ระบบจัดกลุ่มผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น ในเครือข่ายใยแมงมุม

CLUSTERING OF SEARCH ENGINE RESULTS IN WORLD WIDE WEB

ภาณุพงศ์ ชวะวิทย์
PANUPONG CHAVAVIT

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
บัณฑิตวิทยาลัย
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2547
ISBN 974-456-448-2

CLUSTERING OF SEARCH ENGINE RESULTS IN WORLD WIDE WEB

PANUPONG CHAVAVIT

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2004

ISBN 974-456-448-2

หัวข้อวิทยานิพนธ์	**ระบบจัดกลุ่มผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นในเครือข่ายใยแมงมุม**
ชื่อนักศึกษา	**นาย ภาณุพงศ์ ชาวะวิทย์**
รหัสประจำตัว	42067042
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
พ.ศ.	2547
อาจารย์ผู้ควบคุมวิทยานิพนธ์	ผศ. ดร. โชติพัชร์ ภรณวลัย

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอวิธีการใหม่เพื่อจัดกลุ่มผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นในเครือข่ายใยแมงมุม ในปัจจุบันข้อมูลในเครือข่ายใยแมงมุมมีจำนวนมากมายมหาศาลและยังมีอัตราการเพิ่มขึ้นอย่างรวดเร็ว แม้จะมีเครื่องมือการค้นหาข้อมูลที่เรียกว่า เสิร์ชเอ็นจิ้น (Search Engine) แต่เสิร์ชเอ็นจิ้นที่มีอยู่นั้นมีประสิทธิภาพไม่เพียงพอ เนื่องจากการแสดงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นจะแสดงในรูปแบบลำดับของผลการค้นหาข้อมูล เมื่อมีข้อมูลที่เกี่ยวข้องกับคำที่ใช้ค้นหาข้อมูลในหลาย ๆ เนื้อหาที่ต่างกัน ทำให้ผู้ที่ต้องการค้นหาข้อมูลต้องเสียเวลาในการดูผลการค้นหาข้อมูลทีละลำดับว่าเป็นข้อมูลที่มีเนื้อหาที่ต้องการหรือไม่ จึงมีแนวทางในการแก้ปัญหานี้ด้วยการจัดกลุ่มผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นตามเนื้อหาของผลเหล่านั้น โดยการจัดกลุ่มนี้จะใช้คำและ URL ที่อยู่ในแต่ละผลการค้นหาข้อมูล รวมถึง URL ของแต่ละผลการค้นหาข้อมูลมาเป็นข้อมูลเพื่อแสดงความคล้ายระหว่างแต่ละผลการค้นหา เมื่อแสดงผลการค้นหาข้อมูลเป็นกลุ่มจะช่วยให้ผู้ที่ต้องการค้นหาข้อมูล สามารถเลือกกลุ่มของข้อมูลที่มีเนื้อหาตรงตามความต้องการ เป็นการประหยัดเวลา และยังช่วยให้ได้ข้อมูลซึ่งมีประโยชน์แต่อาจจะอยู่ในอันดับหลัง ๆ ของผลการค้นหาข้อมูลด้วย จากผลการทดลองของวิธีการจัดกลุ่มผลการค้นหาข้อมูลโดยวิธีการที่พัฒนาขึ้นแสดงให้เห็นว่าสามารถทำการจัดกลุ่มให้ถูกต้องและครอบคลุมไปถึงข้อมูลที่ไม่มีคำอยู่หรือข้อมูลที่ไม่มี URL อยู่ได้ ซึ่งสามารถทำการจัดกลุ่มได้ดีและมีประสิทธิภาพมากกว่าวิธีการจัดกลุ่มโดยใช้คำแต่เพียงอย่างเดียว

Thesis Title	Clustering of Search Engine result in World Wide Web
Student	Mr. Panupong Chavavit
Student ID.	42067042
Degree	Master of Science
Program	Information Technology
Year	2004
Thesis Advisor	Asst. Prof. Dr. Chotipat Pornavalai

ABSTRACT

This thesis presents the new way for cluster the result of search engine in World Wide Web. At present, there is a vast amount of information on World Wide Web and the growth rate is also extremely high. Although, there is a tool for search the information which is called search engine. But these search engines can not provide the adequate performance due to the search results of search engine are in long list of results. When there are info relevant to search query in many different topics, users have to take a lot of time to check the list of search result one by one for the interested topic. The solution for this situation is to cluster the search result into group according to the topic of each result. This clustering algo uses word and URL in each result and URL of each result as the information to calculate the similarity between each result. When the result of search engine is displayed in the group of results, users can choose the group of interested topic. This will save a lot of time and user can get the useful information which are in the very low order in original search result. The results of the experiment show that the clustering algorithm can cluster result correctly and it also cover the no word or no URL results. And it has the better results and more performance than the clustering algorithm that use only word.

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จได้ด้วยความกรุณาจากอาจารย์ที่ปรึกษา ผศ. ดร. โชติพัชร์ ภรณวลัย ที่ได้ให้ความช่วยเหลือ ให้คำชี้แนะและให้คำปรึกษาในการแก้ปัญหา ช่วยจัดหาอุปกรณ์เพื่อนำมาใช้ในการทดลอง ตลอดจนให้ความรู้และประสบการณ์ที่ดีแก่ข้าพเจ้า

ขอขอบพระคุณกรรมการสอบหัวข้อวิทยานิพนธ์และโครงร่างวิทยานิพนธ์ทุกท่าน ที่ได้กรุณาให้คำชี้แนะตลอดจนแนะนำ จนในที่สุดทำให้วิทยานิพนธ์เล่มนี้สำเร็จได้อย่างสมบูรณ์

ขอขอบคุณบุคคลอีกหลายท่านที่ข้าพเจ้าไม่ได้เอ่ยชื่อไว้ ที่ท่านได้ให้ความช่วยเหลือ ให้ข้อมูลที่มีประโยชน์ และเป็นกำลังใจให้ข้าพเจ้าในทุก ๆ ส่วนของการทำวิทยานิพนธ์เล่มนี้ จนสำเร็จลุล่วงไปด้วยดี

สำหรับคุณความดีอันใดที่เกิดขึ้นจากวิทยานิพนธ์นี้ ข้าพเจ้าขอมอบให้กับบิดามารดาซึ่งเป็นที่เคารพและรักยิ่ง ตลอดจนครูอาจารย์ที่เคารพรักทุกท่านที่ได้ประสิทธิประสาทวิชาความรู้ สั่งสอนอบรม และถ่ายทอดประสบการณ์ที่ดีให้แก่ข้าพเจ้า

ภาณุพงศ์ ชาวะวิทย์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ หน้า

สารบัญรูป

สารบัญรูป (ต่อ)

รูปที่ หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

1.1.1 สาเหตุและลักษณะของปัญหา

เครือข่ายใยแมงมุมเป็นแหล่งข้อมูลขนาดใหญ่ มีจำนวนข้อมูลในระดับหลายพันล้านเว็บเพจขึ้นไป (จากข้อมูลในเว็บไซท์ของ Google [12] ได้ทำการเก็บข้อมูลเว็บเพจไว้ในฐานข้อมูลมากกว่า 3,300,000,000 เว็บเพจ) และคาดกันว่าอัตราการเพิ่มขึ้นของข้อมูลใน เครือข่ายใยแมงมุม มีมากกว่า 1,500,000 หน้า/วัน [8] ถึงแม้จำนวนข้อมูลที่มากมายมหาศาลจะเป็นสิ่งที่ดี แต่ขณะเดียวกันก็ทำให้เกิดปัญหาสำหรับผู้ที่ต้องการค้นหาข้อมูล แม้จะมีเครื่องมือที่ใช้ในการค้นหาข้อมูลซึ่งในปัจจุบันเรียกว่า เสิร์ชเอ็นจิ้น (Search Engine) ปัญหาในการค้นหาข้อมูลนี้ไม่ใช่ปัญหาจากการค้นหาข้อมูลไม่พบ แต่เป็นปัญหาเนื่องจากผลการค้นหาข้อมูลโดยเสิร์ชเอ็นจิ้นที่ได้มีมากเกินไป และข้อมูลจำนวนมากก็ไม่ใช่ข้อมูลที่ผู้ใช้ต้องการจะได้ทั้งหมด

สาเหตุของปัญหาที่เกิดขึ้นนี้มี 2 ประการหลัก ๆ ประการแรก คือ เสิร์ชเอ็นจิ้นที่มีอยู่ในปัจจุบันส่วนใหญ่มีการพัฒนาขึ้นมาใช้งานเป็นเวลานานพอสมควร ซึ่งในช่วงเวลาที่พัฒนาขึ้นมานั้น เป็นการพัฒนาขึ้นมาเพื่อใช้ค้นหาข้อมูลในระดับร้อยล้านเว็บ แต่เมื่อใช้งานมาจนถึงปัจจุบันจำนวนข้อมูลได้เพิ่มมากขึ้นอย่างมหาศาล ทำให้เสิร์ชเอ็นจิ้นซึ่งแสดงผลการค้นหาข้อมูลในรูปแบบลำดับของผล โดยเรียงลำดับจากผลที่มีความสัมพันธ์กับคำที่ใช้ในการค้นหาจากมากที่สุดไปจนถึงน้อยที่สุด เมื่อมีข้อมูลอยู่เป็นจำนวนมากในเครือข่ายใยแมงมุม ทำให้ผลการค้นหาข้อมูลที่ได้จากเสิร์ชเอ็นจิ้นมีจำนวนมากตามไปด้วย ซึ่งผลการค้นหาข้อมูลนี้จะมีความสัมพันธ์กับคำที่ใช้ในการค้นหาในหลากหลายเนื้อหาแตกต่างกันไป เช่น การค้นหาข้อมูลโดยคำที่ใช้ในการค้นหาคือ SEA GAMES จะได้เอกสารที่สัมพันธ์กับคำว่า SEA GAMES ในหลากหลายเนื้อหาได้แก่ SEA GAMES ซึ่งเป็นการแข่งขันกีฬาของภูมิภาคเอเชียตะวันออกเฉียงใต้, SEA GAMES ซึ่งเป็นกีฬาทางทะเล เช่น เรือใบ, กระดานโต้คลื่น, SEA GAMES ซึ่งเป็นเกมคอมพิวเตอร์ที่มีทะเลเข้ามาเกี่ยวข้อง หรือแม้แต่ การตกปลาในทะเลก็เป็นเนื้อหาที่เกี่ยวข้องกับ SEA GAMES ได้เช่นกัน เมื่อผลการค้นหาที่มีเนื้อหาแตกต่างกันถูกเสิร์ชเอ็นจิ้นนำมาแสดงผลรวมกัน ในรูปแบบของลำดับของผลการค้นหาจะทำให้เกิดความยุ่งยากแก่ผู้ที่ค้นหาข้อมูลเพราะผู้ที่ค้นหาข้อมูลจะต้องมองหาเฉพาะข้อมูลที่เป็นเนื้อหาที่ต้องการจากลำดับของผลการค้นหาข้อมูลที่แสดงโดยเสิร์ชเอ็นจิ้น จะเห็นได้ว่าเป็นการค้นหาที่ซ้ำซ้อนและเป็นการเสียเวลาของผู้ที่ค้นหาข้อมูล แทนที่จะช่วยประหยัดเวลาอย่างที่ควรจะเป็น

อีกประการหนึ่ง ผู้ที่ค้นหาข้อมูลมักจะไม่ได้ดูไปถึงผลการค้นหาข้อมูลที่อยู่ลำดับหลัง ๆ จึงทำให้ผลการค้นหาในลำดับหลัง ๆ ไม่ถูกนำไปใช้แม้จะเกี่ยวข้องกับคำที่ใช้ค้นหาและมีเนื้อหาที่ผู้ใช้ต้องการก็ตาม การที่ผลการค้นหาที่มีเนื้อหาแตกต่างกัน ถูกเสิร์ชเอ็นจิ้นนำมาแสดงผลรวมกันในรูปแบบของลำดับผลการค้นหานั้น ส่วนหนึ่งเป็นผลมาจากปัญหาประการที่สอง ซึ่งก็คือ คำที่ใช้ในการค้นหาข้อมูล โดยส่วนใหญ่แล้วจำนวนคำที่มักจะถูกใช้ในการค้นหาข้อมูลจะอยู่ในช่วงจำนวน 1-3 คำ ด้วยจำนวนคำเพียงเท่านี้ ปัญหาที่กล่าวมาข้างต้นจึงมักจะเกิดขึ้น เหตุผลหนึ่งที่คนทั่วไปมักจะใช้คำจำนวนน้อย ๆ ในการค้นหาข้อมูลเพราะเนื่องจากผู้ที่ค้นหาข้อมูลอาจยังไม่แน่ใจกับข้อมูลที่ต้องการจะค้นหา หรือไม่สามารถกำหนดคำที่จะใช้ค้นหาได้อย่างมีประสิทธิภาพ คำที่ใช้จึงมักมีความหมายกว้าง หรือมีความหมายกำกวมนั่นเอง

ถึงแม้เสิร์ชเอ็นจิ้นต่าง ๆ จะพยายามทำฟังก์ชันช่วยเหลือเพิ่มเติมเพื่อให้ผู้ที่ต้องการค้นหาข้อมูลได้ใช้ค้นหาข้อมูลให้ตรงตามความต้องการมากที่สุด และแสดงเฉพาะข้อมูลที่ต้องการจากการค้นหาจริง ๆ เท่านั้น เช่น การใช้เครื่องหมาย "+" เพื่อแสดงว่าต้องมีคำที่นำหน้าโดยเครื่องหมาย "+" ในผลค้นหาข้อมูล การใช้เครื่องหมาย "-" เพื่อแสดงว่าในผลค้นหาข้อมูลที่จะแสดงออกมาต้องไม่มีคำที่ถูกนำหน้าโดยเครื่องหมายลบ "-" อยู่ หรือการใช้เครื่องหมายคำพูด " " คร่อมกลุ่มของคำ เพื่อแสดงว่ากลุ่มของคำนั้น ๆ จะต้องอยู่ในผลการค้นหาข้อมูล ในลักษณะเรียงต่อเนื่องกันตามที่กำหนดในเครื่องหมายคำพูด โดยไม่แยกกันอยู่ในคนละส่วนของผลการค้นหา ซึ่งจะเป็นการใช้ประโยชน์จากความหมายของกลุ่มคำที่เกิดขึ้น ถ้ากลุ่มคำนั้นแยกกันอยู่ในผลการค้นหาอาจไม่มีความหมายเหมือนกับการที่อยู่รวมกัน รวมถึงการใช้คำ "and", "or" หรือ "not" ในกลุ่มคำที่ใช้ ค้นหาข้อมูล เพื่อแสดงการมีอยู่ทุกคำ การมีอยู่ของคำใดคำหนึ่ง หรือการที่จะต้องไม่มีคำนั้นในผลการค้นหาข้อมูลที่นำมาแสดง ตามลำดับ ตัวอย่างการใช้ฟังก์ชันช่วยเหลือร่วมกับกลุ่มคำที่ใช้ในการค้นหาข้อมูล "thailand+travel-hotel" มีความหมายในการค้นหาว่าต้องการค้นหาเว็บไซท์ที่เกี่ยวกับการท่องเที่ยว ซึ่งเกี่ยวข้องกับประเทศไทย โดยจะต้องมีคำว่า thailand และ travel อยู่ในผลการค้นหาข้อมูลทุกผล แต่ผลการค้นหาข้อมูลเหล่านี้ต้องไม่เกี่ยวกับโรงแรม คือไม่มีคำว่า "hotel" อยู่ด้วย เมื่อกำหนดคำที่ใช้ในการค้นหาดังตัวอย่างที่แสดง ก็จะได้ผลค้นหาข้อมูลซึ่งไม่มีการแสดงผลการค้นหาที่เกี่ยวกับการท่องเที่ยวประเทศไทยซึ่งมีข้อมูลโรงแรมที่พัก อันจะช่วยให้เราไม่ต้องเสียเวลาเข้าไปดูข้อมูลซึ่งไม่ต้องการเหล่านั้น เห็นได้ว่าฟังก์ชันที่มีการจัดไว้โดยเสิร์ชเอ็นจิ้นต่าง ๆ เป็นฟังก์ชันที่มีประโยชน์ต่อการค้นหาข้อมูลมาก แต่ในความเป็นจริงแล้วมีคนจำนวนน้อยเท่านั้นที่รู้ถึงวิธีการใช้ฟังก์ชันเหล่านี้อย่างถูกต้อง โดยคนส่วนใหญ่ยังใช้ไม่เป็นหรือบางคนอาจไม่รู้ด้วยซ้ำว่าในเสิร์ชเอ็นจิ้นต่าง ๆ มีฟังก์ชันที่มีประโยชน์เหล่านี้ให้ จึงมีความจำเป็นอย่างยิ่งที่จะต้องมีการพัฒนาวิธีการใหม่ ๆ เพื่อนำมาช่วยปรับปรุงการค้นหาข้อมูล หรือการแสดงผลการค้นหาข้อมูลให้มีประสิทธิภาพมากขึ้น

1.1.2 แนวทางที่น่าสนใจเพื่อใช้แก้ปัญหา

แนวทางหนึ่งที่สามารถนำมาแก้ปัญหานี้ได้ก็คือ การจัดผลการค้นหาข้อมูลให้อยู่ในรูปของกลุ่มของผลตามเนื้อหาที่ผลนั้น ๆ เกี่ยวข้อง แล้วจึงแสดงลำดับของกลุ่มของผลการค้นหาข้อมูลตามความสัมพันธ์กับคำที่ใช้ในการค้นหาข้อมูล ซึ่งจะเป็นการช่วยให้การแสดงผลการค้นหาข้อมูลนั้นมีประสิทธิภาพมากยิ่งขึ้น อันจะช่วยประหยัดเวลาของผู้ที่ค้นหาข้อมูลโดยเลือกกลุ่มของเนื้อหาที่สนใจแทนที่จะเลือกดูผลการค้นหาข้อมูลทีละผล และยังเพิ่มโอกาสที่ผลการค้นหาในลำดับหลัง ๆ จะถูกนำไปใช้ เพราะถูกนำมาแสดงในกลุ่มเดียวกับผลอื่น ๆ ที่มีเนื้อหาด้านเดียวกัน

วิธีการจัดกลุ่มเอกสาร (Document Clustering) เป็นงานวิจัยอีกแนวหนึ่งที่มีผู้สนใจมากขึ้นในปัจจุบัน ส่วนหนึ่งมาจากการที่แนวทางนี้สามารถนำไปใช้ได้กับคนหมู่มากขึ้น โดยการนำไปใช้ร่วมกับอินเตอร์เนท ทั้งนี้เนื่องจากในปัจจุบันอินเตอร์เนทเริ่มเข้าไปมีบทบาทในชีวิตประจำวันของผู้คนทั่วโลก ทำให้มีการสร้างเว็บไซท์เพื่อให้ข้อมูล และให้บริการในรูปแบบต่าง ๆ มากมาย เนื่องจากข้อมูลในอินเตอร์เนทมีจำนวนมากมายมหาศาล ทำให้คนที่ใช้อินเตอร์เนทต้องมีเครื่องมือเพื่อใช้ในการค้นหาข้อมูลที่ต้องการ แต่ข้อมูลมีจำนวนมากเกินกว่าที่เครื่องมือค้นหาข้อมูลเพียงอย่างเดียวจะแสดงผลการค้นหาได้อย่างมีประสิทธิภาพ จึงมีนักวิจัยหลายคนทำการวิจัยเพื่อนำวิธีการจัดกลุ่มเอกสารนี้มาช่วยในการจัดกลุ่มผลการค้นหาข้อมูล เพื่อให้ผลการค้นหาข้อมูลที่แสดงให้ผู้ใช้มีประสิทธิภาพมากยิ่งขึ้น ซึ่งในปัจจุบันการจัดกลุ่มเอกสาร นั้นแบ่งออกเป็น 2 แนวทางหลัก ๆ คือ การจัดกลุ่มเอกสารโดยมีการเรียนรู้จากตัวอย่างก่อน (Supervised Clustering) และการจัดกลุ่มเอกสารโดยไม่ต้องมีการเรียนรู้ก่อน (Unsupervised Clustering)

การจัดกลุ่มเอกสารโดยมีการเรียนรู้จากตัวอย่างก่อนนั้น จะต้องมีการเตรียมตัวอย่างเอกสารของกลุ่มเอกสารต่าง ๆ เพื่อใช้สอนให้ระบบการจัดกลุ่มเรียนรู้ว่า มีกลุ่มเอกสารกลุ่มใดบ้าง มีลักษณะอย่างไรถึงจะรวมอยู่ในกลุ่มนั้น ๆ ในกรณีของการจัดกลุ่มผลการค้นหาข้อมูลก็ต้องมีการทำกลุ่มตัวอย่างของเอกสารในเว็บไซท์ หรือที่เรียกว่า เว็บเพจ โดยต้องมีจำนวนเอกสารที่มากพอ รวมถึงจำนวนกลุ่มที่หลากหลายและครอบคลุมกลุ่มเอกสารทั้งหมดที่มีอยู่ในเครือข่ายใยแมงมุม ซึ่งจะเห็นได้ว่าไม่ใช่เรื่องง่ายเลยที่จะทำได้ ทั้งยังจะต้องมีคนจำนวนมากที่มีความชำนาญเพื่อมาทำกลุ่มตัวอย่างนี้ ที่สำคัญยังเป็นการยากที่จะทำให้มีความน่าเชื่อถือในความถูกต้อง และความครอบคลุมต่อกลุ่มตัวอย่างที่ทำขึ้น สิ่งสำคัญอีกอย่างคือ การที่จะต้องมีการปรับปรุงและเพิ่มข้อมูลให้กับกลุ่มตัวอย่างอย่างต่อเนื่อง เพราะว่าเอกสารในเครือข่ายใยแมงมุม มีการเพิ่มขึ้นอย่างต่อเนื่องตลอดเวลา ซึ่งเป็นไปได้ว่าจะมีเอกสารที่เป็นกลุ่มเอกสารใหม่ ๆ เกิดขึ้น จึงเป็นเรื่องจำเป็นที่จะต้องมีการปรับปรุงกลุ่มตัวอย่าง เพื่อสอนให้ระบบจัดกลุ่มเอกสารเรียนรู้ว่ามีกลุ่มเอกสารใหม่ ๆ เกิดขึ้น เพื่อจะได้จัดกลุ่มเอกสารได้อย่างถูกต้อง จะเห็นได้ว่าการจัดกลุ่มเอกสารโดยต้องมีการ

เรียนรู้จากตัวอย่างก่อน มีความยุ่งยากในหลาย ๆ ขั้นตอน ดังนั้นงานวิจัยนี้จะมุ่งเน้นที่แนวทางการจัดกลุ่มเอกสารโดยไม่ต้องมีการเรียนรู้ก่อน (Unsupervised Clustering)

จากที่ได้ทำการศึกษามา การจัดกลุ่มเอกสารในเครือข่ายใยแมงมุมที่ไม่ต้องมีการเรียนรู้ก่อน (Unsupervised Clustering) ซึ่งต่อไปจะเรียกสั้น ๆ ว่า การจัดกลุ่มเอกสาร มีอยู่ 2 แนวทางหลัก ๆ คือ การจัดกลุ่มเอกสารโดยใช้คำ และ การจัดกลุ่มเอกสารโดยใช้ URL ดังนี้

1. การจัดกลุ่มเอกสารโดยใช้คำ เป็นการนำคำที่เกี่ยวข้องกับเอกสารมาใช้บอกความสัมพันธ์ระหว่างเอกสาร เมื่อรู้ว่าเอกสารใดมีความสัมพันธ์กัน ก็จะสามารถบอกได้ว่าเอกสารเหล่านั้นสามารถถูกนำมารวมกลุ่มกันได้หรือไม่ คำที่เกี่ยวกับเอกสารนั้นมาได้จากหลายที่ เช่น อาจเป็นคำอธิบายเอกสารที่แสดงอยู่ในผลการค้นหาของเสิร์ชเอ็นจิ้น หรือคำที่อยู่ในเอกสารนั้น ๆ เป็นต้น โดยคำในเอกสารที่นำมาใช้นั้น ถ้าเป็นเอกสารทั่วไปก็มักจะใช้คำทั้งหมดที่อยู่ในเอกสาร แต่ถ้าเป็นเอกสารในเครือข่ายใยแมงมุม หรือ เอกสารในรูปแบบ HTML แล้วจะสามารถแบ่งคำเหล่านี้ออกเป็นหลายส่วน เนื่องจากโครงสร้างของเอกสารใน เครือข่ายใยแมงมุม สามารถแบ่งออกเป็นส่วน ๆ ตามแท็ก (tag) คือ แท็กหัวข้อเอกสาร (<title>...</title>), แท็กคำอธิบายเอกสาร (<meta>... </meta >) และแท็กเนื้อความเอกสาร (<body>...</body>) ซึ่งในส่วนของแท็กคำอธิบายเอกสารนั้น ยังสามารถแบ่งออกเป็นชนิดย่อย ๆ อีกหลายชนิด เช่น ชนิดคำหลัก (keyword), ชนิดคำอธิบาย (description), ชนิดผู้เขียน (author) และอื่น ๆ ซึ่งข้อความในแท็กคำอธิบายเอกสาร (<meta>) บางชนิดจะมีประโยชน์สำหรับการจัดกลุ่มเอกสาร แต่บางชนิดก็ไม่มีประโยชน์

การที่เอกสารในเครือข่ายใยแมงมุม มีโครงสร้างที่แบ่งเอกสารออกเป็นส่วน ๆ อย่างชัดเจนนี้ ทำให้เราสามารถเลือกใช้คำเฉพาะบางส่วนของเอกสาร โดยไม่จำเป็นต้องใช้คำทุกคำในเอกสารได้ เป็นข้อดีที่นำมาใช้กับการจัดกลุ่มเอกสารในเครือข่ายใยแมงมุมได้ การที่เราเลือกคำจากบางส่วนของเอกสารมาใช้ในการจัดกลุ่มเอกสาร เนื่องจากคำที่อยู่คนละส่วนของเอกสารจะให้ผลในการจัดกลุ่มเอกสารที่ต่างกัน โดยคำที่อยู่ในแท็กเนื้อความของเอกสาร (<body>... </body>) อาจเพิ่มความกำกวมให้กับการจัดกลุ่มของเอกสารนั้น ๆ ได้ เพราะคำในส่วนนี้จะมีอยู่หลากหลายไม่ได้จำกัดอยู่เฉพาะสิ่งที่เอกสารนั้นเกี่ยวข้องเท่านั้น ในขณะที่คำที่อยู่ในแท็กคำอธิบายเอกสาร (<meta>...</meta>) จะเป็นคำที่ผู้สร้างเอกสาร มีจุดประสงค์เพื่อใช้บอกข้อมูลที่เกี่ยวข้องกับเอกสาร (แต่ข้อมูลในส่วนนี้จะถูกต้องมากน้อยแค่ไหนก็ขึ้นอยู่กับความตั้งใจของผู้สร้างเอกสาร) เช่น คำที่อยู่ในแท็กคำอธิบาย ซึ่งกำหนดชนิดย่อยให้เป็นชนิดคำหลัก (keyword) จะเป็นคำที่เกี่ยวข้องกับเอกสารนั้น ๆ เท่านั้น โดยคำที่ไม่เกี่ยวข้องกับเอกสารจะไม่ถูกใส่ไว้ในส่วนนี้ การเลือกคำเฉพาะบางส่วนของเอกสาร ยังมีผลต่อความรวดเร็วในการจัดกลุ่มเอกสารด้วย เพราะการเลือกคำจากบางส่วนของเอกสารย่อมมีจำนวนคำที่น้อยกว่าการนำคำทั้งเอกสารมาใช้ หรือการใช้คำเฉพาะคำที่อยู่ในแท็กคำอธิบายเอกสาร (<meta>...</meta>) ก็มักจะมีจำนวนคำที่

น้อยกว่าคำที่อยู่ในแท็กเนื้อความเอกสาร (<body>...</body>) แม้จำนวนคำที่น้อยกว่าจะมีความรวดเร็วในการจัดกลุ่มมากกว่า แต่จำนวนคำที่น้อยก็หมายถึงข้อมูลสำหรับการจัดกลุ่มที่น้อยลงด้วย จะเห็นได้ว่าการเลือกที่จะนำคำจากส่วนใดของเอกสารมาใช้สำหรับการจัดกลุ่มเอกสารในเครือข่ายใยแมงมุม ย่อมมีผลโดยตรงต่อคุณภาพของผลการจัดกลุ่มเอกสารและความเร็วในการจัดกลุ่มเอกสาร ทั้งนี้การจะเลือกใช้คำจากส่วนใด ไม่มีข้อกำหนดที่แน่นอนว่าจะต้องใช้คำจากส่วนหนึ่งส่วนใดหรือทั้งหมด แต่ขึ้นกับกระบวนการที่นำมาใช้ในการจัดกลุ่มเอกสาร เช่น กระบวนการจัดกลุ่มเอกสารหนึ่งอาจใช้คำทุกคำในเอกสารในการจัดกลุ่มเอกสารได้ ถ้ามีวิธีที่ดีที่จะใช้ควบคุมและจัดการกับคำที่จะทำให้เกิดความกำกวมได้ เป็นต้น

2. การจัดกลุ่มเอกสารโดยใช้ URL จากงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มเอกสารซึ่งได้ทำการศึกษามา ประเภทแรกจะเป็นวิธีการจัดกลุ่มเอกสารที่นำ URL ของเอกสารมาใช้ในการจัดกลุ่ม โดยการจัดกลุ่มเอกสารนั้นจะเป็นการจัดกลุ่มตามโดเมนของเอกสาร [3] [4] [6] ตัวอย่าง URL ของเอกสาร เช่น หน้าแรกของเว็บไซท์ Yahoo! [9] จะมี URL ของหน้าเอกสารนี้เป็น www.yahoo.com หรืออีกตัวอย่างหนึ่ง หน้าแรกของเว็บไซท์คณะเทคโนโลยีสารสนเทศ สถาบันพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จะมี URL ของหน้าเอกสารนี้เป็น www.it.kmitl.ac.th เป็นต้น โดยมีความเป็นไปได้ที่เอกสารซึ่งอยู่ในโดเมนเดียวกันจะสัมพันธ์กับเนื้อหาในด้านเดียวกัน แต่แนวความคิดนี้ก็ไม่ถูกต้องเสมอไป เพราะยังมีเว็บไซท์ขนาดใหญ่อยู่จำนวนมาก ที่มีเนื้อหาซึ่งเกี่ยวข้องกับหลายเรื่องอยู่ภายในเว็บไซท์เดียว ทำให้การจัดกลุ่มตามโดเมนของเอกสารอาจได้ผลการจัดกลุ่มเอกสารที่ไม่ถูกต้องนัก โดยการจัดกลุ่มเอกสารตามโดเมนนี้ ไม่ได้คำนึงถึงความสัมพันธ์ ระหว่างเนื้อหาภายในเอกสาร กลุ่มของเอกสารที่ได้จากการจัดกลุ่มโดยวิธีการนี้จึงมีขนาดเล็ก เนื่องจากโดยส่วนมากเอกสารซึ่งเป็นผลการค้นหาข้อมูลที่มาจากโดเมนเดียวกันมีอยู่น้อย ทำให้ได้ผลการจัดกลุ่มเอกสารที่ไม่ค่อยมีประโยชน์นัก ในตารางที่ 1.1 เป็นการแสดงให้เห็นถึงจำนวนกลุ่มของเอกสารที่ได้จากจัดกลุ่มเอกสารโดยใช้ URL ของเอกสาร โดยทำการจัดกลุ่มเอกสาร 30 อันดับแรก ซึ่งเป็นผลการค้นหาข้อมูลจากเสิร์ชเอ็นจิ้นต่าง ๆ การที่ใช้เอกสารเพียง 30 อันดับแรกของผลการค้นหาข้อมูลมาแสดงข้อมูลไว้ในตาราง เนื่องจากผู้ที่ค้นหาข้อมูลส่วนใหญ่มักไม่เสียเวลาในการดูผลเกินจากอันดับที่ 30 ลงไป ในแต่ละแถวของตาราง ได้แสดงถึงจุดมุ่งหมายในการค้นหาของคำที่ใช้ค้นหาข้อมูล (จุดมุ่งหมาย), คำที่ให้เสิร์ชเอ็นจิ้นใช้ในการค้นหาข้อมูล (คำที่ใช้ค้นหา), เสิร์ชเอ็นจิ้นที่รับคำที่ใช้ค้นหาข้อมูลเพื่อนำไปค้นหาข้อมูลมาให้เรา (เสิร์ชเอ็นจิ้นที่ใช้), จำนวนเอกสารที่เสิร์ชเอ็นจิ้นสามารถค้นพบจากการค้นหาข้อมูล (จำนวนเอกสารที่พบ), จำนวนเอกสารที่ตรงกับจุดมุ่งหมายในการค้นหาข้อมูลของเราในผลการค้นหาข้อมูล 30 อันดับแรก (จำนวนเอกสารที่ตรงกับจุดมุ่งหมายใน 30 เอกสารแรก), อันดับของผลการค้นหาข้อมูลอันดับแรก

ตารางที่ 1.1 ตัวอย่างคำที่ใช้ในการค้นหาข้อมูลและผลที่ได้จากการค้นหาข้อมูล

จุดมุ่งหมาย	คำที่ใช้ค้นหาข้อมูล	เสิร์ชเอ็นจิ้นที่ใช้	จำนวนเอกสารที่พบ	จำนวนเอกสารที่ตรงกับจุดมุ่งหมายใน 30 เอกสารแรก	อันดับของผลการค้นหาอันดับแรกที่ตรงกับจุดมุ่งหมาย	จำนวนกลุ่มของเอกสารที่มาจากโดเมนเดียวกันใน 30 เอกสารแรก
ข้อมูลเว็บไซท์ของไมโครซอฟท์วินโดวส์	microsoft windows	Alta Vista	13,155,684	3	อันดับที่ 3	26
หาข้อมูลการลักพาตัวโดยเอเลียน	alien	Alltheweb	8,601,785	6	อันดับที่ 14	30
หาข้อมูลเพื่อซื้อไม้ตีเทนนิส	tennis	Google	20,900,000	2	อันดับที่ 26	30
หาเว็บไซท์ที่เกี่ยวกับนางแบบหรือนายแบบ	Model	Hotbot	20,861,602	3	อันดับที่ 6	29

สุดที่ตรงกับจุดมุ่งหมายในการค้นหาข้อมูล (อันดับของผลการค้นหาอันดับแรกที่ตรงกับจุดมุ่งหมาย), และจำนวนกลุ่มของเอกสารที่มาจากโดเมนเดียวกันใน 30 เอกสารแรก ในคอลัมน์สุดท้ายจะแสดงให้เห็นถึงความจริงที่ว่า จำนวนกลุ่มของเอกสารที่มาจากโดเมนเดียวกันมีอยู่น้อยมาก ทำให้ได้ผลซึ่งเป็นกลุ่มเอกสารจำนวนมาก จึงเป็นสิ่งที่ยืนยันได้ว่าการจัดกลุ่มเอกสารตามโดเมนของเอกสารโดยใช้ URL ของเอกสารนั้นยังไม่มีประสิทธิภาพเท่าที่ควร จากตารางที่ 1.1 ยังมีหลายสิ่งที่ควรจะกล่าวถึง ประการแรก จำนวนเอกสารที่พบดังแสดงในคอลัมน์ที่ 4 ย่อมเป็นสิ่งยืนยันถึงสิ่งที่ได้กล่าวไปแล้วถึงจำนวนผลการค้นหาที่มากทำให้ไม่สามารถดูผลที่อยู่ในลำดับหลัง ๆ ได้ ประการที่สอง จำนวนเอกสารที่ตรงกับจุดมุ่งหมายในผลการค้นหา 30 อันดับแรก (คอลัมน์ที่ 5) ซึ่งมีอยู่น้อยจะทำให้เราเสียเวลาถ้าเราต้องเข้าไปดูในแต่ละผลเพื่อครวจสอบว่าเกี่ยวข้องกับสิ่งที่เรากำลังค้นหาหรือไม่ โดยถ้าเราใช้เวลา 10 วินาทีต่อ 1 เอกสาร การดูเอกสาร 30 เอกสาร ต้องใช้เวลา 300 วินาที ซึ่งเท่ากับ 5 นาที โดยส่วนใหญ่เราก็จะได้เอกสารที่เกี่ยวข้องกับสิ่งที่เราต้องการค้นหาไม่มากนัก เป็นการเสียเวลา 5 นาทีที่ไม่คุ้มค่านัก ประการที่สาม อันดับของผลการค้นหาอันดับแรกสุดที่ตรงกับจุดมุ่งหมายในการค้นหาข้อมูล (คอลัมน์ที่ 6) ซึ่งจะเห็นได้ว่าจากตารางผลส่วนใหญ่จะไม่อยู่ในอันดับแรก ทำให้เสียเวลาในการหาผลการค้นหาที่ต้องการ แม้จะเป็นเพียงผลในอันดับแรก ซึ่งผลในอันดับแรกนั้นก็อาจไม่มีข้อมูลที่ต้องการมากนัก ทำให้ต้องเสียเวลาเพิ่มขึ้นในการหาผลต่อไปที่ตรงกับจุดมุ่งหมายที่ต้องการค้นหา

ส่วนประเภทที่สอง จะเป็นการจัดกลุ่มเอกสารโดยใช้ URL ของเอกสาร และ URL ในเอกสารมาร่วมกันในการจัดกลุ่มเอกสาร [7] โดยมีสมมุติฐานว่าถ้าเอกสาร 2 เอกสารถูกลิงค์ (Link) มาโดยเอกสารหนึ่ง ๆ แสดงว่าเอกสาร 2 เอกสารนั้นน่าจะเป็นเอกสารที่มีเนื้อหาในด้านเดียวกัน และในขณะเดียวกันการที่เอกสาร 2 เอกสารลิงค์ไปหาเอกสารเดียวกันก็แสดงว่าเอกสารทั้ง 2 นั้นก็น่าจะเป็นเอกสารที่มีเนื้อหาในด้านเดียวกันเช่นกัน แต่มีเงื่อนไขว่าเอกสารที่ลิงค์เข้ามาและเอกสารที่ถูกลิงค์ไปหานั้นจะต้องไม่อยู่ในโดเมนเดียวกับเอกสารที่กำลังพิจารณาอยู่ เนื่องจากต้องการจะหลีกเลี่ยงลิงค์ที่นำมาใช้เพื่อจุดประสงค์ในการนำทางไปสู่เอกสารอื่น ๆ ในโดเมนเดียวกัน (Navigational Link) จะเห็นว่าแนวความคิดนี้เป็นแนวความคิดที่เหมือนกับแนวความคิดใน [5] ซึ่งเรียกเอกสารที่ ถูกลิงค์มาหาและเอกสารที่มีลิงค์ไปหาเอกสารอื่นว่า Authority และ Hub ตามลำดับ ในการจัดกลุ่มเอกสารของ [7] จะต้องมีข้อมูล URL การลิงค์เข้าและลิงค์ออกของแต่ละเอกสาร ทั้งนี้ URL เหล่านี้จะมาจากเอกสารใด ๆ ก็ได้ไม่มีการจำกัดว่าจะต้องมาจากเอกสารที่อยู่ในกลุ่มที่จะทำการจัดกลุ่มเท่านั้น เพื่อนำข้อมูลเหล่านั้นมาใช้คำนวณหาความสัมพันธ์ระหว่างเอกสาร โดยนำ K-Means มาใช้เป็นกระบวนการในการจัดกลุ่มเอกสาร แต่มีการเปลี่ยนแปลงกระบวนการบางส่วนเพื่อปรับปรุงข้อด้อยที่ฟังก์ชัน K-Means เดิมมีอยู่ ในส่วนของการหาความคล้ายระหว่างเอกสารกับกลุ่มเอกสารนั้นใช้ฟังก์ชันการวัดค่า Cosine ในการคำนวณ ผลที่ได้จาก

การจัดกลุ่มเอกสารด้วยวิธีการนี้ จะได้กลุ่มเอกสารซึ่งเอกสารหนึ่ง ๆ สามารถอยู่ในกลุ่มเอกสารได้มากกว่าหนึ่งกลุ่ม เนื่องจากงานวิจัยนี้มองว่าเอกสารหนึ่ง ๆ สามารถมีได้หลายเนื้อหาอยู่ภายในหน้าเอกสาร

จากผลของการจัดกลุ่มเอกสารโดยใช้ URL ของเอกสารซึ่งเป็นการจัดกลุ่มตามโดเมนของเอกสารที่ได้นำเสนอไป จึงมีแนวคิดที่จะใช้ URL ของเอกสารร่วมกับ URL ในเอกสารมาใช้ในการจัดกลุ่มเอกสาร ทั้งนี้เอกสารจะอยู่ในเว็บเพจเดียวกันหรือไม่ก็ได้ หลักการคือ การใช้เอกสารที่มีคำที่ใช้ในการค้นหาข้อมูลอยู่ในเอกสารเหมือนกัน แล้วนำ URL มาใช้คำนวณหาความสัมพันธ์ระหว่างเอกสารเพื่อหาว่าเอกสารใดมีความสัมพันธ์กันมากพอที่จะรวมเป็นกลุ่มเอกสารเดียวกันได้ แนวคิดในการนำ URL ในเอกสารและ URL ของเอกสารมาใช้ในการจัดกลุ่มเอกสารคือ ถ้าเอกสาร 2 เอกสารมี URL ที่ลิงค์ถึงกัน หรือเอกสารใดเอกสารหนึ่งมี URL ที่ลิงค์ไปยังเอกสารหนึ่ง หรือเอกสาร 2 เอกสารมี URL ที่ชี้ไปยังเอกสารอื่นเหมือนกัน แสดงว่าเอกสารนั้นน่าจะมีเนื้อหาที่เกี่ยวข้องในด้านเดียวกัน ตัวอย่างเช่น A และ B เป็นเอกสารที่ได้จากผลค้นหาข้อมูลเดียวกัน ถ้า A มีลิงค์ไปยัง B หรือ B มีลิงค์ไปยัง A แสดงว่า A และ B น่าจะมีความสัมพันธ์กัน หรือ A และ B มีลิงค์ไปยัง C เหมือนกันโดยที่ C จะอยู่หรือไม่อยู่ในผลการค้นหาเดียวกันกับ A และ B ก็ได้ จะแสดงว่า A และ B น่าจะมีความสัมพันธ์กันได้เช่นกัน จะเห็นได้ว่า URL ในเอกสารสามารถทำให้การรวมกลุ่มเอกสารมีความถูกต้องและครอบคลุมได้มากขึ้น แต่จากการทดลองใช้ URL ในเอกสารและ URL ของเอกสาร เพื่อทำการจัดกลุ่มเอกสาร จะได้ผลการจัดกลุ่มเอกสารซึ่งมีเอกสารที่ถูกจัดกลุ่มไม่มากนัก จากการสังเกตข้อมูลที่ทำการจัดกลุ่มจึงทราบว่า URL ที่เชื่อมโยงระหว่างเอกสารนั้น เชื่อมโยงเอกสารจำนวนไม่เกิน 50% ของเอกสารทั้งหมด จึงทำให้เหลือเอกสารที่ไม่ถูกจัดกลุ่มไม่น้อยกว่า 50% ดังนั้นเพื่อให้การจัดกลุ่มมีประสิทธิภาพมากยิ่งขึ้นจึงนำคำที่อยู่ในเอกสารมาใช้ร่วมกับ URL ในเอกสารและ URL ของเอกสาร ทำให้การจัดกลุ่มเอกสารครอบคลุมเอกสารส่วนใหญ่ที่ต้องการจัดกลุ่มได้

1.1.3 บทสรุป

จากปัญหาต่าง ๆ ในระบบการค้นหาข้อมูลในเครือข่ายใยแมงมุมที่ได้กล่าวมาข้างต้นและแนวทางการแก้ปัญหาที่ได้นำเสนอไป จึงทำให้เกิดความสนใจในการพยายามที่จะทำการวิจัยเพื่อปรับปรุงการค้นหาข้อมูลในเครือข่ายใยแมงมุม ซึ่งหนทางหนึ่งที่สามารถนำมาแก้ปัญหานี้ได้ก็คือ การจัดผลการค้นหาข้อมูลให้อยู่ในรูปของกลุ่มของผลการค้นหาข้อมูลตามเนื้อหาที่ผลนั้น ๆ เกี่ยวข้อง แล้วจึงแสดงลำดับของกลุ่มของผลการค้นหาข้อมูลตามความสัมพันธ์กับคำที่ใช้ในการค้นหาข้อมูล ซึ่งจะเป็นการช่วยให้การแสดงผลการค้นหาข้อมูลนั้นมีประสิทธิภาพมากยิ่งขึ้น อันจะช่วยประหยัดเวลาของผู้ที่ค้นหาข้อมูลโดยให้เลือกกลุ่มของเนื้อหาที่สนใจ แทนที่จะเลือกดูทีละผลการค้นหาข้อมูล และยังเพิ่มโอกาสที่ผลการค้นหาข้อมูลในลำดับหลัง ๆ จะถูกนำไปใช้ เพราะถูก

นำมาแสดงในกลุ่มเดียวกับผลอื่น ๆ ที่มีเนื้อหาด้านเดียวกัน ในขณะที่จำนวนผู้ใช้อินเตอร์เนททั่วโลกมีอยู่หลายร้อยล้านคน และจำนวนผู้เริ่มใช้อินเตอร์เนทก็เพิ่มมากขึ้นทุกวัน จุดประสงค์หลักของทุกคนที่เข้ามาใช้อินเตอร์เนทก็คือ ข้อมูลอันมหาศาลที่อยู่ในอินเตอร์เนท ดังนั้นการพัฒนาระบบค้นหาข้อมูลในอินเตอร์เนทซึ่งเป็นสิ่งจำเป็นสำหรับผู้ที่ต้องการใช้ข้อมูลในอินเตอร์เนท ก็จะยังประโยชน์ให้แก่ผู้ใช้อินเตอร์เนททั่วโลก

ในส่วนต่าง ๆ ของวิทยานิพนธ์เล่มนี้ เมื่อกล่าวถึงผลการค้นหาข้อมูล จะหมายถึงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น และจะใช้คำว่า เอกสาร หรือ เว็บเพจ สลับกันไปเมื่อต้องการจะกล่าวถึงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น

1.2 จุดมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อศึกษา รูปแบบ ลักษณะการทำงานและการแสดงผลของระบบค้นหาข้อมูลในเครือข่ายใยแมงมุม เพื่อให้เกิดความเข้าใจ และรู้ถึงจุดเด่น จุดด้อย ของระบบที่มีอยู่ในปัจจุบัน ซึ่งจะนำไปสู่แนวทางในการแก้ปัญหาและการพัฒนาให้ดียิ่งขึ้น

2. ศึกษาลักษณะการทำงานของการจัดกลุ่มเอกสาร (Clustering) วิธีต่าง ๆ เนื่องจากได้วิเคราะห์แล้วว่า การจัดกลุ่มเอกสารจะสามารถนำมาใช้แก้ปัญหาที่เป็นวัตถุประสงค์ของงานวิจัย การเข้าใจถึงวิธีการจัดกลุ่มเอกสารที่มีอยู่ จะเป็นแนวทางในการพัฒนาวิธีการจัดกลุ่มเอกสารที่มีประสิทธิภาพได้

3. ศึกษาแนวทางการเพิ่มประสิทธิภาพการจัดกลุ่มเอกสาร เพื่อวิจัยหาวิธีพัฒนาการจัดกลุ่มเอกสารวิธีใหม่ที่มีประสิทธิภาพ และสามารถนำไปงานได้จริง

4. พัฒนาระบบการจัดกลุ่มเอกสารให้มีประสิทธิภาพมากขึ้น โดยนำแนวทางที่ได้ทำการวิจัยและศึกษา มาพัฒนาเพื่อใช้ทดลองการจัดกลุ่มให้เห็นว่าผลการจัดกลุ่มที่ได้มีประสิทธิภาพดีขึ้นจริง และเปรียบเทียบประสิทธิภาพของวิธีที่พัฒนาขึ้นกับวิธีที่มีอยู่

1.3 แนวทางการศึกษา

แนวทางการศึกษาในงานวิจัยนี้คือการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำในเอกสาร เป็นวิธีการที่มีประสิทธิภาพและสามารถใช้ปรับปรุงรูปแบบ การแสดงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น ทำให้ผู้ใช้ผลการค้นหาข้อมูลเกิดความสะดวกในการเลือกดูผลการค้นหาข้อมูลและได้ผลที่ต้องการอย่างรวดเร็ว เป็นการสนองตอบวัตถุประสงค์หลักของงานวิจัยนี้ได้เป็นอย่างดี จากการศึกษางานวิจัยที่ทำการจัดกลุ่มเอกสาร [2] [3] [4] และ [7] ทำให้เห็นแนวโน้มที่เป็นไปได้ในการนำการจัดกลุ่มเอกสารมาปรับปรุงการแสดงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น โดยเฉพาะ [2] ซึ่งใช้คำในเอกสารเป็นข้อมูลในการจัดกลุ่ม โดยมีอัลกอริทึมในการจัดกลุ่มที่เรียกว่า Hierarchical

Agglomerative Clustering (HAC) และใช้ฟังก์ชัน Global Quality Function เป็นเงื่อนไขในการจัดกลุ่ม จากการศึกษาพบว่าอัลกอริทึม HAC เป็นอัลกอริทึมสำหรับการจัดกลุ่มที่มีคุณภาพ จึงนำมาใช้เป็นอัลกอริทึมสำหรับการจัดกลุ่มในงานวิจัยนี้ ในการศึกษางานวิจัย [5] และ [7] ซึ่งใช้ประโยชน์จากลิงค์ หรือ URL ระหว่างเอกสารในการจัดลำดับหรือจัดกลุ่มเอกสาร ทำให้เห็นแนวทางในการนำลิงค์และคำในเอกสารมาใช้คำนวณหาความคล้ายระหว่างเอกสาร เพื่อทำการจัดกลุ่มเอกสาร

1.4 ขอบเขตการวิจัย

เป็นการวิจัยเพื่อปรับปรุงการแสดงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น แต่ไม่รวมถึงระบบการค้นหาข้อมูลใน WWW ของเสิร์ชเอ็นจิ้น โดยทำการจัดกลุ่มผลการค้นหาข้อมูลก่อนที่จะแสดงให้ผู้คนหาข้อมูลดู ผลการค้นหาข้อมูลในที่นี้จะได้มาจากเสิร์ชเอ็นจิ้นใดก็ได้ ที่แสดงผลการค้นหาข้อมูลในรูปของลำดับของผลการค้นหาข้อมูล การจัดกลุ่มนี้จะได้ผลดีเมื่อใช้กับเอกสารในรูปแบบ HTML ซึ่งเป็นเอกสารในเครือข่ายใยแมงมุม ถ้านำมาใช้จัดกลุ่มเอกสารที่เป็นข้อความทั่วไปที่ไม่มีลิงค์อยู่ภายในเอกสารก็สามารถทำการจัดกลุ่มได้โดยใช้คำเพียงอย่างเดียว แต่ก็จะไม่ได้ใช้ประสิทธิของระบบจัดกลุ่มอย่างเต็มที่ เนื่องจากระบบจัดกลุ่มจะใช้ประโยชน์จากโครงสร้างของเอกสารในรูปแบบ HTML โดยใช้คำที่อยู่ใน แท็กหัวข้อเอกสาร (<title>...</title>) และแท็กคำอธิบายเอกสาร (<meta>...</meta>) ร่วมกับลิงค์ที่อยู่ในเอกสาร เพื่อเป็นข้อมูลในการคำนวณหาความคล้ายระหว่างเอกสารในขณะที่ทำการจัดกลุ่มเอกสาร ดังนั้นระบบนี้จะสามารถทำการจัดกลุ่มได้ ก็ต่อเมื่อเอกสารนั้นมีคำในแท็กหัวข้อเอกสารหรือแท็กคำอธิบายเอกสาร หรือมีลิงค์ในเอกสารเท่านั้น ข้อจำกัดอีกประการหนึ่งในงานวิจัยนี้ก็คือ ระบบที่พัฒนาขึ้นจะใช้จัดกลุ่มโดยใช้*คำที่เป็นภาษาไทย*ไม่ได้ เนื่องจากไม่ได้พัฒนาฟังก์ชันการตัดคำที่สามารถตัดคำภาษาไทยได้ แต่ระบบจัดกลุ่มที่พัฒนาขึ้นนี้จะสามารถจัดกลุ่มเอกสารได้*ทุกภาษา*โดยใช้ลิงค์ที่อยู่ในเอกสาร ดังนั้นถ้าเอกสารเหล่านั้นมีลิงค์ในเอกสารก็มีโอกาสที่จะถูกจัดกลุ่มได้

1.5 ขั้นตอนของการศึกษา

1. ศึกษาลักษณะการทำงานและการแสดงผลของเสิร์ชเอ็นจิ้น (Search Engine)
2. ศึกษาบทความและผลงานวิจัยต่าง ๆ ที่มีความเกี่ยวข้องกับการค้นหาข้อมูล เพื่อหาแนวทางในการปรับปรุงการค้นหาข้อมูล
3. ศึกษาบทความและผลงานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มเอกสาร
4. ทดลองเขียนโปรแกรมการจัดกลุ่มเอกสารตามผลงานวิจัยที่ศึกษาเพื่อให้เห็นถึงข้อดีและข้อเสีย

5. เขียนโปรแกรมการจัดกลุ่มโดยใช้ลิงค์เป็นข้อมูลเพื่อใช้ในการจัดกลุ่ม แล้วเปรียบเทียบผลลัพธ์ที่ได้

6 เขียนโปรแกรมการจัดกลุ่มโดยใช้ลิงค์และคำเป็นข้อมูลเพื่อใช้ในการจัดกลุ่ม แล้วเปรียบเทียบผลลัพธ์ที่ได้

7 สรุปผลการดำเนินการ และรวบรวมนำจัดทำเอกสารนำเสนอเป็นงานวิจัย

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 Fast and Intuitive Clustering of Web Documents [2]

เป็นวิธีในการจัดกลุ่มผลที่ได้จากการสืบค้นของเสิร์ชเอ็นจิ้น ก่อนที่จะแสดงให้ผู้ใช้ดู ซึ่งจะช่วยให้ผู้ใช้สามารถดูผลจากการสืบค้นได้ง่ายขึ้น โดยจะต้องมีความรวดเร็วพอที่จะจัดกลุ่มข้อมูลจำนวนมากได้ โดยวิธีการจัดกลุ่มนั้นจะใช้คำที่เป็นตัวแทนของแต่ละเอกสารมาสร้างความสัมพันธ์เพื่อใช้ในการจัดกลุ่ม ซึ่งคำที่เป็นตัวแทนของเอกสารนี้ได้มาจากคำอธิบายสั้น ๆ ในผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น และการจัดกลุ่มโดยวิธีการนี้จะแบ่งออกเป็น 2 วิธีตามการใช้คำเพื่อใช้ในการจัดกลุ่ม คือ

1. Word-Intersection Clustering (Word-IC) จะทำการจัดกลุ่มเอกสารโดยใช้กลุ่มของคำที่ใช้ร่วมกันในทุกเอกสารของกลุ่มเอกสาร ซึ่งจะนำอัลกอริทึมสำหรับการจัดกลุ่มเอกสารที่เรียกว่า Hierarchical Agglomerative Clustering (HAC) [2] มาประยุกต์ใช้ โดย HAC จะเริ่มจากการให้แต่ละเอกสาร เป็นกลุ่มของเอกสาร 1 กลุ่ม แล้วทำการรวมกลุ่มเอกสาร ที่มีเนื้อความใกล้เคียงกันที่สุด เป็นกลุ่มเดียวกันไปเรื่อย ๆ จนกระทั่งถึงเงื่อนไขที่ต้องการ ซึ่งเรียกว่า Halting Criterion โดย Pseudo code ของ HAC เป็นดังรูปที่ 2.1

```
Initialize all documents as singleton cluster
        Until (Halting Criterion) do {
        Find two most similar cluster.
        Merge them
        }
```

รูปที่ 2.1 Pseudo code ของอัลกอริทึม Hierarchical Agglomerative Clustering

การหาความคล้ายระหว่างเอกสารของวิธีการนี้ ใช้วิธีการที่เรียกว่า Group Average ซึ่งเป็นความคล้ายเฉลี่ยระหว่าง 2 กลุ่มเอกสาร เนื่องจากเป็นวิธีที่มีคุณภาพในเวลาที่รวดเร็ว ในส่วนของเงื่อนไข หรือ Halting Criterion นั้นจะใช้ฟังก์ชันที่เรียกว่า Global Quality Function (GQF) ดังนี้

$$GQF(c) = \frac{f(C)}{g(|C|)} \sum_{c \in C} S(c) \qquad (2.1)$$

- จากฟังก์ชันนี้ f(c) คือ ฟังก์ชันที่แสดงจำนวนของ cluster ทั้งหมดใน c

- ส่วน g(|c|) เป็นฟังก์ชันที่เพิ่มค่าตามจำนวนของ cluster ที่มีขนาดของตั้งแต่ 2 ขึ้นไป โดยจะมีการเพิ่มในรูปแบบ square root

- $\sum_{c \in C} S(c)$ เป็นผลรวมคะแนนของ cluster ทุก ๆ cluster

- โดย s(c) เป็นคะแนนของแต่ละ cluster ซึ่งคำนวณได้จาก

$$S(c) = |c| \cdot \frac{1 - e^{-\beta h(c)}}{1 + e^{-\beta h(c)}} \qquad (2.2)$$

- | c | เป็นจำนวนเอกสารในกลุ่ม

- h(c) เป็นจำนวนคำที่ทุก ๆ เอกสารใน cluster มีเหมือนกัน

- β เป็นความชันของ SIGMOID FUNCTION ซึ่งจะกำหนดค่าที่ดีที่สุดของขนาดและความสัมพันธ์ในแต่ละ cluster

ดังนั้น Pseudo code ของ Word-IC โดยใช้ HAC และ GQF จะเป็นดังรูปที่ 2.2

```
Initialize all documents as singleton cluster
Until (GQF cannot be increased) do {
    Find two clusters whose merge increase GQF the most.
    Merge them.
    }
```

รูปที่ 2.2 Pseudo code ของอัลกอริทึม Hierarchical Agglomerative Clustering ร่วมกับ Gobal Quality Function

2. Phase-Intersection Clustering (Phase-IC) จะทำการจัดกลุ่มเอกสารโดยใช้วลีที่ยาวที่สุดที่ใช้ร่วมกันในทุกเอกสารของกลุ่มเอกสาร เนื่องจากวลีจะให้ข้อมูลด้านความหมายที่ดีกว่าคำหลายคำที่แยกกัน Phase-IC จะมีอยู่ 2 วิธี คือ Phase-IC ที่ใช้ GQF ซึ่งจะมีกระบวนการทุกอย่างเหมือนกับ Word-IC เพียงแต่เปลี่ยนการเทียบจากการใช้กลุ่มคำมาเป็นวลีเพื่อใช้ในการ

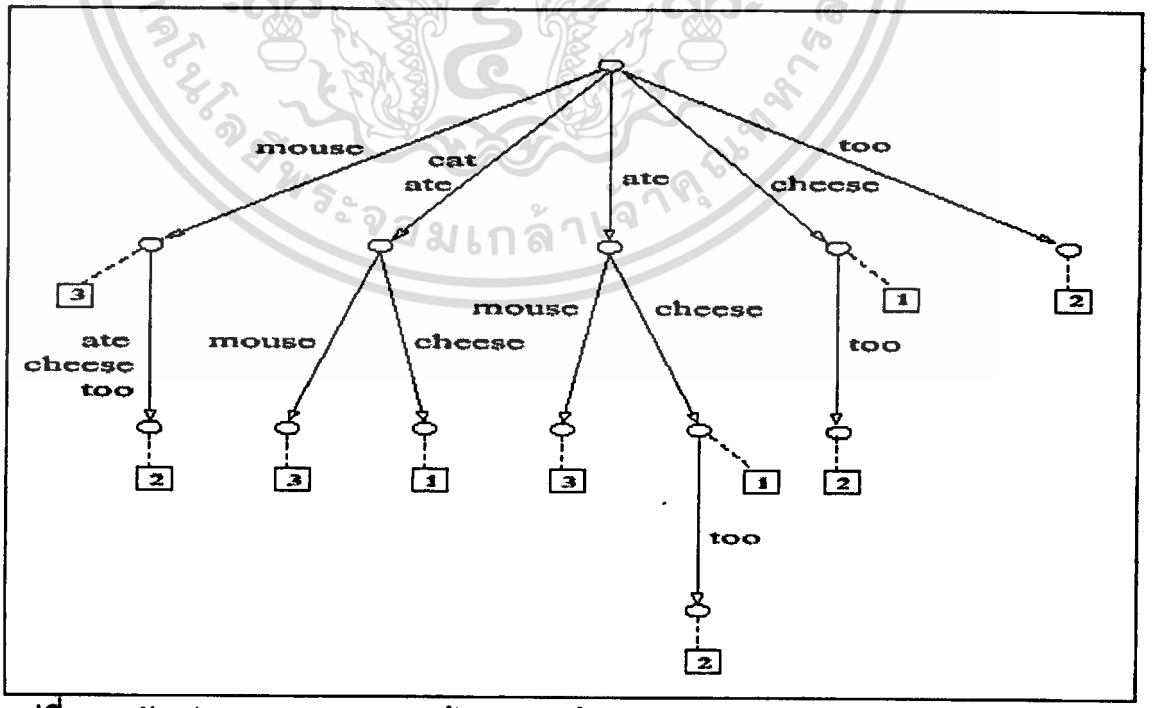
จัดกลุ่มเอกสาร และอีกวิธีคือ Phase-IC ที่ใช้ Suffix Trees โดย Suffix Trees จะประกอบด้วย edge และ node โดย edge จะต้องมีการกำหนด label ซึ่งเป็นคำที่มาจากวลี ซึ่ง edge ที่มาจาก node เดียวกันจะมี label ต่างกันและ label ของ node จะได้มาจาก label ของ edge จาก root มาที่ node ของ suffix trees ดังรูปที่ 2.3 โดยจะมีการให้คะแนนซึ่งหาได้จากฟังก์ชัน

$$S(n) = |n| \cdot \sum_{w \in W} TFIDF(W) \tag{2.3}$$

โดย n คือ node ต่าง ๆ |n| คือ จำนวนเอกสารใน node n, W เป็นกลุ่มของคำในวลีที่เป็น label ของ node n และ

$$TFIDF(w) = TF(w)/DF(w) \tag{2.4}$$

ซึ่ง TF(w) คือ ความน่าจะเป็นที่จะเกิดคำ w จากคำทั้งหมด และ DF(w) คือ ค่าความน่าจะเป็นที่คำ w จะอยู่ในเอกสารต่าง ๆ จากเอกสารทั้งหมด โดย Phase-IC ที่ใช้ Suffix Trees นี้จะทำการจัดกลุ่มเอกสารโดยการสร้าง treeซึ่งจะเป็นการจัดกลุ่มเอกสารในขณะเดียวกัน โดยวลีที่ใช้แทนกลุ่มเอกสารจะได้จาก label ของ edge ทั้งหมดจาก root มาที่ node ที่แทนกลุ่มเอกสารนั้น ๆ

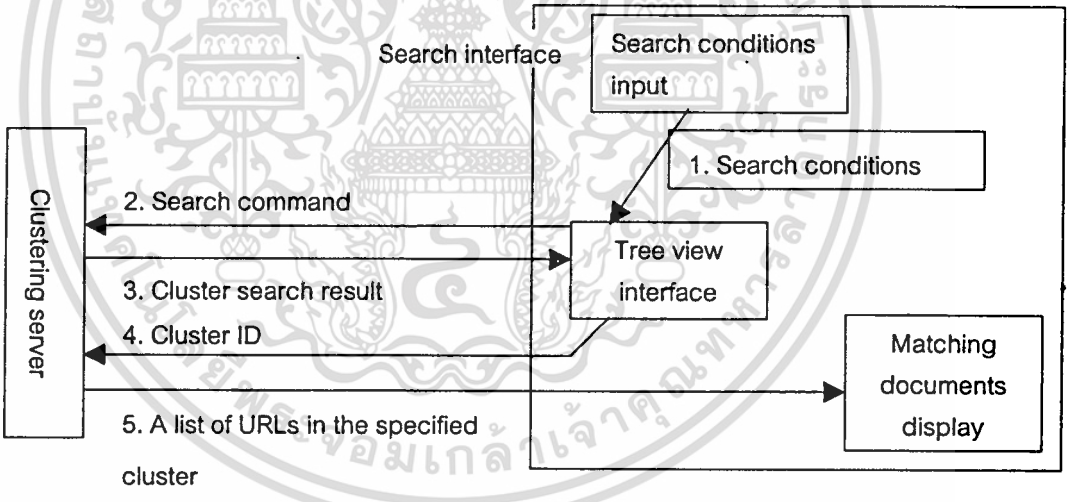


รูปที่ 2.3 ตัวอย่าง Suffix Trees ของข้อความ 3 ข้อความ : (1) "cat ate cheese", (2) "mouse ate cheese too", และ (3) "cat ate mouse"

จากผลของการทดลอง Phase-IC ที่ใช้ Suffix Trees จะมีความรวดเร็วในการจัดกลุ่มข้อมูลมากที่สุด และ Phase-IC ที่ใช้ GQF จะให้คุณภาพของการจัดกลุ่มของข้อมูลดีที่สุด แต่การจัดกลุ่มโดยวิธีนี้ใช้ snippet (คำอธิบายสั้น ๆ ที่ search engine แสดงเพื่อเป็นคำอธิบายของผลการค้นหา เมื่อแสดงผลของการค้นหา) ของผลการค้นหามาคำนวณหาความคล้ายระหว่างเอกสาร ในบางครั้ง snippet ของผลที่มาจากเสิร์ชเอ็นจิ้นอาจไม่สามารถอธิบายได้ดีนัก (มีจุดประสงค์เพื่อให้คนดู มากกว่าให้คอมพิวเตอร์ใช้ในการจัดกลุ่ม) หรือในผลการค้นหาบางผลอาจไม่มี snippet ให้ ทำให้ไม่สามารถจัดกลุ่มได้

2.2 A Domain Cluster Interface for WWW Search [4]

ทำการจัดกลุ่มผมการค้นหาข้อมูล โดยกลุ่มของข้อมูลที่จัดกลุ่มนี้จะจัดตาม ชื่อขององค์กร (Organization name) ของเอกสารนั้น โดยมีสมมุติฐานว่า ชื่อองค์กรจะช่วยให้ผู้ใช้แยกแยะได้ว่าเว็บไซท์นั้นเกี่ยวข้องกับอะไร ซึ่งจะแสดงผลในรูปของต้นไม้ (tree) ดังนั้นวิธีการนี้จึงเป็นวิธีการหนึ่งที่ใช้ประโยชน์จาก URL ของเอกสารมาช่วยในการจัดกลุ่ม โดยมีกระบวนการดังรูปที่ 2.4



รูปที่ 2.4 Data flow in the search interface

1. คำสั่งการค้นหาจะสั่งจาก search interface ไปยัง Keyword search module แล้ว Keyword search module จะติดต่อกับฐานข้อมูลที่เก็บข้อมูลของเว็บ เพื่อทำการค้นหา
2. ตาม keyword และทำการจัดเรียงตาม URL domain name

3. Search result clustering module จะทำการแปลง URL domain name เป็นชื่อองค์กร (Organization name) และทำการจัดกลุ่มเอกสารตาม ชื่อองค์กร แล้วจึงส่งไปแสดงผลให้ผู้ใช้
4. ผู้ใช้เลือกกลุ่มขององค์กรที่ต้องการ แล้วระบบจึงส่งหมายเลขประจำกลุ่ม (Cluster-ID) ไปยัง keyword search module
5. ระบบจะส่ง URL ในกลุ่มที่กำหนดไปยังผู้ใช้

เมื่อสิ้นสุดกระบวนการจะเห็นว่า ผู้ใช้จะได้รับผลการจัดแยกกลุ่มตามชื่อขององค์กรที่เลือก ซึ่งวิธีนี้จะได้ผลดีก็ต่อเมื่อ ผู้ใช้รู้ว่าชื่อขององค์กรนั้น ๆ เกี่ยวข้องกับคำที่ใช้ค้นหา (keyword) ในด้านใด หรือมีความรู้เกี่ยวกับองค์กรนั้น ๆ นั่นเอง ซึ่งในความเป็นจริงแล้วจะเป็นไปได้ยาก ที่จะนำมาใช้ในเครือข่ายใยแมงมุมเพราะองค์กรในเครือข่ายใยแมงมุมนั้นมีจำนวนมหาศาล ทำให้ผู้ใช้ไม่รู้จักองค์กรส่วนใหญ่ที่มีอยู่ และด้วยจำนวนขององค์กรที่มีจำนวนมากทำให้การเก็บฐานข้อมูลขององค์กรเพื่อนำมาใช้เทียบกับ URL domain name นั้นครอบคลุมองค์กรได้จำนวนไม่มากนัก จะเห็นว่าวิธีการนี้เหมาะที่จะนำไปใช้กับบริษัทขนาดใหญ่ที่มีหน่วยงานหลายหน่วย มากกว่าที่จะนำมาใช้กับเครือข่ายใยแมงมุมที่มีจำนวนข้อมูลมากมายมหาศาล และการรวมกลุ่มของเอกสารที่อยู่ในองค์กรเดียวกันนั้น ก็อาจทำให้เอกสารที่เกี่ยวกับคำที่ใช้ค้นหา (keyword) ในหลาย ๆ แง่มุมมารวมกันได้ เพราะเอกสารในองค์กรเดียวกันอาจเกี่ยวข้องกับคำที่ใช้ค้นหา (keyword) หนึ่ง ๆ ในแง่มุมที่ต่างกันได้ และยังเปลี่ยนแปลงได้ตลอดเวลา

2.3 On Ranking and Organizing Web Query Results [3]

เป็นอีกแนวทางหนึ่งในการจัดกลุ่มเอกสาร โดยใช้ URL ของเอกสารมาช่วยในการจัดกลุ่ม ซึ่งได้เสนอแนวทางการจัดกลุ่มเป็น 2 แนวทางหลัก ๆ คือ

1. จัดกลุ่มตามชื่อโดเมน (domain name) การจัดกลุ่มตามชื่อโดเมนนี้จะเป็นการรวมเอกสารที่อยู่ในชื่อโดเมนเดียวกันเข้าด้วยกัน ซึ่งเป็นวิธีที่ง่ายแต่มีข้อเสีย ในกรณีที่ชื่อโดเมนนั้นเป็นเว็บไซท์ที่มีขนาดใหญ่ เช่น Geocities [9] ซึ่งจะไม่เกิดประโยชน์กับผู้ใช้ เพราะ Geocities เป็นผู้ให้บริการใช้พื้นที่ เพื่อนำพื้นที่นั้นไปจัดสร้างโฮมเพจของสมาชิกแต่ละคน ซึ่งสมาชิกของ Geocities แต่ละคนมีสิทธิที่จะสร้างโฮมเพจที่มีเนื้อหาเกี่ยวข้องกับเรื่องใดก็ได้ตามต้องการ ดังนั้นในเว็บของ Geocities จึงมีข้อมูลที่เกี่ยวข้องกับด้านต่าง ๆ อยู่ภายในอย่างมากมาย การจะนำข้อมูลด้านต่าง ๆ มารวมกันเพราะอยู่ในเว็บไซท์เดียวกันจึงเป็นการกระทำที่ไม่สมควร จึงได้ปรับปรุงการจัดกลุ่มตาม ชื่อโดเมนเป็นการจัดกลุ่มตาม logical domain การกำหนด logical domain นั้นสามารถกำหนดได้จาก โครงสร้างของไดเร็กทอรีและ โครงสร้างของลิงค์ระหว่างเอกสาร ซึ่งจะทำให้กลุ่มเอกสารที่มีขนาดใหญ่ถูกแบ่งให้เล็กลง แต่การหา logical domain ที่ดีนั้นก็ทำได้ยาก จึงใช้วิธี

กำหนดจำนวนของเอกสารต่อหนึ่งกลุ่ม แล้วจึงทำการแบ่งกลุ่มของเอกสารนั้นเป็นกลุ่มย่อย ถ้าจำนวนของเอกสารในกลุ่มเกินที่กำหนดไว้ โดยเริ่มจัดกลุ่มตามโดเมนแล้วจึงแบ่งกลุ่มตามไดเร็กทอรีลึกลงไปทีละขั้น ก็จะทำให้กลุ่มของเอกสารขนาดใหญ่ถูกแบ่งให้เล็กลงได้ แต่ยังคงเกิด ปัญหากับเว็บไซท์ขนาดใหญ่ที่มีโครงสร้างไดเร็กทอรี่ที่ลึกจะทำให้เกิดกลุ่มของเอกสารขนาดเล็กที่มีเอกสารจำนวนน้อยหลาย ๆ กลุ่ม

หลังจากจัดกลุ่มแล้วจึงทำกระบวนการที่เรียกว่า Locality Analysis ซึ่งเป็นกระบวนการปรับคะแนนของเอกสารไปยังเอกสารต่าง ๆ ที่อยู่ในกลุ่มเดียวกัน โดยใช้ฟังก์ชันดังนี้

$$adj\_score(i) = \sum_{j \in c} orig\_score(j)\alpha^{D(i,j)} \qquad (2.5)$$

โดย i,j คือเอกสาร i และ j ตามลำดับ adj_score(i) คือ คะแนนที่ถูกปรับใหม่ของเอกสาร i และ orig_score(j) คือคะแนนเริ่มต้นของเอกสาร j D(i,j) เป็นระยะห่างระหว่าง เอกสาร i กับ j ส่วน α เป็นฟังก์ชันลดที่มีค่าระหว่าง 0-1 เป็นตัวที่ใช้กำหนดระยะระหว่างเอกสารที่จะมีผลดึงกันได้ เหตุผลที่ต้องมีการปรับคะแนนของเอกสารใหม่นั้น เพราะการที่เอกสารที่ดีมารวมอยู่ด้วยกัน จะดีกว่าเอกสารที่ดีแต่กระจายกันอยู่ ดังนั้นเมื่อมีการปรับคะแนนแล้วจะทำให้เอกสารที่อยู่ในกลุ่มที่รวมเอกสารดี ๆ ไว้ จะมีคะแนนเพิ่มขึ้น ซึ่งจะมีผลดีกับขั้นตอนต่อมาคือการจัดลำดับของโดเมนเพราะการจัดลำดับโดเมนนี้จะเป็นการคิดคะแนนเฉลี่ยของเอกสารใน domain ดังนั้นถ้าเอกสารที่ถูกปรับคะแนนใหม่มีคะแนนดีขึ้นก็จะทำให้โดเมนนั้นอยู่ในอันดับที่ดี

2. จัดกลุ่มตามประเภทของเอกสาร (Category) ในการจัดกลุ่มเอกสารตามประเภทนี้ระบบนี้จะใช้ตัวจัดประเภท (Classifier) จากภายนอก ซึ่งก็คือ Yahoo! [9] รายละเอียดของกระบวนการแยกประเภท โดย Yahoo! นี้จะอยู่ใน [9] แล้วจึงแสดงผลการจัดกลุ่มตามประเภทให้ผู้ใช้ดู แต่ก็ยังมีข้อเสียในกรณีที่มีเอกสารจำนวนมากอยู่ในแต่ละ Category ทำให้ผู้ใช้จะต้องใช้เวลาในการหาเอกสารที่ต้องการอยู่ จึงเสนอวิธีการแบ่งประเภท (Category) ให้เป็นประเภทย่อย (Sub-Category) เพื่อแบ่งเอกสารออกเป็นกลุ่มที่เล็กลง โดยทั้งประเภท และประเภทย่อยนั้นจะแบ่งตามที่ Yahoo! กำหนด

จากการแบ่งกลุ่มทั้ง 2 แบบ จะเห็นได้ว่าในแบบแรกจะเป็นการจัดกลุ่มตาม URL ของเอกสารทำให้การแบ่งกลุ่มนี้ไม่ได้แบ่งตามความสัมพันธ์ของเอกสาร เช่นเดียวกับการแบ่งกลุ่มด้วยวิธี [4] ส่วนในแบบที่สองนั้น จะครอบคลุมเอกสารได้จำนวนน้อย เพราะการจัดประเภทของ Yahoo! นั้นใช้คนเป็นผู้จัดจึงจัดได้ช้า

2.4 Authoritative Sources in a Hyperlinked Environment [5]

เป็นกระบวนการที่นำ URL ในเอกสาร (link) มาใช้ เพื่อหาและจัดเรียงลำดับผลที่ดี โดยเน้นที่จะนำไปใช้กับ คำที่ใช้ค้นหาที่มีลักษณะเป็นคำทั่วไปไม่มีการเฉพาะเจาะจง (Broad-topic queries) ซึ่งจะทำให้ได้ผลการค้นหาข้อมูลจำนวนมาก แล้วจึงนำผลเหล่านั้นมาประมวลผลเพื่อหาลิงค์ในเอกสารที่ชี้ไปยังผลอันอื่น หลังจากนั้นก็จะนำลิงค์เหล่านั้นมาเป็นข้อมูลใช้ในการคำนวณคะแนนของผลเหล่านั้น ซึ่งแบ่งเป็น 2 อย่าง ได้แก่ Authority คือคะแนนที่ผลนั้นถูกผลอื่นชี้มาหา และ Hub คือคะแนนของการที่ผลนั้น ๆ มีชี้หรือลิงค์ไปยังผลอื่น ๆ แล้วจึงเรียงลำดับผลการค้นหาข้อมูลตามคะแนนของ Authority

กระบวนการของวิธีการนี้จะเริ่มจากนำผลการค้นหาของเสิร์ชเอ็นจิ้น เช่น Altavista [10] มาประมาณ 200 ผล ซึ่งจะถือว่าเป็นผลที่จะต้องถูกนำมาเรียงลำดับ แล้วจึงนำลิงค์ในผลทั้ง 200 ผลนี้มาใช้ แต่เนื่องจากลิงค์ที่ชี้ไปหาผลอื่น ๆ ใน 200 ผลนี้ยังมีน้อยเกินไป ไม่เพียงพอต่อการนำ

SubGraph (σ,ε,t,d)

σ : a query string

ε : a text-based search engine

t,d : natural number

Let R_σ denote the top t results of ε on σ

Set $S_\sigma = R_\sigma$

For each page $p \in R_\sigma$

Let $\Gamma^+(p)$ denote the set of all pages p points to

Let $\Gamma^-(p)$ denote the set of all pages pointing to p

Add all pages in $\Gamma^+(p)$ to S_σ

If $|\Gamma^+(p)| \leq d$ then

Add all pages in $\Gamma^-(p)$ to S_σ

Else

Add an arbitrary set of d pages from $\Gamma^-(p)$ to S_σ

End

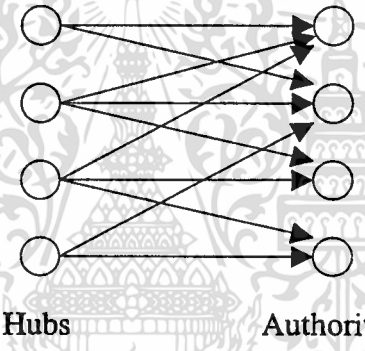
Return S_σ

รูปที่ 2.5 กระบวนการสำหรับการเพิ่มเอกสารเข้าสู่เซทของเอกสารที่ต้องการจัดกลุ่ม

มาคำนวณคะแนน จึงมีการเพิ่มเอกสารเข้ามายังกลุ่มของผลเหล่านี้ โดยเอกสารเหล่านั้นก็คือเอกสารที่ถูกลิงค์ในผล 200 ผลแรกชี้ไปหาและเอกสารที่มีลิงค์ชี้มาหา 200 ผลแรกนั้น เนื่องจากมองว่าเอกสารเหล่านี้เป็นเอกสารที่มีความสัมพันธ์กับ 200 ผลแรก โดยกระบวนการทั้งหมดที่กล่าวมาจะเป็นดังรูปที่ 2.5

การเพิ่มเอกสารที่สัมพันธ์กับผลการค้นหาเข้ามานี้จะทำให้มีข้อมูลของลิงค์ที่จะมาช่วยในการคำนวณเพิ่มนั้น แต่จะไม่ใช้ลิงค์ในส่วนที่ชี้ไปยังเอกสารที่อยู่ใน domain เดียวกัน เนื่องจากลิงค์เหล่านั้นไม่ได้ช่วยบอกความสัมพันธ์ระหว่างเอกสาร

แล้วจึงมาทำการแยกหาเอกสารที่สัมพันธ์กับคำที่ใช้ค้นหา ออกจากกลุ่มของเอกสารที่รวมกันอยู่ โดยใช้ทฤษฎีที่ว่าเอกสารที่สัมพันธ์กันน่าจะถูกชี้โดยเอกสารอื่น ๆ เหมือน ๆ กัน นั่นคือ Authority ที่เกี่ยวข้องกับสิ่งเดียวกันจะถูกชี้โดยกลุ่มของ Hub เหมือนกัน ดังรูป



รูปที่ 2.6 รูปแสดงตัวอย่าง Authority และ Hubs

ดังนั้น Authority กับ Hub จะทำให้เกิดความสัมพันธ์ในลักษณะส่งเสริมกันที่เรียกว่า mutually reinforcing relationship คือ Hub ที่ดีจะชี้ไปยัง Authority ที่ดี และ Authority ที่ดีจะถูกชี้โดย Hub ที่ดี เราจะใช้ความสัมพันธ์ระหว่าง Hub และ Authority ที่ได้นี้ผ่าน กระบวนการ Iterative Algorithm เพื่อทำการหาคะแนน Authority ของเอกสาร p ($x^{<p>}$) และ คะแนน Hub ของเอกสาร p ($y^{<p>}$)

$$x^{<p>} = \sum_{q:(q,p)\in E} y^{<q>} \quad (2.6)$$

$$y^{<p>} = \sum_{q:(q,p)\in E} x^{<q>} \quad (2.7)$$

โดยมี Pseudo code ของ Iterative Algorithm เป็นดังรูปที่ 2.7 จะเห็นว่ากระบวนการของ Iterative Algorithm นี้จะเป็นการทำไปเรื่อย ๆ เพื่อหาจุดดุลยภาพของค่า x และ y ซึ่งจาก code ก็คือจำนวน k รอบ จากกระบวนการ Iterative Algorithm นี้เราสามารถนำไปทำการกรองเพื่อหาเอกสารที่มีค่า Authority และ Hub สูงสุดอย่างละ c เอกสาร โดยมีกระบวนการดังรูปที่ 2.8 โดยกำหนดค่า c ที่ประมาณ 5-10 และการกำหนดค่า k ที่เหมาะสมนั้นสามารถดูได้จาก [5] จากการทดลองแล้วได้ผลที่ดีนั้นแสดงให้เห็นว่า การพิจารณาเพียงลิงค์อย่างเดียวก็เพียงพอ แต่การพิจารณาข้อความร่วมกับลิงค์นั้น ก็น่าจะมาช่วยทำให้ผลดีขึ้น

กระบวนการที่กล่าวมานี้ยังนำมาใช้หาเอกสารที่คล้ายกับเอกสารที่กำหนดได้ด้วย และยังสามารถนำมาใช้กับคำที่มีหลายความหมาย โดยถ้าสามารถแยกกลุ่มของ Authority และ Hub ที่เรียงกันตามคะแนน ออกเป็นกลุ่มย่อย ๆ หลาย ๆ กลุ่ม โดยการแยกตามช่วงคะแนน ซึ่งจะได้กลุ่มย่อยแต่ละกลุ่มนั้นเป็นตัวแทนความหมายแต่ละความหมายของคำได้

```
Iterate(G,k)
    G: a collection of n linked pages
    k: a natural number
    Let z denote the vector (1,1,1,...,1) ∈ R^n
        Set x_0 := z
        Set y_0 := z
        For i = 1,2,...,k
            Apply the I operation to (x_{i-1}, y_{i-1}), obtaining new x-weights x'_i
            Apply the O operation to (x'_i, y_{i-1}), obtaining new y-weights y'_i
            Normalize x'_i, obtaining x_i
            Normalize y'_i, obtaining y_i
        End
        Return (x_k, y_k)
```

รูปที่ 2.7 Pseudo code ของ Iterative Algorithm

จากกระบวนการทั้งหมดจะเห็นได้ว่า แม้ว่าผลการจัดลำดับจะทำได้ดี แต่กระบวนการที่ซับซ้อนรวมไปถึงการใช้ข้อมูลจำนวนมากเพื่อคำนวณคะแนน Authority กับ Hub ทำให้ใช้เวลามากไม่เหมาะกับการใช้งานจริงนัก การแสดงผลยังไม่ได้แสดงเป็นกลุ่มของผลที่เกี่ยวข้องในด้าน

Filter (G,k,c)

G: a collection of n linked pages

k,c: natural numbers

(x_k, y_k) := Iterate(G,k)

Report the pages with the c largest coordinates in x_k as authorities.

Report the pages with the c largest coordinates in y_k as hubs.

รูปที่ 2.8 กระบวนการกรองเอกสารที่มีค่า Authority และ Hub สูงสุดอย่างละ c เอกสาร

เดียวกัน แต่แสดงเพียงเอกสารลำดับต้น ๆ 5-10 เอกสาร ซึ่งน่าจะไม่เพียงพอต่อความต้องการของผู้ใช้ เพราะเอกสารที่แสดงเหล่านั้นอาจไม่ใช่ความหมายของคำที่ผู้ใช้ต้องการค้นหาก็ได้

2.5 Link Based Clustering of Web Search Results [7]

งานวิจัยนี้มีแนวความคิดที่ใกล้เคียงกับวิทยานิพนธ์ฉบับนี้ คือเห็นว่าผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นไม่มีประสิทธิภาพเพียงพอ ทำให้ผู้ที่ต้องการค้นหาข้อมูลต้องเสียเวลากับผลการค้นหาข้อมูลที่ได้มาจากเสิร์ชเอ็นจิ้นจึงเสนอแนวทางแก้ไขโดยทำการจัดกลุ่มผลการค้นหาข้อมูลที่ได้มาจากเสิร์ชเอ็นจิ้น โดยใช้วิธีการจัดกลุ่มที่เรียกว่า การวิเคราะห์ลิงค์ (Link Analysis) ซึ่งมีพื้นฐานบนลิงค์ที่ร่วมกันระหว่างผลการค้นหาข้อมูลหรือเรียกได้อีกอย่างว่าเอกสาร โดยใช้การวิเคราะห์การมีลิงค์ออก (Co-citation) เหมือนกัน และการมีลิงค์เข้า (Comlping) เหมือนกัน และยังมีการปรับ อัลกอริทึม K-means เพื่อให้อัลกอริทึมนี้จัดการกับข้อมูลที่ไม่ต้องการ และใช้กับผลการค้นหาข้อมูลในเว็บได้ด้วย โดยการจัดกลุ่มนี้จะทำเฉพาะหน้าเอกสารที่มีคุณภาพโดยกรองหน้าเอกสารที่ไม่มีคุณภาพออก โดยงานวิจัยนี้ยังได้แนวความคิดมาจาก [1] ซึ่งนำลิงค์มาใช้ในการจัดลำดับผลการค้นหาข้อมูล [1] เสนอว่ามีเว็บเพจอยู่ 2 ชนิด คือ ฮับ (hub) และ ออโทริตี้ (authority) ซึ่งเว็บเพจทั้ง2ชนิดนี้จะส่งเสริมซึ่งกันและกัน ถ้านำแนวคิดนี้มาใช้ในการจัดกลุ่มเอกสารจึงน่าจะให้ผลที่ดีได้ แม้วิธีการจัดกลุ่มส่วนใหญ่จะใช้คำเป็นปัจจัยสำคัญ แต่งานวิจัยนี้ได้พยายามมองหาคุณสมบัติอื่นในเว็บเพจมาใช้เป็นข้อมูลเพื่อใช้จัดกลุ่มเอกสาร สิ่งที่ได้คือ

1. ไฮเปอร์ลิงค์ (Hyperlink) ระหว่างเว็บเพจซึ่งเป็นความแตกต่างที่สำคัญระหว่างเอกสารทั่วไปกับเว็บเพจและสิ่งนี้น่าจะเป็นข้อมูลที่สำคัญเพื่อนำมาใช้จัดกลุ่มเว็บเพจที่เกี่ยวเนื่องกัน

2. เว็บเพจที่ได้จากเสิร์ชเอ็นจิ้นส่วนใหญ่จะเป็นหน้าแรก ๆ ของเว็บไซท์ ซึ่งโดยปกติเว็บเพจหน้าแรก ๆ เหล่านี้มักจะประกอบด้วยลิงค์และรูปมากกว่าคำที่เป็นเนื้อหาจำนวนมาก

3. เว็บเพจสามารถทำขึ้นในภาษาที่ต่างกัน ซึ่งเป็นปัญหากับวิธีการจัดกลุ่มที่ใช้คำเมื่อต้องทำการจัดกลุ่มเอกสารที่ไม่ใช่ภาษาอังกฤษ

ดังนั้น แนวทางของงานวิจัยนี้จะทำการจัดกลุ่มผลการค้นหาในเว็บ โดยใช้วิธีการที่เรียกว่า การวิเคราะห์ลิงค์ (Link Analysis) ซึ่งวิธีการนี้มีแนวความคิดคือ หน้าเอกสารที่มีลิงค์เข้าและลิงค์ออก เหมือนกันน่าจะมีความสัมพันธ์กันอย่างเหนียวแน่น โดยถ้าพิจารณาหน้าเอกสาร p และ q คือหน้าเอกสารที่ p และ q ชี้ไปหาเหมือนกัน และหน้าเอกสารที่ชี้มาหา p และ q เหมือนกัน โดยลิงค์ระหว่างหน้าเอกสารเหล่านี้ต้องเป็นลิงค์คนละเว็บไซท์กัน ซึ่งงานวิจัยนี้จะนำ การวิเคราะห์ลิงค์นี้ไปใช้ร่วมกับ K-means ที่ได้ทำการปรับเพื่อใช้กับการจัดกลุ่มเอกสารเว็บเพจโดยเฉพาะ

2.5.1 การจัดกลุ่มโดยวิธี Link Analysis

โดยใช้การมีลิงค์ออกเหมือนกัน (Co-citation) และการมีลิงค์เข้าเหมือนกัน (Coulping) เพื่อแสดงว่าหน้าเอกสารหนึ่งจะรวมกับอีกหน้าเอกสารหนึ่งได้ ถ้ามีจำนวนลิงค์เข้า-ออกเหล่านี้เหมือนกันมากพอ ดังนั้น สำหรับแต่ละ URL P ในผลการค้นหาข้อมูล R จะถูกเก็บลิงค์ที่ชี้ออกจากเอกสาร รวมถึงการหาลิงค์ที่ชี้เข้าสู่เอกสาร ซึ่งการหาลิงค์ที่เข้าสู่เอกสารนี้จะใช้ Alta Vista [10] เป็นเครื่องมือในการหา ดังนั้น จะได้ข้อมูลของทั้งลิงค์เข้าและลิงค์ออกสำหรับแต่ละ URL P

2.5.1.1 ข้อกำหนด

- ตัวแทนของแต่ละหน้าเอกสาร P ใน R แต่ละหน้าเอกสาร P ใน R จะถูกแทนด้วยเวคเตอร์ 2 เวคเตอร์ คือ เวคเตอร์ P_{out} (มี N มิติ) และ เวคเตอร์ P_{in} (มี M มิติ) โดยมิติที่ i ในเวคเตอร์ P_{out} จะแสดงว่า P มีลิงค์ออกที่ i จากลิงค์ออกทั้งหมด N ลิงค์ และเวคเตอร์ P_{in} ก็เช่นกัน

- การวัดความคล้าย งานวิจัยนี้ปรับฟังก์ชัน Cosine ให้สามารถวัดความคล้ายได้จากลิงค์เข้าและลิงค์ออกที่มีเหมือนกัน ระหว่างหน้าเอกสาร P และ Q โดยลักษณะของฟังก์ชัน Cosine เป็นดังรูป 2.10

$$Cosine\ (P,Q) = (P \bullet Q)/(||P||\ ||Q||) = ((P_{out} \bullet Q_{out}) + (P_{in} \bullet Q_{in}) /(||P||\ ||Q||) \quad (2.8)$$

โดย $||P||^2 = (\sum_{1}^{N} P_{Out_i}^2 + \sum_{1}^{M} P_{In_j}^2)$ จำนวนลิงค์ออกและลิงค์เข้าทั้งหมดของเอกสาร P,

$||P||^2 = (\sum_{1}^{N} P_{Out_i}^2 + \sum_{1}^{M} P_{In_j}^2)$ จำนวนลิงค์ออกและลิงค์เข้าทั้งหมดของเอกสาร P

$(P_{out} \bullet Q_{out})$ เป็นดอทโปรดักซ์ของเวคเตอร์ P_{out} และ Q_{out} เพื่อวัดลิงค์ออกที่มีร่วมกันระหว่าง P และ Q ซึ่ง $(P_{in} \bullet Q_{in})$ ก็เช่นเดียวกัน ขณะที่ ||P|| เป็นความยาวของเวคเตอร์ P

- จุดศูนย์กลางของกลุ่ม (Center Point of Cluster) ให้จุดศูนย์กลาง C เป็นตัวแทนของกลุ่มเอกสาร S เมื่อต้องการคำนวณความคล้ายระหว่าง P และกลุ่มเอกสาร S โดย |S| เป็นจำนวนเอกสารในกลุ่ม เมื่อจุดศูนย์กลางเป็นจุดสมมุติจึงควรมีขนาดเล็กกว่า 1 ดังนั้นจึงได้

$$C_{Out} = 1/|S| \sum_{P_i \in S} P_{i_{Out}} \quad (2.9)$$

$$C_{In} = 1/|S| \sum_{P_i \in S} P_{i_{In}} \quad (2.10)$$

$$Similarity\ (P,S) = Cosine\ (P,C) \quad (2.11)$$

- ใกล้ลิงค์ทั่วไปของกลุ่ม (Near-Common Link of Cluster) คือลิงค์ที่มีอยู่ในสมาชิกส่วนใหญ่ของกลุ่ม ถ้าลิงค์หนึ่งมีอยู่ใน 50% ของสมาชิกในกลุ่ม ก็จะเรียกลิงค์นั้นว่า 50% ใกล้ลิงค์ทั่วไปของกลุ่ม ถ้ามีอยู่ในสมาชิกทุกตัวของกลุ่ม จะเรียกว่า ลิงค์ทั่วไปของกลุ่ม

2.5.1.2 วิธีการจัดกลุ่ม

งานวิจัยนี้ได้เสนอวิธีการจัดกลุ่มวิธีใหม่โดยปรับปรุง K-means ปกติให้สามารถจัดกลุ่มผลการค้นหาข้อมูลบนเว็บ และแก้ข้อเสียของ K-means โดยวิธีการดังนี้

- กรองหน้าที่ไม่สัมพันธ์ออก โดยจะทำการจัดกลุ่มโดยเลือกเฉพาะหน้าเอกสารที่มีคุณภาพ (โดยเลือกเฉพาะหน้าที่มีผลรวมของลิงค์เข้าและลิงค์ออกอย่างน้อย 2) ซึ่งจะเป็นการเพิ่มความแม่นยำให้กับผลการจัดกลุ่ม

- การกำหนดความคล้ายขั้นต่ำ (Similarity Threshold) ความคล้ายขั้นต่ำถูกกำหนดไว้เพื่อวัดว่าหน้าเอกสารหนึ่งสามารถรวมเข้ากับกลุ่มเอกสารหนึ่งได้หรือไม่ โดยความคล้ายเป็นการหาจำนวนลิงค์ทั่วไป (Common Link) ที่มีเหมือนกันระหว่าง 2 หน้าเอกสารที่ต่างกัน จึงไม่ยากที่จะกำหนดค่าตรงนี้

- ใช้ใกล้ลิงค์ทั่วไปของกลุ่ม เพื่อแสดงความรวมกันภายในกลุ่ม จากการทดลองพบว่า 30% ใกล้ลิงค์ทั่วไปเหมาะสมที่สุด โดยทุกกลุ่มต้องมีอย่างน้อย 1 ลิงค์ ในกลุ่มที่มีอยู่ในเอกสารอย่างน้อย 30% ของกลุ่ม

- การใส่หน้าเอกสารหนึ่ง ๆ ไปอยู่ในกลุ่ม ก็ต่อเมื่อ ก) ความคล้ายระหว่างหน้าเอกสารและกลุ่มเอกสารมีมากกว่าความคล้ายขั้นต่ำและ ข) หน้าเอกสารต้องมีลิงค์ที่เหมือนกับใกล้ลิงค์ทั่วไปของกลุ่มนั้น ๆ ถ้าไม่มีกลุ่มใดตรงตามเกณฑ์ หน้าเอกสารก็จะกลายเป็นกลุ่มใหม่ เวคเตอร์ศูนย์กลางจะถูกใช้เพื่อคำนวณความคล้ายและจะถูกคำนวณใหม่ถ้ามีสมาชิกใหม่เข้ามา โดยหน้าเอกสารหนึ่งจะเป็นสมาชิกได้ใน 10 กลุ่มที่มีความคล้ายสูงสุดเท่านั้น ทุกหน้าเอกสารจะถูกนำไปทำการจัดกลุ่มตามลำดับจะไม่มีการเปลี่ยนแปลงกับเวคเตอร์ศูนย์กลางของแต่ละกลุ่ม

- ทำกลุ่มผลลัพธ์โดยการรวมกลุ่มพื้นฐาน (Base Cluster) กลุ่มผลลัพธ์ทำขึ้นโดยการรวมกลุ่มพื้นฐาน 2 กลุ่ม ที่มีสมาชิกในกลุ่มเหมือนกันมาที่สุด ซึ่งควบคุมการรวมโดย Merging Threshold

2.5.2 **การทดลองและวัดผล**

1. สภาพแวดล้อมในการทดลอง

- การเก็บข้อมูล เพื่อวัดประสิทธิภาพ ประสิทธิผล และความยืดหยุ่น จึงทำการทดลองกับคำที่ใช้ค้นหาที่ต่างกัน โดยเสิร์ชเอ็นจิ้นที่ต่างกันและจำนวนผลการค้นหาข้อมูลที่ต่างกัน เนื่องจากวิธีการนี้มีพื้นฐานอยู่ที่การวิเคราะห์ลิงค์ (Link Analysis) จึงใช้จำนวนลิงค์เข้าที่ต่างกันด้วย เก็บข้อมูลโดยการดาวน์โหลดผลการค้นหาข้อมูล แล้วเก็บลิงค์ออก(out-links)ออกจากแต่ละหน้าเอกสารและหาลิงค์เข้า(in-links) โดยใช้ Alta Vista [10]

- การทำความสะอาดข้อมูล เนื่องจากในเว็บอาจมีข้อมูลที่ซ้ำกันซึ่งจะทำให้การจัดกลุ่มผิดพลาดจึงมีการกำจัดออกโดยดูว่าถ้าหน้าเอกสาร p และ q จะซ้ำกันเมื่อ 1) มี 8 ลิงค์ออก (out-links) ในแต่ละหน้าเอกสารและ 2) มีลิงค์ทั่วไป (link in common) อย่างน้อยหน้าเอกสารที่มีเปอร์เซ็นของลิงค์ที่ซ้ำมากกว่าจะถูกกำจัด

- ใช้อัลกอริทึมที่พัฒนาเพื่อสร้างกลุ่มพื้นฐาน

- ทำกลุ่มผลลัพธ์ ซึ่งจะเกิดโดยการรวมกลุ่มพื้นฐานทีละ 2 กลุ่ม ซึ่งมีจำนวนสมาชิกในกลุ่มส่วนใหญ่เหมือนกัน (เช่น 75%)

2. ผลการทดลอง

มีเพียง 60 - 70% ของผลการค้นหาข้อมูล ที่ถูกนำมาจัดกลุ่ม หลังจากการทำความสะอาดข้อมูล ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอสามารถทำกลุ่มที่มีเนื้อหาหลัก ๆ เกี่ยวข้องกับคำที่ใช้ค้นหาข้อมูลได้ และมีเอกสาร 30-50% ไม่ถูกจัดกลุ่ม

3. การวัดผลการจัดกลุ่มตามค่าเอ็นโทรฟี (Entropy)

ใช้เอ็นโทรฟี (Entropy) เพื่อวัดคุณภาพการจัดกลุ่ม เอ็นโทรฟีทำการวัดโดยเทียบกลุ่มที่ได้จากการจัดกลุ่มโดยวิธีที่พัฒนาขึ้น เทียบกับกลุ่มที่ถูกต้อง ในที่นี้กลุ่มที่ถูกต้องได้จากผู้วิจัยทำการตรวจสอบแต่ละเอกสาร และกำหนดเองว่าแต่ละเอกสารควรอยู่กลุ่มใด ซึ่งอาจเกิดความผิดพลาดในการกำหนดกลุ่มซึ่งเกิดจากความเอนเอียงได้ ให้ CS แทนกลุ่มที่ได้จากการทดลอง เอ็นโทรฟีของกลุ่ม j คือ

$$E(j) = -\sum_{i} p_{ij} \log(p_{ij}) \qquad (2.12)$$

โดย P_{ij} เป็นความน่าจะเป็นที่สมาชิกของกลุ่ม j ควรอยู่ในประเภท i เอ็นโทรฟีเฉลี่ยของกลุ่มที่ถูกจัด คำนวณจากผลรวมของเอ็นโทรฟีของแต่ละกลุ่มซึ่งถ่วงน้ำหนักโดยขนาดของแต่ละกลุ่ม ดังนี้

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E(j)}{n} \qquad (2.13)$$

โดย n_j = ขนาดของกลุ่ม j, m = จำนวนกลุ่มและ n = จำนวนข้อมูลทั้งหมด

จากการวัดผลการจัดกลุ่มของคำ Jaguar ที่ใช้ในการค้นหาข้อมูล แสดงให้เห็นว่ากลุ่มขนาดเล็กมีความเสถียรมากกว่า และมีบางกลุ่มทั้งเล็กและใหญ่ที่ไม่สามารถอธิบายได้ยังต้องการการพัฒนาต่อไป

2.5.3 บทสรุป

ผลงานวิจัยนี้ได้นำเสนอวิธีการใหม่ในการจัดกลุ่มผลการค้นหาข้อมูล โดยใช้วิธีการที่เรียกว่า การวิเคราะห์ลิงค์ ซึ่งวิธีการนี้ยังใช้การกรองผลการค้นหาข้อมูลที่ไม่เกี่ยวข้องออกด้วย เพื่อทำการจัดกลุ่มผลการค้นหาข้อมูลที่มีคุณภาพสูงเท่านั้น การจัดกลุ่มผลการค้นหาข้อมูลนี้ก็เพื่อให้ผู้ใช้ดูข้อมูลได้สะดวก โดยการทดลองได้ทำการจัดกลุ่มผลการค้นหาข้อมูลของคำที่ใช้ในการค้นหาเพียง 3 คำ คือ Jaguar, Data mining และ Java ซึ่งไม่เพียงพอในการยืนยันว่าคุณภาพของการจัดกลุ่มโดยวิธีการนี้มีประสิทธิภาพมากน้อยเพียงไร และยังไม่มีการเปรียบเทียบกับวิธีการอื่น ๆ จึงไม่สามารถบอกได้ว่าประสิทธิภาพเมื่อเทียบกับวิธีอื่นเป็นอย่างไร และในการวัดคุณภาพการจัดกลุ่มของวิธีการวิเคราะห์ลิงค์นี้ ทำการวิเคราะห์โดยเทียบกับกลุ่มข้อมูลที่ถูกต้อง ซึ่งผู้ทำการวิจัยเป็นคนจัดเอง จึงอาจมีข้อผิดพลาดเกิดขึ้นในการจัดทำกลุ่มข้อมูลที่ถูกต้อง จนเป็นเหตุให้อาจเกิดความผิดพลาดในการคำนวณคุณภาพของการจัดกลุ่มได้ เนื่องจากการจัดกลุ่มด้วยวิธีการนี้จำเป็นต้องใช้ข้อมูลของผลการค้นหาข้อมูลทั้งหมดก่อนที่จะเริ่มการจัดกลุ่มผลการค้นหาข้อมูล ในขณะที่งานวิจัยนี้ต้องการทำการจัดกลุ่มทันทีเมื่อได้ผลการค้นหาข้อมูลจากเสิร์ชเอ็นจิ้น จึงทำให้วิธีการจัดกลุ่มนี้ไม่เหมาะต่อการทำการจัดกลุ่มออนไลน์ ในการเปรียบเทียบคุณภาพการจัดกลุ่มระหว่างวิธีการนี้กับวิธีการที่พัฒนาขึ้นในวิทยานิพนธ์เล่มนี้ ไม่สามารถทำได้เนื่องจากกลุ่มข้อมูลที่ใช้ในการทำลองของวิทยานิพนธ์เล่มนี้ไม่เหมาะในการนำมาใช้กับวิธีการนี้ โดยกลุ่มข้อมูลของวิทยานิพนธ์นี้มีลิงค์เข้ามาหาแต่ละข้อมูลน้อยมาก ทำให้ข้อมูลในส่วนนี้ขาดหายไป ซึ่งเป็นข้อมูลที่สำคัญอย่างยิ่งในการจัดกลุ่มโดยวิธีการวิเคราะห์ลิงค์ (Link Analysis)

2.6 A Large Benchmark Dataset for Web Document Clustering [8]

งานวิจัยนี้มีจุดประสงค์หลักที่จะเสนอกลุ่มของเอกสารเพื่อใช้ในการทดลองกับการจัดกลุ่มเอกสาร หรืองานที่ใกล้เคียง กลุ่มของเอกสารนี้อาจจะไม่เหมาะกับการจัดกลุ่มเอกสารทุกวิธี แต่ก็ได้ออกแบบให้ยืดหยุ่นและมีประโยชน์ที่สุดเท่าที่จะเป็นไปได้ ซึ่งงานวิจัยที่เน้น จะเป็นการจัดกลุ่มที่ไม่ต้องทำกลุ่มตัวอย่างให้ระบบเรียนรู้ก่อนการจัดกลุ่มเอกสาร เนื่องจากการทำกลุ่มตัวอย่างเป็นไปได้ยาก และต้องมีการทำอย่างต่อเนื่องเพราะในเครือข่ายใยแมงมุมย่อมมีข้อมูลซึ่งอยู่ในกลุ่มใหม่ ๆ เกิดขึ้นอยู่ตลอดเวลา ในงานวิจัยนี้ยังได้ทำการทดลองพื้นฐานเพื่อให้วิธีการจัดกลุ่มเอกสารอื่น ๆ ใช้เปรียบเทียบ โดยใช้ฟังก์ชัน K-means มาทำการจัดกลุ่มเอกสาร

2.6.1 **กลุ่มข้อมูลที่ใช้วัดประสิทธิภาพการจัดกลุ่ม**

กลุ่มข้อมูลที่ใช้ในการทดลองในงานวิจัยต่าง ๆ มีหลากหลายขนาด น้อยบ้างมากบ้าง ซึ่งขนาดของกลุ่มข้อมูลนี้มีผลต่อวิธีการจัดกลุ่มเอกสารโดยไม่มีการเรียนรู้เป็นอย่างยิ่ง การจัดกลุ่มเอกสารเองเพื่อนำมาใช้ในการทดลองนั้นต้องใช้เวลามาก และยังขาดความน่าเชื่อถือในความถูกต้องด้วย ดังนั้นนักวิจัยส่วนใหญ่จึงใช้ข้อมูลที่ถูกจัดกลุ่มไว้แล้วโดยคนซึ่งเป็นผู้เชี่ยวชาญ และกลุ่มข้อมูลเหล่านี้ยังได้รับความมั่นใจในความถูกต้องโดยผู้ที่ค้นหาข้อมูลทั่วไปในเครือข่ายใยแมงมุม แหล่งข้อมูลเหล่านี้มาจากเว็บไซต์ประเภทที่เรียกว่า ดัชนีเว็บ (Web Directory) เช่น Open Directory Project [15], Yahoo! Categories [9] และ Look Smart [16] ปัจจัยอีกประการที่สำคัญ คือ จำนวนกลุ่มในกลุ่มข้อมูล เป็นสิ่งที่เห็นได้ชัดว่า การจัดกลุ่มเอกสารที่มาจากกลุ่มเอกสาร 2 กลุ่มรวมกัน ย่อมง่ายกว่ากลุ่มเอกสารที่มาจาก 10 กลุ่มเอกสารรวมกัน แนวทางของงานวิจัยนี้ในการทำกลุ่มข้อมูลคือจะใช้กลุ่มข้อมูลจำนวนมากโดยมีเอกสารในจำนวนที่เพียงพอในแต่ละกลุ่ม ปัจจัยอีกอย่างของกลุ่มข้อมูลที่ใช้ทดลองซึ่งมีความแตกต่างกันในการทดลองของแต่ละงานวิจัย คือ เนื้อหาของเอกสารในกลุ่มทดลอง ซึ่งเนื้อหาที่ต่างกันนี้ย่อมมีผลทำให้การจัดกลุ่มแตกต่างกันไปด้วย ทำให้ความต้องการกลุ่มข้อมูลกลางที่มีมาตราฐานเพื่อใช้วัดความแตกต่างระหว่างวิธีการจัดกลุ่มต่าง ๆ จึงมีความชัดเจนยิ่งขึ้น

2.6.1.1 การออกแบบกลุ่มข้อมูล

ในการเลือกข้อมูลมาใส่ในกลุ่มข้อมูลนั้น สิ่งสำคัญเกี่ยวกับเอกสารที่จะนำมาใส่ในกลุ่มข้อมูลคือ เอกสารแต่ละเอกสารต้องมีข้อมูลให้ระบบการจัดกลุ่มสามารถนำไปใช้ทำการจัดกลุ่มได้ ถ้ามีเอกสารที่มีข้อมูลน้อยหรือไม่มีข้อมูลให้ระบบจัดกลุ่มเอกสารนำไปใช้ คงทำให้การจัดกลุ่มเป็นไปได้ยาก ดังนั้นจึงอาจมองได้ว่า กลุ่มข้อมูลนี้อาจไม่ใช่ตัวแทนของเอกสารในเครือข่ายใยแมงมุมตามความเป็นจริง เนื่องจากต้องการกลุ่มข้อมูลที่มีเอกสารจำนวนมาก เพื่อให้สามารถใช้ทำการทดลองการจัดกลุ่มได้อย่างมีประสิทธิภาพ และทำการทดลองได้หลากหลายจุดประสงค์ จึงเลือกใช้กลุ่มของเอกสารที่มีการจัดกลุ่มไว้แล้วโดยคน ซึ่งได้มาจาก Open Directory Project [15] และ

Yahoo! Categories [9] ส่วนการเลือกกลุ่มประเภทของข้อมูลนั้นมีทั้งกลุ่มประเภทที่แตกต่างกัน และกลุ่มประเภทที่ใกล้เคียงกัน เพื่อให้สามารถนำกลุ่มข้อมูลที่ทำขึ้นมาใช้ทดลองการจัดกลุ่มข้อมูลได้อย่างหลากหลาย

วิธีการเลือกกลุ่มประเภทข้อมูลเพื่อนำมาใส่รวมในกลุ่มข้อมูลที่จะจัดทำขึ้นมีวิธีการ ดังนี้ เลือกกลุ่มประเภทข้อมูลจำนวน 2 กลุ่ม ที่มีเนื้อหากว้าง ๆ แต่ทั้ง 2 กลุ่มนี้ต้องมีเนื้อหาต่างกันมาก ๆ คือ "Banking & Finance" และ "Programming Languages" แล้วจึงเลือกกลุ่มย่อยจำนวน 3 กลุ่มจากแต่ละกลุ่มประเภทข้อมูล 2 กลุ่มนี้ จึงได้กลุ่มข้อมูลจำนวน 6 กลุ่ม คือ "Commercial Banks", "Building Societies", "Insurance Agencies", "Java", "C/C+" และ "Visual Basic" จะสังเกตได้ว่าการพยายามที่จะทำการจัดกลุ่มข้อมูลทั้ง 6 กลุ่มนี้ให้แยกเป็นกลุ่มข้อมูล 2 กลุ่ม (Finance และ Programming Language) เป็นสิ่งที่ทำได้ง่ายกว่าการพยายามจัดกลุ่มข้อมูลประเภท Programming Language ทั้งหมดให้แยกออกเป็น 3 กลุ่ม ซึ่งเมื่อมองดูแล้วก็ยังง่ายกว่าการพยายามจัดกลุ่มข้อมูลทั้ง 6 กลุ่มที่นำมารวมกันให้แยกออกเป็น 6 กลุ่มที่ถูกต้อง จะเห็นได้ว่ากลุ่มข้อมูลทั้ง 6 ที่เลือกมานี้ สามารถรองรับจุดมุ่งหมายของเราในการทำกลุ่มข้อมูลที่สามารถนำมาใช้ทดลองการจัดกลุ่มได้หลากหลายในระดับความยากง่ายที่ต่างกัน แล้วจึงทำการเลือกกลุ่มประเภทข้อมูลอีก 4 กลุ่ม โดย 2 กลุ่มเป็นกลุ่มที่มีเนื้อหาใกล้เคียงกับกลุ่มข้อมูล 6 กลุ่มที่เลือกมาแล้ว คือ "Astronomy" และ "Biology" ซึ่งมีเนื้อหาในหมวดหมู่วิทยาศาสตร์ที่ใกล้เคียงกับ "Programming Languages" ส่วนอีก 2 กลุ่มจะเลือกกลุ่มข้อมูลที่แตกต่างจาก 8 กลุ่มที่เลือกไว้

ตารางที่ 2.1 ประเภทของข้อมูลในกลุ่มข้อมูล รวมถึงเรื่องที่แต่ละประเภทเกี่ยวข้อง

Dataset Id	Dataset Category	Associated Theme
A	Commercial Banks	Banking & Finance
B	Building Societies	Banking & Finance
C	Insurance Agencies	Banking & Finance
D	Java	Programming Languages
E	C / C++	Programming Languages
F	Visual Basic	Programming Languages
G	Astronomy	Science
H	Biology	Science
I	Soccer	Sport
J	Motor Sport	Sport
K	Sport	Sport

แล้ว ซึ่งได้กลุ่ม "Soccer" และ "Motor Sport" ท้ายที่สุดมีการเพิ่มกลุ่มข้อมูลอีก 1 กลุ่ม ซึ่งเป็นกลุ่มหลักที่มีกลุ่มข้อมูลย่อย 2 กลุ่มที่ได้เลือกมาแล้วอยู่ในกลุ่มหลักนี้ เพื่อใช้ทดสอบการจัดกลุ่มข้อมูลที่เป็นลำดับชั้น (Hierarchical) ซึ่งกลุ่มที่เลือกมาคือ "Sport" แต่ลบข้อมูลที่อยู่ในกลุ่ม "Soccer" และ "Motor Sport" ออก ทำให้เราได้กลุ่มข้อมูลที่มีความซับซ้อนมากขึ้นเพื่อใช้ทดสอบการจัดกลุ่ม รายละเอียดของกลุ่มประเภทข้อมูลที่เลือกมาทั้งหมดแสดงในตารางที่ 2.1 โดยมีกลุ่มข้อมูลทั้งหมด 11 กลุ่ม แต่ละกลุ่มมีเอกสาร 1,000 เอกสาร

2.6.1.2 การเลือกเอกสารที่จะนำมาใช้

ในกลุ่มเอกสาร 10 กลุ่มแรกที่อยู่ใน ตารางที่ 2.1 ข้อมูลเกี่ยวกับทุกเว็บไซท์จะถูกเก็บขึ้นมาจากกลุ่มต่าง ๆ ที่แสดงอยู่ใน Open Directory Project [15] และ Yahoo! [9] แล้วจึงนำมารวมกัน จะมีเพียงข้อมูลเกี่ยวกับหน้าแรกของเว็บไซท์เท่านั้นที่จะถูกเก็บ เพื่อให้เว็บสไปเดอร์ (Web Spider) ตามเข้าไปเก็บข้อมูลของทุก ๆ เว็บเพจ (webpage) ในเว็บไซท์นั้น ๆ ทั้งหมด โดยจะบันทึกขนาดและ URL ของแต่ละเว็บเพจ ซึ่งขนาดของเว็บเพจนั้น จะเป็นขนาดซึ่งหลังจากการลบสคริปต์ (scripts), สไตล์ชีต (style-sheets) หรือ คอมเมนท์ (comment) ออกแล้ว แล้วจึงจัดเรียง URL ของแต่ละเว็บไซท์ตามจำนวนหน้าที่ถูกอ้างอิงจากมากไปน้อย โดยเว็บไซท์ที่มีจำนวนหน้าที่ถูกอ้างอิงน้อยกว่า 10 หน้าจะถูกตัดออก ถ้าเว็บเพจใดมีโครงสร้างในลักษณะเฟรมระบบจะนำเฟรมแต่ละหน้ามารวมอยู่ในที่เดียวกันเพื่อให้แน่ใจว่า ข้อมูลในเว็บเพจนั้นมีอยู่ครบถ้วนตามที่ผู้ที่ท่องเว็บเห็นเมื่อเข้ามาดู

2.6.2 **การทดลองพื้นฐานด้วยการจัดกลุ่มโดยใช้** K-means

การทดลองพื้นฐานนี้ทำขึ้นเพื่อใช้เป็นผลการทดลองที่ใช้เปรียบเทียบกับวิธีการจัดกลุ่มอื่น ๆ โดยการทดลองนี้ใช้วิธีที่เป็นพื้นฐานที่สุด คือ K-means และยังใช้วิธีทำเวคเตอร์คุณสมบัติที่ใช้เป็นตัวแทนเอกสารจากความถี่ของคำ ซึ่งเป็นวิธีที่ธรรมดามาก จุดประสงค์ของการเลือกวิธีการเหล่านี้มี 2 ประการ คือ ประการแรก เป็นสิ่งที่ใช้ยืนยันว่าวิธีการที่ซับซ้อนยิ่งกว่าจะสามารถทำการทดลองกับกลุ่มข้อมูลนี้ได้ ถ้าวิธีการที่เป็นพื้นฐานธรรมดายังสามารถทำการจัดกลุ่มได้ ประการที่สอง แม้จะมีงานวิจัยของวิธีการจัดกลุ่มโดยไม่ต้องมีการเรียนรู้ก่อน (unsupervised clustering) เกิดขึ้นมากมาย แต่เมื่อเทียบกันแล้วงานวิจัยยังห่างไกลจากงานวิจัยของวิธีการจัดกลุ่มข้อมูลโดยมีการเรียนรู้ก่อน (supervised clustering) จึงมีความจำเป็นที่จะทำกลุ่มข้อมูลมาตราฐานเพื่อใช้ในการทดลองซึ่งยังขาดแคลนอยู่

เพื่อที่จะทำความเข้าใจประสิทธิภาพการจัดกลุ่มข้อมูลที่มาจากกลุ่มที่มีเนื้อหาใกล้เคียงกันของฟังก์ชัน K-means จึงทำการทดลองกับข้อมูล 2 กลุ่ม โดยกลุ่มแรกเป็นการทดลองใช้ K-means เพื่อแยกกลุ่มข้อมูลที่มาจากการรวมของข้อมูล 2 กลุ่มที่มีความคล้ายกันมาก คือ กลุ่ม B และ C (ดูตารางที่ 2.1) การทดลองอีกกลุ่มเป็นการใช้ฟังก์ชัน K-Means แยกกลุ่มข้อมูลที่มาจาก

การรวมกันของข้อมูล 2 กลุ่มที่ต่างกัน คือ กลุ่ม A และ I (ดูตารางที่ 2.1) ในทุกกรณีจะใช้เอกสารจำนวน1,000 เอกสารในแต่ละกลุ่มข้อมูล โดยทำการจัดกลุ่ม 10 รอบ สำหรับแต่ละเวคเตอร์คุณสมบัติของเอกสารที่ต่างกัน 16 ลักษณะ ซึ่งเกิดขึ้นจากการที่มีการปรับเปลี่ยนวิธีการทำเวคเตอร์คุณสมบัติของเอกสารดังนี้ อย่างแรก เป็นเงื่อนไขในการที่จะใช้หรือไม่ใช้การเปลี่ยนคำให้อยู่ในรูปรากของคำ (word-stemming) ก่อนการทำเวคเตอร์คุณสมบัติของเอกสารอย่างที่สอง เป็นเงื่อนไขในการเลือกที่จะกำจัด คำที่ไม่มีความหมายต่อเนื้อหาของเอกสาร (stop-words) ออกหรือไม่ และอย่างสุดท้ายเป็นการเลือกใช้คำที่มีความถี่สูงสุด h% เท่านั้น มาใช้ทำเวคเตอร์คุณสมบัติของเอกสารโดยค่า h ที่นำมาใช้ในการทดลอง คือ 0.5, 1, 1.5 และ 2

2.6.2.1 การเก็บลักษณะสำคัญของเอกสาร

การทำในขั้นตอนนี้จะเน้นความง่ายจึงจะใช้คำจากส่วนที่ถูกแสดงในหน้าเว็บเบราเซอร์ (web browser) เท่านั้น ซึ่งก็คือคำที่ผู้ใช้เว็บสามารถมองเห็นได้ โดยไม่มีการให้น้ำหนักใด ๆ เพิ่มจากคุณลักษณะที่แตกต่างของคำเช่น ตัวหนา, ตัวเอียง หรือสีที่ต่างกัน วิธีการเก็บลักษณะสำคัญของเอกสารทุก ๆ เอกสารเป็นดังนี้

- เก็บทุกคำในเอกสารที่แสดงในหน้าจอ
- ถ้าใช้การกำจัดคำที่ไม่มีความหมายต่อเนื้อหาของเอกสาร ก็จะทำการกำจัดคำที่อยู่ในรายการคำที่ไม่มีความหมายต่อเนื้อหาออก
- ถ้าใช้การเปลี่ยนคำให้อยู่ในรูปรากของคำ จะทำการรวมคำที่มีรากของคำเดียวกันเข้าด้วยกัน โดยให้นับความถี่ของทุกคำเป็น 1
- แล้วจึงทำการเก็บรวบรวมความถี่ของคำแต่ละคำในเอกสารนั้น ๆ

เมื่อทำการประมวลผลทุกเอกสารในกลุ่มที่ต้องการเสร็จแล้ว จึงสร้างรายการคำหลักที่มีคำทุกคำในกลุ่มข้อมูลที่จะทำการจัดกลุ่ม รวมถึงข้อมูลความถี่ของคำแต่ละคำด้วย แล้วจึงตัดคำในรายการคำหลักให้เหลือเฉพาะคำที่มีความถี่สูงสุด h% ของคำทั้งหมด โดยค่า h จะเปลี่ยนแปลงไปในระหว่างการทดลอง สุดท้ายก็จะสร้างเวคเตอร์คุณสมบัติสำหรับแต่ละเอกสาร (v_i) ซึ่งคุณสมบัติที่ j ใน v_i คือ w_{ij}/s_i โดย w_{ij} เป็นจำนวนความถี่ของคำที่มีความถี่ในลำดับที่ j ในเอกสาร i, s_i เป็นจำนวนคำทั้งหมดในเอกสาร i

2.6.2.2 การจัดกลุ่ม

ทำโดยใช้วิธีการจัดกลุ่มแบบ K-means เป็นแบบมาตราฐาน ซึ่งเริ่มโดยการสุ่มเวคเตอร์มาเป็นศูนย์กลางของกลุ่มเอกสารเริ่มต้น โดยการสุ่มเลือกจากเวคเตอร์คุณสมบัติของเอกสาร

2.6.3 ผลการทดลอง

สำหรับแต่ละการปรับเปลี่ยนค่าต่าง ๆ ในการทดลอง จะมีการวัดค่า 3 อย่าง คือ mean, median และ ค่าความถูกต้องที่ดีที่สุดจากการจัดกลุ่มทั้ง 10 รอบ สำหรับแต่ละการตั้งค่า ถ้า

เอกสาร 2 กลุ่มถูกจัดกลุ่มได้เป็น 2 กลุ่มอย่างถูกต้องแสดงว่ามีความถูกต้องในการจัดกลุ่ม 100% แต่ถ้าผลการจัดกลุ่มได้กลุ่มเอกสารเพียงกลุ่มเดียวแสดงให้เห็นถึง ผลเสียที่เกิดจากผลกระทบของการสุ่มเลือกจุดศูนย์กลางของกลุ่มเอกสาร ทำให้ค่า mean และ median ต่ำ (ใกล้เคียง 50%) ซึ่งแสดงให้เห็นว่ามากกว่าครึ่งหนึ่งของจำนวน 10 รอบที่ได้ผลการจัดกลุ่มเอกสารเพียง 1 กลุ่ม แต่การทดลองได้แสดงให้เห็นว่าการจัดกลุ่มให้แก่กลุ่มเอกสารที่ได้มาจากการรวมเอกสาร 2 กลุ่มที่มีเนื้อหาต่างกัน ได้ผลดีกว่าการจัดกลุ่มให้แก่กลุ่มเอกสารที่ได้มาจากการรวมเอกสาร 2 กลุ่มที่มีเนื้อหาใกล้เคียงกัน ในส่วนของกลุ่มเอกสารที่เนื้อหาต่างกันนั้นการใช้การเปลี่ยนรูปคำให้อยู่ในรูปรากของคำ (word-stemming) ทำให้ผลการจัดกลุ่มดีกว่าไม่ใช่ แต่การใช้การเปลี่ยนรูปคำให้เป็นรากของคำร่วมกับการกำจัดคำที่ไม่มีผลต่อความหมายของเนื้อหาในเอกสาร กลับทำให้ผลการจัดกลุ่มแย่กว่าการใช้การกำจัดคำที่ไม่มีผลต่อความหมายของเนื้อหาในเอกสาร (Stopword Removal) เพียงอย่างเดียว ในส่วนกลุ่มเอกสารที่มีเนื้อหาใกล้เคียงกันนั้น จะได้ผลการจัดกลุ่มดีที่สุดเมื่อใช้การเปลี่ยนรูปคำให้เป็นรากของคำร่วมกับการกำจัดคำที่ไม่มีผลต่อความหมายของเนื้อหาในเอกสาร สำหรับขนาดของเวคเตอร์คุณสมบัติของเอกสารนั้น ไม่มีข้อสรุปที่แน่นอน เนื่องจากแม้เวคเตอร์คุณสมบัติที่ยาวกว่าจะให้ผลที่ดีกว่า แต่ผลที่เที่ยงตรงและแน่นอนจะได้จากเวคเตอร์คุณสมบัติที่สั้นกว่า จากการทดลองเวคเตอร์คุณสมบัติที่มีความยาว 40 เพียงพอสำหรับการจัดกลุ่มเอกสารให้ได้ผลการจัดกลุ่มที่ดีได้

2.6.4 **บทสรุป**

ผู้ทำการวิจัยนี้เห็นว่า การจัดกลุ่มเอกสารเว็บเพจเป็นแนวทางในการค้นคว้าที่น่าสนใจและมีอนาคต โดยเฉพาะการจัดกลุ่มเอกสารที่ไม่มีการเรียนรู้ก่อน (unsupervised clustering) จึงได้ทำกลุ่มข้อมูลที่สามารถใช้ในการทดลองด้านนี้ขึ้น เพื่อเป็นกลุ่มข้อมูลให้งานวิจัยต่าง ๆ ที่เกี่ยวข้องสามารถนำไปใช้จะได้เกิดความสะดวกในการเปรียบเทียบคุณภาพที่ได้จากการจัดกลุ่มโดยวิธีต่าง ๆ ได้ง่าย จากการอ่านเอกสารงานวิจัยนี้และทำความเข้าใจในคุณสมบัติของกลุ่มข้อมูลที่ทำขึ้น เห็นว่ากลุ่มข้อมูลนี้น่าจะสามารถนำมาใช้ในการทดลองด้านการจัดกลุ่มเอกสารได้ดีพอสมควร แต่เนื่องจากไม่สามารถดาวน์โหลดกลุ่มข้อมูลนี้มาใช้ทดลองการจัดกลุ่มได้ เพราะเว็บไซท์ที่เก็บกลุ่มข้อมูลนี้ปิดปรับปรุงชั่วคราวในขณะที่พยายามจะเข้าไปนำกลุ่มข้อมูลมาทดลอง จึงไม่สามารถแสดงผลให้ดูได้ว่า การจัดกลุ่มที่พัฒนาขึ้นในวิทยานิพนธ์เล่มนี้สามารถทำการจัดกลุ่มข้อมูลที่มีอยู่ในกลุ่มข้อมูลมาตราฐานกลุ่มข้อมูลสำหรับการทดลองนี้เป็นผลอย่างไร

2.7 STOP WORD & WORD STEMMING

2.7.1 **การตัดคำที่ไม่มีความหมายกับเนื้อหาหรือคำที่ใช้อยู่ทั่วไปในทุกเอกสาร** (Stop word)

ในเอกสารภาษาอังกฤษแต่ละเอกสารย่อมมีคำอยู่มากบ้างน้อยบ้างแตกต่างกันไป คำที่อยู่ในเอกสารเหล่านี้แสดงให้เห็นถึงเนื้อหาที่เอกสารนั้น ๆ แสดงแก่ผู้อ่านหรือเกี่ยวข้อง เมื่อนำคำที่อยู่ในเอกสารเหล่านี้มาวิเคราะห์ด้วยทฤษฎีและวิธีการที่เหมาะสมย่อมสามารถนำมาใช้ประโยชน์ได้ โดยเฉพาะในด้านการวิเคราะห์เอกสารเพื่อการจัดกลุ่ม ไม่ว่าจะเป็นการจัดกลุ่มโดยมีการกำหนดประเภทและจำนวนกลุ่มที่แน่นอนไว้แล้ว (Test Classification) หรือการจัดกลุ่มที่ไม่มีการกำหนดประเภทและจำนวนกลุ่มไว้ก่อน (Test Clustering) การนำคำทุกคำที่มีอยู่ในเอกสารมาใช้ในการจัดกลุ่มเอกสารนั้น มีข้อดีในกรณีที่เรามีข้อมูลที่เราจะนำมาใช้ในการจัดกลุ่มเอกสารมากกว่าการที่เราจะเลือกใช้คำบางคำหรือบางส่วนของเอกสาร แต่ก็มีข้อเสียเนื่องจากคำที่นำมาใช้นั้นบางคำเป็นคำที่ไม่มีความหมายต่อเนื้อหาของเอกสาร หรือเป็นคำที่ใช้อยู่ทั่วไปในเอกสารต่าง ๆ เช่น he, they, or, if, here, the, in, on, an, yes, no, yesterday เป็นต้น แต่เมื่อนำคำเหล่านี้มาวิเคราะห์เพื่อการจัดกลุ่มเอกสารจะทำให้เกิดความกำกวม และทำให้การจัดกลุ่มเอกสารเกิดความผิดพลาดได้ และการนำคำที่ไม่มีความหมายต่อเนื้อหาของเอกสาร หรือคำที่ใช้อยู่ทั่วไปในเอกสารต่าง ๆ มาใช้ในการจัดกลุ่มเอกสาร จะทำให้การจัดกลุ่มเอกสารใช้เวลานานกว่าการไม่นำคำเหล่านี้มาใช้ เนื่องจากจำนวนคำทุกคำในเอกสารย่อมมากกว่าจำนวนคำที่ได้ทำการคัดเลือกคำที่ไม่จำเป็นต่อการจัดกลุ่มเอกสารออกไป เมื่อจำนวนคำมากกว่าการวิเคราะห์ก็ย่อมใช้เวลามากขึ้นไปด้วย ดังนั้นการจัดกลุ่มเอกสารในงานวิจัยนี้จึงนำการตัดคำที่ไม่มีความหมายกับเนื้อหามาใช้เพื่อแยกคำที่ไม่ต้องการออกจากเอกสารก่อนที่จะทำการจัดกลุ่มเอกสาร ซึ่งส่วนหนึ่งรายการของคำที่ไม่ใช้ในการจัดกลุ่มในงานวิจัยนี้ เป็นส่วนหนึ่งของรายการของคำใน [1] แสดงอยู่ในรูปที่ 2.9 รายการของคำทั้งหมดที่ไม่ใช้ในการจัดกลุ่มในงานวิจัยนี้ สามารถดูได้ที่ภาคผนวก ก.

about	across	after
all	allow	also
although	always	and
any	are	around
because	become	been
before	between	but
could	does	...

รูปที่ 2.9 ตัวอย่างคำทั่วไปที่อยู่ในรายการคำที่ไม่มีความหมายกับเนื้อหา

เนื่องจากงานวิจัยนี้เป็นการจัดกลุ่มเอกสารในเครือข่ายใยแมงมุม ซึ่งมีคำที่ใช้อยู่ทั่วไปในเอกสารซึ่งอยู่ในเครือข่ายใยแมงมุม แต่ไม่เป็นคำที่ใช้อยู่ทั่วไปในเอกสารปกติ จึงได้เพิ่มคำเหล่านี้ เพื่อให้การจัดกลุ่มเอกสารในเครือข่ายใยแมงมุม มีผลที่ดีขึ้น ส่วนหนึ่งของรายการของคำที่ใช้อยู่

ทั่วไปในเอกสารซึ่งอยู่ในเครือข่ายใยแมงมุม แสดงอยู่ในรูปที่ 2.10 รายการทั้งหมดของคำที่ใช้อยู่ทั่วไปในเอกสารซึ่งอยู่ในเครือข่ายใยแมงมุม สามารถดูได้ที่ภาคผนวก ก.

web	page	click
next	link	browse
back	previous	home
www	top	url
enter	login	logout
site	sites	main
copyright	online	...

รูปที่ 2.10 ตัวอย่างคำในเว็บที่อยู่ในรายการคำที่ไม่มีความหมายกับเนื้อหา

2.7.2 การเปลี่ยนรูปของคำให้อยู่ในรูปรากของคำ (Word stemming)

คำในภาษาอังกฤษไม่เหมือนคำในภาษาไทย ตรงที่คำในภาษาอังกฤษจะมีการเปลี่ยนรูปได้หลายลักษณะ เช่นการเปลี่ยนรูปจากรากของคำ เป็นรูปของคำกริยาในเหตุการณ์ที่กำลังกระทำหรือกำลังเกิดขึ้น เช่น run เปลี่ยนรูปเป็น running, sleep เปลี่ยนรูปเป็น sleeping, move เปลี่ยนรูปเป็น moving, laugh เปลี่ยนรูปเป็น laughing เป็นต้น การเปลี่ยนรูปจากรากของคำเป็นรูปของคำกริยาในอดีตการณ์ เช่น catch เปลี่ยนรูปเป็น caught, shoot เปลี่ยนรูปเป็น shot, walk เปลี่ยนรูปเป็น walked, rob เปลี่ยนรูปเป็น robbed เป็นต้น แต่คำกริยาบางคำมีการเปลี่ยนรูปตามกาลเวลาได้ 2 รูป ซึ่งเป็นการเปลี่ยนจากรากของคำเป็นรูปของคำกริยาในอดีตการณ์ และรูปของคำกริยาในอดีตการณ์ที่มีการกระทำต่อเนื่องมาจนถึงปัจจุบัน เช่น drive เปลี่ยนรูปเป็น drove และเปลี่ยนรูปเป็น driven, write เปลี่ยนรูปเป็น wrote และเปลี่ยนรูปเป็น written, go เปลี่ยนรูปเป็น went และเปลี่ยนรูปเป็น gone, see เปลี่ยนรูปเป็น saw และเปลี่ยนรูปเป็น seen เป็นต้น การเปลี่ยนรูปที่มักจะเกิดขึ้นบ่อยครั้งอีกอย่างหนึ่งของคำในภาษาอังกฤษ คือ การเปลี่ยนรูปจากคำนามที่เป็นเอกพจน์เป็นคำนามที่เป็นพหูพจน์ ซึ่งก็มีการเปลี่ยนรูปได้หลากหลายลักษณะ ด้วย เช่น pen เปลี่ยนรูปเป็น pens, umbrella เปลี่ยนรูปเป็น umbrellas, class เปลี่ยนรูปเป็น classes, church เปลี่ยนรูปเป็นchurches, dish เปลี่ยนรูปเป็น dishes, university เปลี่ยนรูปเป็น universities, butterfly เปลี่ยนรูปเป็น butterflies เป็นต้น จะเห็นได้ว่าการเปลี่ยนรูปของคำภาษาอังกฤษในรูปแบบต่าง ๆ ที่ได้แสดงมานี้ ทำให้เกิดความยุ่งยากในการนำคำมาใช้ในการจัดกลุ่มเอกสาร เนื่องจากคำที่เปลี่ยนรูปไปเมื่อนำมาใช้ในการจัดกลุ่มเอกสาร ก็จะเหมือนเป็นคนละคำกับคำเดิมซึ่งยังไม่มีการเปลี่ยนรูป จึงไม่สามารถนำคำที่มีความหมายเดียวกันแต่อยู่คนละรูปมาแสดงความสัมพันธ์ระหว่างเอกสารหรือกลุ่มเอกสารได้ จึงได้นำฟังก์ชันที่ใช้สำหรับเปลี่ยนรูปของคำให้กลับมาอยู่ในรูปเดิม ก่อนที่จะนำคำเหล่านั้นไปใช้ในการจัดกลุ่มเอกสาร โดยในงานวิจัยนี้จะใช้

ฟังก์ชันในการเปลี่ยนรูปเฉพาะคำพหูพจน์เป็นคำเอกพจน์เท่านั้น เนื่องจาก จากการสังเกตุคำที่มีการเปลี่ยนรูปจากฐานข้อมูลที่ใช้ในการทดลองการจัดกลุ่มเอกสาร จะเป็นการเปลี่ยนรูปจากคำเอกพจน์เป็นคำพหูพจน์เป็นส่วนใหญ่ จากงานวิจัยหลาย ๆ งานก่อนหน้านี้ ที่เกี่ยวข้องกับการจัดกลุ่มเอกสาร ก็แสดงให้เห็นว่าการเปลี่ยนรูปของคำให้อยู่ในรูปรากของคำ จะช่วยทำให้ผลการจัดกลุ่มเอกสารมีคุณภาพดีขึ้น

บทที่ 3
การจัดกลุ่มเอกสารโดยใช้ลิงค์และคำ

3.1 วิธีดำเนินการวิจัย

งานวิจัยนี้ดำเนินการวิจัยโดยการตั้งสมมุติฐานเริ่มแรก คือลิงค์ที่อยู่ในเอกสารที่อยู่ในเครือข่ายใยแมงมุม (เอกสาร html) สามารถนาใช้ในการจัดกลุ่มเอกสารได้ โดยได้แนวความคิดมาจากการศึกษางานวิจัย [5] ซึ่งเป็นงานวิจัยที่นำลิงค์ในเอกสารมาใช้ในการจัดลำดับเอกสาร จากนั้นจึงทำการศึกษางานวิจัยที่เกี่ยวข้องกับการจัดกลุ่มเอกสารโดยวิธีต่าง ๆ ซึ่งมีทั้งการจัดกลุ่มเอกสารโดยใช้คำจากคำอธิบายสั้น ๆ ในผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นเป็นข้อมูลในการจัดกลุ่มเอกสาร โดยใช้อัลกอริทึมที่เรียกว่า Hierarchical Agglomerative Clustering (HAC) เป็นอัลกอริทึมในการจัดกลุ่ม และงานวิจัยที่ใช้ URL ของเอกสารเป็นข้อมูล [3] [4] [6] ในการจัดกลุ่มเอกสารที่อยู่ในโดเมนเดียวกันเข้าเป็นกลุ่มเดียวกัน รวมถึงงานวิจัยที่ใช้ลิงค์ในเอกสารเพียงอย่างเดียวเพื่อทำการจัดกลุ่มเอกสาร [7] จากงานวิจัยนี้ทำให้เห็นแนวโน้มในการใช้ลิงค์เพื่อการจัดกลุ่มเอกสารที่มีประสิทธิภาพ และจากงานวิจัยในการจัดกลุ่มเอกสารทั้งหมด ทำให้เห็นข้อดีข้อเสียและจุดบกพร่องต่าง ๆ ในงานวิจัยด้านการจัดกลุ่มเอกสาร

จึงเริ่มทำการพัฒนาระบบจัดกลุ่มเอกสารโดยใช้ลิงค์เป็นข้อมูลในการจัดกลุ่มเอกสาร และใช้อัลกอริทึม HAC เป็นอัลกอริทึมสำหรับการจัดกลุ่มเอกสาร ผลการจัดกลุ่มจากระบบที่พัฒนาขึ้นคือสามารถทำการจัดกลุ่มเอกสารที่มีลิงค์ไปยังเอกสารอื่น ๆ ในกลุ่มเอกสารทดลองได้ดี แต่ไม่สามารถจัดกลุ่มให้แก่เอกสารที่ไม่มีลิงค์ไปยังเอกสารอื่น ๆ ในกลุ่มเอกสารที่ใช้ทดลองได้ ทำให้รู้ว่าในระบบที่พัฒนาขึ้นนี้ข้อมูลลิงค์เพียงอย่างเดียวไม่สามารถทำการจัดกลุ่มเอกสารที่ครอบคลุมเอกสารส่วนใหญ่ได้ จึงเริ่มทำการพัฒนาระบบการจัดกลุ่มเอกสารที่ใช้ข้อมูลลิงค์ร่วมกับคำทุกคำในเอกสาร จากการทดลองพบว่าสามารถเพิ่มความครอบคลุมเอกสารที่อยู่ในกลุ่มท่ลองได้ แต่ประสิทธิภาพที่ได้ยังไม่ดีนัก จึงทำการวิเคราะห์ผลการจัดกลุ่มที่ได้ รวมถึงคำที่เป็นตัวแทนกลุ่มเอกสาร จึงพบความกำกวมของคำที่นำมาใช้ในการจัดกลุ่มเอกสาร ทำให้ต้องนำฟังก์ชันการเปลี่ยนรูปคำจากคำพหูพจน์มาเป็นตำเอกพจน์ (รากของคำ) หรือ Word Stemming, การตัดคำที่ไม่มีผลต่อความหมายของคำ หรือ Stop Word Removal, และการเลือกใช้คำเฉพาะที่จากแท็กหัวข้อเอกสารและแท็กคำอธิบายเอกสาร มาลดความกำกวมของคำ เมื่อทำการทดลองอีกครั้งโดยใช้ระบบที่ลดความกำกวมของคำแล้ว จึงได้ผลการจัดกลุ่มที่มีประสิทธิภาพ และเป็นผลการจัดกลุ่มที่ดีกว่าการจัดกลุ่มโดยใช้คำเพียงอย่างเดียวในงานวิจัย [2]

3.2 แนวความคิดและทฤษฎีที่ใช้ในงานวิจัย

เนื่องจากเห็นความสำคัญของลิงค์ในเอกสาร html ซึ่งสามารถบอกถึงความสัมพันธ์ทางอ้อมระหว่างเอกสาร html ได้ เพราะเอกสารที่มีการอ้างถึงกัน (มีลิงค์ไปหากันหรือมีลิงค์ไปยังเอกสารอื่นเหมือนกัน) แสดงว่ามีความเกี่ยวเนื่องกัน ซึ่งโดยส่วนมากมักจะมีเนื้อหาในเอกสารในด้านเดียวกัน ดังนั้นจึงนำข้อมูลของลิงค์ในเอกสารต่าง ๆ มาเป็นข้อมูลในการจัดกลุ่มเอกสาร เพื่อแสดงเป็นผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น แต่เนื่องจากข้อมูลจากลิงค์ในเอกสารเพียงอย่างเดียว จะทำให้เกิดการจัดกลุ่มของเอกสารได้ไม่เกินร้อยละ 40 - 50 ของจำนวนเอกสารทั้งหมด ทำให้เหลือเอกสารอีกประมาณครึ่งหนึ่งที่ไม่ได้ถูกจัดกลุ่ม เพื่อให้การจัดกลุ่มครอบคลุมเอกสารส่วนใหญ่จึงนำคำในส่วนของแท็กหัวข้อเอกสาร (<title>...</title>) และแท็กคำอธิบายเอกสาร (<meta>...</meta>) ของเอกสารมาใช้ช่วยในการจัดกลุ่ม

ในการรวมผลจากการค้นหาของเสิร์ชเอ็นจิ้น ให้เป็นกลุ่มที่มีเนื้อหาในด้านเดียวกันได้นั้น จะช่วยอำนวยความสะดวกให้ผู้ใช้ในการดูผลการค้นหาข้อมูล ถ้ากลุ่มใดเป็นเนื้อหาที่ไม่ตรงกับความต้องการ ก็ไม่ต้องดูเอกสารทั้งกลุ่มนั้นได้ หรือถ้าพบกลุ่มที่ต้องการก็จะสามารถดูเอกสารในกลุ่มนั้นได้ทั้งหมด จะเห็นได้ว่าเป็นการช่วยประหยัดเวลาให้ผู้ใช้ โดยเป็นการเลือกดูทีละกลุ่มเอกสาร แทนการเลือกดูทีละเอกสาร

โดยการจับกลุ่มของเอกสารนี้จะใช้อัลกอริทึม ที่เรียกว่า Hierarchical Agglomerative Clustering (HAC) ร่วมกับเงื่อนไขในการจัดกลุ่ม ซึ่งเรียกว่า Gobal Quality Funciton (GQF) ดังรูปที่ 3.1

```
Initialize all documents as singleton cluster
    Until (GQF cannot be increased) do {
        Find 2 clusters whose increase GQF the most
        Merge them.
    }
```

รูปที่ 3.1 อัลกอริทึม HAC ในการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำ

โดยฟังก์ชันที่ใช้ในการคำนวณหาคะแนนของกลุ่มที่รวมกันใหม่นั้น จะประกอบด้วยคะแนนจาก 2 ส่วน คือ คะแนนความคล้ายระหว่างกลุ่มที่ได้จากลิงค์ และ คะแนนความคล้ายระหว่างกลุ่มที่ได้จากคำ แล้วจึงนำมารวมกัน โดยใช้ฟังก์ชัน ดังนี้

$$S(c) = \gamma * s1(c) + (1-\gamma) * s2(c) \quad (3.1)$$

โดย γ เป็นค่าระหว่าง 0–1 ใช้เพื่อกำหนดผลที่คะแนน s1(c) และ s2(c) จะมีต่อ S(c)

s1(c) เป็นฟังก์ชันที่ใช้คำนวณหาคะแนนความคล้ายระหว่างกลุ่มที่ได้จากคำ

s2(c) เป็นฟังก์ชันที่ใช้คำนวณหาคะแนนความคล้ายระหว่างกลุ่มที่ได้จากลิงค์

$$s1(c) = |c| \cdot \frac{1 - e^{-\beta h(c)}}{1 + e^{-\beta h(c)}} \tag{3.2}$$

โดย | c | เป็นจำนวนเอกสารในกลุ่ม c

h(c) เป็นจำนวนคำที่มีเหมือนกันระหว่าง 2 cluster

β เป็นความชันของ SIGMOID FUNCTION ซึ่งจะกำหนดค่าที่ดีที่สุดของขนาดและความสัมพันธ์ในแต่ละ cluster ซึ่งช่วยกำหนดความแตกต่างของจำนวนคำ h(c) ที่ต่างกัน

$$s2(c) = \frac{\sum_{x \in c} \alpha^{D(x,y)}}{|c|} \tag{3.3}$$

โดย x เป็นเอกสารที่เป็นสมาชิกของ cluster c

y เป็นเอกสารที่มีลิงค์ไปยัง cluster c หรือถูกลิงค์มาจาก cluster c

α เป็นค่าระหว่าง 0 – 1 ใช้เพื่อกำหนดระยะห่างที่เอกสาร x และ y ยังมีผลต่อกัน

D(x,y) เป็นระยะลิงค์ ระหว่างเอกสาร x กับเอกสาร y

| c | เป็นจำนวนเอกสารในกลุ่ม cluster c

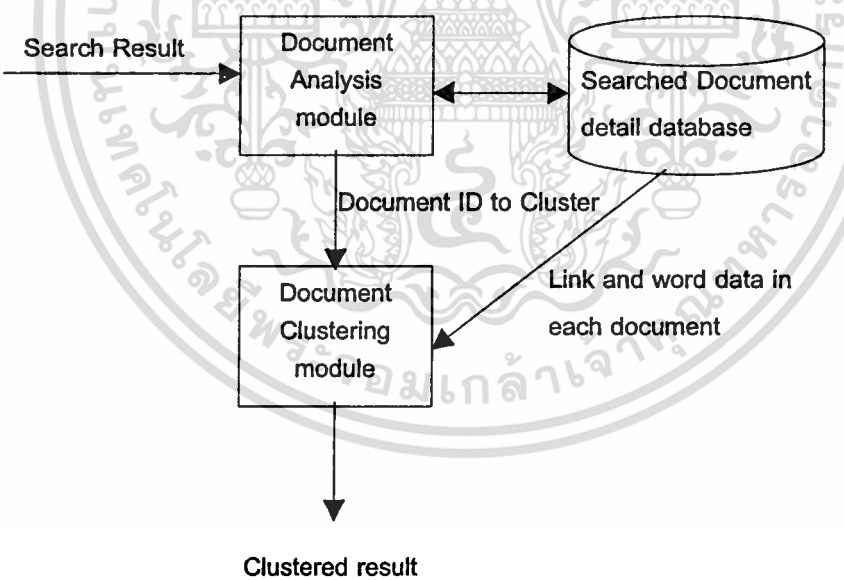
โดยระยะทางระหว่างเว็บเพจ X และเว็บเพจ Y ซึ่งเป็นเว็บเพจที่เป็นผลในการค้นหาข้อมูลของเสิร์ชเอ็นจิ้น จะมีระยะทางเท่ากับ 1 ถ้าเว็บเพจ X มีลิงค์ไปยังเว็บเพจ Y หรือ เว็บเพจ Y มีลิงค์ไปยังเว็บเพจ X หรือเว็บเพจ X และเว็บเพจ Y มีลิงค์ไปยังเว็บเพจ Z เหมือนกัน โดยที่เว็บเพจ Z จะเป็นผลจากการค้นหาข้อมูลของ เสิร์ชเอ็นจิ้นเช่นเดียวกับเว็บเพจ X และเว็บเพจ Y หรือไม่ก็ได้ แต่ระยะทางระหว่างเว็บเพจ X และเว็บเพจ Y จะมีค่าเป็น 0.5 (มีระยะใกล้กว่า 1) ถ้าเว็บเพจ X มีลิงค์ไปยังเว็บเพจ Y และเว็บเพจ Y ก็มีลิงค์กลับมายังเว็บเพจ X ด้วย ซึ่งเป็นสิ่งที่บ่งบอกว่าเว็บเพจ X และเว็บเพจ Y มีความเกี่ยวข้องกันมาก

และฟังก์ชัน Global Quality Function คือ

$$GQF(C) = \frac{f(C)}{g(|C|)} \sum_{c \in C} S(c) \tag{3.4}$$

จากฟังก์ชันนี้ f(C) คือ ฟังก์ชันของอัตราส่วนของเอกสารที่ถูกจัดกลุ่มต่อเอกสารทั้งหมด ส่วน g(|C|) เป็นฟังก์ชันที่เพิ่มค่าตามจำนวนของ cluster ที่มีขนาดของตั้งแต่ 2 ขึ้นไปแล้วยกกำลังด้วย 0.5 และ $\sum_{c \in C} S(c)$ เป็นผลรวมคะแนนของ cluster ทุกๆ cluster c ซึ่งเป็นสมาชิกใน C โดย S(c) เป็นคะแนนของแต่ละ cluster c จะเห็นได้ว่า f(C) จะเป็นตัวชักจูงให้มีเอกสารที่ถูกรวมกลุ่มมากขึ้นเรื่อยๆ เนื่องจากเมื่อเอกสารถูกรวมกลุ่มมากขึ้นค่าของ f(C) ก็จะมากขึ้นทำให้คะแนนที่คำนวณได้จาก GQF เพิ่มขึ้นด้วย ในขณะที่ g(|C|) จะพยายามทำให้กลุ่มเอกสารที่มีขนาดเล็กมีจำนวนน้อยลง โดยเมื่อจำนวนกลุ่มที่ได้จากการจัดกลุ่มมีมาก ค่าของ g(|C|) ก็จะมากตามทำให้ค่า GQF ลดลง แต่ถ้ากลุ่มเอกสารที่ได้จากการจัดกลุ่มมีการรวมกัน จะทำให้จำนวนกลุ่มเอกสารลดลง มีผลให้ค่าของ g(|C|) ลดลง ดังนั้นคะแนน GQF ก็จะเพิ่มขึ้น

เมื่อจะทำการรวมกลุ่มของเอกสาร 2 กลุ่มเข้าด้วยกันเป็นกลุ่มเอกสารกลุ่มใหม่นั้น คำที่เป็นตัวแทนของกลุ่มเอกสารทั้ง 2 กลุ่มจะได้เป็นตัวแทนของเอกสารกลุ่มใหม่ก็ต่อเมื่อ คำๆนั้นมีอยู่ในเอกสาร จำนวน 50% ของเอกสารในกลุ่มใหม่ ทั้งนี้เพื่อไม่ให้เสียคำที่เป็นตัวแทนกลุ่มที่ดีแต่ไม่มีอยู่ในทุกเอกสาร ซึ่งจะช่วยเพิ่มโอกาสในการจัดกลุ่มได้มากกว่ากลุ่มเอกสารที่มีคำจำนวนน้อยๆ เป็นตัวแทนกลุ่ม กระบวนการของการจัดกลุ่มโดยใช้ลิงค์และคำแสดงอยู่ในรูปที่ 3.2



รูปที่ 3.2 แผนผังแสดงขั้นตอนการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำ

3.3 การเตรียมข้อมูลสำหรับการทดลอง

3.3.1 ลักษณะข้อมูล การเลือกข้อมูล และเหตุผลในการเลือกข้อมูล

งานวิจัยนี้เป็นงานวิจัยเพื่อการจัดกลุ่มเอกสารในเครือข่ายใยแมงมุม ดังนั้นการเตรียมข้อมูลสำหรับการทดลองจึงต้องมีการเตรียมเอกสาร ซึ่งเป็นเอกสารที่อยู่ในรูปแบบของเอกสารในเครือข่ายใยแมงมุม หรือเรียกอีกอย่างว่า เอกสารในรูปแบบ html โดยเอกสารในรูปแบบ html นี้สามารถหาได้ทั่วไปในเว็บไซท์ที่อยู่ในเครือข่ายใยแมงมุม ซึ่งก็คือ เว็บเพจแต่ละหน้าในเว็บไซท์นั่นเอง แต่การนำเว็บเพจหรือเอกสาร html ใด ๆ ในเครือข่ายใยแมงมุมมาใช้ทำการทดลองการจัดกลุ่มเอกสาร จะทำให้การวิเคราะห์ผลการทดลอง และการวัดคุณภาพการจัดกลุ่มเอกสารทำได้ยาก เนื่องจากเราไม่รู้ว่าเอกสาร html เหล่านี้เอกสารใดบ้างควรจะอยู่กลุ่มเดียวกัน และเอกสารใดอยู่คนละกลุ่มกัน แม้ว่าเราจะทำการกำหนดกลุ่มให้เอกสารเหล่านี้ด้วยตัวเอง โดยการตรวจสอบเอกสารแต่ละเอกสาร เพื่อดูเนื้อหาและกำหนดกลุ่มให้แก่เอกสารเหล่านี้ แต่การทำเช่นนี้ก็จำเป็นต้องใช้เวลามาก เนื่องจากเอกสารที่นำมาทดลองนั้นต้องมีจำนวนมากพอสมควร ซึ่งมีจำนวนหลายร้อยหรือหลายพันเอกสาร การใช้เอกสารจำนวนมากก็เพื่อให้กลุ่มเอกสารที่ใช้ในการทดลองมีความหลากหลาย และทำให้แน่ใจว่าวิธีการจัดกลุ่มเอกสารที่พัฒนาขึ้นในงานวิจัย จะสามารถนำไปใช้จัดกลุ่มเอกสารในเครือข่ายใยแมงมุมได้จริง นอกจากเวลาที่ต้องใช้แล้ว ความถูกต้องของการกำหนดกลุ่มให้เอกสารแต่ละเอกสารก็มีความสำคัญ เพราะการกำหนดกลุ่มให้เอกสารนี้จะมีผลกระทบไปถึงการวิเคราะห์และการวัดคุณภาพของผลการจัดกลุ่มเอกสารของระบบที่ทำการวิจัยและพัฒนาขึ้นด้วย ถ้าทำการกำหนดกลุ่มให้เอกสารผิด แม้การจัดกลุ่มเอกสารจะได้ผลที่ถูกต้อง แต่ก็อาจจะตีความไปว่าการจัดกลุ่มเอกสารของระบบที่พัฒนาขึ้นนั้นทำการจัดกลุ่มผิดก็ได้ เนื่องจากเราไม่รู้ว่าเราทำการกำหนดกลุ่มให้เอกสารผิด หรือถ้าเรากำหนดกลุ่มให้เอกสารผิด แล้วระบบของเราก็ทำการจัดกลุ่มเอกสารผิดตามที่เรากำหนดกลุ่มเอกสาร ก็จะทำให้เราไม่รู้ว่าระบบของเรานั้นมีการจัดกลุ่มเอกสารที่ผิดพลาด ซึ่งเรื่องการกำหนดกลุ่มให้แก่เอกสารด้วยตัวเองของผู้ที่ทำการวิจัยเรื่องการจัดกลุ่มเอกสารนี้ จะเป็นสาเหตุสำคัญประการหนึ่งที่ผู้อื่นอาจตั้งข้อสงสัยในความถูกต้องของข้อมูลที่ใช้ในการทำการทดลองการจัดกลุ่ม และระบบที่ใช้ในการจัดกลุ่มด้วย ซึ่งจะเป็นเหตุให้ความน่าเชื่อถือของงานวิจัยมีน้อยลงไปด้วย

ด้วยเหตุผลที่ได้แสดงมาข้างต้นและความสำคัญของข้อมูลที่นำมาใช้สำหรับทดลองการจัดกลุ่มเอกสาร งานวิจัยนี้จึงนำเอกสารที่ได้รับการจัดกลุ่มแล้วโดย Yahoo! [9] มาใช้เป็นข้อมูลเพื่อใช้ในการทดลองการจัดกลุ่มเอกสาร เนื่องจาก Yahoo! [9] เป็นเว็บไซท์ที่มีการจัดทำฐานข้อมูลของเว็บไซท์ต่างๆในเครือข่ายใยแมงมุม โดยแยกเป็นหมวดหมู่ ซึ่งเรียกว่า Web Directory โดยการแยกเอกสารหรือเว็บไซท์เป็นหมวดหมู่หรือเรียกอีกอย่างว่าการกำหนดกลุ่มให้แก่เอกสารหรือเว็บไซท์ของ Yahoo! นั้นจะทำการจัดหมวดหมู่โดยผู้เชี่ยวชาญ จึงมีความน่าเชื่อถือในความถูกต้อง และยังมีกลุ่มเอกสารที่ครอบคลุมเนื้อหาในด้านต่าง ๆ อย่างครบถ้วน อีกทั้ง Yahoo! ก็เป็นเว็บไซท์ที่ได้รับความนิยม เพื่อใช้ในการค้นหาข้อมูลบนอินเตอร์เนท จึงเป็นแหล่งข้อมูลที่น่าเชื่อถือ

ที่จะนำมาใช้ในการทดลองการจัดกลุ่มเอกสารได้ดี โดยการจัดกลุ่มของ Yahoo! [9] จะทำการเก็บ URL ของแต่ละเว็บเพจไว้ในกลุ่มที่กำหนดขึ้น ซึ่งในแต่ละกลุ่มก็จะมีการแบ่งเป็นกลุ่มย่อยอยู่ภายใน และบางกลุ่มย่อยก็เป็นการเชื่อมโยงไปยังกลุ่มย่อยอื่น ซึ่งหมายความว่า กลุ่มย่อยบางกลุ่มสามารถเป็นสมาชิกในหลายๆกลุ่มใหญ่ได้ กลุ่มลำดับแรกที่ Yahoo! กำหนดนั้นมีอยู่ 14 กลุ่ม ประกอบด้วย Business & Economy, Computer & Internet, News & Media, Entertainment, Recreation & Sports, Health, Government, Regional, Society & Culture, Education, Art & Humanities, Science, Social Science และ Reference จากกลุ่มหลักทั้ง 14 กลุ่มนี้แสดงให้เห็นว่า Yahoo! [9] สามารถจัดหมวดหมู่เพื่อครอบคลุมเอกสาร ได้แทบทุกประเภท

3.3.2 **ขั้นตอนในการรวบรวมข้อมูล**

การนำข้อมูลเอกสาร html ที่ได้รับการจัดกลุ่มโดย Yahoo! มาใช้นั้น เราทำโดยเข้าไปยังกลุ่มที่ Yahoo! [9] กำหนดขึ้นและเลือกกลุ่มที่มีขนาดใหญ่ซึ่งก็คือ มีเอกสารจำนวนมาก อยู่ในกลุ่ม แล้วจึงทำการโหลด เอกสารเว็บเพจที่เป็นสมาชิกในกลุ่มนั้น ตาม URL ของเอกสารเว็บเพจนั้นมาเก็บไว้ในเครื่องคอมพิวเตอร์ที่จะใช้ทำการทดลอง การโหลดเอกสารนั้นจะหลีกเลี่ยงการโหลดเอกสารที่อยู่ในกลุ่มของเอกสารที่เป็นการลิงค์ไปยังกลุ่มเอกสารอื่น เพื่อหลีกเลี่ยงความสับสนในการจัดกลุ่มเอกสาร และให้เกิดความแน่ใจว่าเอกสารหนึ่ง ๆ จะเป็นสมาชิกของกลุ่มเอกสารเดียวเท่านั้น อันจะทำให้การวิเคราะห์ผลการจัดกลุ่มเอกสารของระบบที่พัฒนาขึ้นทำได้ง่ายและถูกต้องเนื่องจากจะสามารถบอกได้แน่นอนว่าเอกสารเว็บเพจนั้นๆเป็นสมาชิกของกลุ่มใด โดยประเภทและประเภทย่อยของเอกสารที่ดาวน์โหลดมาใช้ในกลุ่มข้อมูลสำหรับการจัดกลุ่มประกอบด้วย

1. กลุ่มข้อมูลประเภท ARTS โดยข้อมูลประเภทย่อยที่เก็บคือ design ·arts และ museums galleries and centers

2. กลุ่มข้อมูลประเภท BUSINESS AND ECONOMY โดยข้อมูลประเภทย่อยที่เก็บคือ employment and work, finance and investment, และ trade

3. กลุ่มข้อมูลประเภท COMPUTERS AND INTERNET โดยข้อมูลประเภทย่อยที่เก็บคือ communications and networking, hardware, internet, multimedia, และ software

4. กลุ่มข้อมูลประเภท ENTERTAINMENT โดยข้อมูลประเภทย่อยที่เก็บคือ amusement and theme parks, comics and animation และ virtual cards

5. กลุ่มข้อมูลประเภท RECREATION โดยข้อมูลประเภทย่อยที่เก็บคือ sports และ travel

การเก็บเอกสารที่ดาวน์โหลดตาม URL ที่อยู่ในกลุ่มต่าง ๆ ของเว็บ Yahoo! นั้น เมื่อดาวน์โหลดมาแล้วก็จะทำการเก็บแยกไดเร็กทอรี่ของแต่ละกลุ่ม การไม่เก็บเอกสารที่อยู่คนละกลุ่ม

มารวมอยู่ในไดเร็กทอรี่เดียวกันนั้น ก็เพื่อให้เกิดความสะดวกเมื่อต้องการนำเอกสารเหล่านั้นมาใช้ทดลองการจัดกลุ่มเอกสาร ข้อมูล URL ของแต่ละเอกสารเว็บเพจก็มีการเก็บไว้เพื่อเป็นข้อมูลที่สำคัญส่วนหนึ่งสำหรับการจัดกลุ่มเอกสาร เมื่อทำการดาวน์โหลดเอกสารเว็บเพจมาเก็บไว้ในเครื่องคอมพิวเตอร์แล้วจึงทำการเก็บคำในส่วนของ title, คำในส่วนของ meta, URL ของแต่ละเอกสารเว็บเพจ และ URL ในแต่ละเอกสารเว็บเพจ เข้าสู่ฐานข้อมูล จะเห็นได้ว่าคำในส่วนของแท็กหัวข้อเอกสาร (<title>) และคำในส่วนของแท็กคำอธิบายเอกสาร (<meta>) เท่านั้นที่ถูกเก็บไว้ในฐานข้อมูล เนื่องจากเราเห็นว่าคำจากสองส่วนนี้เพียงพอที่จะใช้ในการจัดกลุ่มเอกสาร และเป็นการลดความกำกวมของคำที่จะใช้จัดกลุ่มเอกสาร โดยหลีกเลี่ยงการใช้คำในส่วนของแท็กเนื้อความของเอกสาร (<body>) เมื่อไม่ใช้คำจากส่วนของแท็กเนื้อความของเอกสาร ทำให้จำนวนคำน้อยลงมีผลให้การจัดกลุ่มเอกสารมีความรวดเร็วขึ้น ข้อมูลของคำในแต่ละเว็บเพจที่เก็บเข้าฐานข้อมูลนั้นจะเก็บจำนวนที่คำๆนั้นมีอยู่ในเว็บเพจนั้นๆและยังมีการตัดคำที่ไม่มีความหมายกับเนื้อหาในเอกสาร (Stop Word) ออก รวมถึงการเปลี่ยนรูปของคำพหูพจน์ให้อยู่ในรูปของคำเอกพจน์ (Word Stemming) เพื่อลดความผันผวนของคำที่อยู่ในเอกสาร ทำให้คำที่มีอยู่ในฐานข้อมูล มีคุณภาพเพียงพอสำหรับนำไปใช้ในการจัดกลุ่มเอกสาร ในส่วนของฐานข้อมูล URL ในแต่ละเว็บเพจจะมีการตัด URL ที่ไม่มีความหมายต่อเว็บเพจออกด้วยเช่นกัน เช่น URL ที่เป็นการโฆษณา และ URL ของเว็บไซท์สำหรับให้บริการนับการเข้าใช้เว็บไซท์แก่เว็บไซท์ต่าง ๆ (counter) เป็นต้น

บทที่ 4
ผลการทดลอง

4.1 ขั้นตอนของการทดลองจัดกลุ่มเอกสาร

1. เตรียมข้อมูลในฐานข้อมูลของคำและลิงค์สำหรับกลุ่มเอกสารที่จะทำการจัดกลุ่ม
2. เตรียมข้อมูลในฐานข้อมูลหมายเลขเอกสารที่จะทำการจัดกลุ่มเอกสาร
3. คำนวณระยะทางที่สั้นที่สุดระหว่างเอกสาร จากข้อมูลการลิงค์กันระหว่างอกสาร
4. เตรียมพื้นที่สำหรับเก็บผลจากการจัดกลุ่มเอกสาร
5. กำหนดวิธีการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำ หรือใช้คำอย่างเดียว
6. ทำการจัดกลุ่มเอกสาร

4.2 การเปรียบเทียบประสิทธิภาพการจัดกลุ่มเอกสาร

4.2.1 ฟังก์ชันสำหรับวัดคุณภาพการจัดกลุ่มเอกสาร

ในการทดลองเพื่อทดสอบคุณภาพการจัดกลุ่มของวิธีการที่พัฒนาขึ้นนั้น กลุ่มของเอกสารที่นำมาจัดกลุ่มจะได้มาจาก การผสมกันระหว่างเอกสารคนละประเภท ซึ่งแบ่งประเภทโดย *Yahoo!* [9] ประเภทละ 50 หรือ 100 เอกสารเท่า ๆ กัน และจะเป็นการผสมเอกสาร ตั้งแต่ 2 ถึง 3 ประเภท เมื่อทำการจัดกลุ่มเสร็จแล้ว จึงต้องมีการคำนวณหาคุณภาพของการจัดกลุ่มเพื่อใช้ในการเปรียบเทียบคุณภาพ โดยฟังก์ชันที่ใช้ในการคำนวณคุณภาพนี้ จะคำนวณคุณภาพของการจัดกลุ่มจากตัวแทนที่ใหญ่ที่สุดของแต่ละกลุ่มเอกสารที่ได้จากการจัดกลุ่ม ดังนี้

$$Q(C)=\frac{\sum_{c\in C}\left(\sqrt{t(c)}-\sqrt{w(c)}\right)}{\sum_{c\in B}\sqrt{\binom{|c|}{2}}} \qquad (4.1)$$

โดย B เป็นเซตของกลุ่มเอกสารซึ่งอยู่ในข้อมูลที่นำมาทำการจัดกลุ่ม

C เป็นเซตของตัวแทนที่ใหญ่ที่สุดของแต่ละกลุ่มเอกสาร ที่ได้จากการจัดกลุ่ม

t(c) เป็นจำนวนคู่ของเอกสารที่ถูกจับอยู่ในกลุ่มเดียวกันอย่างถูกต้อง

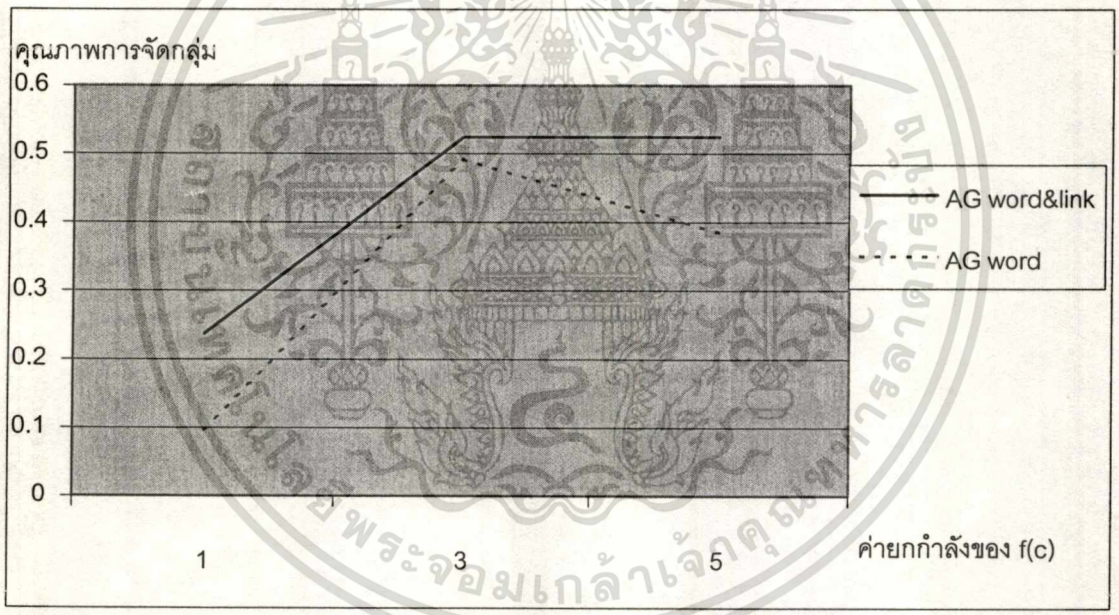
w(c) เป็นจำนวนคู่ของเอกสารที่ถูกจับกลุ่มผิดพลาดมาอยู่ในกลุ่มเดียวกัน

|c| เป็นจำนวนเอกสารในกลุ่มเอกสาร c

4.2.2 ผลการจัดกลุ่มเอกสาร

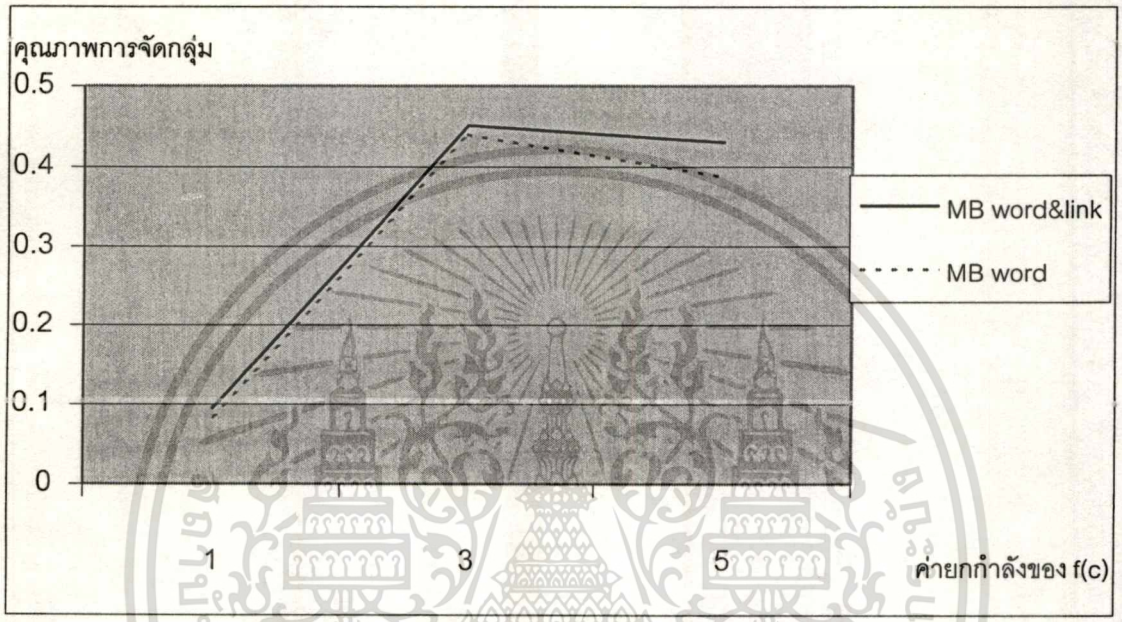
จากผลการทดลอง แสดงให้เห็นว่าผลการจัดกลุ่มเอกสารโดยใช้คำและลิงค์ จะให้ผลที่ดีกว่าการจัดกลุ่มเอกสารโดยใช้คำแต่เพียงอย่างเดียว และคุณภาพการจัดกลุ่มจะดีที่สุดเมื่อให้ฟังก์ชัน f(c) เป็นฟังก์ชันยกกำลัง 3 ซึ่ง f(c) เป็นฟังก์ชันย่อยในฟังก์ชัน GQF(C) โดยดูได้จากกราฟแสดงคุณภาพการจัดกลุ่มเอกสารที่ได้จากผลการจัดกลุ่มเอกสาร

โดยชุดข้อมูลในรูปที่ 4.1 เป็นชุดข้อมูล AG100 ซึ่งมาจากกลุ่มข้อมูล 2 กลุ่มคือ กลุ่มข้อมูล A คือกลุ่มข้อมูลด้านสถาปัตยกรรม (ARCHITECTURE) และกลุ่มข้อมูล G คือกลุ่มข้อมูลด้านการออกแบบกราฟฟิก (GRAPHIC DESIGN) โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร ซึ่งทั้ง 2 กลุ่มข้อมูลนี้เป็นกลุ่มข้อมูลย่อยในกลุ่มข้อมูลหลักด้านศิลปะ (ART) ที่ทำการจัดกลุ่มโดย *Yahoo!* [9] การจัดกลุ่มเอกสารในชุดข้อมูล AG นี้แสดงให้เห็นว่า การจัดกลุ่มเอกสารโดยใช้คำและลิงค์สามารถจัดกลุ่มเอกสารที่มีเนื้อหาในด้านเดียวกันได้ดี



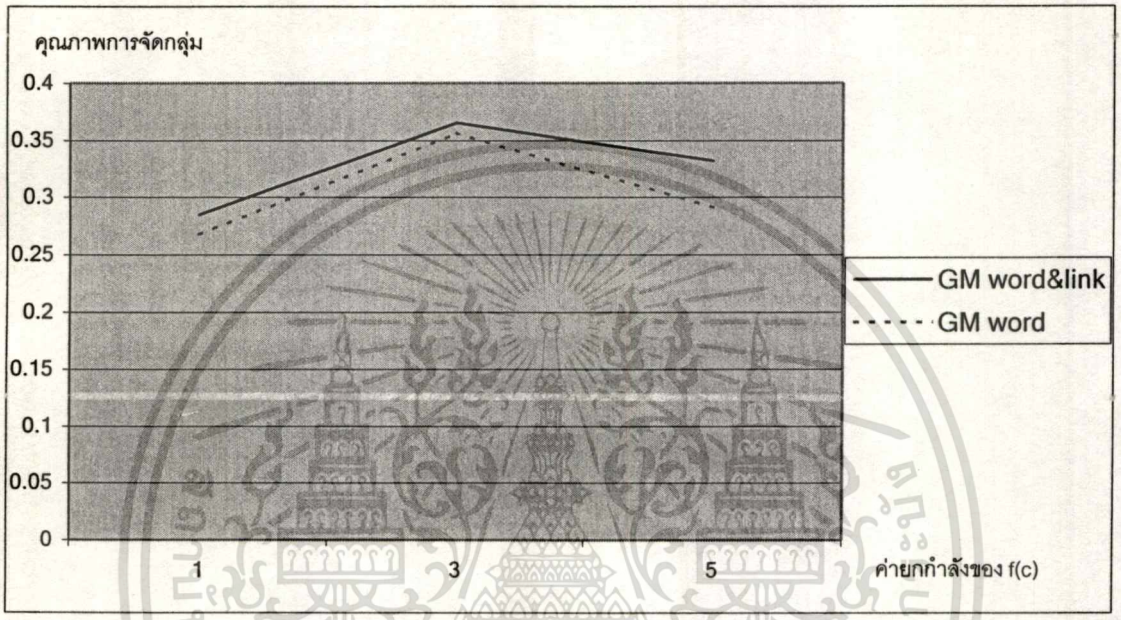
รูปที่ 4.1 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล AG100 จำนวน 100 เอกสาร

ชุดข้อมูลในรูปที่ 4.2 เป็นชุดข้อมูล MB100 ซึ่งมาจากกลุ่มข้อมูล 2 กลุ่มคือ กลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านศิลปะ (ART) และกลุ่มข้อมูล B คือกลุ่มข้อมูลด้านตำแหน่งงาน (JOB) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านธุรกิจ (BUSINESS) ชุดข้อมูลนี้จะแสดงให้เห็นถึงคุณภาพการจัดกลุ่มเอกสารที่มาจากกลุ่มข้อมูลที่มีเนื้อหาคนละด้านกัน โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร



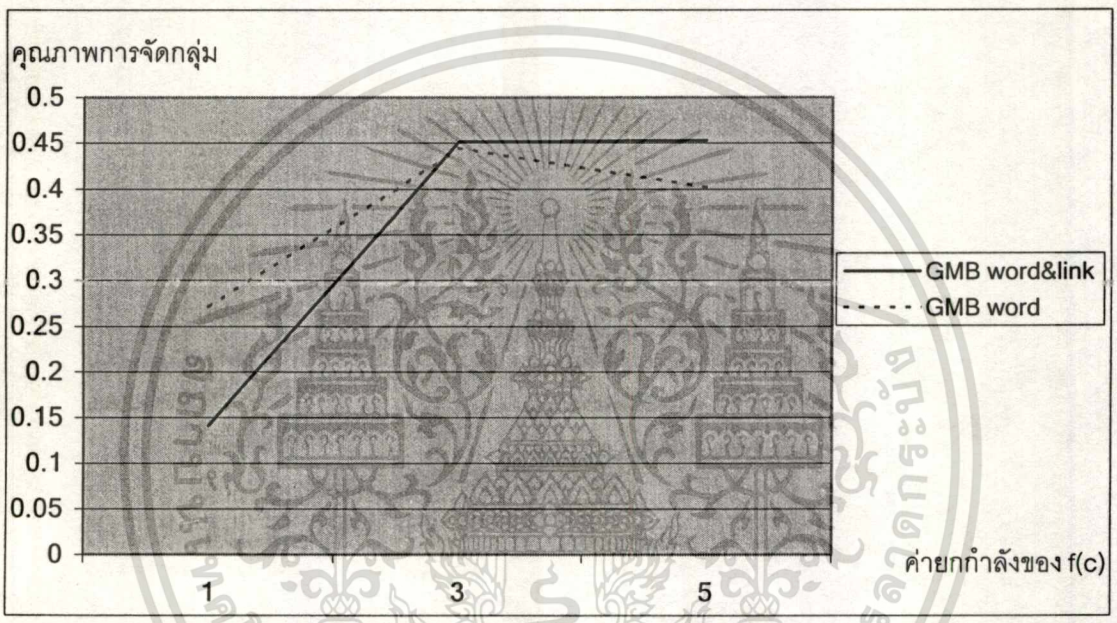
รูปที่ 4.2 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล MB100 จำนวน 100 เอกสาร

ชุดข้อมูลในรูปที่ 4.3 เป็นชุดข้อมูล GM100 ซึ่งมาจากกลุ่มข้อมูล 2 กลุ่มคือ กลุ่มข้อมูล G คือกลุ่มข้อมูลด้านการออกแบบกราฟฟิก (GRAPHIC DESIGN) และกลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านศิลปะ (ART) ทั้งสองกลุ่ม ดังนั้นชุดข้อมูลนี้เป็นอีกชุดหนึ่งที่จะแสดงให้เห็นถึงคุณภาพการจัดกลุ่มเอกสารที่มาจากกลุ่มข้อมูลที่มีเนื้อหาในด้านเดียวกันด้านกัน โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร



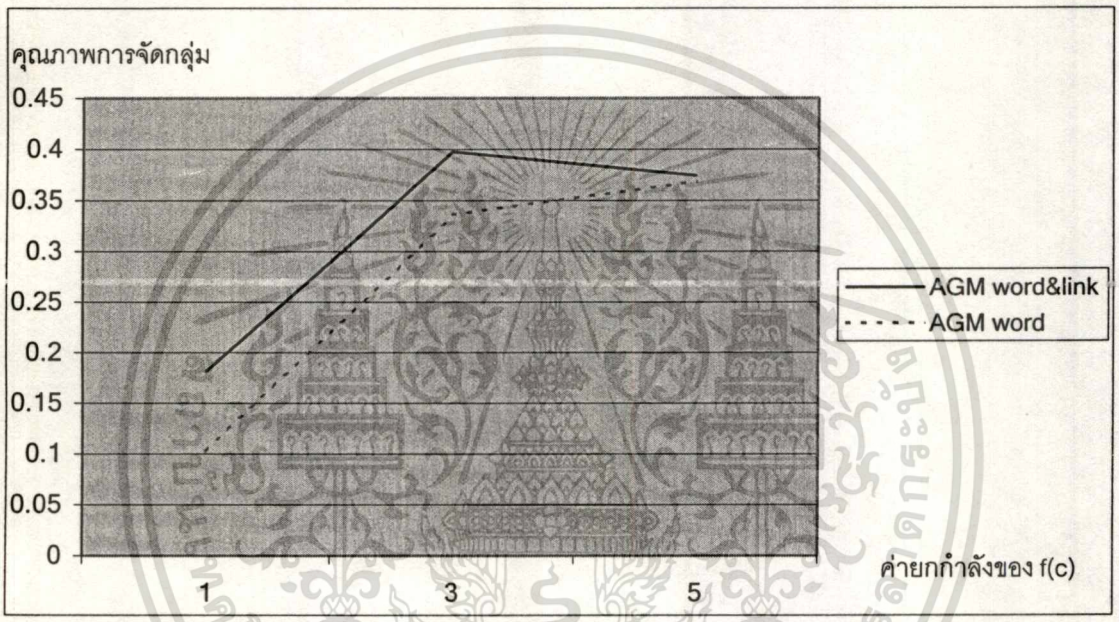
รูปที่ 4.3 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล GM100 จำนวน 100 เอกสาร

ชุดข้อมูลในรูปที่ 4.4 เป็นชุดข้อมูล GMB150 ซึ่งมาจากกลุ่มข้อมูล 3 กลุ่มคือ กลุ่มข้อมูล G คือกลุ่มข้อมูลด้านการออกแบบกราฟฟิก (GRAPHIC DESIGN) กลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านศิลปะ (ART) ทั้งสองกลุ่มข้อมูล และกลุ่มข้อมูล B คือกลุ่มข้อมูลด้านตำแหน่งงาน (JOB) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านธุรกิจ (BUSINESS) ชุดข้อมูลนี้เป็นชุดข้อมูลที่นำกลุ่มข้อมูลที่มีเนื้อหาในด้านเดียวกันด้านกันสองกลุ่มมารวมกับกลุ่มข้อมูลที่มีเนื้อหาด้านอื่น เพื่อทดสอบการจัดกลุ่มข้อมูลที่มีความซับซ้อนมากขึ้น โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร



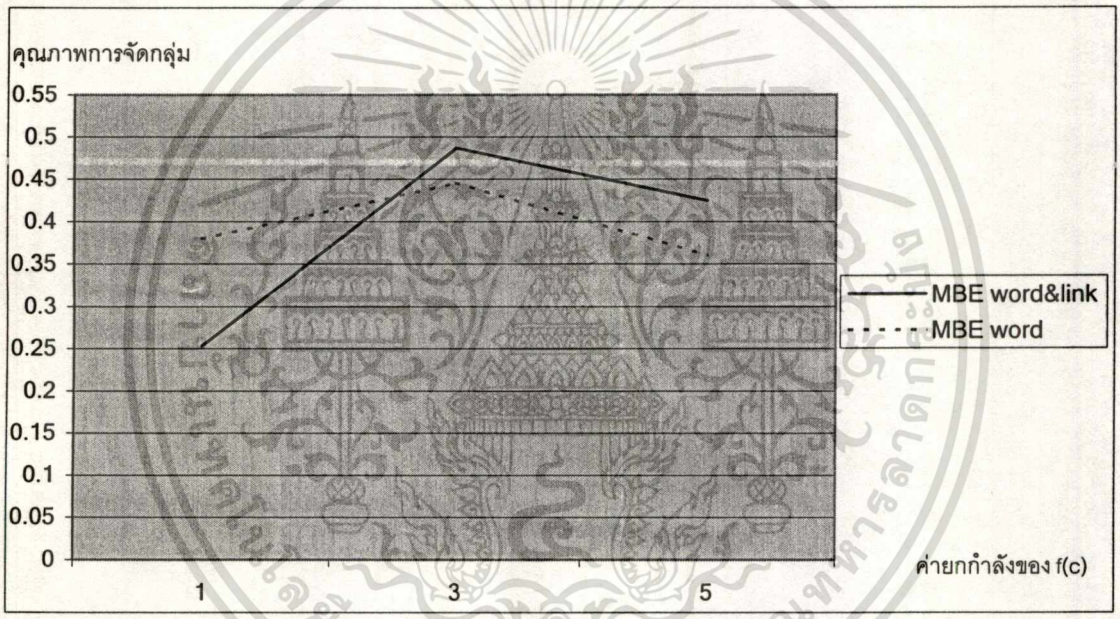
รูปที่ 4.4 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล GMB150 จำนวน 150 เอกสาร

ชุดข้อมูลในรูปที่ 4.5 เป็นชุดข้อมูล AGM150 ซึ่งมาจากกลุ่มข้อมูล 3 กลุ่มคือ กลุ่มข้อมูล A คือกลุ่มข้อมูลด้านสถาปัตยกรรม (ARCHITECTURE) กลุ่มข้อมูล G คือกลุ่มข้อมูลด้านการออกแบบกราฟฟิก (GRAPHIC DESIGN) และกลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) ซึ่งอยู่ในกลุ่มข้อมูลหลักด้านศิลปะ (ART) ทั้งหมด ชุดข้อมูลนี้เป็นชุดข้อมูลที่นำกลุ่มข้อมูลที่มีเนื้อหาในด้านเดียวกันด้านกันสองกลุ่ม มารวมกับกลุ่มข้อมูลที่มีเนื้อหาด้านอื่น เพื่อทดสอบการจัดกลุ่มข้อมูลที่มีความซับซ้อนยิ่งกว่าชุดข้อมูล GMB150 เพราะต้องทำ การจัดกลุ่มข้อมูลที่มีเนื้อหาในด้านเดียวกันทั้งกลุ่ม โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร



รูปที่ 4.5 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล AGM150 จำนวน 150 เอกสาร

ชุดข้อมูลในรูปที่ 4.6 เป็นชุดข้อมูล MBE150 ซึ่งมาจากกลุ่มข้อมูล 3 กลุ่มคือ กลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) กลุ่มข้อมูล B คือกลุ่มข้อมูลด้านตำแหน่งงาน (JOB) และกลุ่มข้อมูล E คือกลุ่มข้อมูลด้านบัตรอวยพร (VIRTUAL CARDS) ซึ่งทั้ง 3 กลุ่มข้อมูลอยู่ในกลุ่มข้อมูลหลักคนละด้านกันดังนี้ ด้านศิลปะ (ART) ด้านธุรกิจ (BUSINESS) และด้านความบันเทิง (ENTERTAINMENT) ตามลำดับ จะเห็นได้ว่าชุดข้อมูลนี้เป็นชุดข้อมูลที่นำกลุ่มข้อมูลที่มีเนื้อหาคนละด้านกันทั้งสามกลุ่มมารวมกัน เพื่อทดสอบการจัดกลุ่มข้อมูลที่มีความซับซ้อนมากกว่าชุดข้อมูล MB100 เนื่องจากมีจำนวนข้อมูลมากกว่าถึงแม้จะมีเนื้อหาในด้านเดียวกันทุกกลุ่มข้อมูล และมีความซับซ้อนน้อยกว่าการจัดกลุ่มข้อมูลของชุดข้อมูล AGM150 และ GMB150 โดยมีเอกสารในแต่ละกลุ่มจำนวน 50 เอกสาร

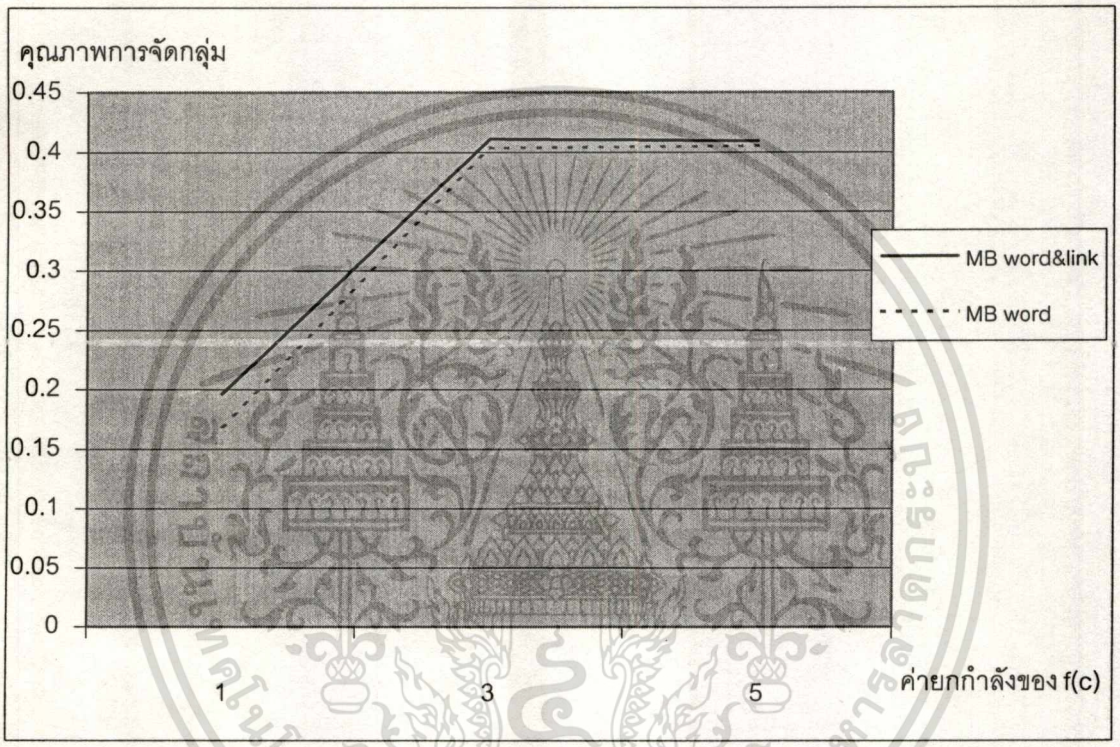


รูปที่ 4.6 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล MBE150 จำนวน 150 เอกสาร

โดยชุดข้อมูลในรูปที่ 4.7 เป็นชุดข้อมูล AG200 ซึ่งมาจากกลุ่มข้อมูล 2 กลุ่มคือ กลุ่มข้อมูล A คือกลุ่มข้อมูลด้านสถาปัตยกรรม (ARCHITECTURE) และกลุ่มข้อมูล G คือกลุ่มข้อมูลด้านการออกแบบกราฟฟิก (GRAPHIC DESIGN) เช่นเดียวกับชุดข้อมูล AG100 แต่ต่างกันที่มีเอกสารในแต่ละกลุ่มจำนวน 100 เอกสาร ซึ่งทั้ง 2 กลุ่มข้อมูลนี้เป็นกลุ่มข้อมูลย่อยในกลุ่มข้อมูลหลักด้านศิลปะ (ART) จะเห็นได้ว่าคุณภาพการจัดกลุ่มที่ได้ น้อยกว่าชุดข้อมูล AG100 ซึ่งมีจำนวนเอกสารน้อยกว่า

รูปที่ 4.7 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล AG200 จำนวน 200 เอกสาร

โดยชุดข้อมูลในรูปที่ 4.8 เป็นชุดข้อมูล MB200 ซึ่งมาจากกลุ่มข้อมูล 2 กลุ่มคือ กลุ่มข้อมูล M คือกลุ่มข้อมูลด้านพิพิธภัณฑ์ (MUSEUM) และกลุ่มข้อมูล B คือกลุ่มข้อมูลด้านตำแหน่งงาน (JOB) เช่นเดียวกับชุดข้อมูล MB100 แต่ต่างกันที่มีเอกสารในแต่ละกลุ่มจำนวน 100 เอกสาร ซึ่งทั้ง 2 กลุ่มข้อมูลนี้เป็นกลุ่มข้อมูลที่มีเนื้อหาคนละด้านกัน คือด้านศิลปะ (ART) และด้านธุรกิจ (BUSINESS) ตามลำดับ จะเห็นได้ว่าคุณภาพการจัดกลุ่มที่ได้ น้อยกว่าชุดข้อมูล MB100 ซึ่งมีจำนวนเอกสารน้อยกว่า



รูปที่ 4.8 กราฟเปรียบเทียบคุณภาพการจัดกลุ่มเอกสารเฉลี่ยของชุดข้อมูล MB200 จำนวน 200 เอกสาร

4.3 การวิเคราะห์ผลการจัดกลุ่มเอกสาร

เมื่อนำผลที่ได้จากการทดลองจัดกลุ่มเอกสารมาวิเคราะห์ ทำให้ทราบถึงปัจจัยที่มีผลต่อคุณภาพการจัดกลุ่มเอกสาร และรู้ถึงเหตุผลที่คุณภาพการจัดกลุ่มต่างกันแม้จะทำการจัดกลุ่มในชุดข้อมูลมีขนาดเท่ากัน ดังนี้

1. เมื่อเทียบคุณภาพการจัดกลุ่มเอกสารของชุดข้อมูลที่มีขนาด 100 เอกสารเท่าๆกัน คือ AG100, GM100, และ MB100 จะเห็นได้ว่า ชุดข้อมูล AG100 มีคุณภาพการจัดกลุ่มเอกสารดีที่สุดแม้จะเป็นชุดข้อมูลที่ได้มาจากการรวมกลุ่มข้อมูลย่อยที่อยู่ในกลุ่มข้อมูลหลัก ART เหมือนกัน ซึ่งมากกว่าชุดข้อมูล MB100 ที่มีคุณภาพเป็นอันดับสอง แม้ชุดข้อมูล MB100 จะเป็นการรวมกลุ่ม

ข้อมูลย่อยที่มาจากคนละกลุ่มข้อมูลหลัก ซึ่งทำการจัดกลุ่มได้ง่ายกว่า และ GM100 มีคุณภาพสูงเป็นอันดับที่สาม เหตุผลก็คือ เมื่อทำการตรวจสอบจำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กันภายในชุดข้อมูล จะได้ผลดังตารางที่ 4.1

ตารางที่ 4.1 ข้อมูลจำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กันในชุดข้อมูลจำนวน 100 เอกสาร

ชุดข้อมูล	จำนวนคำเฉลี่ย	จำนวนเอกสารที่ลิงค์กัน
AG100	22 คำ / 1เอกสาร	49 เอกสาร
GM100	21 คำ / 1เอกสาร	25 เอกสาร
MB100	27 คำ / 1เอกสาร	5 เอกสาร

ทำให้รู้ว่า ชุดข้อมูล AG100 เมื่อเทียบกับชุดข้อมูล MB100 ที่จัดกลุ่มได้ง่ายกว่าแล้ว มีจำนวนคำเฉลี่ยที่ใกล้เคียงกัน ในขณะที่จำนวนเอกสารที่ลิงค์กันของชุดข้อมูล AG100 มีมากกว่าชุดข้อมูล MB100 ถึง 44 เอกสาร ซึ่งก็หมายความว่ามีข้อมูลสำหรับการจัดกลุ่มเอกสารที่มากกว่า ทำให้ได้คุณภาพของผลการจัดกลุ่มที่ดีกว่า แม้ข้อมูลในชุดข้อมูล AG100 จะทำการจัดกลุ่มได้ยากกว่า ส่วนเหตุผลที่ชุดข้อมูล MB100 มีคุณภาพการจัดกลุ่มที่สูงกว่าชุดข้อมูล GM100 เนื่องจากเป็นชุดข้อมูลที่จัดกลุ่มได้ง่ายกว่าและยังมีจำนวนคำเฉลี่ยมากกว่า ส่วนชุดข้อมูล GM100 มีจำนวนคำเฉลี่ยที่ไม่มากนักและจำนวนเอกสารที่ลิงค์ที่ไม่มากพอทำให้มีคุณภาพการจัดกลุ่มเอกสารต่ำที่สุด

อีกสิ่งหนึ่งที่น่าสนใจเมื่อดูรูปที่ 4.1, 4.2 และ 4.3 เทียบกันแล้ว คือความแตกต่างของคุณภาพการจัดกลุ่มเอกสารระหว่าง การจัดกลุ่มเอกสารด้วยลิงค์ร่วมกับคำ และการจัดกลุ่มเอกสารด้วยคำอย่างเดียว เมื่อดูที่จุด f(c)=3 จะเห็นว่ากราฟของชุดข้อมูล AG100 มีความต่างมาก ทั้งนี้ชุดข้อมูล AG100 มีจำนวนเอกสารที่ลิงค์กัน 49 เอกสาร ส่วนชุดข้อมูล GM100 มีความต่างน้อยลง โดยมีจำนวนเอกสารที่ลิงค์กัน 25 เอกสาร และชุดข้อมูล MB100 มีความต่างที่น้อยที่สุด และมีจำนวนเอกสารที่ลิงค์เพียง 5 เอกสาร จากสิ่งที่กล่าวมานี้แสดงให้เห็นว่า จำนวนเอกสารที่ลิงค์กันมีผลทำให้คุณภาพการจัดกลุ่มเอกสารโดยใช้ลิงค์ร่วมกับคำ ต่างจากคุณภาพการจัดกลุ่มเอกสารโดยใช้คำอย่างเดียว จึงเป็นสิ่งที่ยืนยันได้ว่าการนำลิงค์มาร่วมในการจัดกลุ่มเอกสารทำให้คุณภาพการจัดกลุ่มเอกสารดีกว่าการใช้คำเพียงอย่างเดียว

2...เมื่อเทียบคุณภาพการจัดกลุ่มเอกสารของชุดของมูลที่มีขนาด 150 เอกสาร ได้แก่ AGM150, GMB150, และ MBE150 ซึ่งได้มาจากการรวมกลุ่มเอกสารขนาด 50 เอกสารจำนวน 3 กลุ่ม โดยข้อมูลจำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กัน แสดงอยู่ในตารางที่ 4.2 จากการสังเกตคุณภาพการจัดกลุ่มเอกสารของชุดข้อมูลขนาด 150 เอกสารทั้งสามชุด ในรูปที่ 4.4, 4.5 และ 4.6 ร่วมกับข้อมูลในตารางที่ 4.2 ทำให้สามารถสรุปวิเคราะห์ผลการจัดกลุ่มเอกสารได้ดังนี้

ตารางที่ 4.2 ข้อมูลจำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กันในชุดข้อมูลจำนวน 150 เอกสาร

ชุดข้อมูล	จำนวนคำเฉลี่ย	จำนวนเอกสารที่ลิงค์กัน
AGM150	25 คำ / 1เอกสาร	51 เอกสาร
GMB150	32 คำ / 1เอกสาร	29 เอกสาร
MBE150	36 คำ / 1เอกสาร	11 เอกสาร

ชุดข้อมูล AGM150 มีคุณภาพการจัดกลุ่มเอกสารต่ำสุด เนื่องจากชุดข้อมูลนี้เป็นการรวมกันของกลุ่มข้อมูลย่อยที่มีเนื้อหาอยู่ในกลุ่มข้อมูลหลัก ARTS เหมือนกัน ทำให้การจัดกลุ่มทำได้ยาก ประกอบกับมีจำนวนคำเฉลี่ยที่น้อยจึงทำให้คุณภาพการจัดกลุ่มต่ำ แม้จะมีจำนวนเอกสารที่ลิงค์กันมากที่สุดใน 3 ชุดข้อมูล ก็ยังไม่เพียงพอที่จะทำให้คุณภาพเพิ่มขึ้นเนื่องจากขนาดของชุดข้อมูลเพิ่มขึ้น แต่ผลของจำนวนเอกสารที่ลิงค์กันทำให้การจัดกลุ่มเอกสารโดยใช้ลิงค์และคำมีคุณภาพที่มากกว่าการจัดกลุ่มโดยใช้คำพอสมควร

ชุดข้อมูล GMB150 มีคุณภาพการจัดกลุ่มเอกสารมากกว่า ชุดข้อมูล AGM150 เนื่องจากข้อมูลในชุดข้อมูล GMB150 ประกอบด้วยกลุ่มข้อมูลย่อย G และ M ซึ่งอยู่ในกลุ่มข้อมูลหลัก ARTS เหมือนกัน และกลุ่มข้อมูลย่อย B ที่มาจากกลุ่มข้อมูลหลัก BUSINESS จึงทำการจัดกลุ่มได้ง่ายกว่า อีกทั้งมีจำนวนคำเฉลี่ยที่มากกว่า จึงทำให้คุณภาพการจัดกลุ่มมากกว่า แต่ความแตกต่างของคุณภาพการจัดกลุ่มน้อยเพราะจำนวนเอกสารที่ลิงค์กันมีน้อย

ชุดข้อมูล MBE150 มีความแตกต่างระหว่างคุณภาพการจัดกลุ่มโดยใช้ลิงค์และคำกับการจัดกลุ่มโดยใช้คำมากพอสมควรแม้จะมีจำนวนเอกสารที่ลิงค์กันน้อย ซึ่งความแตกต่างนี้มากกว่า GMB150 แต่น้อยกว่า AGM150 เหตุผลน่าจะมาจากการที่กลุ่มข้อมูลแต่ละกลุ่มในชุดข้อมูลนี้มาจากคนละชุดข้อมูลหลัก ทำให้สามารถทำการจัดกลุ่มได้ง่าย เมื่อจำนวนคำเฉลี่ยในชุดข้อมูลมีมากจึงทำให้คุณภาพการจัดกลุ่มของชุดข้อมูลนี้ดีที่สุดในกลุ่มของชุดข้อมูลขนาด 150 เอกสาร

เมื่อเทียบคุณภาพการจัดกลุ่มระหว่าง ชุดข้อมูลที่ได้มาจากการรวมกันของกลุ่มข้อมูล 2 กลุ่ม (ชุดข้อมูลขนาด 2 กลุ่ม) กับชุดข้อมูลที่ได้มาจากการรวมกันของกลุ่มข้อมูล 3 กลุ่ม (ชุดข้อมูลขนาด 3 กลุ่ม) โดยการเทียบนี้จะเทียบชุดข้อมูลขนาด 2 กลุ่มกับชุดข้อมูลขนาด 3 กลุ่มที่มีสมาชิกเป็นกลุ่มข้อมูลทั้งหมดในชุดข้อมูลขนาด 2 กลุ่ม กล่าวคือ เทียบ AG100 กับ AGM150, GM100 กับ GMB150 และ MB100 กับ MBE150 ซึ่งเปรียบเสมือนการเปรียบเทียบการจัดกลุ่มเอกสารที่มีการเพิ่มจำนวนกลุ่มในชุดข้อมูลขณะที่ขนาดของแต่ละกลุ่มข้อมูลไม่เปลี่ยนแปลง (ในที่นี้ขนาดของแต่ละกลุ่มคือ 50 เอกสาร) พบว่า

ก.) มีเพียงชุดข้อมูล AGM150 ที่มีคุณภาพการจัดกลุ่มน้อยกว่าชุดข้อมูล AG100 ทั้งนี้มีสาเหตุมาจาก จำนวนคำเฉลี่ยใน AGM150 มีเพียง 25 คำ ทำให้มีข้อมูลที่จะนำไปใช้ในการจัดกลุ่มน้อยเกินไป และสาเหตุอีกประการหนึ่ง กลุ่มข้อมูล A, G และ M มาจากกลุ่มข้อมูลหลักเดียวกันคือ ART ดังนั้น AG100 ซึ่งเป็นการรวมกันของกลุ่มข้อมูลในด้านเดียวกัน 2 กลุ่มย่อมทำการจัดกลุ่มได้ง่ายกว่า AGM150 ซึ่งเป็นการรวมกันของกลุ่มข้อมูลในด้านเดียวกันถึง 3 กลุ่ม

ข.) ชุดข้อมูล GMB150 มีคุณภาพการจัดกลุ่มที่ดีกว่าชุดข้อมูล GM100 เนื่องจากกลุ่มข้อมูล B เป็นกลุ่มข้อมูลคนละด้านกับกลุ่มข้อมูล G และ M จึงมีโอกาสในการที่จะจัดกลุ่มข้อมูลได้ถูกต้องเพิ่มขึ้น แม้ GMB150 จะมีจำนวนเอกสารมากกว่า GM100 ก็ตาม

ค.) ชุดข้อมูล MBE150 มีคุณภาพการจัดกลุ่มที่ดีกว่าชุดข้อมูล MB100 เนื่องจากกลุ่มข้อมูล E เป็นกลุ่มข้อมูลคนละด้านกับกลุ่มข้อมูล M และ B จึงมีโอกาสในการที่จะจัดกลุ่มข้อมูลได้ถูกต้องเพิ่มขึ้น แม้ MBE150 จะมีจำนวนเอกสารมากกว่า MB100 ก็ตาม

3. สุดท้ายเป็นการเทียบคุณภาพการจัดกลุ่มในชุดข้อมูลขนาด 200 เอกสาร ได้แก่ ชุดข้อมูล AG200 และ MB200 เมื่อนำข้อมูลจากตารางที่ 4.3 มาร่วมกับผลการจัดกลุ่มของข้อมูลทั้ง 2 ชุดในรูปที่ 4.7 และ 4.8 ทำให้สามารถวิเคราะห์ผลการจัดกลุ่มได้ดังนี้

ตารางที่ 4.3 ข้อมูลจำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กันในชุดข้อมูลจำนวน 200 เอกสาร

ชุดข้อมูล	จำนวนคำเฉลี่ย	จำนวนเอกสารที่ลิงค์กัน
AG200	42 คำ / 1เอกสาร	63 เอกสาร
MB200	25 คำ / 1เอกสาร	35 เอกสาร

AG200 มีคุณภาพการจัดกลุ่มที่ด้อยกว่า MB200 เนื่องจาก AG200 เป็นชุดข้อมูลที่ได้มาจากการรวมกลุ่มข้อมูล A และ G ซึ่งเป็นกลุ่มข้อมูลในด้านเดียวกัน ทำให้การจัดกลุ่มทำได้ยากกว่า MB200 ซึ่งได้มาจากการรวมกลุ่มข้อมูลคนละด้านเข้าด้วยกัน แม้ AG200 จะมีจำนวนเอกสารที่ลิงค์กันจำนวนมากถึง 63 เอกสาร แต่ยังไม่มากพอเมื่อคิดเป็นสัดส่วนต่อขนาดของชุดข้อมูล (200 เอกสาร) จึงทำให้จำนวนเอกสารที่ลิงค์กันไม่สามารถช่วยเพิ่มคุณภาพการจัดกลุ่มได้เหมือนในกรณีของชุดข้อมูล AG100

จุดประสงค์อีกประการของการทำการทดลองกับชุดข้อมูล AG200 และ MB200 เพื่อดูผลที่เกิดขึ้นจากการจัดกลุ่มในกรณีที่ขนาดของกลุ่มข้อมูลเพิ่มจาก 50 เอกสาร (ในชุดข้อมูล AG100 และ MB100) เป็น 100 เอกสาร ในขณะที่จำนวนกลุ่มเท่าเดิม จากผลของคุณภาพการจัด

กลุ่มจะเห็นได้ว่า คุณภาพการจัดกลุ่มของ AG100 และ MB100 น้อยกว่า AG200 และ MB200 ดังนั้นจึงสรุปได้ว่า ขนาดของกลุ่มข้อมูลที่เพิ่มขึ้นทำให้การจัดกลุ่มทำได้ยากขึ้น

4.4 สรุปผลการทดลอง

บทที่ 4 นี้เป็นการนำเสนอข้อมูลที่เกี่ยวข้องกับการทดลองในงานวิจัยนี้ โดยเริ่มการนำเสนอด้วยขั้นตอนการทำการทดลอง ซึ่งเริ่มตั้งแต่การเตรียมข้อมูลต่าง ๆ ที่จำเป็นสำหรับการทดลองจัดกลุ่มเอกสาร ทั้งในฐานข้อมูลและนอกฐานข้อมูล แล้วจึงเริ่มทำการทดลอง สิ่งต่อมาที่นำเสนอคือวิธีการเปรียบเทียบประสิทธิการจัดกลุ่มเอกสาร ซึ่งได้กล่าวถึงจำนวนข้อมูลและจำนวนกลุ่มข้อมูลที่มีในแต่ละชุดข้อมูลที่ใช้ในการทดลอง รวมถึงได้แสดงฟังก์ชันการวัดคุณภาพการจัดกลุ่มเอกสาร ดังสมการที่ 4.1 พร้อมกับอธิบายคุณสมบัติของตัวแปรต่าง ๆ ในสมการ แล้วจึงนำเสนอผลการทดลอง ซึ่งใช้ฟังก์ชันการวัดคุณภาพการจัดกลุ่มในการคำนวณคุณภาพการจัดกลุ่มข้อมูลในแต่ละชุดข้อมูล แล้วแสดงเป็นกราฟเพื่อความสะดวกในการดูเทียบผล โดยได้แสดงคุณภาพการจัดกลุ่มของวิธีการจัดกลุ่มโดยใช้ลิงค์ร่วมกับคำเปรียบเทียบกับวิธีการจัดกลุ่มโดยใช้คำ จากการวิเคราะห์คุณภาพการจัดกลุ่มได้ข้อสรุปดังนี้

- การจัดกลุ่มโดยใช้ลิงค์ร่วมกับคำให้ผลการจัดกลุ่มดีกว่าการจัดกลุ่มโดยใช้คำเพียงอย่างเดียว
- จำนวนเอกสารที่ลิงค์กันเป็นปัจจัยสำคัญทำให้คุณภาพการจัดกลุ่มโดยใช้ลิงค์และคำดีกว่าการใช้คำเพียงอย่างเดียว
- ข้อมูลที่มารวมกลุ่มกันมีเนื้อหาในด้านเดียวกัน ทำให้การจัดกลุ่มทำได้ยากกว่า ข้อมูลที่มีเนื้อหาคนละด้านกัน
- คุณภาพของการจัดกลุ่มข้อมูลลดลง เมื่อจำนวนกลุ่มข้อมูลที่มาร่วมกันในชุดข้อมูลเพิ่มขึ้น
- คุณภาพของการจัดกลุ่มข้อมูลลดลง เมื่อจำนวนเอกสารในกลุ่มข้อมูลเพิ่มขึ้น

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

วิทยานิพนธ์เล่มนี้นำเสนองานวิจัยเพื่อการจัดกลุ่มเอกสารในเครือข่ายใยแมงมุม ซึ่งมีจุดประสงค์ในการปรับปรุงการแสดงผลการค้นหาข้อมูลของเสิร์ชเอ็นจิ้นที่ไม่มีประสิทธิภาพ การจัดกลุ่มนี้ใช้ลิงค์และคำในเอกสารเป็นข้อมูลในการจัดกลุ่ม คำที่ใช้มาจากแท็กหัวข้อเอกสาร (title) และแท็กคำอธิบายเอกสาร (meta) เท่านั้น เพื่อหลีกเลี่ยงคำที่มีความกำกวมจากส่วนเนื้อความเอกสาร (body) โดยใช้กระบวนการจัดกลุ่มเอกสารที่เรียกว่า Hierarchical Agglomerative Clustering (HAC) เช่นเดียวกับงานวิจัย [2] ซึ่งเริ่มโดยให้แต่ละเอกสารเป็นกลุ่มเอกสาร แล้วจึงทำการรวมกลุ่มเอกสารที่มีความคล้ายกันมากที่สุดเข้าด้วยกันเป็นกลุ่มใหม่ จนกว่าจะถึงเงื่อนไขที่ทำให้การจัดกลุ่มหยุดลง ในส่วนของเงื่อนไขที่ใช้ควบคุมการหยุดการจัดกลุ่มนั้นใช้ฟังก์ชันที่เรียกว่า Global Quality Function (GQF) แต่งานวิจัยนี้ยังมีความแตกต่างกับงานวิจัย [2] ตรงที่การจัดกลุ่มในงานวิจัยนี้ใช้ลิงค์และคำในการหาความคล้ายระหว่างเอกสาร และใช้ฟังก์ชันในการคำนวณความคล้ายระหว่างเอกสารที่พัฒนาขึ้นใหม่ ในขณะที่งานวิจัย [2] ใช้เพียงคำในการหาความคล้ายระหว่างเอกสารเท่านั้น เมื่อได้ทำการทดลองเพื่อวัดประสิทธิภาพการจัดกลุ่มและเปรียบเทียบกับการจัดกลุ่มโดยใช้คำอย่างเดียว ซึ่งทำข้อมูลที่นำมาใช้ในการทดลองขึ้นเอง โดยเก็บรวบรวมข้อมูลที่มีการจัดกลุ่มแล้วจากเว็บไซท์ Yahoo! [9] เพื่อให้ทราบถึงกลุ่มที่ถูกต้องของข้อมูลที่นำมาทดลอง และนำข้อมูลนี้มาใช้ตรวจสอบผลการจัดกลุ่มของระบบที่พัฒนาขึ้น ผลการจัดกลุ่มแสดงให้เห็นว่า การจัดกลุ่มเอกสารโดยใช้ลิงค์และคำซึ่งเป็นระบบที่พัฒนาขึ้น ทำการจัดกลุ่มได้ดีกว่าการจัดกลุ่มโดยใช้เพียงคำอย่างเดียว และลิงค์ยังเป็นปัจจัยสำคัญที่ทำให้คุณภาพการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำแตกต่างจากการจัดกลุ่มเอกสารโดยใช้คำมากขึ้น ซึ่งคุณภาพการจัดกลุ่มเอกสารที่ดีที่สุดเมื่อค่ายกกำลังของ f(c) มีค่าเท่ากับ 3 สาเหตุของการเปลี่ยนค่ายกกำลังของ f(c) เพื่อทำให้คะแนนที่ได้จากฟังก์ชัน GQF ในแต่ละครั้งของการจัดกลุ่มมีอัตราการเพิ่มขึ้นที่น้อยลง มีผลให้โอกาสที่การจัดกลุ่มครั้งต่อไปจะทำให้คะแนนของฟังก์ชัน GQF เพิ่มขึ้นมีความเป็นไปได้มากขึ้น ในกรณีที่ค่ายกกำลังของ f(c) เป็น 1 มีคุณภาพการจัดกลุ่มที่ไม่ดีนั้น เมื่อตรวจสอบจากผลการจัดกลุ่มเอกสารจึงพบว่า มีสาเหตุมาจากจำนวนเอกสารที่ถูกจัดกลุ่มมีน้อย ทำให้เมื่อคำนวณคะแนนคุณภาพการจัดกลุ่มจึงมีคะแนนที่น้อยด้วย ในขณะที่การเพิ่มค่ายกกำลังของ f(c) เป็น 5 มีผลให้คุณภาพการจัดกลุ่มของชุดข้อมูลที่ใช้ทดลองส่วนใหญ่มีคุณภาพที่น้อยกว่าเมื่อค่ายกกำลังของ f(c) เป็น 3 สาเหตุของคุณภาพการจัดกลุ่มที่ลดลงนั้น เมื่อดูผลการจัดกลุ่มจึงพบว่า

จำนวนเอกสารที่ถูกจัดกลุ่มมีมากขึ้น แต่มีจำนวนเอกสารที่ถูกจัดกลุ่มผิดพลาดมากขึ้นด้วย ซึ่งจุดนี้เองเป็นผลให้คุณภาพการจัดกลุ่มเอกสารเมื่อค่ายกกำลังของ f(c) เท่ากับ 5 มีคุณภาพที่น้อยกว่าค่ายกกำลังของ f(c) เป็น 3 คุณภาพการจัดกลุ่มเอกสารของแต่ละชุดข้อมูลที่แสดงเป็นกราฟนั้นยังแสดงว่า ชุดข้อมูลที่มีจำนวนเอกสารที่ลิงค์กันมากหรืออีกนัยหนึ่งก็คือมีข้อมูลการลิงค์กันของเอกสารจำนวนมาก จะทำให้คุณภาพการจัดกลุ่มเอกสารโดยใช้ลิงค์และคำดีกว่าคุณภาพการจัดกลุ่มเอกสารโดยใช้คำอย่างเดียวมากยิ่งขึ้น

ในการเปรียบเทียบผลการจัดกลุ่มเอกสารจากหลาย ๆ ชุดข้อมูลที่ใช้ทำการทดลองยังแสดงให้เห็นว่า การเพิ่มขึ้นของจำนวนกลุ่มเอกสารในชุดข้อมูลจาก 2 กลุ่มเป็น 3 กลุ่มในขณะที่จำนวนเอกสารในแต่ละกลุ่มเท่าเดิม ทำให้คุณภาพการจัดกลุ่มเอกสารลดลง ซึ่งแสดงให้เห็นถึงแนวโน้มของการจัดกลุ่มเอกสารที่ทำได้ยากขึ้นเมื่อชุดข้อมูลมาจากการรวมกันของกลุ่มข้อมูลจำนวนมากขึ้น และจำนวนเอกสารในกลุ่มข้อมูลที่เพิ่มจาก 50 เอกสารในแต่ละกลุ่มเป็น 100 เอกสารในแต่ละกลุ่ม ก็ทำให้คุณภาพการจัดกลุ่มเอกสารลดลงเช่นกัน ซึ่งการเพิ่มของจำนวนเอกสารในกลุ่มนี้ก็แสดงให้เห็นแนวโน้มความยากในการจัดกลุ่มเอกสารที่เพิ่มขึ้นเมื่อขนาดของกลุ่มเอกสารที่จะต้องทำการจัดกลุ่มมีขนาดใหญ่ขึ้น สิ่งสำคัญอีกประการที่มีผลต่อคุณภาพการจัดกลุ่มก็คือ ข้อมูลสำหรับใช้ในการจัดกลุ่ม ซึ่งก็คือจำนวนคำเฉลี่ยในแต่ละเอกสารและจำนวนเอกสารที่ลิงค์กัน จากผลการทดลองที่ได้พบว่าการจัดกลุ่มจะมีคุณภาพดีขึ้นเมื่อมีข้อมูลในการจัดกลุ่มที่มากขึ้น จึงสรุปได้ว่าทั้งการเพิ่มของจำนวนกลุ่มข้อมูลที่นำมาใช้ในการทดลองและการเพิ่มของจำนวนเอกสารในกลุ่มข้อมูลแต่ละกลุ่ม มีผลทำให้การจัดกลุ่มข้อมูลทำได้ยากขึ้นและทำให้คุณภาพการจัดกลุ่มด้อยลง จำนวนคำเฉลี่ยและจำนวนเอกสารที่ลิงค์กันก็เป็นปัจจัยสำคัญที่ทำให้คุณภาพการจัดกลุ่มดีขึ้น

5.2 ข้อเสนอแนะเพื่องานวิจัยในอนาคต

1. เพื่อการพัฒนาให้ระบบนี้สามารถจัดกลุ่มได้ดียิ่งขึ้น ควรจะหาทางแก้ปัญหาการขาดแคลนข้อมูลทั้งลิงค์และคำ ที่จะนำมาใช้ในการจัดกลุ่มเอกสาร เนื่องจากเป็นสาเหตุของผลการจัดกลุ่มที่มีคุณภาพไม่ค่อยดี

2. แนวทางการใช้ลิงค์นี้ สามารถนำไปประยุกต์ใช้จัดกลุ่มเอกสารด้านงานวิจัยได้ โดยใช้เอกสารอ้างอิงในงานวิจัยเป็นเสมือนลิงค์ที่อยู่ในเอกสาร HTML

เอกสารอ้างอิง

[1] G. Salton. **Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer**, Reading, MA : Addison-Wesley. 1989.

[2] O. Zamir, O. Etzioni, O. Madani, and R. Karp. "Fast and Intuitive Clusturing of Web Documents." **Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining** 1997. Pp. 287-290

[3] Okan Kolak, Web-Syan Li. "On Ranking and Organizing Web Query Results" **Proceedings of Knowledge and Data Engineering Exchange** 1999. Pp. 26-33

[4] Hisashi S., Hajime T., Tomonari K., and Yoshiyuki K. "A domain cluster Interface for WWW Search" **Proceedings of 9th International Workshop on Database and Expert Systems Applications,** 1998. Pp. 455-460

[5] Jon M. Kleinberg "Authoritative Sources in a Hyperlinked Environment, " **Proceedings of the ACM-SIAM Symposium on Discrete Algorithms** 1998.

[6] Quoc Vu, Web-Syan Li, and Edward Chang. "On Constructing Personalized Navigaion Trees for Web Documents" **Proceedings of the 8 th World Wide Web Conference, Toronto, Canada, May,** 1999. Pp. 94-95

[7] Yitong W., Masaru K. "Link Based Clustering of Web Search Results" **Lecture Notes in Computer Science,** vol. 2118, 2001. Pp. 255-266

[8] Mark P. Sinka, David W. Corne. "A Large Benchmark Dataset for Web Document Clustering" **Soft Computing Systems : Design, Management and Applications, of Frontiers in Artificial Intelligence and Applications,** vol. 87, 2002. Pp. 881-890

[9] Yahoo! Inc., *Yahoo!*, http://www.yahoo.com

[10] Alta Vista Company, Alta Vista, http://www.altavista.com

[11] Yahoo! Inc., Geocities, http://www.geocities.com

[12] Google Search Engine, Google, http://www.google.com

[13] Overture Services, Inc., AlltheWeb :: find it all, http://www.alltheweb.com

[14] Lycos, Inc., Hotbot, http://www.hotbot.com

[15] Open Directory Project, http://www.dmoz.org

[16] LookSmart, http://www.looksmart.com

ภาคผนวก ก

รายการคำที่ไม่มีความหมายกับเนื้อหา

รายการคำที่ไม่มีความหมายกับเนื้อหา

ในงานวิจัยนี้คำที่ไม่มีความหมายกับเนื้อหา (Stop Word) แบ่งเป็น 2 ประเภท

1. คำทั่วไปที่ไม่มีความหมายกับเนื้อหา

about	every	might	that	why
across	everybody	mine	the	will
after	everyday	more	their	with
all	everyone	most	them	within
allow	few	neither	then	would
also	for	never	there	yet
although	from	none	therefore	you
always	had	not	these	your
and	has	now	they	anywhere
any	have	often	this	awesome
are	her	only	those	else
around	here	other	though	fast
because	him	our	thus	free
become	his	over	until	great
been	how	same	used	list
before	however	seldom	was	lot
between	into	shall	were	make
but	its	she	what	need
could	just	should	when	new
does	last	since	where	out
either	less	some	which	without
entire	let	someone	while	
even	many	sometime	whole	
ever	may	than	whom	

2. คำในเครือข่ายใยแมงมุมที่ไม่มีความหมายกับเนื้อหา

back	links	point
click	login	previous
copyright	logout	site
enter	main	sites
home	meta	top
homepage	navigation	url
html	next	web
internet	online	website
link	page	www

ภาคผนวก ข.

ผลงานวิจัยที่เกี่ยวข้องกับการทำวิทยานิพนธ์และได้รับการตีพิมพ์

C. Pornavalai, and P. Chavavit, "Clustering of Search Engine Results in World Wide Web", ISCIT, Songkhla, Thailand, 3-5 September, 2003.

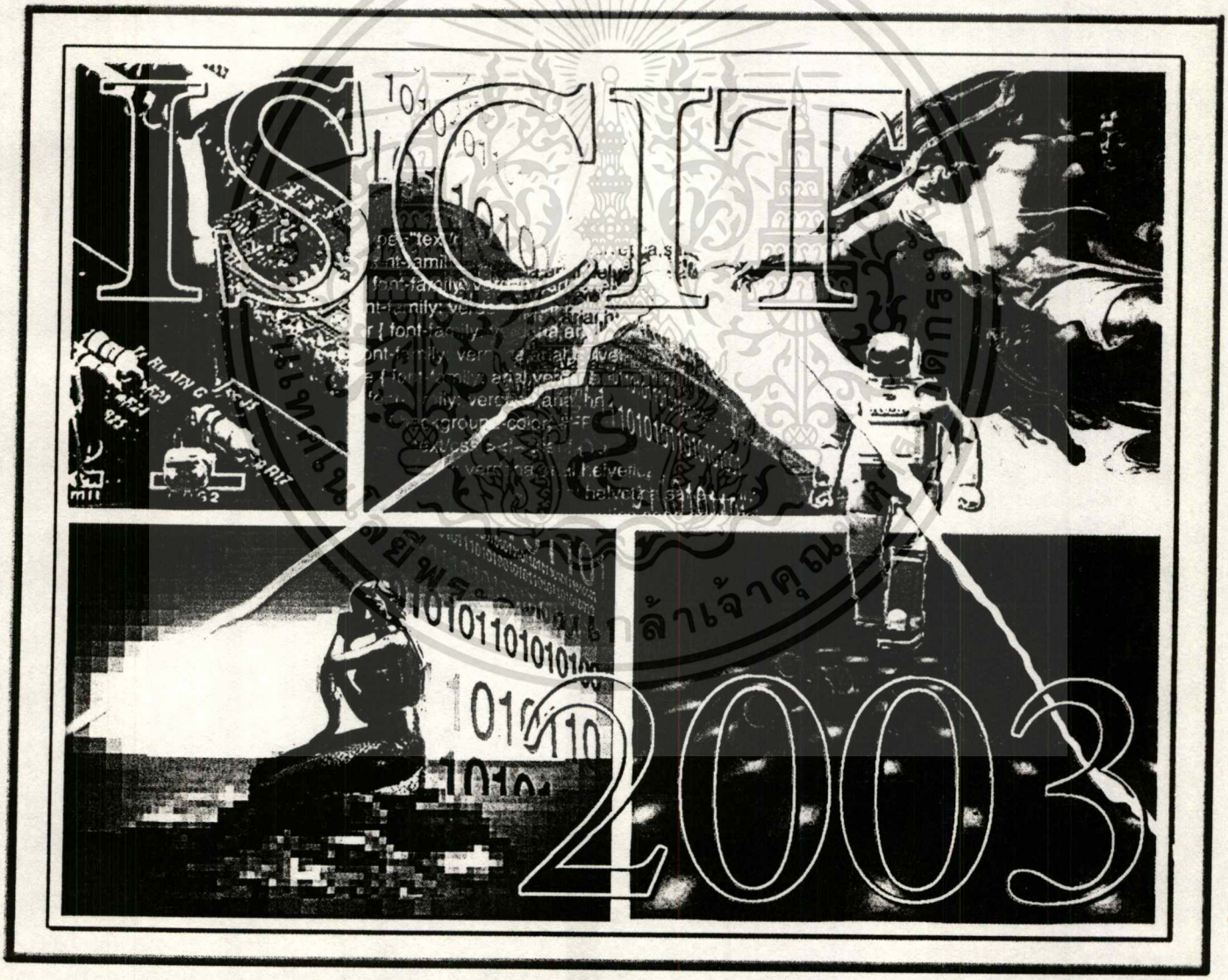
Volume I

Proceedings

The Third International Symposium on Communications and Information Technologies

September 3-5, 2003

BP Samila Beach Hotel and Resort, Songkhla, Thailand



ISCIT 2003

ISBN 974-644-437-9

CLUSTERING OF SEARCH ENGINE RESULTS IN WORLD WIDE WEB

Chotipat Pornavalai and Panupong Chavavit

Faculty of Information Technology
King Mongkut's Institute of Technology
Ladkrabang, THAILAND
Phone:+66-2-737-2551-4, Fax:+66-2-326-4332
Email: chotipat@it.kmitl.ac.th

Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang,
THAILAND
Phone:+66-2-737-2551-4, Fax:+66-2-326-4332
Email: cpanupong@hotmail.com

ABSTRACT

The vast amount and the extremely high growth of WWW documents cause the less effectiveness of web search engine. Because the traditional web search engine returns result in the long list format, therefore document clustering was introduced to show the list of results from search in groups of similarity. Users can save their time by finding relevant group of results instead of the long list of results. We have developed the clustering method to make a group of results by introducing the new concept of combining both words and the URL links in the documents. The simulation results show that it has better performance than using only words from the documents, such as proposed in other papers.

KEYWORDS: WWW, search engine, word and link clustering

1. INTRODUCTION

The vast amount and the extremely high growth of WWW documents cause the less effectiveness of web search engine in the present. Especially, with results of search engine in long list format, there are a lot of documents in response to user queries which relevant to user queries in many different aspects. So users have to go through each of results to find the corresponding result which could waste much time. Nowadays, there are mainly 2 methods of clustering. First method is clustering documents by using words in document to find the relevant of two documents which contain same words.[2] Second method is to use URL *of* document to cluster documents in the same domain (cluster by position of document) [3],[4] but do not cluster by the relevance of content in document. From the advantage and disadvantage of both methods, we propose the clustering method that uses word and URL *in* document, which is relevant to the user's queries. The URL in documents will indicate the relevance of document for clustering if both document A and document B have URL of document C in their documents (in other words, both document A and document B link to document C), or document A has URL of document B and document B has URL of document A. So URL in document will make clustering more correctly and better coverage.

The rest of paper is organized as follows. Section 2 presents the related works. The proposed hyperlink structure for document clustering is then presented in Section 3. The results are in section 4. The conclusions and future works are in the last section.

2. RELATED WORKS

There are many papers related to this topic. We show some that are most relevant to our work in the following subsections.

2.1 Fast and Intuitive Clustering of Web Documents [2]

In their paper, listed results of search engine were clustered before showing to the users. This will allow users to be able to browse results of search engine easier. The algorithms are also fast enough to be applied to thousands of documents. This clustering algorithm uses words or phases for cluster the documents by intersect the documents in a cluster to determine the set of words (or phases) shared by all documents in the clusters. Several clustering algorithms were evaluated on collections of snippets returned from web search engines. First method of this algorithm is called Word-Intersection clustering (Word-IC) which characterizes clusters by the set of words shared by every document in the cluster. Second method is Phrase-Intersection Clustering (Phase-IC). Phrase-IC looks at the phrases that are common to a group of documents, as an indication of the group's cohesion. Word-IC is a HAC algorithm that uses GQF to quantify the quality of a clustering. Each iteration of the HAC algorithm, the two clusters which merged results with the highest increase of the GQF are merged. The algorithm terminates when no merge can increases the GQF. Our proposed here in this paper is very much similar to their work (using Word-IC). However, we proposed to use URL links, not only words, *in* documents, as the information for clustering. We also propose two scoring functions that use URL for calculating the relevance between two clusters.

2.2 On Ranking and Organizing Web Query Results [3]

This is another document clustering method that uses URL *of* document that presents two clustering techniques, Clustering by domain name, and Clustering by Category. From these two techniques, first technique clusters document according to URL *of* document which is not cluster by relevant of document so clustering by this technique will decrease only a few list of result because there is not much result come from the same domain. The second technique can cover only a few documents compare to all documents in WWW because information about category of documents from WWW cover not much document as categorization in *Yahoo!* done by humans.

2.3 A Domain Cluster Interface for WWW Search [4]

This algorithm clusters the documents in the search result by the organization name which is derived from its URL domain name and display results in hierarchical tree view form. So this is another algorithm that clusters documents by using URLs of documents. This method is effective for the users who know the relation of search query to the organization name or know about that organization. So it is not suitable to use in WWW, because there are large amounts of organizations in WWW which users might not know. And it is hard to make database of organization names to compare with URL domain name that cover most of WWW.

2.4 Authoritative Sources in a Hyperlinked Environment [5]

This is not a clustering method technique but it is a good technique of using URL *in* the document. This is a distillation technique that finds and ranks the best document relevant to a broad search topic. The documents from the search result were analyzed to find links in the documents then scores of documents will be calculated by using links information. This means scores of documents come from relevant between documents. There are two kinds of score for each document, which are Authority and Hub. Authority of document comes from linked by other document so high authority score comes from linked by many good documents. Hub of document comes from link to other documents so high score of Hub comes from link to many good documents. And result will be ranked according to Authority score. As results are not groups of result then user has to spend their time look for result in their interested aspect.

3. THE PROPOSED HYPERLINK STRUCTURE FOR DOCUMENT CLUSTERING

We consider the URL *in* document as important information to show the relevance between documents. If documents have link to each other or two documents have some links to the same documents, these two documents are related. So URL *in* document will be useful for clustering result return from web search engine. But less than fifty percent of all documents will be cluster by use only URL. To increase coverage of clustering, words in title tag and meta tag of documents are used for clustering too.

Clustering results from search engine into groups of related document will ease the user in looking for the results. And users can save their time by finding relevant group of results instead of the long list of results. This clustering uses the algorithm called Hierarchical Agglomerative Clustering (HAC) [2] which starts with each document in a cluster of its own, iterating by merging the two closest clusters whose merge results in the highest increase in the Global Quality Function (GQF), and terminate when no merge increase the GQF. This is the psuedo-code for HAC:

```
1>  Initialize all documents as singleton cluster
2>  Until (GQF cannot be increased) do {
3>      Find two clusters whose merge increase GQF the most.
4>      Merge them.
5>  }
```

From the HAC algorithm, we use Global Quality Function (GQF) as a halting criterion to quantify the quality of a clustering. Thus

$$GQF(C) = \frac{f(C)}{g(|C|)} \sum_{c \in C} s(c) \qquad (1)$$

where f(C) is the function of ratio of the number of clustered document in the data set to the overall number of document in the same data set.
g(|C|) is the number of clusters of size two or more raised to the power of 0.5
s(c) is the score of cluster c

The proposed scoring function s(c) of each cluster composes of 2 sub-functions. First sub-function is the score from word in the document. Next sub-function is the score from URL *in* the document. Weighting factor (γ) controls the weight of URL sub-functions in scoring function. Thus

$$s(c) = s1(c) * (1+(\gamma * s2(c))) \qquad (2)$$

where γ is weighting factor value 0-1 to define effect of s2(c) to s1(c).
s1(c) is sub-function that scores from word.
s2(c) is sub-function that scores from link.

Then
$$s1(c) = |c| \cdot \frac{1 - e^{-\beta h(c)}}{1 + e^{-\beta h(c)}} \qquad (3)$$

where | c | is the number of documents in cluster c.

h(c) is number of word that common to all documents in the cluster.

β determines slope of sigmoid function which specifies a trade-off between the size and cohesion of the cluster.

$$s2(c) = \frac{\sum_{x \in c1, y \in c2} \alpha^{D(x,y)}}{|c|} \quad (4)$$

where x is document in cluster c1.
y is document in cluster c2.
α has value between 0 and 1.
|c| is the number of document in cluster c.

All of the scoring function above will be used in our proposed algorithm to make a good result of clustering. Psuedo-code of our algorithm will show how we use each scoring function in our algorithm as follow:

```
1>  Parse word in title tag and meta tag and link from
    documents into data base.
2>  Cut stop words and stop links out of database.
3>  Change pural word into root word.
4>  Find shortest path between each document in the test group.
5>  Set each document as a single cluster.
6>  While (new GQF more than current GQF) do {
7>    For (each pair of cluster) {
8>      Calculate word score and link score of the pair of
       cluster.
9>      Calculate score of new cluster from word and link
       score.
10>     Calculate new GQF when the new cluster occurs.
11>     Record id of both clusters that give the maximum GQF.
12>   }
13>   If (new GQF more than current GQF) {
14>     Merge both clusters.
15>   }
16> }
```

4. EXPERIMENTAL RESULTS

In our experiment, we use information in web categorized from *Yahoo!* [7] as a standard to compare with our clustering results. Because web category of Yahoo! is classified by human experts then it is reliable in correctness. Yahoo! has a wide variety of categories that are suitable for testing the clustering. Then we download the html documents for clustering from links in many categories in Yahoo! and parse each document for words in title tag, meta tag and the URLs *in* the document into document detail database. The words and URLs in the database will be the information for clustering. The average number of words in each document is 25 words per document. But some words and URLs that are meaningless for the content of documents will be deleted to reduce variation of content in documents. And we also change pural word to root word to reduce variation.

After two clusters merge into a cluster, the information (word, URL) of each cluster will be merged and become information of the new cluster. In merging word from two clusters, we merge neither all words from both clusters into the new cluster. Because some words may not describe the cluster only by words that common in both clusters, we may lose some good uncommon words and decrease the chance of clustering for the new cluster. So in this paper, we merge the words that occur in 50% or more of documents in the new cluster.

To test quality of our cluster algorithm, we create group of documents as test groups by merging documents from different categories. The numbers of documents from each category is 50, 100, 150 and 200 and merge 2 to 5 categories for each test groups. After clustering documents with our algorithm, we need scoring function to compare with the original cluster. The scoring function of quality is

$$Q(C) = \frac{\sum_{c \in C} (\sqrt{t(c)} - \sqrt{f(c)})}{\sum_{c \in B} \sqrt{\binom{|c|}{2}}} \quad (5)$$

where B = group of documents for clustering
C = group of largest clustering documents for each test group
t(c) = number of true-positive pairs in cluster c∈C
f(c) = number of false-positive pairs in cluster c∈C
|c| = size of ciuster c

From our preliminary experiment from three test groups, we found that using word and link get better quality than clustering by using only word. And we got the best quality of clustering when f(c) in (1) be the function that raised to the power 3 which can be see in Figure 1, 2, and 3. These test groups compose of 2 groups of documents with 50 documents in each group.

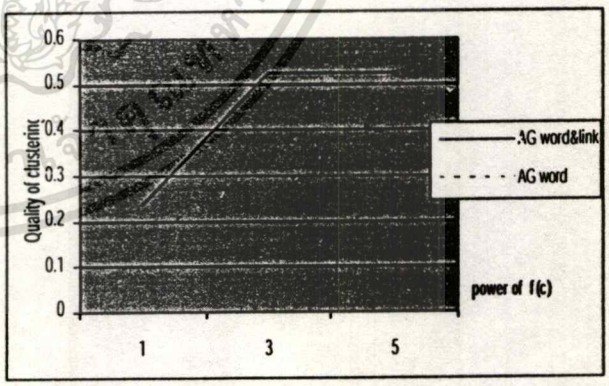


Figure 1. Comparison of average quality in test group AG

In figure 1, A is a data set of Architect and G is a data set of Graphic Design. Both Architect and Graphic Design are sub-categories in the Art category, which are categorized by *Yahoo!* In AG test group, there are average 22 words in each document and there are 49 documents that link with other documents in the test group. From the result of clustering test group AG, our purposed clustering algorithm can make a good result of clustering even though documents are from the same category but in different sub-category.

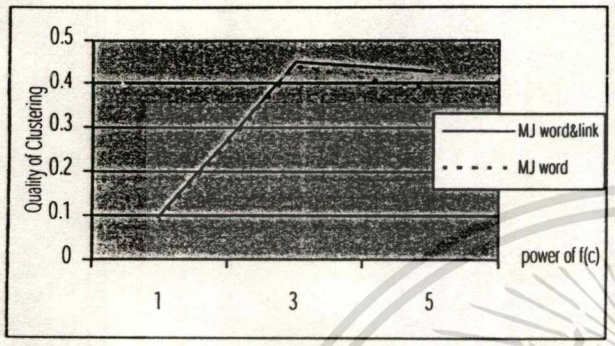


Figure 2. Comparison of average quality in test group MJ

In figure 2, M is a data set of Museum which is a sub-category in the Art category. And J is a data set of Job which is a sub-category in the Business category. In MJ test group, there are average 27 words in each document and there are 5 documents that link with other documents in the test group. As the less link information in this test group, the result of clustering with word is nearly the same as result of clustering with word and link. But clustering with word and link still make a better result.

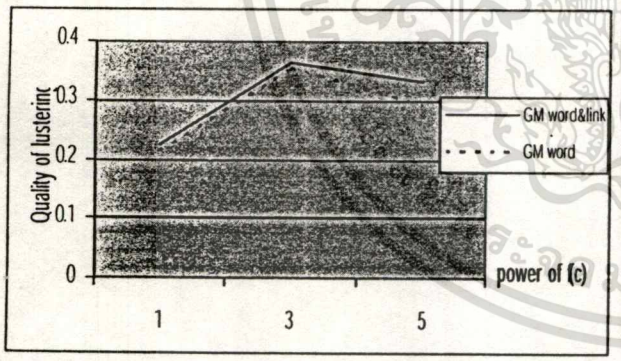


Figure 3. Comparison of average quality in test group GM

In GM test group, G is a data set of Graphic Design and M is a data set of Museum, both data sets are sub-categories in the Art category. There are average 21 words in each document and there are 25 documents that link with other documents in the test group. This is another example of clustering documents from the same category but in different sub-category. The average quality of clustering result in this test group is in figure 3.

From the results of the experiment, our word and URL clustering method got better quality score than word clustering method that proposed in [2], in three different test groups. The weight of URL score that add to word score has some effect to the score of the cluster. Clustering with word and link make better result than clustering with only word when more link information in the test group.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed the algorithm for clustering the web documents lists from the search engine. Our algorithm based on the combination with the words in the documents and URL links. We control document variation by cut out stop word and stop link and change pural word into root word. The preliminary simulation results show that the proposed algorithm could find better cluster than the existing related works on this topic.

In our future work, we observe that the relevance between queries, which users use to search in search engine, and document is information that useful for clustering. Because each document does not relevance to a query in the same level then the most relevance document should have the relevance score more than the other documents. And this relevance score will make the better clustering result.

REFERENCES

[1] G. Salton, "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*", Reading, MA: Addison-Wesley, 1989.

[2] O. Zamir, O. Etzioni, O. Madani, and R. Karp. "Fast and Intuitive Clusturing of Web Documents", in *Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining*, 1997.

[3] Okan Kolak, and Web-Syan Li, "On Ranking and Organizing Web Query Results", in *Proceedings of Knowledge and Data Engineering Exchange*, pp.26-33, 1999. (KDEX'99)

[4] Hisashi Shimamura, Hajime Takano, Tomonari Kamba, and Yoshiyuki Koseki, "A domain cluster Interface for WWW Search", in *Proceedings of 9th International Workshop on Database and Expert Systems Applications*, pp.455-460, 1998.

[5] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.

[6] Yahoo! Inc., *Yahoo!*, http://www.yahoo.com

[7] Alta Vista Company, Alta Vista, http://www.altavista.com

[8] Yahoo!.Inc.,.Geocities, http://www.geocites.com

ประวัติผู้เขียน

ชื่อ-นามสกุล	นายภาณุพงศ์ ชวะวิทย์
วัน เดือน ปีเกิด	28 เมษายน 2520 ที่กรุงเทพมหานคร
ที่อยู่	314/227-8 ซอยรามคำแหง 76 ถนนรามคำแหง แขวงหัวหมาก เขตบางกะปิ กรุงเทพฯ 10240 โทร.0-2376-3547
ประวัติการศึกษา	2541 วิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยอัสสัมชัญ