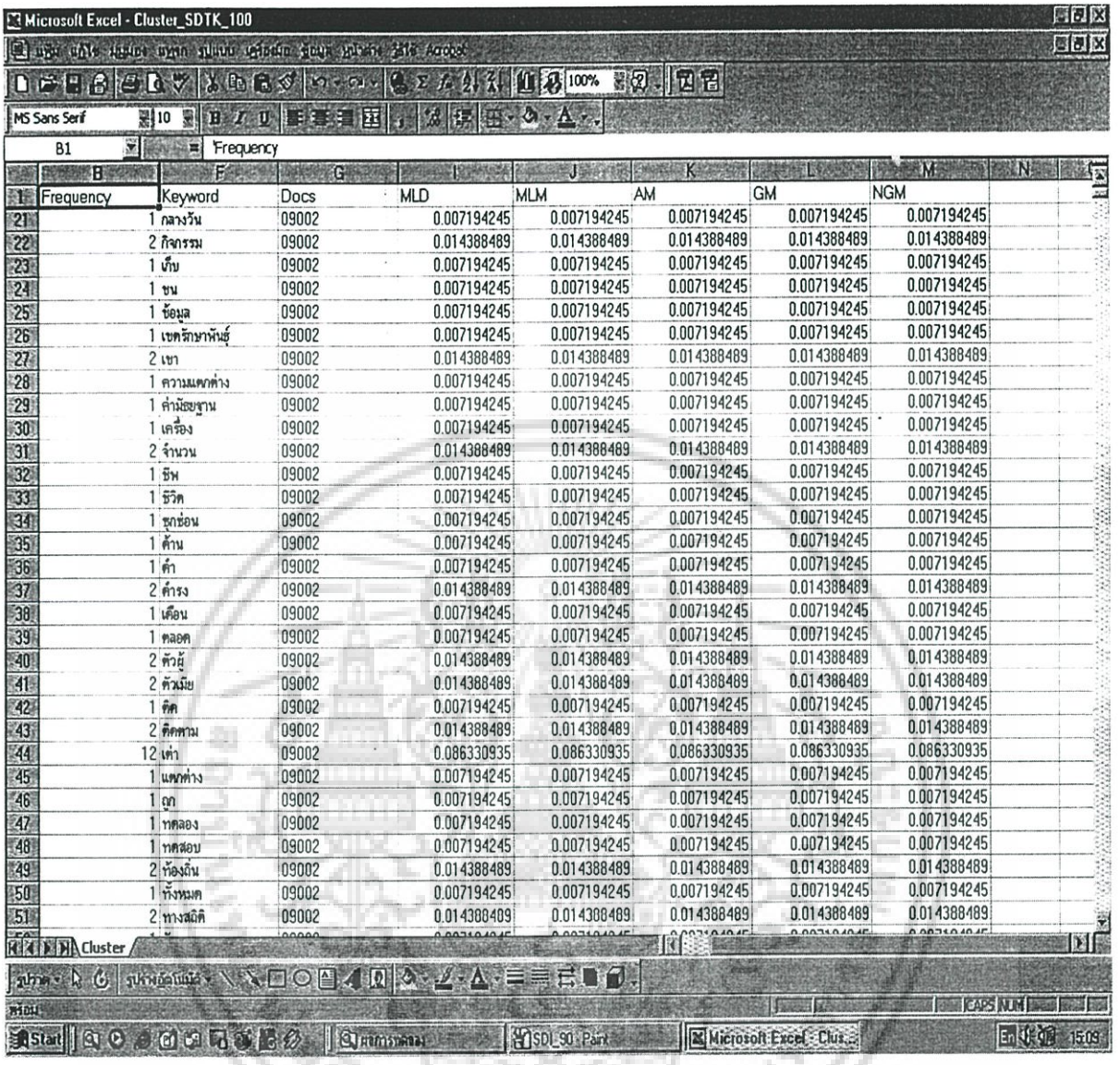
รูปที่ ข.7 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำสำคัญร่วมกับการหาค่าระยะห่าง ของคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 60



Microsoft Excel - Cluster_SDI_90

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21		1 กลางวัน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
22		2 กิจกรรม	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
23		1 เก็บ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
24		1 ขน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
25		1 ข้อมูล	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
26		1 เขตรักษาพันธุ์	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
27		2 เขา	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
28		1 ความแตกต่าง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
29		1 ค่านัยยะสำคัญ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
30		1 เครื่อง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
31		2 จำนวน	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
32		1 ชีพ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
33		1 ชีวิต	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
34		1 ซุกซ่อน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
35		1 ด้าน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
36		1 ดำ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
37		2 ดำรง	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
38		1 เดือน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
39		1 ตลอด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
40		2 ตัวผู้	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
41		2 ตัวเมีย	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
42		1 ติด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
43		2 ติดตาม	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
44		12 เต่า	09002	0.086330935	0.086330935	0.086330935	0.086330935	0.086330935	
45		1 แตกต่าง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
46		1 ถูก	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
47		1 ทดลอง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
48		1 ทดสอบ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
49		2 ท้องถิ่น	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
50		1 ทั้งหมด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
51		2 ทางสถิติ	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	

รูปที่ ข.8 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำสำคัญร่วมกับการหาค่าระยะห่าง ของคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 90



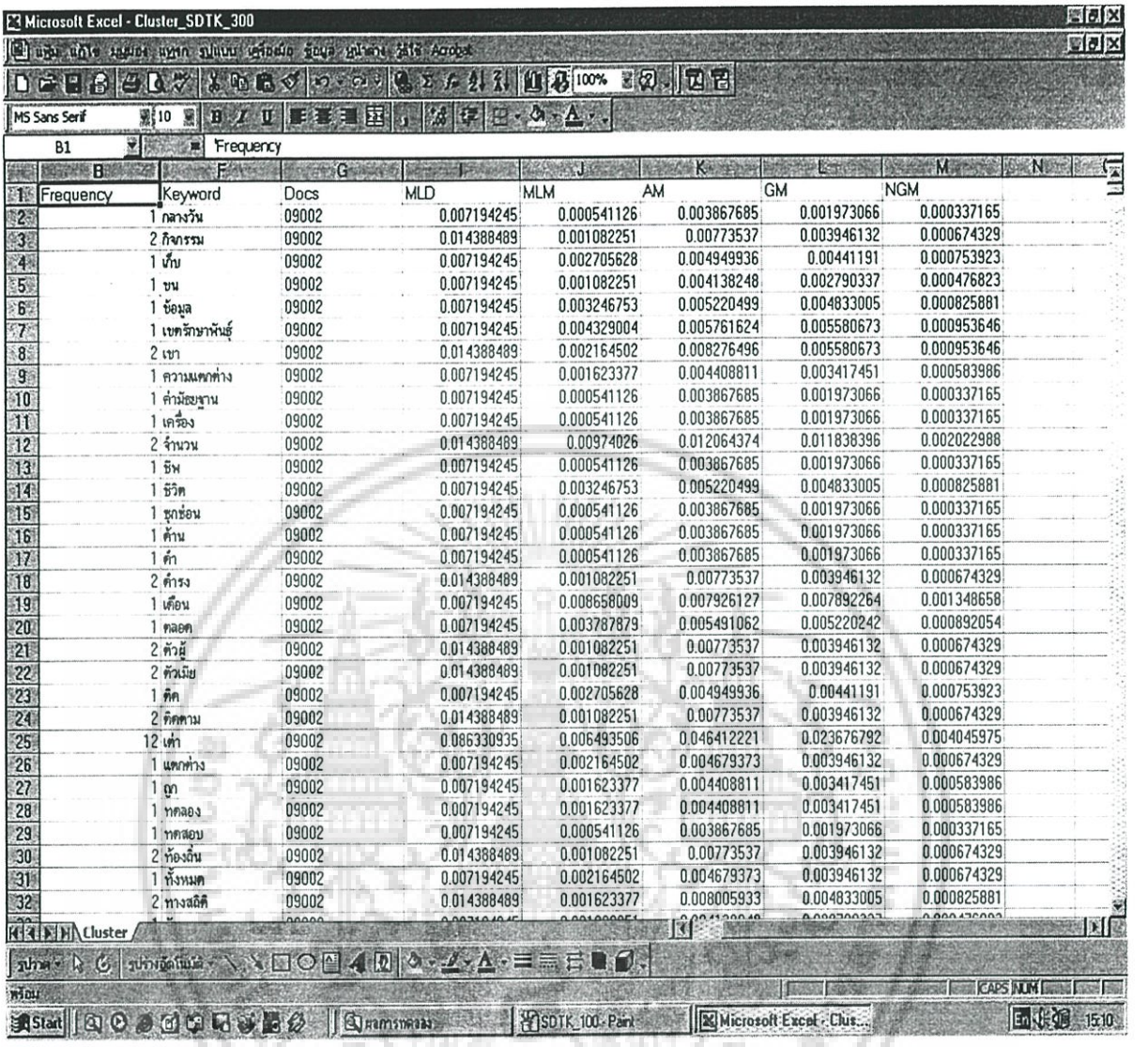
Microsoft Excel - Cluster_SDTK_100

B1 = 'Frequency

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21	1	กลางวัน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
22	2	กิจกรรม	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
23	1	เก็บ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
24	1	ชน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
25	1	ข้อมูล	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
26	1	เขตรักษาพันธุ์	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
27	2	เขา	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
28	1	ความแตกต่าง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
29	1	ค่ามัธยฐาน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
30	1	เครื่อง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
31	2	จำนวน	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
32	1	ชีพ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
33	1	ชีวิต	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
34	1	ซุกซ่อน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
35	1	ด้าน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
36	1	ดำ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
37	2	ดำรง	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
38	1	เดือน	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
39	1	ตลอด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
40	2	ตัวผู้	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
41	2	ตัวเมีย	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
42	1	ติด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
43	2	ติดตาม	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
44	12	เต่า	09002	0.086330935	0.086330935	0.086330935	0.086330935	0.086330935	
45	1	แตกต่าง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
46	1	ถูก	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
47	1	ทดลอง	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
48	1	ทดสอบ	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
49	2	ท้องถิ่น	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	
50	1	ทั้งหมด	09002	0.007194245	0.007194245	0.007194245	0.007194245	0.007194245	
51	2	ทางสถิติ	09002	0.014388489	0.014388489	0.014388489	0.014388489	0.014388489	

Cluster

รูปที่ ข.9 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำสำคัญร่วมกับการหาค่าระยะห่าง ของคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 100



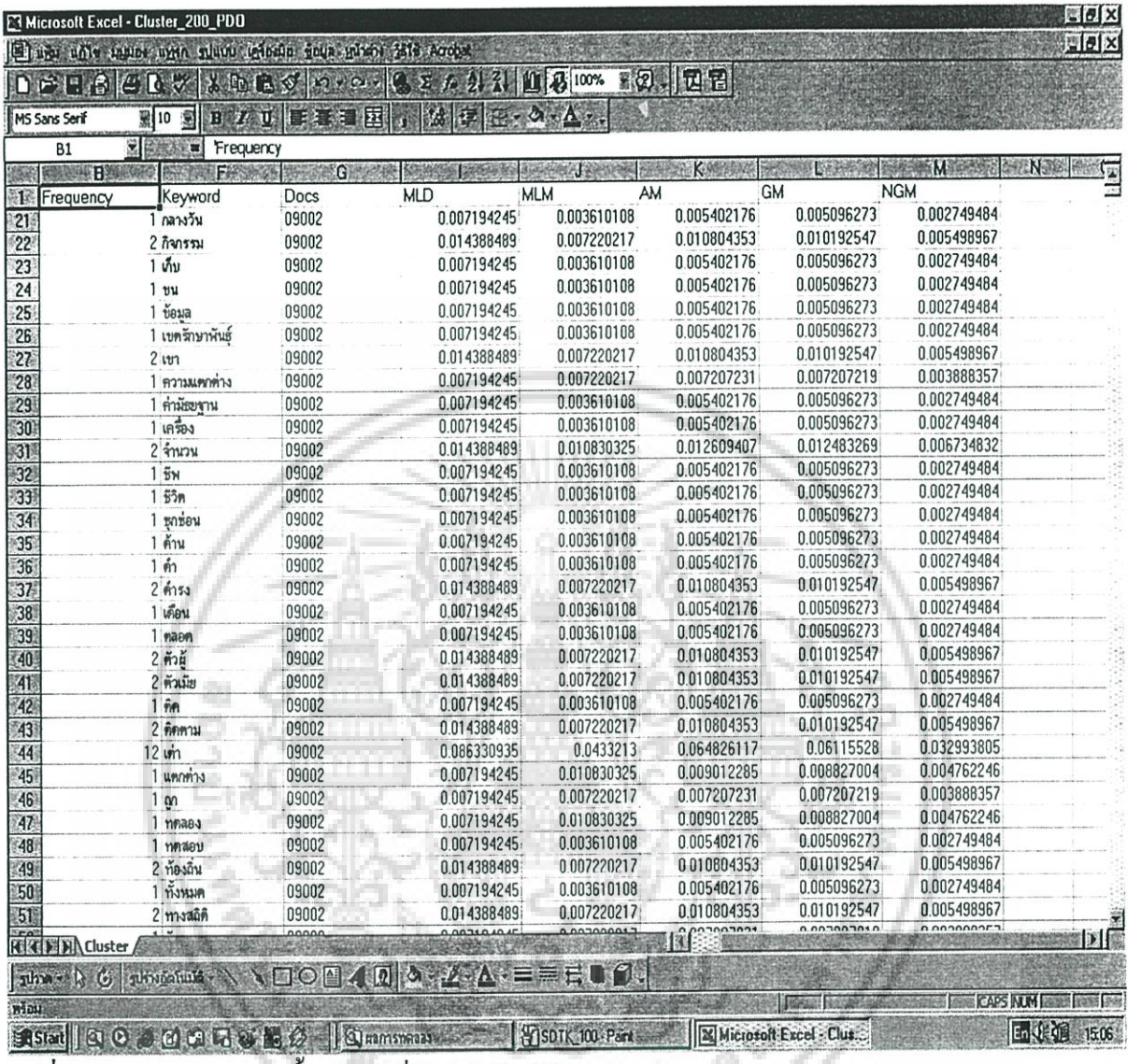
Microsoft Excel - Cluster_SDTK_300

B1: Frequency

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
2	1	กลางวัน	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
3	2	กิจกรรม	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
4	1	เก็บ	09002	0.007194245	0.002705628	0.004949936	0.00441191	0.000753923	
5	1	ขน	09002	0.007194245	0.001082251	0.004138248	0.002790337	0.000476823	
6	1	ข้อมูล	09002	0.007194245	0.003246753	0.005220499	0.004833005	0.000825881	
7	1	เขตรักษาพันธุ์	09002	0.007194245	0.004329004	0.005761624	0.005580673	0.000953646	
8	2	เขา	09002	0.014388489	0.002164502	0.008276496	0.005580673	0.000953646	
9	1	ความแตกต่าง	09002	0.007194245	0.001623377	0.004408811	0.003417451	0.000583986	
10	1	ค่ามัธยฐาน	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
11	1	เครื่อง	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
12	2	จำนวน	09002	0.014388489	0.00974026	0.012064374	0.011838396	0.002022988	
13	1	ชีพ	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
14	1	ชีวิต	09002	0.007194245	0.003246753	0.005220499	0.004833005	0.000825881	
15	1	ซุกซ่อน	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
16	1	ด้าน	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
17	1	ดำ	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
18	2	ดำรง	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
19	1	เดือน	09002	0.007194245	0.008658009	0.007926127	0.007892264	0.001348658	
20	1	ตลอด	09002	0.007194245	0.003787879	0.005491062	0.005220242	0.000892054	
21	2	ตัวผู้	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
22	2	ตัวเมีย	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
23	1	ติด	09002	0.007194245	0.002705628	0.004949936	0.00441191	0.000753923	
24	2	ติดตาม	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
25	12	เต่า	09002	0.086330935	0.006493506	0.046412221	0.023676792	0.004045975	
26	1	แตกต่าง	09002	0.007194245	0.002164502	0.004679373	0.003946132	0.000674329	
27	1	ถูก	09002	0.007194245	0.001623377	0.004408811	0.003417451	0.000583986	
28	1	ทดลอง	09002	0.007194245	0.001623377	0.004408811	0.003417451	0.000583986	
29	1	ทดสอบ	09002	0.007194245	0.000541126	0.003867685	0.001973066	0.000337165	
30	2	ท้องถิ่น	09002	0.014388489	0.001082251	0.00773537	0.003946132	0.000674329	
31	1	ทั้งหมด	09002	0.007194245	0.002164502	0.004679373	0.003946132	0.000674329	
32	2	ทางสถิติ	09002	0.014388489	0.001623377	0.008005933	0.004833005	0.000825881	

Cluster

รูปที่ ข.10 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำสำคัญร่วมกับการหาค่าระยะห่าง ของคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 300



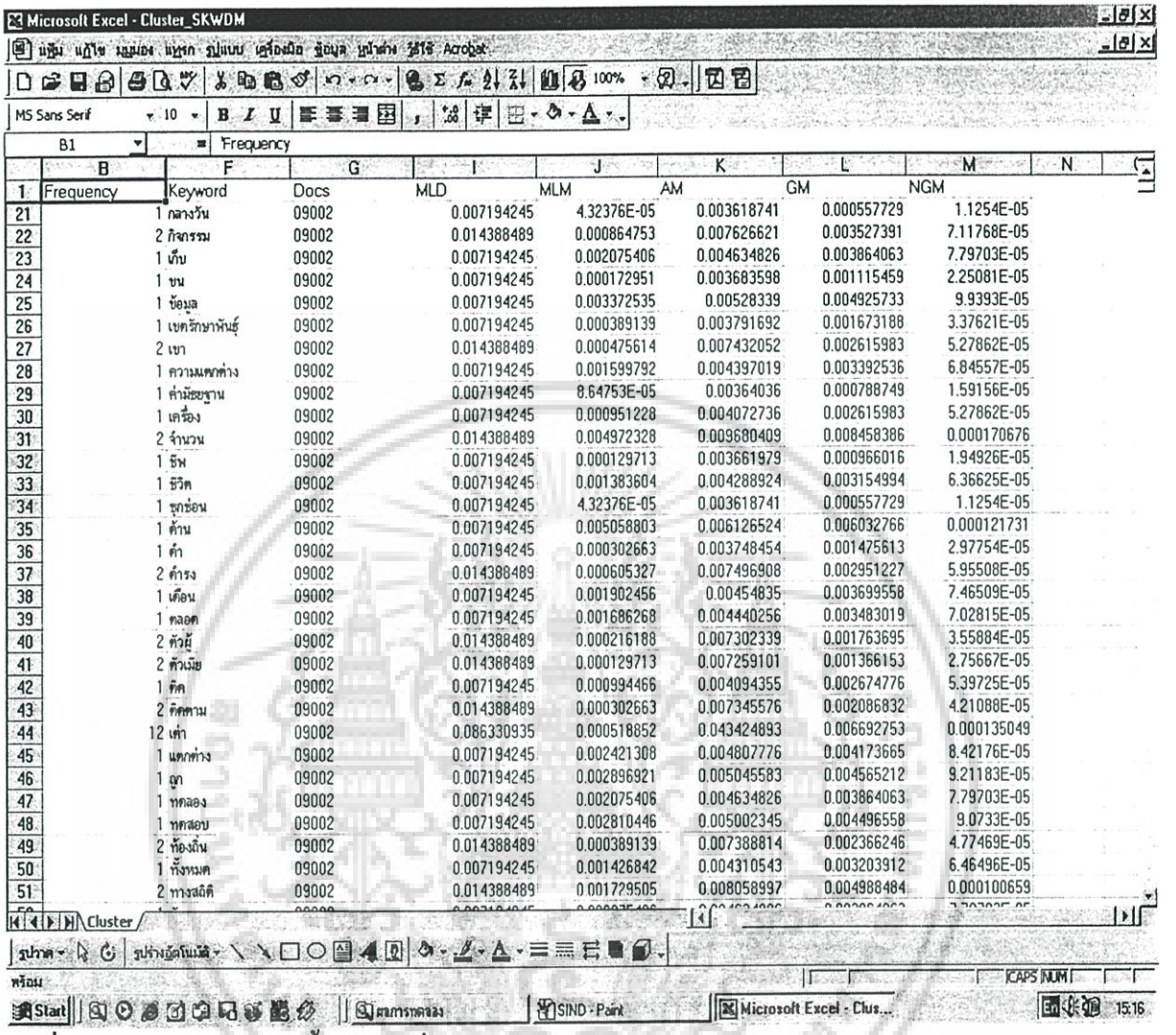
Microsoft Excel - Cluster_200_PDO

B1 Frequency

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21	1	กลางวัน	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
22	2	กิจกรรม	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
23	1	เก็บ	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
24	1	ขน	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
25	1	ข้อมูล	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
26	1	เขตรักษาพันธุ์	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
27	2	เขา	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
28	1	ความแตกต่าง	09002	0.007194245	0.007220217	0.007207231	0.007207219	0.003888357	
29	1	ค่ามัธยฐาน	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
30	1	เครื่อง	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
31	2	จำนวน	09002	0.014388489	0.010830325	0.012609407	0.012483269	0.006734832	
32	1	ชีพ	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
33	1	ชีวิต	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
34	1	ชุกชุม	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
35	1	ด้าน	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
36	1	ดำ	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
37	2	ดำรง	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
38	1	เดือน	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
39	1	ตลอด	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
40	2	ตัวผู้	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
41	2	ตัวเมีย	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
42	1	ติด	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
43	2	ติดตาม	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
44	12	เต่า	09002	0.086330935	0.0433213	0.064826117	0.06115528	0.032993805	
45	1	แตกต่าง	09002	0.007194245	0.010830325	0.009012285	0.008827004	0.004762246	
46	1	ถูก	09002	0.007194245	0.007220217	0.007207231	0.007207219	0.003888357	
47	1	ทดลอง	09002	0.007194245	0.010830325	0.009012285	0.008827004	0.004762246	
48	1	ทดสอบ	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
49	2	ท้องถิ่น	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	
50	1	ทั้งหมด	09002	0.007194245	0.003610108	0.005402176	0.005096273	0.002749484	
51	2	ทางสถิติ	09002	0.014388489	0.007220217	0.010804353	0.010192547	0.005498967	

Cluster

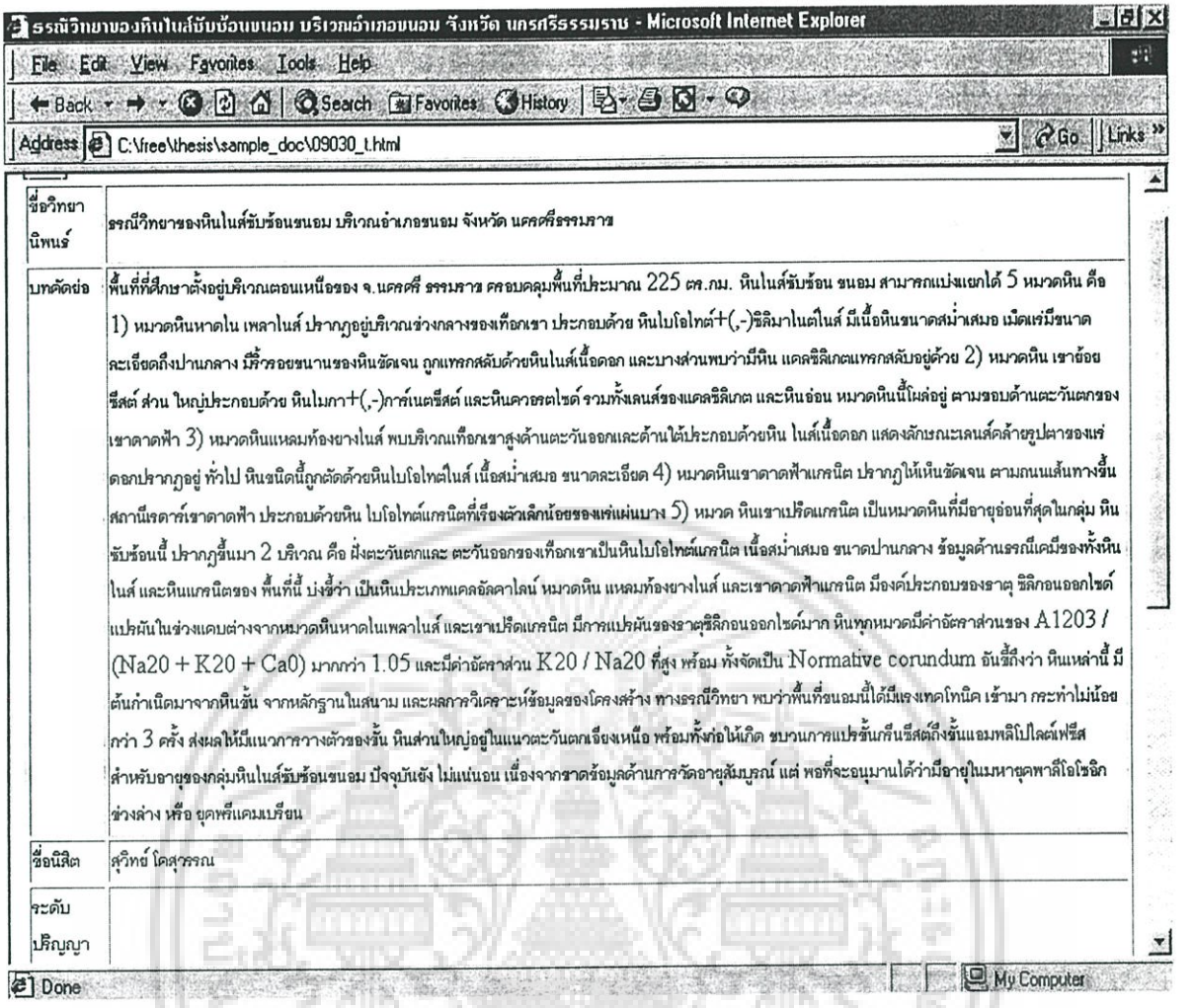
รูปที่ ข.11 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธี PDO



Microsoft Excel - Cluster_SKWDM

	B	F	G	I	J	K	L	M	N
1	Frequency	Keyword	Docs	MLD	MLM	AM	GM	NGM	
21	1	กลางวัน	09002	0.007194245	4.32376E-05	0.003618741	0.000557729	1.1254E-05	
22	2	กิจกรรม	09002	0.014388489	0.000864753	0.007626621	0.003527391	7.11768E-05	
23	1	เก็บ	09002	0.007194245	0.002075406	0.004634826	0.003864063	7.79703E-05	
24	1	ขน	09002	0.007194245	0.000172951	0.003683598	0.001115459	2.25081E-05	
25	1	ข้อมูล	09002	0.007194245	0.003372535	0.00528339	0.004925733	9.9393E-05	
26	1	เขตรักษาพันธุ์	09002	0.007194245	0.000389139	0.003791692	0.001673188	3.37621E-05	
27	2	เขา	09002	0.014388489	0.000475614	0.007432052	0.002615983	5.27862E-05	
28	1	ความแตกต่าง	09002	0.007194245	0.001599792	0.004397019	0.003392536	6.84557E-05	
29	1	ค่ามัธยฐาน	09002	0.007194245	8.64753E-05	0.00364036	0.000788749	1.59156E-05	
30	1	เครื่อง	09002	0.007194245	0.000951228	0.004072736	0.002615983	5.27862E-05	
31	2	จำนวน	09002	0.014388489	0.004972328	0.009680409	0.008458386	0.000170676	
32	1	ชัพ	09002	0.007194245	0.000129713	0.003661979	0.000966016	1.94926E-05	
33	1	ชีวิต	09002	0.007194245	0.001383604	0.004288924	0.003154994	6.36625E-05	
34	1	ซุกซ่อน	09002	0.007194245	4.32376E-05	0.003618741	0.000557729	1.1254E-05	
35	1	ด้าน	09002	0.007194245	0.005058803	0.006126524	0.006032766	0.000121731	
36	1	ดำ	09002	0.007194245	0.000302663	0.003748454	0.001475613	2.97754E-05	
37	2	ดำรง	09002	0.014388489	0.000605327	0.007496908	0.002951227	5.95508E-05	
38	1	เดือน	09002	0.007194245	0.001902456	0.00454835	0.003699558	7.46509E-05	
39	1	ทลอด	09002	0.007194245	0.001686268	0.004440256	0.003483019	7.02815E-05	
40	2	ตัวผู้	09002	0.014388489	0.000216188	0.007302339	0.001763695	3.55884E-05	
41	2	ตัวเมีย	09002	0.014388489	0.000129713	0.007259101	0.001366153	2.75667E-05	
42	1	ติด	09002	0.007194245	0.000994466	0.004094355	0.002674776	5.39725E-05	
43	2	ติดตาม	09002	0.014388489	0.000302663	0.007345576	0.002086832	4.21088E-05	
44	12	เต่า	09002	0.086330935	0.000518852	0.043424893	0.006692753	0.000135049	
45	1	แตกต่าง	09002	0.007194245	0.002421308	0.004807776	0.004173665	8.42176E-05	
46	1	ถูก	09002	0.007194245	0.002896921	0.005045583	0.004565212	9.21183E-05	
47	1	ทดลอง	09002	0.007194245	0.002075406	0.004634826	0.003864063	7.79703E-05	
48	1	ทดสอบ	09002	0.007194245	0.002810446	0.005002345	0.004496558	9.0733E-05	
49	2	ท้องถิ่น	09002	0.014388489	0.000389139	0.007388814	0.002366246	4.77469E-05	
50	1	ทั้งหมด	09002	0.007194245	0.001426842	0.004310543	0.003203912	6.46496E-05	
51	2	ทางสถิติ	09002	0.014388489	0.001729505	0.008058997	0.004988484	0.000100659	

รูปที่ ข.12 ผลการคำนวณน้ำหนักคำที่จัดกลุ่มด้วยวิธี DF



ธรณีวิทยาของหินไนส์ชับซ้อนขนอม บริเวณอำเภอขนอม จังหวัด นครศรีธรรมราช - Microsoft Internet Explorer

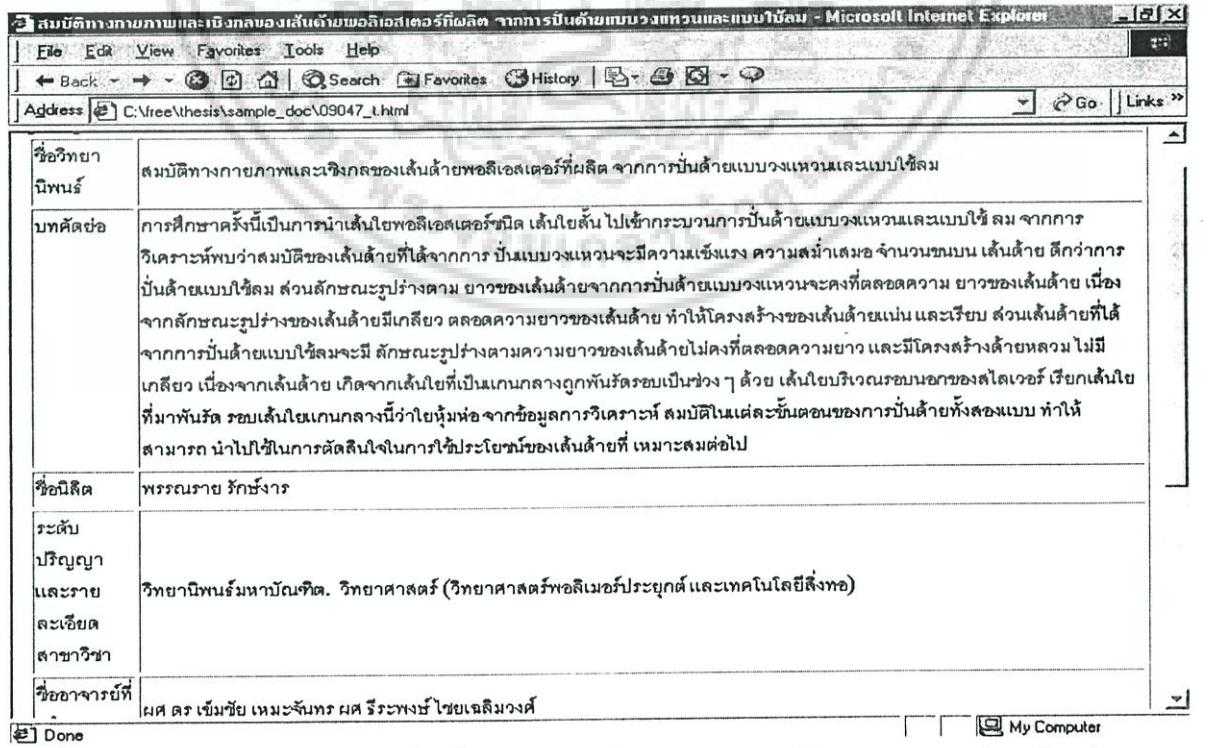
File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09030_t.html

ชื่อวิทยานิพนธ์	ธรณีวิทยาของหินไนส์ชับซ้อนขนอม บริเวณอำเภอขนอม จังหวัด นครศรีธรรมราช
บทคัดย่อ	พื้นที่ที่ศึกษาตั้งอยู่บริเวณตอนเหนือของ จ.นครศรี ธรรมราช ครอบคลุมพื้นที่ประมาณ 225 ตร.กม. หินไนส์ชับซ้อน ขนอม สามารถแบ่งแยกได้ 5 หมวดหิน คือ 1) หมวดหินหาดไน เพลาไนส์ ปรากฏอยู่บริเวณช่วงกลางของเทือกเขา ประกอบด้วย หินไบโอไทต์+(,-)ซิลิมาไนต์ไนส์ มีเนื้อหินขนาดสม่ำเสมอ เม็ดแร่มีขนาดละเอียดถึงปานกลาง มีริ้วรอยขนานของหินชัดเจน ถูกแทรกสลับด้วยหินไนส์เนื้อตอก และบางส่วนพบว่ามีหิน แคลซิลิเกตแทรกสลับอยู่ด้วย 2) หมวดหิน เขาย้อย ซีสต์ ส่วน ใหญ่ประกอบด้วย หินไมกา+(,-)การ์เนตซีสต์ และหินควอรตไซต์ รวมทั้งเลนส์ของแคลซิลิเกต และหินอ่อน หมวดหินนี้โผล่อยู่ ตามขอบด้านตะวันตกของเขาดาดฟ้า 3) หมวดหินแหลมท้องยางไนส์ พบบริเวณเทือกเขาสูงด้านตะวันออกและด้านใต้ประกอบด้วยหิน ไนส์เนื้อตอก แสดงลักษณะเลนส์คล้ายรูปตาของแร่ตอกปรากฏอยู่ ทั่วไป หินชนิดนี้ถูกตัดด้วยหินไบโอไทต์ไนส์ เนื้อสม่ำเสมอ ขนาดละเอียด 4) หมวดหินเขาดาดฟ้าแกรนิต ปรากฏให้เห็นชัดเจน ตามถนนเส้นทางขึ้นสถานีเรดาร์เขาดาดฟ้า ประกอบด้วยหิน ไบโอไทต์แกรนิตที่เรืองตัวเล็กน้อยของแร่แผ่นบาง 5) หมวด หินเขาเปรียดแกรนิต เป็นหมวดหินที่มีอายุอ่อนที่สุดในกลุ่ม หินชับซ้อนนี้ ปรากฏขึ้นมา 2 บริเวณ คือ ฝั่งตะวันตกและ ตะวันออกของเทือกเขาเป็นหินไบโอไทต์แกรนิต เนื้อสม่ำเสมอ ขนาดปานกลาง ข้อมูลด้านธรณีเคมีของทั้งหินไนส์ และหินแกรนิตของ พื้นที่นี้ บ่งชี้ว่า เป็นหินประเภทแคลอัลคาไลน์ หมวดหิน แหลมท้องยางไนส์ และเขาดาดฟ้าแกรนิต มีองค์ประกอบของธาตุ ซิลิกอนออกไซด์แปรผันในช่วงแคบต่างจากหมวดหินหาดไนเพลาไนส์ และเขาเปรียดแกรนิต มีการแปรผันของธาตุซิลิกอนออกไซด์มาก หินทุกหมวดมีค่าอัตราส่วนของ A1203 / (Na20 + K20 + Ca0) มากกว่า 1.05 และมีค่าอัตราส่วน K20 / Na20 ที่สูง พร้อม ทั้งจัดเป็น Normative corundum อันชี้ถึงว่า หินเหล่านี้ มีต้นกำเนิดมาจากหินชั้น จากหลักฐานในสนาม และผลการวิเคราะห์ข้อมูลของโครงสร้าง ทางธรณีวิทยา พบว่าพื้นที่ขนอมนี้ได้มีแรงเทคโทนิค เข้ามา กระทำไม่น้อยกว่า 3 ครั้ง ส่งผลให้มีแนวการวางตัวของชั้น หินส่วนใหญ่อยู่ในแนวตะวันตกเฉียงเหนือ พร้อมทั้งก่อให้เกิด ขบวนการแปรชั้นกรีนชีสต์ถึงชั้นแอมพลิโบไลต์เฟรีส สำหรับอายุของกลุ่มหินไนส์ชับซ้อนขนอม ปัจจุบันยัง ไม่แน่นอน เนื่องจากขาดข้อมูลด้านการวัดอายุสัมบูรณ์ แต่ พอที่จะอนุมานได้ว่ามีอายุในมหายุคพาลีโอโซอิกช่วงล่าง หรือ ยุคพรีแคมเบรียน
ชื่อนิสิต	สุวิทย์ โคสุวรรณ
ระดับปริญญา	

Done My Computer



สมบัติทางกายภาพและเชิงกลของเส้นด้ายพอลิเอสเตอร์ที่ผลิต จากการปั่นด้ายแบบวงแหวนและแบบใช้ลม - Microsoft Internet Explorer

File Edit View Favorites Tools Help

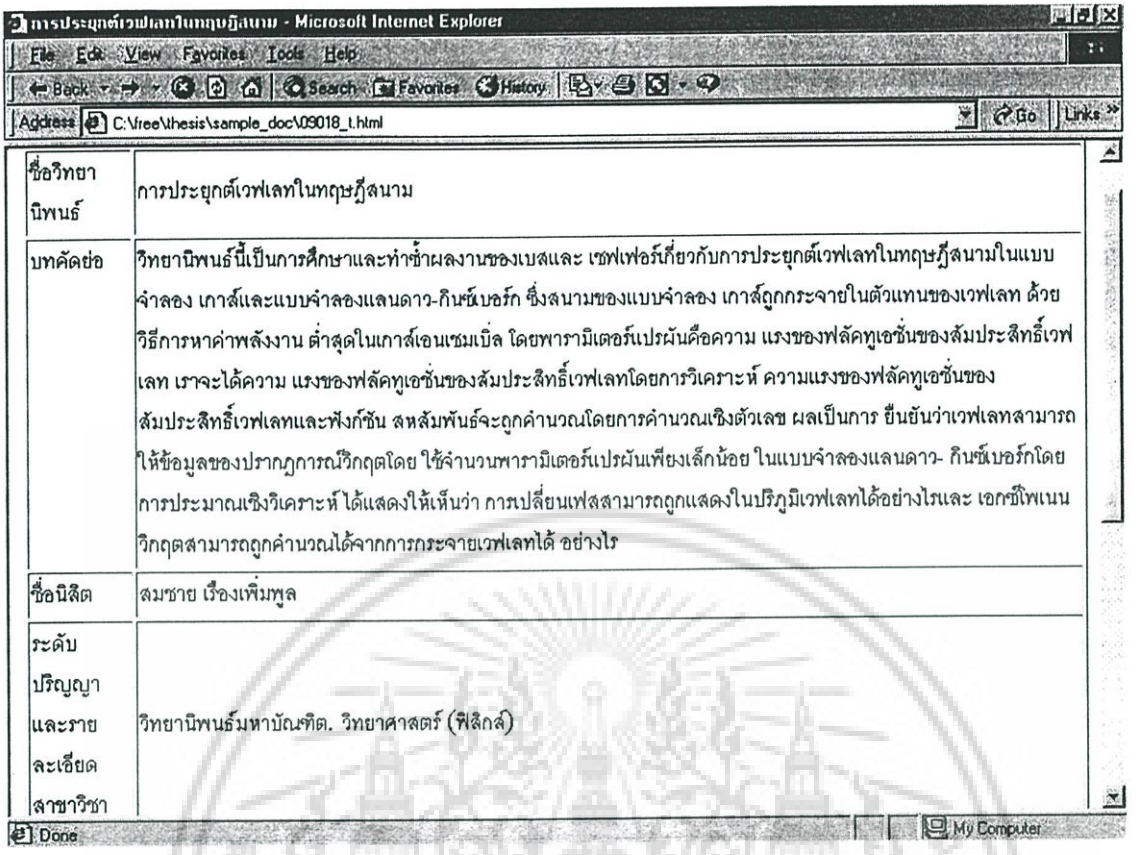
Back Search Favorites History

Address C:\free\thesis\sample_doc\09047_t.html

ชื่อวิทยานิพนธ์	สมบัติทางกายภาพและเชิงกลของเส้นด้ายพอลิเอสเตอร์ที่ผลิต จากการปั่นด้ายแบบวงแหวนและแบบใช้ลม
บทคัดย่อ	การศึกษาครั้งนี้เป็นการนำเส้นใยพอลิเอสเตอร์ชนิด เส้นใยสั้น ไปเข้ากระบวนการปั่นด้ายแบบวงแหวนและแบบใช้ ลม จากการวิเคราะห์พบว่าสมบัติของเส้นด้ายที่ได้จากการ ปั่นแบบวงแหวนจะมีความเข้งแรง ความสม่ำเสมอ จำนวนขนบน เส้นด้าย ดีกว่าการปั่นด้ายแบบใช้ลม ส่วนลักษณะรูปร่างตาม ยาวของเส้นด้ายจากการปั่นด้ายแบบวงแหวนจะคงที่ตลอดความ ยาวของเส้นด้าย เนื่องจากลักษณะรูปร่างของเส้นด้ายมีเกลียว ตลอดความยาวของเส้นด้าย ทำให้โครงสร้างของเส้นด้ายแน่น และเรียบ ส่วนเส้นด้ายที่ได้จากการปั่นด้ายแบบใช้ลมจะมี ลักษณะรูปร่างตามความยาวของเส้นด้ายไม่คงที่ตลอดความยาว และมีโครงสร้างด้ายหลวม ไม่มีเกลียว เนื่องจากเส้นด้าย เกิดจากเส้นใยที่เป็นเกนกลางถูกพันรัดรอบเป็นช่วง ๆ ด้วย เส้นใยบริเวณรอบนอกของสไลเวอร์ เรียกเส้นใยที่มาพันรัด รอบเส้นใยเกนกลางนี้ว่าใยหุ้มห่อ จากข้อมูลการวิเคราะห์ สมบัติในแต่ละขั้นตอนของการปั่นด้ายทั้งสองแบบ ทำให้สามารถ นำไปใช้ในการตัดสินใจในการใช้ประโยชน์ของเส้นด้ายที่ เหมาะสมต่อไป
ชื่อนิสิต	พรรณราย รักษ์งาร
ระดับปริญญาและรายละเอียดสาขาวิชา	วิทยานิพนธ์มหาบัณฑิต. วิทยาศาสตร์ (วิทยาศาสตร์พอลิเมอร์ประยุกต์ และเทคโนโลยีสิ่งทอ)
ชื่ออาจารย์ที่	ผศ ดร เข็มชัย เหมะจันทร ผศ รีระพงษ์ ไชยเฉลิมวงศ์

Done My Computer

รูปที่ ข.13 ตัวอย่างเอกสารในกลุ่มที่ 5ที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำเพียงอย่างเดียว



การประยุกต์เวฟเลทในทฤษฎีสนาม - Microsoft Internet Explorer

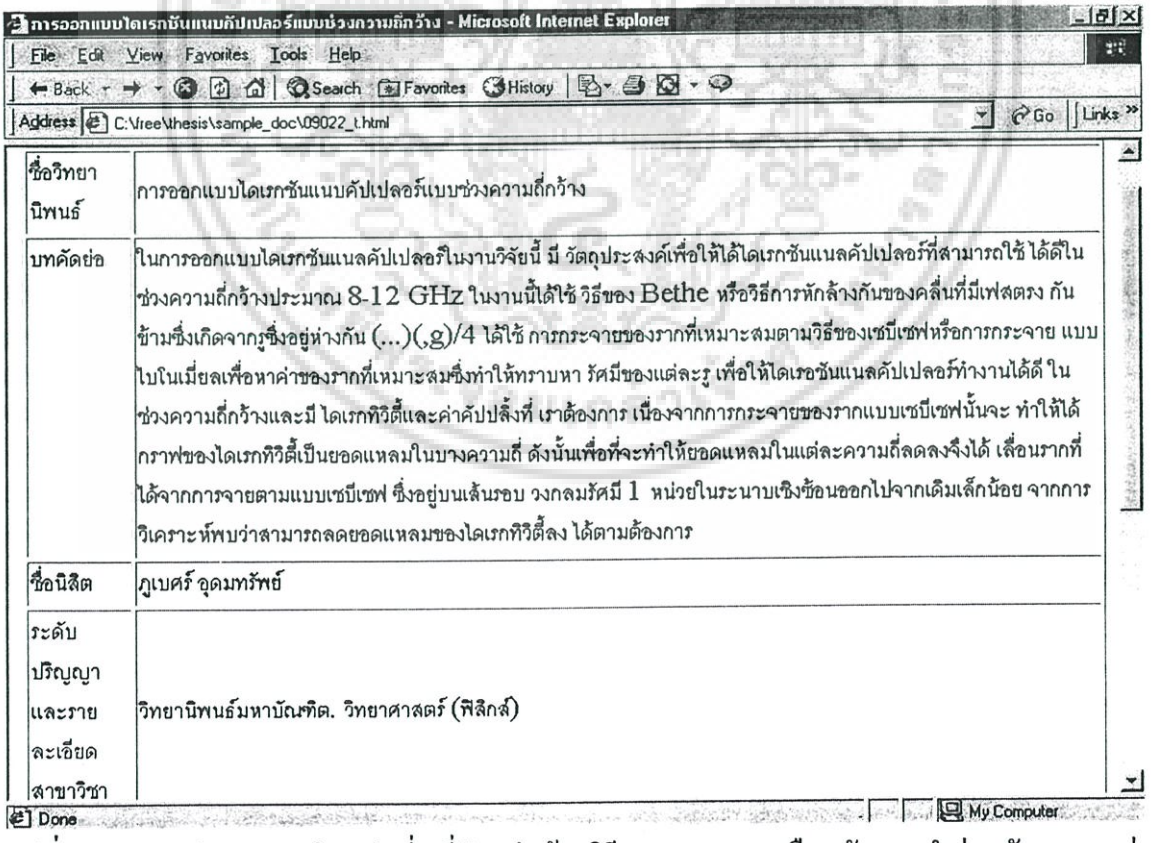
File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09018_t.html Go Links

ชื่อวิทยานิพนธ์	การประยุกต์เวฟเลทในทฤษฎีสนาม
บทคัดย่อ	วิทยานิพนธ์นี้เป็นการศึกษาและทำซ้ำผลงานของเบสและ เซฟเฟอร์เกี่ยวกับการประยุกต์เวฟเลทในทฤษฎีสนามในแบบจำลอง เกาส์และแบบจำลองแลนดาว-กินซ์เบอร์ก ซึ่งสนามของแบบจำลอง เกาส์ถูกกระจายในตัวแทนของเวฟเลท ด้วยวิธีการหาค่าพลังงาน ต่ำสุดในเกาส์เอนเซมเบิล โดยพารามิเตอร์แปรผันคือความ แรงของฟลัคทูเอชั่นของสัมประสิทธิ์เวฟเลท เราจะได้ความ แรงของฟลัคทูเอชั่นของสัมประสิทธิ์เวฟเลทโดยการวิเคราะห์ ความแรงของฟลัคทูเอชั่นของสัมประสิทธิ์เวฟเลทและฟังก์ชัน สหสัมพันธ์จะถูกคำนวณโดยการคำนวณเชิงตัวเลข ผลเป็นการ ยืนยันว่าเวฟเลทสามารถให้ข้อมูลของปรากฏการณ์วิกฤตโดย ใช้จำนวนพารามิเตอร์แปรผันเพียงเล็กน้อย ในแบบจำลองแลนดาว- กินซ์เบอร์กโดยการประมาณเชิงวิเคราะห์ ได้แสดงให้เห็นว่า การเปลี่ยนเฟสสามารถถูกแสดงในปริภูมิเวฟเลทได้อย่างไรและ เอกซ์โพเนนวิกฤตสามารถถูกคำนวณได้จากการกระจายเวฟเลทได้ อย่างไร
ชื่อนิสิต	สมชาย เรืองเพิ่มพูล
ระดับปริญญาและรายละเอียดสาขาวิชา	วิทยานิพนธ์มหาบัณฑิต. วิทยาศาสตร์ (ฟิสิกส์)

Done My Computer



การออกแบบไดเรกชันแนลคัปเปลอร์แบบช่วงความถี่กว้าง - Microsoft Internet Explorer

File Edit View Favorites Tools Help

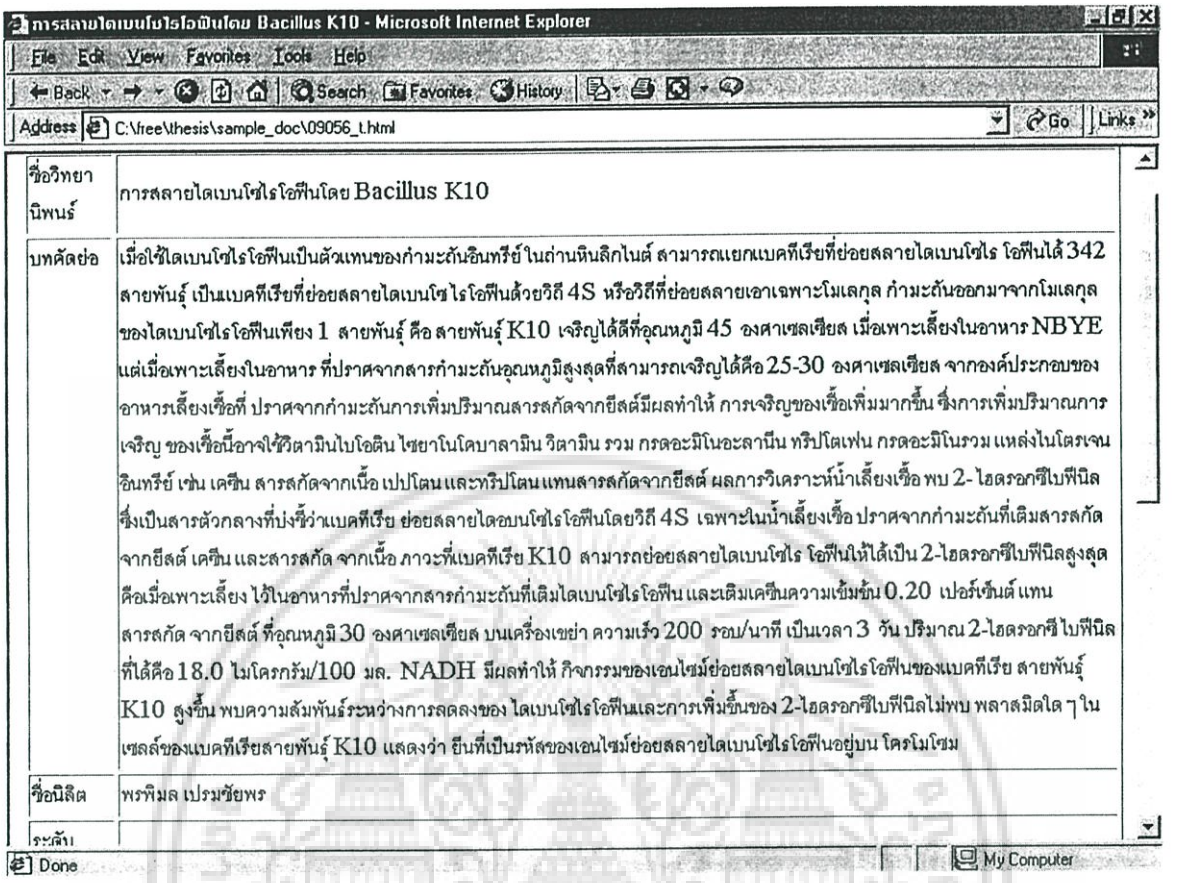
Back Search Favorites History

Address C:\free\thesis\sample_doc\09022_t.html Go Links

ชื่อวิทยานิพนธ์	การออกแบบไดเรกชันแนบคัปเปลอร์แบบช่วงความถี่กว้าง
บทคัดย่อ	ในการออกแบบไดเรกชันแนลคัปเปลอร์ในงานวิจัยนี้ มี วัตถุประสงค์เพื่อให้ได้ไดเรกชันแนลคัปเปลอร์ที่สามารถใช้ ได้ดีในช่วงความถี่กว้างประมาณ 8-12 GHz ในงานนี้ได้ใช้ วิธีของ Bethe หรือวิธีการหักล้างกันของคลื่นที่มีเฟสตรง กันข้ามซึ่งเกิดจากรูซึ่งอยู่ห่างกัน (...)(,g)/4 ได้ใช้ การกระจายของรากที่เหมาะสมตามวิธีของเชบีเชฟหรือการกระจาย แบบไบโนเมี่ยลเพื่อหาค่าของรากที่เหมาะสมซึ่งทำให้ทราบหา รัศมีของแต่ละรู เพื่อให้ไดเรชันแนลคัปเปลอร์ทำงานได้ดี ในช่วงความถี่กว้างและมี ไดเรกทิวิตี้และค่าคัปปลิ้งที่ เราต้องการ เนื่องจากการกระจายของรากแบบเชบีเชฟนั้นจะ ทำให้ได้กราฟของไดเรกทิวิตี้เป็นยอดแหลมในบางความถี่ ดังนั้นเพื่อที่จะทำให้ยอดแหลมในแต่ละความถี่ลดลงจึงได้ เลื่อนรากที่ได้จากการจายตามแบบเชบีเชฟ ซึ่งอยู่บนเส้นรอบ วงกลมรัศมี 1 หน่วยในระนาบเชิงซ้อนออกไปจากเดิมเล็กน้อย จากการวิเคราะห์พบว่าสามารถลดยอดแหลมของไดเรกทิวิตี้ลง ได้ตามต้องการ
ชื่อนิสิต	ภูเบศร์ อุดมทรัพย์
ระดับปริญญาและรายละเอียดสาขาวิชา	วิทยานิพนธ์มหาบัณฑิต. วิทยาศาสตร์ (ฟิสิกส์)

Done My Computer

รูปที่ ข.14 ตัวอย่างเอกสารในกลุ่มที่ 5ที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำร่วมกับ การหาค่าระยะห่าง ของกลุ่มคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 60 เฉพาะคำ intersec



การสลายไดเบนโซไรโอฟินโดย Bacillus K10 - Microsoft Internet Explorer

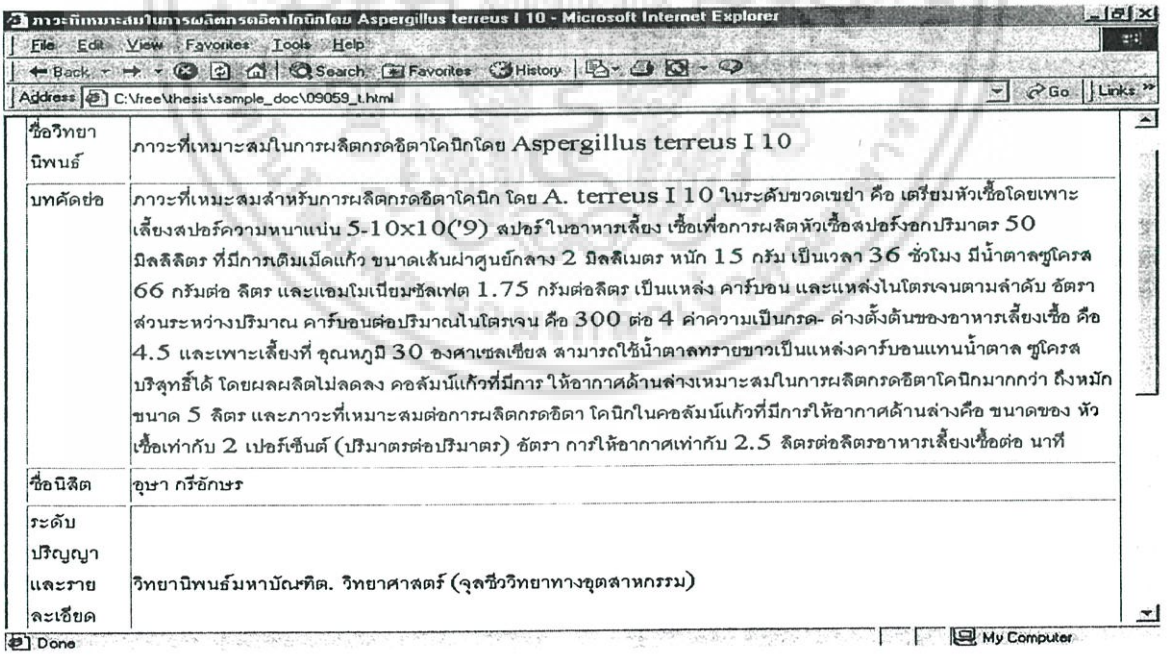
File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09056_t.html Go Links

ชื่อวิทยานิพนธ์	การสลายไดเบนโซไรโอฟินโดย Bacillus K10
บทคัดย่อ	เมื่อใช้ไดเบนโซไรโอฟินเป็นตัวแทนของกำมะถันอินทรีย์ ในถ่านหินลิกไนต์ สามารถแยกแบคทีเรียที่ย่อยสลายไดเบนโซไร โอฟินได้ 342 สายพันธุ์ เป็นแบคทีเรียที่ย่อยสลายไดเบนโซ ไรโอฟินด้วยวิถี 4S หรือวิถีที่ย่อยสลายเอาเฉพาะโมเลกุล กำมะถันออกมาจากโมเลกุล ของไดเบนโซไรโอฟินเพียง 1 สายพันธุ์ คือ สายพันธุ์ K10 เจริญได้ดีที่อุณหภูมิ 45 องศาเซลเซียส เมื่อเพาะเลี้ยงในอาหาร NBYE แต่เมื่อเพาะเลี้ยงในอาหาร ที่ปราศจากสารกำมะถันอุณหภูมิสูงสุดที่สามารถเจริญได้คือ 25-30 องศาเซลเซียส จากองค์ประกอบของ อาหารเลี้ยงเชื้อที่ ปราศจากกำมะถันการเพิ่มปริมาณสารสกัดจากยีสต์มีผลทำให้ การเจริญของเชื้อเพิ่มมากขึ้น ซึ่งการเพิ่มปริมาณการ เจริญ ของเชื้อนี้อาจใช้วิตามินไบโอติน ไซยาโนโคบาลามิน วิตามิน รวม กรดอะมิโนอะลานีน ทริปโตเฟน กรดอะมิโนรวม แหล่งไนโตรเจน อินทรีย์ เช่น เคซีน สารสกัดจากเนื้อ เปปโตน และทริปโตน แทนสารสกัดจากยีสต์ ผลการวิเคราะห์น้ำเลี้ยงเชื้อ พบ 2-ไฮดรอกซีไบฟีนิล ซึ่งเป็นสารตัวกลางที่บ่งชี้ว่าแบคทีเรีย ย่อยสลายไดอบนโซไรโอฟินโดยวิถี 4S เฉพาะในน้ำเลี้ยงเชื้อ ปราศจากกำมะถันที่เติมสารสกัด จากยีสต์ เคซีน และสารสกัด จากเนื้อ ภาวะที่แบคทีเรีย K10 สามารถย่อยสลายไดเบนโซไร โอฟินให้ได้เป็น 2-ไฮดรอกซีไบฟีนิลสูงสุด คือเมื่อเพาะเลี้ยง ไว้ในอาหารที่ปราศจากสารกำมะถันที่เติมไดเบนโซไรโอฟิน และเติมเคซีนความเข้มข้น 0.20 เปอร์เซ็นต์ แทน สารสกัด จากยีสต์ ที่อุณหภูมิ 30 องศาเซลเซียส บนเครื่องเขย่า ความเร็ว 200 รอบ/นาที เป็นเวลา 3 วัน ปริมาณ 2-ไฮดรอกซี ไบฟีนิล ที่ได้คือ 18.0 ไมโครกรัม/100 มล. NADH มีผลทำให้ กิจกรรมของเอนไซม์ย่อยสลายไดเบนโซไรโอฟินของแบคทีเรีย สายพันธุ์ K10 สูงขึ้น พบความสัมพันธ์ระหว่างการลดลงของ ไดเบนโซไรโอฟินและการเพิ่มขึ้นของ 2-ไฮดรอกซีไบฟีนิลไม่พบ พลาสมิดใด ๆ ใน เซลล์ของแบคทีเรียสายพันธุ์ K10 แสดงว่า ยีนที่เป็นรหัสของเอนไซม์ย่อยสลายไดเบนโซไรโอฟินอยู่บน โครโมโซม
ชื่อนิสิต	พรพิมล เปรมชัยพร
ระดับ	

Done My Computer



ภาวะที่เหมาะสมในการผลิตกรดอิตาโคนิกโดย Aspergillus terreus I 10 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

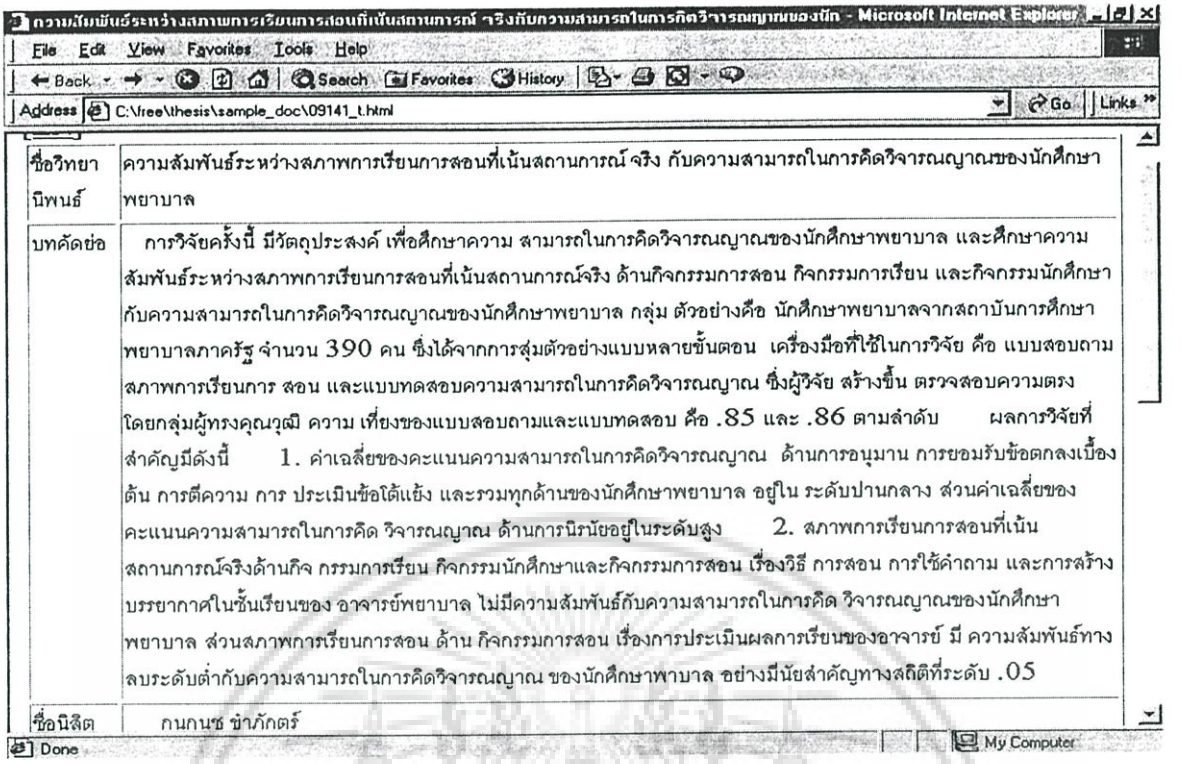
Back Search Favorites History

Address C:\free\thesis\sample_doc\09059_t.html Go Links

ชื่อวิทยานิพนธ์	ภาวะที่เหมาะสมในการผลิตกรดอิตาโคนิกโดย Aspergillus terreus I 10
บทคัดย่อ	ภาวะที่เหมาะสมสำหรับการผลิตกรดอิตาโคนิก โดย A. terreus I 10 ในระดับขวดเขย่า คือ เตรียมหัวเชื้อโดยเพาะเลี้ยงสปอร์ความหนาแน่น 5-10x10('9) สปอร์ในอาหารเลี้ยง เชื้อเพื่อการผลิตหัวเชื้อสปอร์งอกปริมาตร 50 มิลลิลิตร ที่มีการเติมเม็ดแก้ว ขนาดเส้นผ่าศูนย์กลาง 2 มิลลิเมตร หนัก 15 กรัม เป็นเวลา 36 ชั่วโมง มีน้ำตาลซูโครส 66 กรัมต่อ ลิตร และแอมโมเนียมซัลเฟต 1.75 กรัมต่อลิตร เป็นแหล่ง คาร์บอน และแหล่งไนโตรเจนตามลำดับ อัตราส่วนระหว่างปริมาณ คาร์บอนต่อปริมาณไนโตรเจน คือ 300 ต่อ 4 ค่าความเป็นกรด- ด่างตั้งต้นของอาหารเลี้ยงเชื้อ คือ 4.5 และเพาะเลี้ยงที่ อุณหภูมิ 30 องศาเซลเซียส สามารถใช้น้ำตาลทรายขาวเป็นแหล่งคาร์บอนแทนน้ำตาล ซูโครส บริสุทธิ์ได้ โดยผลผลิตไม่ลดลง คอลัมน์แก้วที่มีการ ให้อากาศด้านล่างเหมาะสมในการผลิตกรดอิตาโคนิกมากกว่า ถังหมักขนาด 5 ลิตร และภาวะที่เหมาะสมต่อการผลิตกรดอิตา โคนิกในคอลัมน์แก้วที่มีการให้อากาศด้านล่างคือ ขนาดของ หัวเชื้อเท่ากับ 2 เปอร์เซ็นต์ (ปริมาตรต่อปริมาตร) อัตรา การให้อากาศเท่ากับ 2.5 ลิตรต่อลิตรอาหารเลี้ยงเชื้อต่อ นาที
ชื่อนิสิต	อุษา กรีอักษร
ระดับปริญญาและรายละเอียด	วิทยานิพนธ์มหาบัณฑิต. วิทยาศาสตร์ (จุลชีววิทยาทางอุตสาหกรรม)

Done My Computer

รูปที่ ข.15 ตัวอย่างเอกสารในกลุ่มที่ 14ที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำร่วมกับ การหาค่าระยะห่าง ของกลุ่มคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 90 เฉพาะคำ intersec



ความสัมพันธ์ระหว่างสภาพการเรียนการสอนที่เน้นสถานการณ์ จริงกับความสามารถในการคิดวิจารณญาณของนัก - Microsoft Internet Explorer

File Edit View Favorites Tools Help

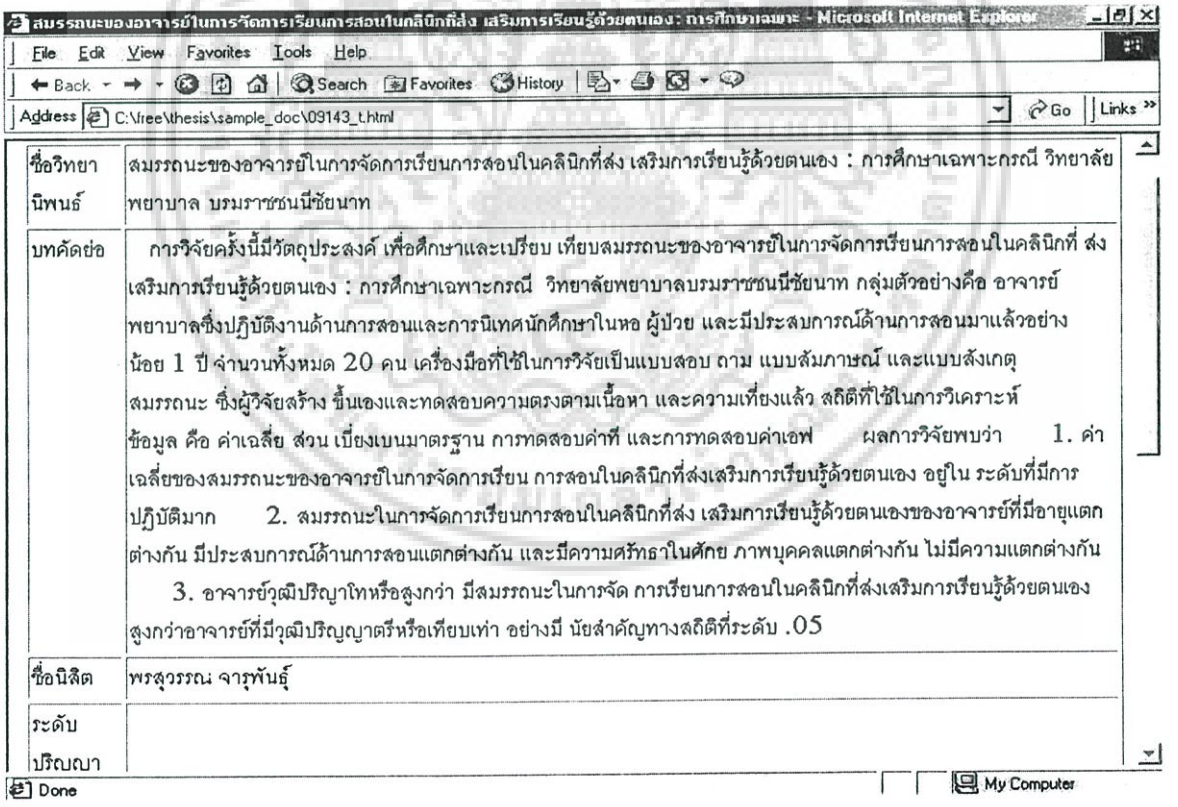
Back Search Favorites History

Address C:\free\thesis\sample_doc\09141_t.html

Go Links

ชื่อวิทยานิพนธ์	ความสัมพันธ์ระหว่างสภาพการเรียนการสอนที่เน้นสถานการณ์ จริง กับความสามารถในการคิดวิจารณญาณของนักศึกษาพยาบาล
บทคัดย่อ	การวิจัยครั้งนี้ มีวัตถุประสงค์ เพื่อศึกษาความ สามารถในการคิดวิจารณญาณของนักศึกษาพยาบาล และศึกษาความสัมพันธ์ระหว่างสภาพการเรียนการสอนที่เน้นสถานการณ์จริง ด้านกิจกรรมการสอน กิจกรรมการเรียน และกิจกรรมนักศึกษากับความสามารถในการคิดวิจารณญาณของนักศึกษาพยาบาล กลุ่ม ตัวอย่างคือ นักศึกษาพยาบาลจากสถาบันการศึกษาพยาบาลภาครัฐ จำนวน 390 คน ซึ่งได้จากการสุ่มตัวอย่างแบบหลายขั้นตอน เครื่องมือที่ใช้ในการวิจัย คือ แบบสอบถามสภาพการเรียนการ สอน และแบบทดสอบความสามารถในการคิดวิจารณญาณ ซึ่งผู้วิจัย สร้างขึ้น ตรวจสอบความตรงโดยกลุ่มผู้ทรงคุณวุฒิ ความ เที่ยงของแบบสอบถามและแบบทดสอบ คือ .85 และ .86 ตามลำดับ ผลการวิจัยที่สำคัญมีดังนี้ 1. ค่าเฉลี่ยของคะแนนความสามารถในการคิดวิจารณญาณ ด้านการอนุมาน การยอมรับข้อตกลงเบื้องต้น การตีความ การ ประเมินข้อโต้แย้ง และรวมทุกด้านของนักศึกษาพยาบาล อยู่ใน ระดับปานกลาง ส่วนค่าเฉลี่ยของคะแนนความสามารถในการคิด วิจารณญาณ ด้านการนิรนัยอยู่ในระดับสูง 2. สภาพการเรียนการสอนที่เน้นสถานการณ์จริงด้านกิจ กรรมการเรียน กิจกรรมนักศึกษาและกิจกรรมการสอน เรื่องวิธี การสอน การใช้คำถาม และการสร้างบรรยากาศในชั้นเรียนของ อาจารย์พยาบาล ไม่มีความสัมพันธ์กับความสามารถในการคิด วิจารณญาณของนักศึกษาพยาบาล ส่วนสภาพการเรียนการสอน ด้าน กิจกรรมการสอน เรื่องการประเมินผลการเรียนของอาจารย์ มี ความสัมพันธ์ทางลบระดับต่ำกับความสามารถในการคิดวิจารณญาณ ของนักศึกษาพาบาล อย่างมีนัยสำคัญทางสถิติที่ระดับ .05
ชื่อนิสิต	กนกนช ข่าภักตร์

Done My Computer



สมรรถนะของอาจารย์ในการจัดการเรียนการสอนในคลินิกที่ส่ง เสริมการเรียนรู้ด้วยตนเอง : การศึกษาเฉพาะ - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

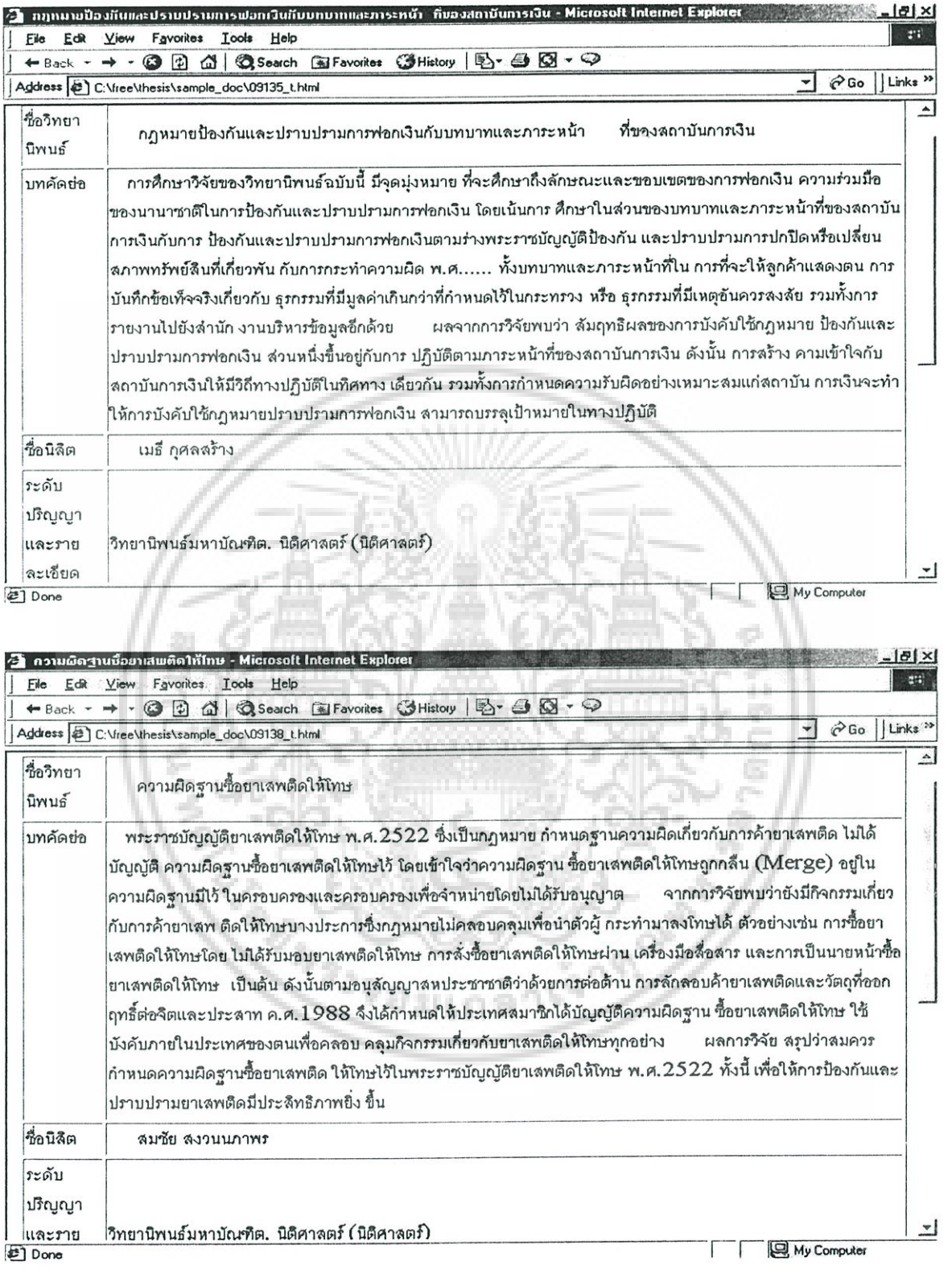
Address C:\free\thesis\sample_doc\09143_t.html

Go Links

ชื่อวิทยานิพนธ์	สมรรถนะของอาจารย์ในการจัดการเรียนการสอนในคลินิกที่ส่ง เสริมการเรียนรู้ด้วยตนเอง : การศึกษาเฉพาะกรณี วิทยาลัยพยาบาล บรมราชชนนีชัยนาท
บทคัดย่อ	การวิจัยครั้งนี้มีวัตถุประสงค์ เพื่อศึกษาและเปรียบ เทียบสมรรถนะของอาจารย์ในการจัดการเรียนการสอนในคลินิกที่ ส่งเสริมการเรียนรู้ด้วยตนเอง : การศึกษาเฉพาะกรณี วิทยาลัยพยาบาลบรมราชชนนีชัยนาท กลุ่มตัวอย่างคือ อาจารย์พยาบาลซึ่งปฏิบัติงานด้านการสอนและการนิเทศนักศึกษาในหอ ผู้ป่วย และมีประสบการณ์ด้านการสอนมาแล้วอย่างน้อย 1 ปี จำนวนทั้งหมด 20 คน เครื่องมือที่ใช้ในการวิจัยเป็นแบบสอบ ถาม แบบสัมภาษณ์ และแบบสังเกตุสมรรถนะ ซึ่งผู้วิจัยสร้าง ขึ้นเองและทดสอบความตรงตามเนื้อหา และความเที่ยงแล้ว สถิติที่ใช้ในการวิเคราะห์ข้อมูล คือ ค่าเฉลี่ย ส่วน เบี่ยงเบนมาตรฐาน การทดสอบค่าที และการทดสอบค่าเอฟ ผลการวิจัยพบว่า 1. ค่าเฉลี่ยของสมรรถนะของอาจารย์ในการจัดการเรียน การสอนในคลินิกที่ส่งเสริมการเรียนรู้ด้วยตนเอง อยู่ใน ระดับที่มีการปฏิบัติมาก 2. สมรรถนะในการจัดการเรียนการสอนในคลินิกที่ส่ง เสริมการเรียนรู้ด้วยตนเองของอาจารย์ที่มีอายุแตกต่างกัน มีประสบการณ์ด้านการสอนแตกต่างกัน และมีความศรัทธาในศักย ภาพบุคคลแตกต่างกัน ไม่มีความแตกต่างกัน 3. อาจารย์วุฒิปริญาโทหรือสูงกว่า มีสมรรถนะในการจัด การเรียนการสอนในคลินิกที่ส่งเสริมการเรียนรู้ด้วยตนเองสูงกว่าอาจารย์ที่มีวุฒิปริญญาตรีหรือเทียบเท่า อย่างมี นัยสำคัญทางสถิติที่ระดับ .05
ชื่อนิสิต	พรสุวรรณ จารุพันธุ์
ระดับปริญญา	

Done My Computer

รูปที่ ข.16 ตัวอย่างเอกสารในกลุ่มที่ 33ที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำร่วมกับ การหาค่าระยะห่างกลุ่มคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 100 เอาทั้งคำ intersec และไม่ intersec



กฎหมายป้องกันและปราบปรามการฟอกเงินกับบทบาทและภาระหน้า ที่ของสถาบันการเงิน - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09135_t.html Go Links

ชื่อวิทยานิพนธ์	กฎหมายป้องกันและปราบปรามการฟอกเงินกับบทบาทและภาระหน้า ที่ของสถาบันการเงิน
บทคัดย่อ	การศึกษาวิจัยของวิทยานิพนธ์ฉบับนี้ มีจุดมุ่งหมาย ที่จะศึกษาถึงลักษณะและขอบเขตของการฟอกเงิน ความร่วมมือของนานาชาติในการป้องกันและปราบปรามการฟอกเงิน โดยเน้นการ ศึกษาในส่วนของบทบาทและภาระหน้าที่ของสถาบันการเงินกับการ ป้องกันและปราบปรามการฟอกเงินตามร่างพระราชบัญญัติป้องกัน และปราบปรามการปกปิดหรือเปลี่ยนสภาพทรัพย์สินที่เกี่ยวพัน กับการกระทำความผิด พ.ศ..... ทั้งบทบาทและภาระหน้าที่ใน การที่จะให้ลูกค้าแสดงตน การบันทึกข้อเท็จจริงเกี่ยวกับ ธุรกรรมที่มีมูลค่าเกินกว่าที่กำหนดไว้ในกระทรวง หรือ ธุรกรรมที่มีเหตุอันควรสงสัย รวมทั้งการรายงานไปยังสำนัก งานบริหารข้อมูลอีกด้วย ผลจากการวิจัยพบว่า สัมฤทธิผลของการบังคับใช้กฎหมาย ป้องกันและปราบปรามการฟอกเงิน ส่วนหนึ่งขึ้นอยู่กับการ ปฏิบัติตามภาระหน้าที่ของสถาบันการเงิน ดังนั้น การสร้าง ความเข้าใจกับสถาบันการเงินให้มีวิถีทางปฏิบัติในทิศทาง เดียวกัน รวมทั้งการกำหนดความรับผิดอย่างเหมาะสมแก่สถาบัน การเงินจะทำให้การบังคับใช้กฎหมายปราบปรามการฟอกเงิน สามารถบรรลุเป้าหมายในทางปฏิบัติ
ชื่อนิสิต	เมธี กุศลสร้าง
ระดับปริญญาและรายละเอียด	วิทยานิพนธ์มหาบัณฑิต. นิติศาสตร์ (นิติศาสตร์)

Done My Computer

ความผิดฐานซื้อยาเสพติดให้โทษ - Microsoft Internet Explorer

File Edit View Favorites Tools Help

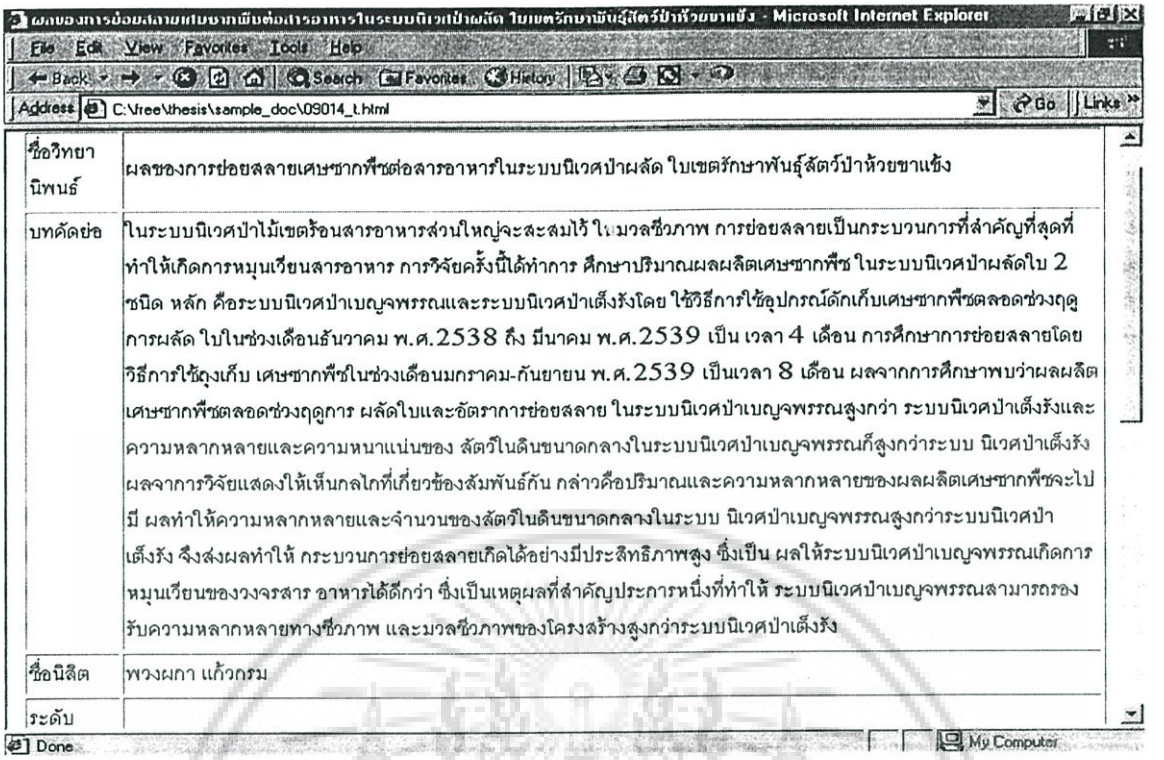
Back Search Favorites History

Address C:\free\thesis\sample_doc\09138_t.html Go Links

ชื่อวิทยานิพนธ์	ความผิดฐานซื้อยาเสพติดให้โทษ
บทคัดย่อ	พระราชบัญญัติยาเสพติดให้โทษ พ.ศ.2522 ซึ่งเป็นกฎหมาย กำหนดฐานความผิดเกี่ยวกับการค้ายาเสพติด ไม่ได้บัญญัติ ความผิดฐานซื้อยาเสพติดให้โทษไว้ โดยเข้าใจว่าความผิดฐาน ซื้อยาเสพติดให้โทษถูกกลืน (Merge) อยู่ในความผิดฐานมีไว้ ในครอบครองและครอบครองเพื่อจำหน่ายโดยไม่ได้รับอนุญาต จากการวิจัยพบว่ายังมีกิจกรรมเกี่ยวกับการค้ายาเสพ ติดให้โทษบางประการซึ่งกฎหมายไม่คลอบคลุมเพื่อนำตัวผู้ กระทำมาลงโทษได้ ตัวอย่างเช่น การซื้อยาเสพติดให้โทษโดย ไม่ได้รับมอบยาเสพติดให้โทษ การสั่งซื้อยาเสพติดให้โทษผ่าน เครื่องมือสื่อสาร และการเป็นนายหน้าซื้อยาเสพติดให้โทษ เป็นต้น ดังนั้นตามอนุสัญญาสหประชาชาติว่าด้วยการต่อต้าน การลักลอบค้ายาเสพติดและวัตถุที่ออกฤทธิ์ต่อจิตและประสาท ค.ศ.1988 จึงได้กำหนดให้ประเทศสมาชิกได้บัญญัติความผิดฐาน ซื้อยาเสพติดให้โทษ ใช้บังคับภายในประเทศของตนเพื่อคลอบ คลุมกิจกรรมเกี่ยวกับยาเสพติดให้โทษทุกอย่าง ผลการวิจัย สรุปว่าสมควรกำหนดความผิดฐานซื้อยาเสพติด ให้โทษไว้ในพระราชบัญญัติยาเสพติดให้โทษ พ.ศ.2522 ทั้งนี้ เพื่อให้การป้องกันและปราบปรามยาเสพติดมีประสิทธิภาพยิ่ง ขึ้น
ชื่อนิสิต	สมชัย สงวนนภาพร
ระดับปริญญาและราย	วิทยานิพนธ์มหาบัณฑิต. นิติศาสตร์ (นิติศาสตร์)

Done My Computer

รูปที่ ข.17 ตัวอย่างเอกสารในกลุ่มที่ 32 ที่จัดกลุ่มด้วยวิธีการหาความเหมือนกันของคำร่วมกับ การหาค่าระยะห่าง ของกลุ่มคำสำคัญระหว่างกลุ่มเอกสารที่ Distance 300 เอาทั้งคำ intersec และไม่ intersec



ผลของการย่อยสลายเศษซากพืชต่อสารอาหารในระบบนิเวศป่าผลัดใบเขตรักษาพันธุ์สัตว์ป่าห้วยขาแข้ง - Microsoft Internet Explorer

File Edit View Favorites Tools Help

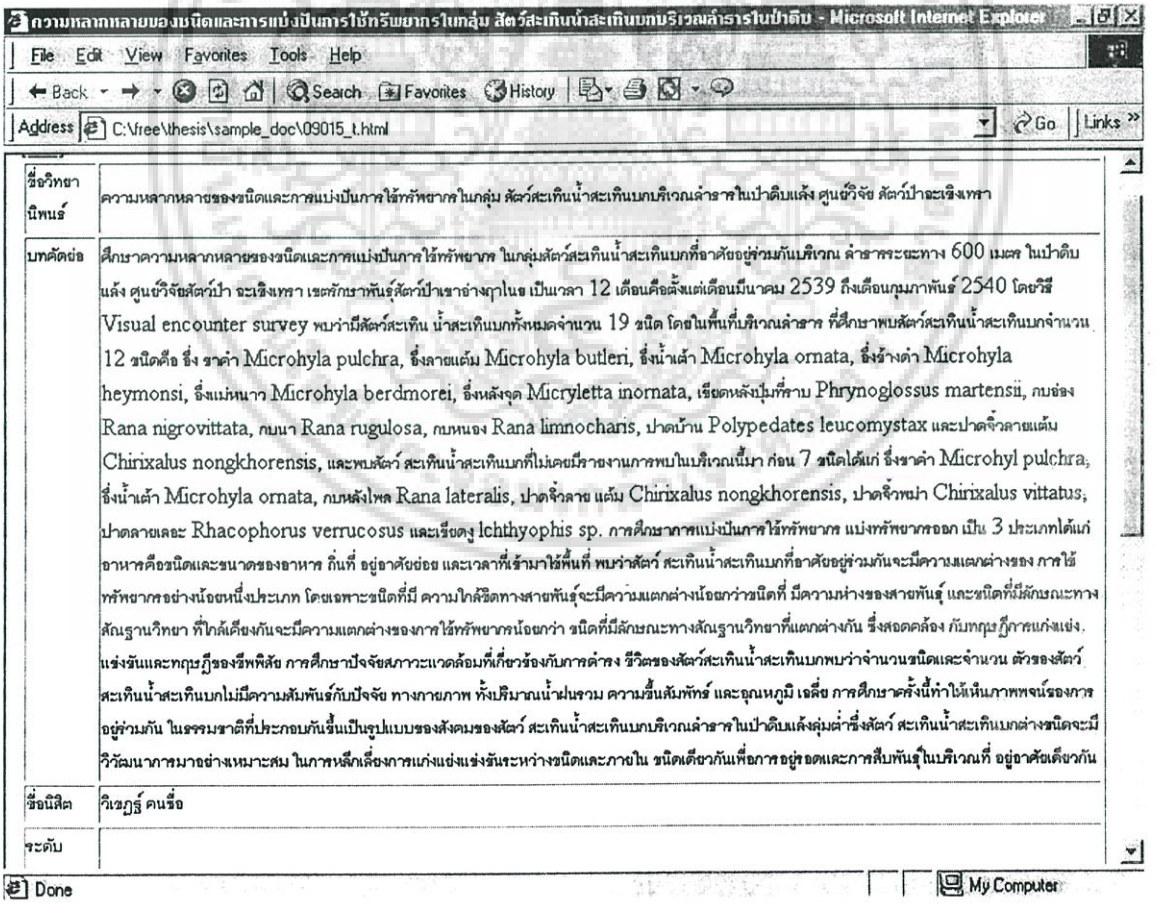
Back Search Favorites History

Address C:\free\thesis\sample_doc\09014_t.html

Go Links

ชื่อวิทยานิพนธ์	ผลของการย่อยสลายเศษซากพืชต่อสารอาหารในระบบนิเวศป่าผลัด ใบเขตรักษาพันธุ์สัตว์ป่าห้วยขาแข้ง
บทคัดย่อ	ในระบบนิเวศป่าไม้เขตร้อนสารอาหารส่วนใหญ่จะสะสมไว้ ในมวลชีวภาพ การย่อยสลายเป็นกระบวนการที่สำคัญที่สุดที่ทำให้เกิดการหมุนเวียนสารอาหาร การวิจัยครั้งนี้ได้ทำการ ศึกษาปริมาณผลผลิตเศษซากพืช ในระบบนิเวศป่าผลัดใบ 2 ชนิด หลัก คือระบบนิเวศป่าเบญจพรรณและระบบนิเวศป่าเต็งรังโดย ใช้วิธีการใช้อุปกรณ์ดักเก็บเศษซากพืชตลอดช่วงฤดูการผลัด ใบในช่วงเดือนธันวาคม พ.ศ.2538 ถึง มีนาคม พ.ศ.2539 เป็น เวลา 4 เดือน การศึกษาการย่อยสลายโดยวิธีการใช้ถุงเก็บ เศษซากพืชในช่วงเดือนมกราคม-กันยายน พ.ศ.2539 เป็นเวลา 8 เดือน ผลจากการศึกษาพบว่าผลผลิตเศษซากพืชตลอดช่วงฤดูการ ผลัดใบและอัตราการย่อยสลาย ในระบบนิเวศป่าเบญจพรรณสูงกว่า ระบบนิเวศป่าเต็งรังและความหลากหลายและความหนาแน่นของ สัตว์ในดินขนาดกลางในระบบนิเวศป่าเบญจพรรณก็สูงกว่าระบบ นิเวศป่าเต็งรัง ผลจากการวิจัยแสดงให้เห็นกลไกที่เกี่ยวข้องสัมพันธ์กัน กล่าวคือปริมาณและความหลากหลายของผลผลิตเศษซากพืชจะไปมี ผลทำให้ความหลากหลายและจำนวนของสัตว์ในดินขนาดกลางในระบบ นิเวศป่าเบญจพรรณสูงกว่าระบบนิเวศป่าเต็งรัง จึงส่งผลทำให้ กระบวนการย่อยสลายเกิดได้อย่างมีประสิทธิภาพสูง ซึ่งเป็น ผลให้ระบบนิเวศป่าเบญจพรรณเกิดการหมุนเวียนของวงจรสาร อาหารได้ดีกว่า ซึ่งเป็นเหตุผลที่สำคัญประการหนึ่งที่ทำให้ ระบบนิเวศป่าเบญจพรรณสามารถรองรับความหลากหลายทางชีวภาพ และมวลชีวภาพของโครงสร้างสูงกว่าระบบนิเวศป่าเต็งรัง
ชื่อนิสิต	พวงผกา แก้วกรม
ระดับ	

Done My Computer



ความหลากหลายของชนิดและการแบ่งปันการใช้ทรัพยากรในกลุ่ม สัตว์สะเทินน้ำสะเทินบกบริเวณลำธารในป่าดิบแล้ง - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

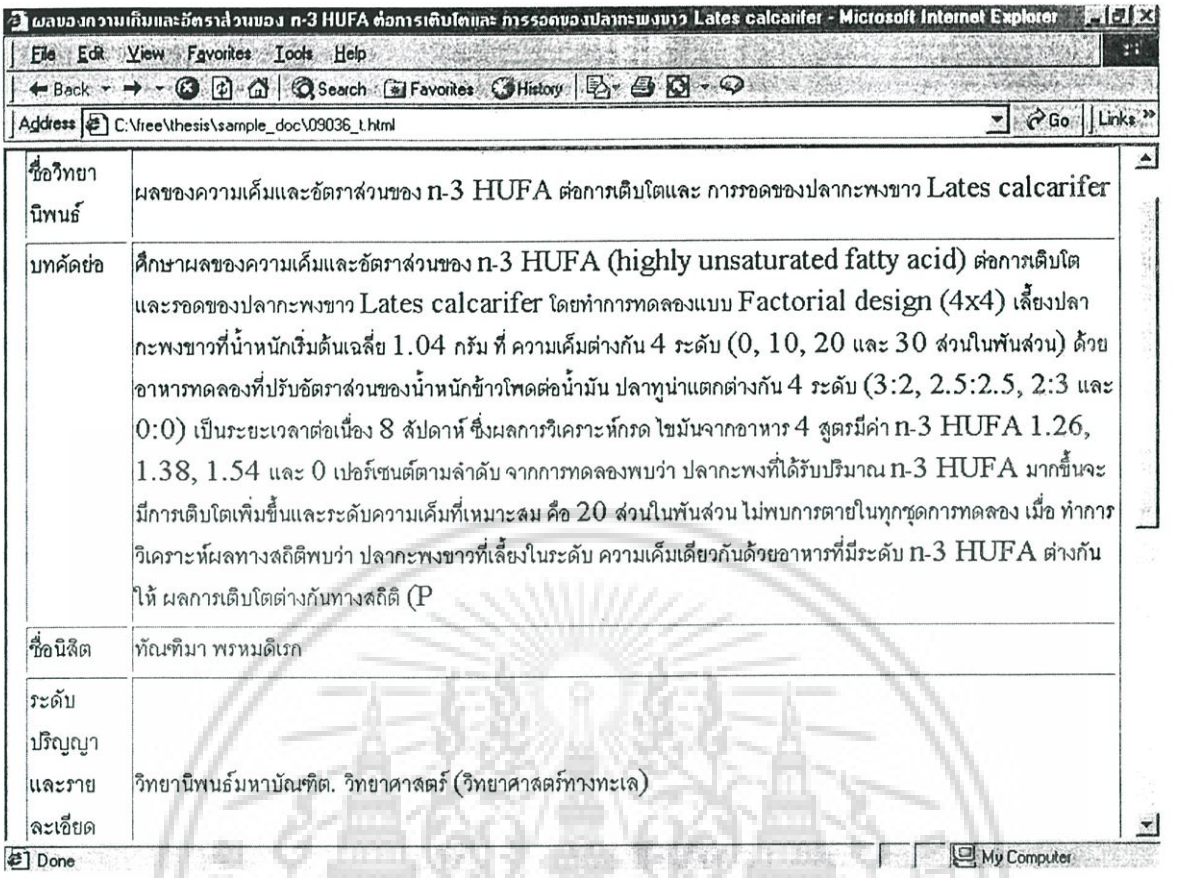
Address C:\free\thesis\sample_doc\09015_t.html

Go Links

ชื่อวิทยานิพนธ์	ความหลากหลายของชนิดและการแบ่งปันการใช้ทรัพยากรในกลุ่ม สัตว์สะเทินน้ำสะเทินบกบริเวณลำธารในป่าดิบแล้ง ศูนย์วิจัย สัตว์ป่าฉะเชิงเทรา
บทคัดย่อ	ศึกษาความหลากหลายของชนิดและการแบ่งปันการใช้ทรัพยากร ในกลุ่มสัตว์สะเทินน้ำสะเทินบกที่อาศัยอยู่ร่วมกันบริเวณ ลำธารระยะทาง 600 เมตร ในป่าดิบแล้ง ศูนย์วิจัยสัตว์ป่า ฉะเชิงเทรา เขตรักษาพันธุ์สัตว์ป่าเขาอ่างฤาไนย เป็นเวลา 12 เดือนคือตั้งแต่เดือนมีนาคม 2539 ถึงเดือนกุมภาพันธ์ 2540 โดยวิธี Visual encounter survey พบว่ามีสัตว์สะเทิน น้ำสะเทินบกทั้งหมดจำนวน 19 ชนิด โดยในพื้นที่บริเวณลำธาร ที่ศึกษาพบสัตว์สะเทินน้ำสะเทินบกจำนวน 12 ชนิดคือ ซึ่ง ขาค่า Microhyla pulchra, ซึ่งลายแต้ม Microhyla butleri, ซึ่งน้ำเต้า Microhyla ornata, ซึ่งร่างดำ Microhyla heymonsi, ซึ่งแม่หนาว Microhyla berdmorei, ซึ่งหลังจุด Micryletta inornata, เขียดหลังปุ่มที่ราบ Phrynoglossus martensii, กบช่อง Rana nigrovittata, กบนา Rana rugulosa, กบหนอง Rana limnocharis, ปาดบ้าน Polypedates leucomystax และปาดจิ้วลายแต้ม Chirixalus nongkhorensis, และพบสัตว์ สะเทินน้ำสะเทินบกที่ไม่เคยมีรายงานการพบในบริเวณนี้มา ก่อน 7 ชนิดได้แก่ ซึ่งขาค่า Microhyl pulchra, ซึ่งน้ำเต้า Microhyla ornata, กบหลังไพล Rana lateralis, ปาดจิ้วลาย แต้ม Chirixalus nongkhorensis, ปาดจิ้วพม่า Chirixalus vittatus, ปาดลายเลอะ Rhacophorus verrucosus และเขียดงู Ichthyophis sp. การศึกษาการแบ่งปันการใช้ทรัพยากร แบ่งทรัพยากรออก เป็น 3 ประเภทได้แก่ อาหารคือชนิดและขนาดของอาหาร ถิ่นที่ อยู่อาศัยย่อย และเวลาที่เข้ามาใช้พื้นที่ พบว่าสัตว์ สะเทินน้ำสะเทินบกที่อาศัยอยู่ร่วมกันจะมีความแตกต่างของ การใช้ทรัพยากรอย่างน้อยหนึ่งประเภท โดยเฉพาะชนิดที่มี ความใกล้ชิดทางสายพันธุ์จะมีความแตกต่างน้อยกว่าชนิดที่ มีความห่างของสายพันธุ์ และชนิดที่มีลักษณะทางสัณฐานวิทยา ที่ใกล้เคียงกันจะมีความแตกต่างของการใช้ทรัพยากรน้อยกว่า ชนิดที่มีลักษณะทางสัณฐานวิทยาที่แตกต่างกัน ซึ่งสอดคล้อง กับทฤษฎีการแก่งแย่ง แย่งชันและทฤษฎีของวีพพิลัย การศึกษาปัจจัยสภาวะแวดล้อมที่เกี่ยวข้องกับการดำรง ชีวิตของสัตว์สะเทินน้ำสะเทินบกพบว่าจำนวนชนิดและจำนวน ตัวของสัตว์สะเทินน้ำสะเทินบกไม่มีความสัมพันธ์กับปัจจัย ทางกายภาพ ทั้งปริมาณน้ำฝนรวม ความชื้นสัมพัทธ์ และอุณหภูมิ เฉลี่ย การศึกษาครั้งนี้ทำให้เห็นภาพชุมชนของการอยู่ร่วมกัน ในธรรมชาติที่ประกอบกันขึ้นเป็นรูปแบบของสังคมของสัตว์ สะเทินน้ำสะเทินบกบริเวณลำธารในป่าดิบแล้งคุ่มต่ำซึ่งสัตว์ สะเทินน้ำสะเทินบกต่างชนิดจะมีวิวัฒนาการมาอย่างเหมาะสม ในการหลีกเลี่ยงการแก่งแย่งแย่งชันระหว่างชนิดและภายใน ชนิดเดียวกันเพื่อการอยู่รอดและการสืบพันธุ์ในบริเวณที่ อยู่อาศัยเดียวกัน
ชื่อนิสิต	วิเชฎฐ์ คนซื่อ
ระดับ	

Done My Computer

รูปที่ ข.18 ตัวอย่างเอกสารในกลุ่มที่ 6 ที่จัดกลุ่มด้วยวิธีการ PDO



ผลของความเค็มและอัตราส่วนของ n-3 HUFA ต่อการเติบโตและ การรอดของปลากะพงขาว Lates calcarifer - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09036_t.html Go Links

ชื่อวิทยานิพนธ์	ผลของความเค็มและอัตราส่วนของ n-3 HUFA ต่อการเติบโตและ การรอดของปลากะพงขาว Lates calcarifer
บทคัดย่อ	ศึกษาผลของความเค็มและอัตราส่วนของ n-3 HUFA (highly unsaturated fatty acid) ต่อการเติบโตและรอดของปลากะพงขาว Lates calcarifer โดยทำการทดลองแบบ Factorial design (4x4) เลี้ยงปลากะพงขาวที่น้ำหนักเริ่มต้นเฉลี่ย 1.04 กรัม ที่ ความเค็มต่างกัน 4 ระดับ (0, 10, 20 และ 30 ส่วนในพันส่วน) ด้วยอาหารทดลองที่ปรับอัตราส่วนของน้ำหนักข้าวโพดต่อน้ำมัน ปลาทูน่าแตกต่างกัน 4 ระดับ (3:2, 2.5:2.5, 2:3 และ 0:0) เป็นระยะเวลาต่อเนื่อง 8 สัปดาห์ ซึ่งผลการวิเคราะห์กรด ไขมันจากอาหาร 4 สูตรมีค่า n-3 HUFA 1.26, 1.38, 1.54 และ 0 เปอร์เซนต์ตามลำดับ จากการทดลองพบว่า ปลากะพงที่ได้รับปริมาณ n-3 HUFA มากขึ้นจะมีการเติบโตเพิ่มขึ้นและระดับความเค็มที่เหมาะสม คือ 20 ส่วนในพันส่วน ไม่พบการตายในทุกชุดการทดลอง เมื่อ ทำการวิเคราะห์ผลทางสถิติพบว่า ปลากะพงขาวที่เลี้ยงในระดับ ความเค็มเดียวกันด้วยอาหารที่มีระดับ n-3 HUFA ต่างกันให้ ผลการเติบโตต่างกันทางสถิติ (P
ชื่อนิสิต	ทัณฑิมา พรหมดิเรก
ระดับปริญญาและรายละเอียด	วิทยานิพนธ์มหาบัณฑิต. วิทยาศาสตร์ (วิทยาศาสตร์ทางทะเล)

Done My Computer

ระดับโปรตีนและไขมันที่เหมาะสมในอาหารเม็ดแบบแห้งสำหรับ ปลากะพงขาว Lates calcarifer วัยรุ่น - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites History

Address C:\free\thesis\sample_doc\09037_t.html Go Links

ชื่อวิทยานิพนธ์	ระดับโปรตีนและไขมันที่เหมาะสมในอาหารเม็ดแบบแห้งสำหรับ ปลากะพงขาว Lates calcarifer วัยรุ่น
บทคัดย่อ	ทำการศึกษาระดับโปรตีนและไขมันที่เหมาะสมของอาหาร ชนิดเม็ดแบบแห้งในการเลี้ยงปลากะพงขาว Lates calcarifer วัยรุ่นให้มีการเติบโตและอัตราการอดสูงสุด ออกแบบการทดลอง แบบ factorial (3 x 4) โดยแบ่งการทดลองออกเป็น 2 ระยะ ระยะที่ 1 ศึกษาการเติบโตของปลากะพงขาวขนาดน้ำหนัก ตัวเฉลี่ย 1.1+(,-)0.1 กรัม เป็นระยะเวลา 8 สัปดาห์ โดยใช้ อาหารที่มีโปรตีน 3 ระดับที่ 35, 40 และ 45 เปอร์เซ็นต์ แต่ละระดับมีไขมัน 10, 15, 20 และ 25 เปอร์เซ็นต์ เลี้ยง ปลาในกระชังขนาด 0.5 x 0.5 x 0.8 ลูกบาศก์เมตร กระชังละ 20 ตัว ทำ 3 ซ้ำต่อ 1 สูตรการทดลอง ผลการทดลองปรากฏว่า ปลาที่เลี้ยงด้วยอาหารสูตร 45/15 (โปรตีน/ไขมัน) ให้การ เติบโตดีที่สุด (น้ำหนักตัวเฉลี่ย 30.0+(,-)1.0 กรัม) แต่ไม่แตกต่างทางสถิติกับปลาที่เลี้ยงด้วยอาหารสูตร 45/25 และ 45/20 (น้ำหนักตัวเฉลี่ย 29.3+(,-)1.4 กรัมและ 28.4 +(,_)1.3 กรัม ตามลำดับ) นอกจากนี้สูตร 45/15 ยังให้อัตรา การเติบโตสัมพัทธ์ต่อวันสูงสุด (0.5) ให้ค่าอัตราการเลก เนื้อต่ำสุด (1.26) ค่าอัตราการบริโภคอาหารและพลังงาน ต่อวันต่ำสุด (4.57 เปอร์เซ็นต์ และ 26.0 กิโลแคลอรี/100 กรัมน้ำหนักตัว ตามลำดับ) และให้ค่าอัตราการบริโภคโปรตีน ต่อวัน 2.05 เปอร์เซ็นต์ เมื่อพิจารณาคุณค่าทางโภชนาการ พบว่าสูตร 45/15 มีค่าการใช้โปรตีนสุทธิ ประสิทธิภาพการ ใช้โปรตีน และประสิทธิภาพการใช้พลังงาน 1.46, 1.78 และ 0.14 กรัม/กิโลแคลอรี ตามลำดับ จากการศึกษาระยะนี้ไม่พบ การตายเกิดขึ้นในทุกชุดการทดลอง ระยะที่ 2 ศึกษาอัตราการย่อยอาหารโดยใช้ปลากะพงขาว ชุดเดิมทำการทดลองต่อจากระยะที่ 1 ในตู้กระจก ขนาด 0.3 x 0.6 x 0.3 ลูกบาศก์เมตร ตู้ละ 10 ตัว ใช้ระบบถ่ายเท น้ำตลอดเวลา ผลการทดลองพบอัตราการย่อยโปรตีนและพลังงาน ที่เวลา 3 ชั่วโมงหลังให้อาหารมีค่าระหว่าง 82.45-91.74 เปอร์เซ็นต์ และ 82.45-92.33 เปอร์เซ็นต์ ตามลำดับ และ ที่ 6 ชั่วโมงหลังให้อาหารมีค่าระหว่าง 85.92-92.33 เปอร์เซ็นต์ และ 85.04-91.90 เปอร์เซ็นต์ ตามลำดับ สัดส่วน ของความสามารถในการย่อยพลังงานต่อกรัมโปรตีนของอาหารทั้ง 12 สูตรมีค่าระหว่าง 10.47-13.52 กิโลแคลอรี/กรัมโปรตีน

Done My Computer

รูปที่ ข.19 ตัวอย่างเอกสารในกลุ่มที่ 9 ที่จัดกลุ่มด้วยวิธีการ DF

ประวัติผู้เขียน

นาย ไพฑูรย์ นุชแจ้ง เกิดเมื่อวันที่ 13 มิถุนายน 2507 ที่จังหวัดนนทบุรี สำเร็จการศึกษาวิทยาศาสตรบัณฑิต (สาขา เทคโนโลยีอิเล็กทรอนิกส์) จากมหาวิทยาลัยรามคำแหง ปีการศึกษา 2540 วิทยาศาสตรบัณฑิต (ประมง) จากมหาวิทยาลัยเกษตรศาสตร์ ปีการศึกษา 2531

ปี พ.ศ. 2532 เข้าทำงานในตำแหน่งนักวิชาการประมง และ พนักงานขายอาหารสัตว์น้ำ บริษัทเครือเจริญโภคภัณฑ์อาหารสัตว์ ปัจจุบันทำธุรกิจส่วนตัว