การสกัดคำสำคัญจากการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์บนเว็บบอร์ด

นายเอกภูมิ ภูมิพันธุ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

EXTRACTING KEYWORDS FROM ELECTRONICS WORD OF MOUTH IN WEBBOAD COMMUNICATION

MR. EKKAPHUM PHUMIPHAN

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Computer Science Program in Computer Science

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

หัวข้อวิทยานิพนธ์	การสกัดคำสำคัญจากการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์บนเว็บบอร์ด
โดย	นายเอกภูมิ ภูมิพันธุ์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

…….. คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

…… ประธานกรรมการ
(ศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา)

……... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ)

…… กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร. วรเศรษฐ สุวรรณิก)

เอกภูมิ ภูมิพันธุ์ : การสกัดคำสำคัญจากการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์บนเว็บบอร์ด. (EXTRACTING KEYWORDS FROM ELECTRONICS WORD OF MOUTH IN WEBBOARD COMMUNITY) **อ. ที่ปรึกษาวิทยานิพนธ์ : ผศ.ดร. สุกรี สินธุภิญโญ, 41 หน้า.**

การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์เป็นสารสนเทศที่มีประโยชน์อย่างมากในการวางแผนการตลาด เพราะข้อความที่ปรากฏอยู่ในการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้น จัดเป็นข้อมูลที่เป็นความเห็นที่แท้จริงของผู้บริโภค ซึ่งหากเราสามารถนำคำสำคัญที่เกิดขึ้นในการสื่อสารแบบนี้มาใช้จะเป็นประโยชน์อย่างสูงต่อการสร้างสรรค์ผลิตภัณฑ์ใหม่ งานวิจัยชิ้นนี้นำเสนอการสกัดกลุ่มคำสำคัญจากข้อมูลที่เป็นการสื่อสารแบบปากต่อปากแบบอิเล็กทรอนิกส์ในแต่ละสายโยงใยของข้อมูล ด้วยวิธีการแบ่งการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ออกเป็นสายโยงใยของข้อความที่แตกต่างกันโดยใช้วิธีการตัดแบ่งเนื้อหา แล้วจึงค้นหากลุ่มสำคัญในแต่ละสายโยงใยนั้น โดยให้น้ำหนักคำสำคัญแต่ละคำด้วยวิธีการทีเอฟไอดีเอฟแบบปรับปรุงเพื่อสายโยงใย Modified for Thread - TFIDF: MT-TFIDF) ซึ่งเป็นวิธีการที่ผู้วิจัยได้นำเสนอ จากผลการทดลองพบว่า ในคำค้นทั้งหมดที่นำมาทดลองในงานวิจัยนี้ วิธีการ MT-TFIDF ให้ผลการค้นหาคำสำคัญที่ดีกว่าวิธีการ TFIDF แบบปกติทั้งสามคำค้น โดยมีระดับความมั่นใจที่ 90% และในการค้นหาคำสำคัญจากคำค้นสองคำจากสามคำนั้น วิธีการที่นำเสนอให้ผลการค้นหาคำสำคัญที่ดีกว่าที่ระดับความมั่นใจที่ 95%

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต.....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก.....

ปีการศึกษา 2554

5371470621 : MAJOR COMPUTER SCIENCE

KEYWORDS : Electronics word of mouth / Text tiling / Thread detection / Word weighting / Keywords extraction

EKKAPHUM PHUMIPHAN : EXTRACTING KEYWORDS FROM ELECTRONICS WORD OF MOUTH IN WEBBOARD COMMUNITY. ADVISOR : ASST.PROF. Sukree Sinthupinyo, Ph.D., 41 pp.

Electronics Word of Mouth (E-WOM) is very useful information for drawing a marketing plan, because text in E-WOM comes from actual opinion of consumers. Keywords hidden in E-WOM could be highly valuable in New Product Development process. Hence, we propose a novel method which can extract such keywords from threads in E-WOM. The proposed method divides the original text from the web site into several threads using a general Text Tiling algorithm. Then weights of each word in each thread are calculated using the Modified for Thread – TFIDF (MT-TFIDF) which is our main contribution. The experimental results show that the keywords sorted by MT-TFIDF are in better order than the original TFIDF. We ran experiments with three words and found that the results obtained from MT-TFIDF are better than the original TFIDF at 90% confidence level. Moreover, the results from two words in all three words are better at 95% confidence level.

Department : Computer Engineering Student's Signature .....

Field of Study : Computer Science Advisor's Signature .....

Academic Year : 2011 .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จได้เพราะได้รับความอนุเคราะห์จากบุคคลหลายท่าน ในลำดับแรกสุดขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ ที่คอยช่วยสละเวลา เพื่อให้คำปรึกษา ชี้แนะแนวทางแก้ไขปัญหา ช่วยจุดประกายความคิดและข้อเสนอแนะต่างๆ สร้างสรรค์องค์ความรู้ใหม่ๆ ตลอดจนการตรวจทานและแนะนำการแก้ไขข้อบกพร่องต่างๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้เป็นรูปเป็นร่างขึ้นมาได้ ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ ซึ่งประกอบด้วย ศาสตราจารย์ ดร. ประภาส จงสถิตย์วัฒนา ประธานกรรมการสอบวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร. วรเศรษฐ สุวรรณิก กรรมการสอบวิทยานิพนธ์ ที่ช่วยสละเวลามาช่วยชี้แนะ และให้คำปรึกษา ในมุมมองที่หลากหลาย เพื่อให้ผลงานมีความสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ คุณดลชนก แก้วสุจริต สำหรับคำแนะนำ ข้อมูลต่างๆ มุมมองที่แตกต่างตอนจนข้อเสนอแนะในการสรุปเรื่องราวต่างๆ ที่ทำให้วิทยานิพนธ์ฉบับนี้สำเร็จขึ้นมาได้ รวมถึงขอขอบพระคุณทีมงานการตลาดของบริษัท CPRAM ที่ช่วยให้ข้อมูลและข้อเสนอแนะ ที่ทำให้ได้ผลลัพธ์ของผลงานที่ดียิ่งขึ้น

สุดท้ายนี้ ขอขอบพระคุณคุณพ่อ คุณแม่ ที่คอยช่วยเหลือในด้านต่างๆทั้งทางตรงและทางอ้อม อย่างมากมาย ทำให้สามารถดำเนินงานมาจนถึงจุดนี้ได้ และขอขอบพระคุณเพื่อนๆ พี่ๆ น้องๆ ตลอดจนทีมงานบุคลากรของทางภาควิชาวิศวกรรมคอมพิวเตอร์ ที่ช่วยเหลือและให้กำลังใจมาโดยตลอด ทำให้สามารถทำทุกอย่างสำเร็จไปได้ด้วยดี

สารบัญ

สารบัญภาพ

สารบัญตาราง

บทที่ 1

บทนำ

ความเป็นมาและความสำคัญของปัญหา

การสื่อสารแบบปากต่อปากจากบุคคลไปสู่บุคคล (WOM - Word of Mouth) เป็นการสื่อสารที่ไม่เป็นทางการ สามารถกระจายของข่าวสารไปได้อย่างรวดเร็ว และได้รับความน่าเชื่อถือค่อนข้างสูง เนื่องจากการสื่อสารแบบปากต่อปากนั้น ผู้ส่งสารมักจะมีการถ่ายทอดอารมณ์ความรู้สึกได้ละเอียด รวมถึงความสัมพันธ์ของผู้ส่งสารและผู้รับสารที่ใกล้ชิดกันก็มีอิทธิพลต่อความน่าเชื่อถือสูงเช่นกัน ทำให้การสื่อสารแบบปากต่อปากนั้นมีความน่าสนใจสำหรับนักการตลาดอย่างยิ่ง เพราะนอกจากจะเป็นช่องทางสื่อสารที่น่าเชื่อถือสูงแล้ว ยังเป็นช่องทางที่ได้ข้อมูลจากกลุ่มลูกค้าอย่างถูกต้องแท้จริงและใช้ต้นทุนต่ำอีกด้วย

การสื่อสารแบบปากต่อปากแบบดั้งเดิมนั้น เป็นการสื่อสารระหว่างบุคคลโดยถ่ายทอดด้วยคำพูดเพียงอย่างเดียวเท่านั้น ประกอบกับรูปแบบการสื่อสารที่ไม่เป็นทางการ ทำให้สามารถรวบรวมข้อมูลการสื่อสารมาใช้ประโยชน์ได้ยาก แต่ในปัจจุบันการสื่อสารออนไลน์ เข้ามามีบทบาทกับชีวิตประจำวันมากขึ้น การสื่อสารแบบปากต่อปากแบบดั้งเดิมนั้น ก็มีการพัฒนามาเป็นการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ (Electronic Word of Mouth Communication: eWOM) ข้อมูลการสนทนาจะเกิดขึ้นบน เครือข่ายสังคมต่างๆ (Social Network) ทำให้การรวบรวมข้อมูลการสื่อสารทำได้ดีกว่าการสื่อสารแบบปากต่อปากแบบดั้งเดิม และในแง่ของการตลาดนั้น ช่วยส่งผลให้เจ้าของกิจการและผู้บริโภคสามารถติดต่อสื่อสารกันได้โดยตรงมากขึ้น เพราะช่องทางการสื่อสารที่ขยายกว้างขึ้น รวมไปถึงสามารถรวบรวมและจัดเก็บข้อมูลเพื่อนำไปใช้ประโยชน์ได้ง่ายขึ้น

งานวิจัยชิ้นนี้ได้นำเสนอการวิเคราะห์และค้นหาคำสำคัญจากข้อมูลส่วนที่เป็นการสื่อสารปากต่อปากแบบอิเล็คทรอนิกส์ (eWOM) ในหัวข้อต่างๆที่เกี่ยวข้องต่อเนื่องกัน ออกมาจากข้อมูลจำนวนมากบนเครือข่ายสังคมต่างๆ โดยแต่เดิมจะใช้วิธีการให้น้ำหนักความสำคัญของคำด้วย TFIDF ซึ่งไม่สามารถแสดงความแตกต่างในด้านของการกระจายตัวของคำในหัวข้อย่อยที่แตกต่างกันได้ ดังนั้นจึงมีการเสนอวิธีการประยุกต์การให้น้ำหนักความสำคัญของคำสำหรับหัวข้อย่อยที่

ต่อเนื่องกัน (MT-TFIDF)เพื่อให้เหมาะสมกับการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์มากขึ้น และเพื่อให้สามารถนำไปใช้ในวิเคราะห์และใช้ในการวางแผนเชิงการตลาดต่อไปได้

วัตถุประสงค์ของการวิจัย

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อสร้างวิธีการค้นพบกลุ่มคำสำคัญของการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ ที่สามารถทำให้ทราบถึงความสำคัญจากหัวข้อของการสื่อสารที่ต่อเนื่องกันในแต่ละหัวข้อ เพื่อให้สามารถนำไปใช้ในการพัฒนาผลิตภัณฑ์ใหม่ได้ (New Product Development-NPD)โดยข้อมูลที่นำมาใช้ทดลองในงานวิจัยจะใช้ข้อมูลจาก เว็บไซต์ www.pantip.com/cafe/food เท่านั้น

ขอบเขตของการวิจัย

1. ศึกษาและค้นพบกลุ่มคำสำคัญจากการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ (eWOM) ที่มีความต่อเนื่องกัน ซึ่งแบ่งตามหัวข้อต่างๆ จากข้อมูลเว็บไซต์ www.pantip.com/cafe/food ได้ โดยข้อมูลที่ใช้จะเป็นข้อมูลในช่วงเดือนมกราคม ปี พ.ศ.2554 ถึง เดือน มกราคม ปี พ.ศ.2555
2. ศึกษาและทดลองรวบรวมข้อมูลและแบ่งสายโยงใย (Threads) รวมถึงความสัมพันธ์ของ ข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ โดยใช้คำสำคัญเป็นเป็นตัวบ่งชี้ความสัมพันธ์
3. ปรับปรุงวิธีการคำนวณหาค่าความสำคัญของคำในเอกสารให้เหมาะสมกับข้อมูลที่มีลักษณะเป็นข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์มากขึ้น
4. ข้อมูลที่นำมาใช้ในการทดลองนั้นจะอยู่ในขอบเขตของอุตสาหกรรมอาหารสร้างสรรค์ และผลลัพธ์ที่ได้จะเป็นข้อมูลที่เหมาะสมสำหรับนำไปใช้ในอุตสาหกรรมอาหารสร้างสรรค์เท่านั้น

ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถนำทฤษฎีและผลการทดลองที่ได้ไปใช้ในการสร้างผลิตภัณฑ์ใหม่ในอุตสาหกรรมอาหารสร้างสรรค์ รวมไปถึงสามารถนำไปใช้ประกอบการตัดสินใจในการวางแผนการตลาดในอุตสาหกรรมอาหารสร้างสรรค์ได้

2. ค้นพบสำคัญของการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ ที่สามารถนำไปประยุกต์ใช้ประโยชน์กับงานทางด้านการตลาดต่อไปได้

ลำดับขั้นตอนในการเสนอผลการวิจัย

1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. ศึกษารูปแบบของข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากแหล่งข้อมูลต่างๆ ภายใต้ขอบเขตการศึกษา
3. ศึกษาวิธีการประยุกต์คำภาษาไทย เพื่อให้มีลักษณะเหมาะสมกับการนำมาใช้ในการทดลองตามทฤษฎีที่เกี่ยวข้อง
4. ศึกษาและวิเคราะห์แนวคิดและทฤษฎีที่นำมาใช้ในด้านของการแบ่งสายโยงใยจากบทความและการระบุหัวข้อด้วยคำสำคัญ เพื่อให้สามารถนำมาประยุกต์ใช้ได้เหมาะสมกับงานที่สุด
5. ศึกษาและเลือกใช้งานเครื่องมือที่เกี่ยวข้องกับงานตามความเหมาะสม
6. วิเคราะห์และทดลองการรวบรวมและค้นหารูปแบบสำคัญของข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ตามทฤษฎีที่ศึกษา
7. วัดผลความถูกต้องที่ได้ รวมถึงการปรับแต่งค่าตัวแปรและวิธีการต่างๆเพื่อให้ได้ผลลัพธ์ที่เหมาะสมและตรงตามจุดประสงค์มากยิ่งขึ้น
8. ทำการทดลองซ้ำ ในข้อมูลที่หลากหลายยิ่งขึ้น และปรับแต่งค่าตัวแปรและวิธีการต่างๆเพื่อให้ได้ผลลัพธ์ที่ดียิ่งขึ้น วิเคราะห์ผลการทดลอง และบันทึกผลที่ได้
9. สรุปผลการวิจัย

โครงสร้างของวิทยานิพนธ์

วิทยานิพนธ์ฉบับนี้จะมีโครงสร้างแบ่งเป็นบทต่างๆ 5 บทหลัก ประกอบด้วย บทนำ เอกสารและทฤษฎีที่เกี่ยวข้อง ระเบียบขั้นตอนวิธีการดำเนินงานวิจัย การทดลองและผลการทดลอง และสรุปผลการวิจัยและแนวทางการพัฒนาต่อ

บทแรกจะนำเสนอความเป็นมาและความสำคัญของปัญหา ขอบเขตของงานวิจัย วัตถุประสงค์ของงานวิจัย ประโยชน์ที่คาดว่าจะได้รับ ลำดับขั้นตอนของการนำเสนองานวิจัย

โครงสร้างวิทยานิพนธ์ และผลงานที่ตีพิมพ์จากวิทยานิพนธ์ ในบทที่ 2 จะนำเสนอ ทฤษฎีที่เกี่ยวข้อง และ งานวิจัยที่เกี่ยวข้อง บทที่ 3 นำเสนอแนวคิดและวิธีการดำเนินการวิจัย บทที่สี่นำจะนำเสนอผลการทดลองที่ได้ การวิเคราะห์และอ่านผลการทดลอง และในบทที่ห้าจะนำเสนอผลสรุปของงานวิจัย รวมถึงแนวทางในการพัฒนางานวิจัยต่อไปในอนาคต

ผลงานที่ตีพิมพ์จากวิทยานิพนธ์

ส่วนหนึ่งของวิทยานิพนธ์ฉบับนี้ได้รับการเผยแพร่ ในหัวข้อเรื่อง “EXTRACTING KEYWORDS FROM ELECTRONICS WORD OF MOUTH IN WEBBOAD” ในงานประชุมวิชาการ “The 8th National Conference on Computing and Information Technology” โดย เอกภูมิ ภูมิพันธุ์ , สุกรี สินธุภิญโญ และ ดลชนก แก้วสุจริต ที่ประเทศไทย ระหว่างวันที่ 9-10 พฤษภาคม 2555

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้อง

1. การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ (Electronics word-of-mouth)

การสื่อสารปากต่อปากแบบดั้งเดิม (Traditional Word of Mouth) คือการสื่อสารที่มีลักษณะไม่เป็นทางการที่สื่อสารระหว่างบุคคลตั้งแต่สองคนเป็นต้นไป โดยมักจะเป็นการสื่อสารแบบเผชิญหน้า เพื่อแลกเปลี่ยนข้อมูลต่างๆโดยมักมีการถ่ายทอดประสบการณ์หรือความคิดเห็นลงไปด้วย

การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ (Electronics Word of Mouth : eWOM) คือ การสื่อสารแบบปากต่อปากที่มีการพัฒนาการขึ้นมาการสื่อสารปากต่อปากแบบดั้งเดิม โดยมีการเปลี่ยนแปลงจากการสื่อสารแบบเผชิญหน้า ไปเป็นการสื่อสารผ่านทางสื่ออิเล็กทรอนิกส์ เช่น จดหมายอิเล็กทรอนิกส์ (e-mail) , กระดานสนทนา (Webboard) หรือ เครือข่ายสังคม (Social Network) เป็นต้น

การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์บนกระดานสนทนา (Webboard)นั้น จะมีลักษณะเป็นการตอบโต้กันผ่านทางสื่ออิเล็กทรอนิกส์ที่เป็นเว็บบอร์ด โดยผู้ตอบนั้นอาจจะมีการรู้จักหรือไม่รู้จักกันมาก่อนก็ได้ โดยลักษณะของเรื่องราวที่มีการสื่อสารนั้นจะเป็นเรื่องราวที่ตอบโต้ โดยอาจมีหัวข้อย่อยหลายหัวข้ออยู่ภายในการโต้ตอบนั้น สามารถแบ่งออกเป็นเรื่องย่อยๆหลายเรื่องที่มีหัวข้อย่อยที่แตกต่างกันได้โดยการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้นจะมีลักษณะทั่วไปดังนี้ [1]

1. มีประสิทธิภาพสูงในเชิงการตลาด และมีต้นทุนต่ำ
2. มีความน่าเชื่อถือสูง เนื่องจากเป็นข้อมูลที่มาจากผู้บริโภคโดยตรง
2. ไม่มีลักษณะจำเพาะที่ชัดเจน
3. มีการโต้ตอบระหว่างผู้สื่อสารสูง

นอกจากนี้การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์มีข้อได้เปรียบเมื่อเทียบกับการสื่อสารปากต่อปากแบบดั้งเดิม คือ สามารถกระจายข้อมูลการสื่อสารไปได้ทั่วโลก ทำให้ประสิทธิภาพของ

การสื่อสารสูงขึ้น สามารถเก็บข้อมูลและค้นหาข้อมูลได้มากขึ้น แม้ว่าบางครั้งผู้สื่อสารไม่ได้รู้จักกันก็สามารถแลกเปลี่ยนข้อมูลได้

การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ในด้านของการตลาดนั้นมีประโยชน์อย่างมาก โดยสามารถนำไปใช้ทำการตลาดแบบไวรัล (Viral Marketing) ด้วยสื่ออิเล็กทรอนิกส์มากมายได้อย่างมีประสิทธิภาพ ทำให้ผลลัพธ์ที่ได้นั้นมีการกระจายข้อมูลที่เป็นไปอย่างรวดเร็วและมีต้นทุนที่ต่ำ ข้อมูลที่ได้จะมีความสำคัญต่อการตัดสินของผู้บริโภคอย่างสูง รวมถึงการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้นยังช่วยให้ผู้ผลิตรับรู้ถึงข้อคิดเห็นของผู้บริโภคและสามารถนำข้อมูลที่ผ่านการกลั่นกรองเพื่อหาส่วนที่สนใจไปวิเคราะห์แล้วนำไปใช้ประโยชน์ในการปรับปรุงผลิตภัณฑ์หรือวางแผนการตลาดต่อไปได้

2. วิธีการตัดแบ่งเนื้อความ (TextTiling Algorithm)

วิธีการตัดแบ่งเนื้อความ (TextTiling algorithm)[2] เป็นวิธีการแบ่งข้อความหรือบทความให้อยู่ในรูปของหัวข้อย่อย โดยคำนวณจากการกำหนดเชิงภาษาของคำในแต่ละบทความ ซึ่งสามารถแบ่งการทำงานเป็นสามขั้นตอน ดังนี้

2.1 การแบ่งหน่วยย่อย (Tokenization)

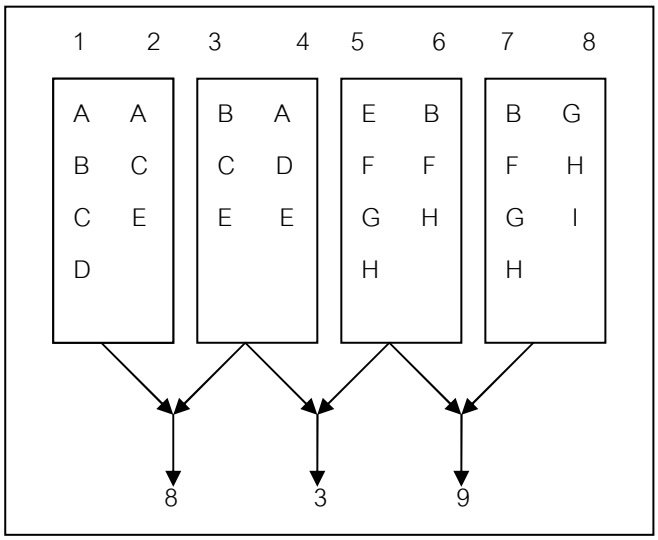
การแบ่งหน่วยย่อย[2] เป็นการแบ่งข้อความทั้งหมด ออกเป็นคำและเป็นประโยคเทียม ที่เหมาะสมแก่การคำนวณและกำหนดคะแนนเชิงภาษาในลำดับต่อไป โดยมีขั้นตอนดังนี้

- แบ่งข้อความออกเป็นคำ หรือเป็นหน่วยย่อยที่มีเอกลักษณ์เฉพาะ (Token) คำที่ถูกแบ่งแล้วจะถูกคัดเลือกเพื่อนำคำที่ไม่มีความสำคัญต่อการเปลี่ยนแปลงของหัวข้อย่อย (Stop word) ออกไป
- แยกข้อความทั้งหมดออกเป็นประโยคเทียม (Token Sentences) โดยประโยคเทียมคือชุดของคำที่ต่อเนื่องกันมีขนาดจำนวนคำเท่ากันทุกชุด

2.2 การกำหนดคะแนนเชิงภาษาเพื่อระบุขอบเขตหัวข้อย่อย (Lexical score determination)

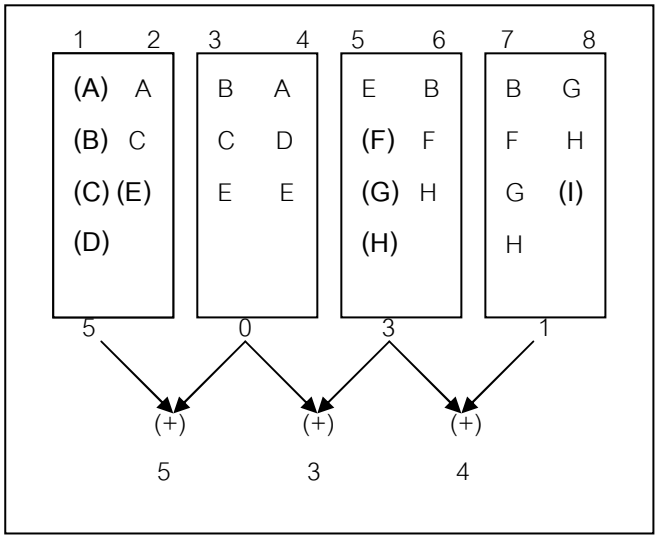
การกำหนดคะแนนเชิงภาษาจะดำเนินการโดยการสร้างบล็อก[2] สำหรับใช้ในการเปรียบเทียบคำในแต่ละประโยค ซึ่งแต่ละบล็อกนั้น จะมีการกำหนดจำนวนประโยคภายในเป็นค่าคงที่ (k) ที่เท่ากันทุกบล็อก แล้วคำนวณคะแนนเชิงภาษาโดยในงานวิจัยชิ้นนี้ได้เสนอการคำนวณแบบต่างๆ ดังนี้

- คำนวณจากการเปรียบเทียบบล็อกที่อยู่ติดกัน (Block comparison algorithm) ด้วยวิธีการหาค่าผลรวมทั้งหมดของผลคูณสเกลาร์ของความถี่ของคำในแต่ละบล็อก ผลลัพธ์ที่ได้นั้นจะแสดงคะแนนเชิงภาษาของสองบล็อกที่อยู่ติดกัน



ภาพ 1 ตัวอย่างการกำหนดคะแนนเชิงภาษาโดยคำนวณจากค่าผลคูณสเกลาร์ของความถี่ของคำในแต่ละบล็อก โดย กำหนดให้อักษรแต่ละตัวแทนค่าของคำหนึ่งคำในประโยค ,k = 2 และมี token sentence = 8

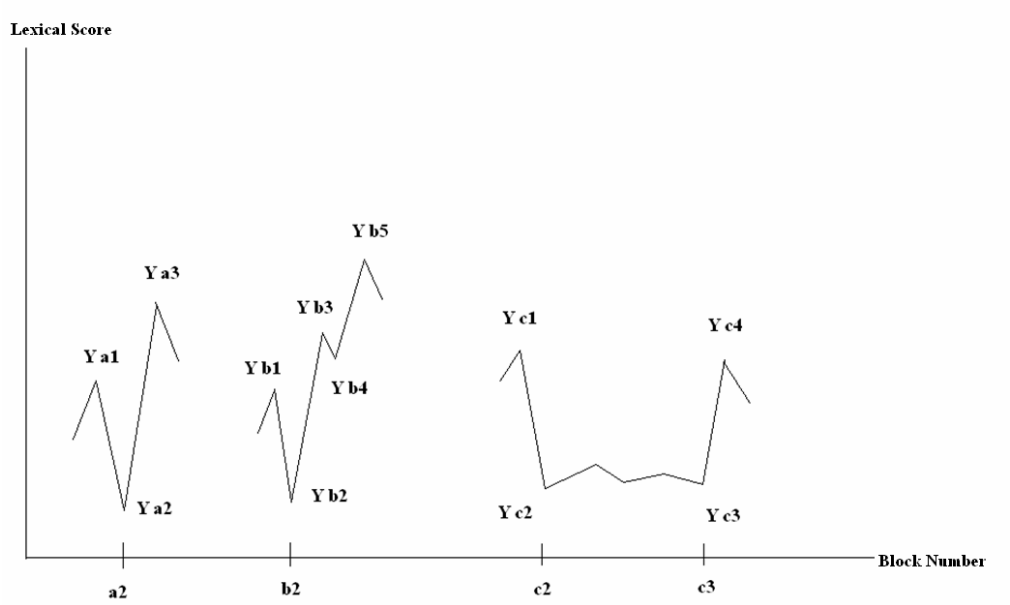
- คำนวณจากคำที่ถูกค้นพบครั้งแรก (Vocabulary Introduction) คำนวณจากผลรวมของการพบจำนวนคำใหม่ที่ถูกค้นพบครั้งแรกระหว่างช่องว่างของแต่ละประโยคที่กำหนด



ภาพ 2 ตัวอย่างการคำนวณค่าผลรวมของการพบจำนวนคำใหม่ที่ถูกพบครั้งแรกระหว่างช่องว่างของแต่ละประโยค โดยกำหนดให้อักษรแต่ละตัวแทนคำหนึ่งคำในประโยค และแทนค่าประโยคด้วยกรอบ

2.3 <u>การระบุขอบเขต (Boundary Identification)</u>

- การระบุขอบเขต [2] จะทำการคำนวนหาค่าคะแนนเชิงลึก (depth score) ของคะแนนเชิงภาษาของบล็อกที่อยู่ติดกัน โดยการนำคะแนนเชิงภาษาของบล็อกแต่ละบล็อกที่ติดกันมาสร้างเป็นกราฟแล้วคำนวนหาตำแหน่งของกราฟที่มีคะแนนเชิงภาษาของบล็อกที่อยู่ติดกันต่ำ เมื่อเปรียบเทียบกับ คะแนนของบล็อกที่อยู่ทางซ้ายและทางขวาของตำแหน่งนั้นๆ ตามตัวอย่างใน ภาพที่ 3



ภาพ 3 ตัวอย่างการสร้างกราฟเพื่อคำนวนหาค่าคะแนนเชิงลึก โดยแกนนอนแสดงลำดับของบล็อกที่อยู่ติดกัน แกนตั้งแสดงระดับคะแนนเชิงภาษาที่ได้ โดย a2 , b2 ,c2 และ c3 แสดงจุดแบ่งหัวข้อย่อยในกรณีที่แตกต่างกัน

การวัดผลของงานวิจัย

งานวิจัยชิ้นนี้แสดงการวัดผลโดยเปรียบเทียบกับการแบ่งหัวข้อย่อยโดยใช้ผู้อ่านเป็นผู้ตัดสิน (Reader Judge) และเปรียบเทียบในเชิงคุณภาพและปริมาณ แล้ววัดผลด้วยการประเมินค่า ความแม่นยำ (Precision) และ ค่าการระลึกได้ (Recall) ที่ได้ ตามเงื่อนไขดังนี้

- การวัดผลโดยใช้ผู้อ่านเป็นผู้ตัดสิน (Reader judge) นั้น ข้อความทั้งหมดจะถูกตัดสินโดยผู้อ่านจำนวนมากกว่าหนึ่งคน ผลลัพธ์ที่ได้จะแตกต่างกันไปตามผู้อ่านแต่ละคน จึงมีการทำคำนวนค่าสัมประสิทธิ์ความสอดคล้อง (Kappa coefficient : K) เพื่อหาความสอดคล้องของผลลัพธ์ที่ได้ออกมา
- การตั้งค่าตัวแปร (Parameter Setting) มีผลต่อผลการวัดผลอย่างมาก โดยค่าตัวแปรที่จำเป็นต่อการวัดผลคือ ความยาวของประโยค (w) , ขนาดของบล็อก (k) , จำนวนรอบในการปรับเรียบ (n), ความกว้างของการปรับเรียบ (s) และ จำนวนขอบเขตทั้งหมดที่กำหนดให้ (B)

- การเปรียบเทียบผลลัพธ์ที่ได้ระหว่างการคำนวณจากคำที่ถูกค้นพบครั้งแรก (Vocabulary Introduction) กับ การคำนวณจากการเปรียบเทียบบล็อกที่อยู่ติดกัน (Block comparison algorithm) โดยเปรียบเทียบจากค่าสัมประสิทธิ์ความสอดคล้อง (Kappa coefficient : K) , ค่าความแม่นยำ (Precision) และ ค่าการระลึกได้ (Recall : R) กำหนดขอบเขตล่างของผลลัพธ์มีค่าอ้างอิงตามค่าพื้นฐาน (Baseline) และค่าขอบเขตบนเปรียบเทียบกับผลลัพธ์ที่ได้จากการตัดสินด้วยผู้อ่าน (Reader Judge) ซึ่งผลลัพธ์ที่ได้จากการเปรียบเทียบนั้นพบว่า ในกรณีที่กำหนดให้มีจำนวนการแบ่งขอบเขตของหัวข้อย่อยเป็นจำนวนมากนั้น ผลลัพธ์ที่ได้จะมีค่าโอกาสที่จะได้ข้อมูลที่ถูกต้องครบถ้วนสูงขึ้น และในทางกลับกันในกรณีที่มีการกำหนดให้จำนวนการแบ่งหัวข้อย่อยต่ำลง จะทำให้ค่าความแม่นยำสูงขึ้น และในทั้งสองกรณีนั้นพบว่า คะแนนของการคำนวณจากการเปรียบเทียบบล็อกที่อยู่ติดกันนั้นให้ผลลัพธ์ที่ดีกว่าการคำนวนจากคำที่ถูกค้นพบครั้งแรกในทุกๆด้าน

Baseline		Tiling(V)						Tiling (B)						Judges		
		LC			HC			LC			HC					
P	R	K	P	R	K	P	R	K	P	R	K	P	R	K	P	R
50	51	23	52	78	32	58	64	46	66	75	47	71	59	65	83	71

ตาราง 1 เปรียบเทียบผลลัพธ์จากข้อมูลทดสอบที่ได้จากการคำนวณด้วยวิธีต่างๆ โดย K คือค่าสัมประสิทธิ์ความสอดคล้อง , P คือค่าความแม่นยำ และ R คือค่าโอกาสที่จะได้ข้อมูลครบถ้วน ส่วนของ Tiling (V) แสดงผลลัพธ์ที่คำนวณได้จากการคำนวณคำที่ถูกค้นพบครั้งแรก ส่วนของ Tiling (B) แสดงผลลัพธ์ที่คำนวณได้จากการคำนวณเปรียบเทียบบล็อกที่อยู่ติดกัน โดยทั้งสองส่วนจะมีการคำนวณโดยกำหนดให้มีทั้งกรณีที่กำหนดให้จำนวนขอบเขตหัวข้อย่อยมีมาก (LC) และจำนวนขอบเขตหัวข้อย่อยมีน้อย (HC) การเปรียบเทียบทั้งหมดจะอยู่เหนือขอบเขตล่างซึ่งเป็นค่าพื้นฐาน (Baseline) และอยู่ภายใต้ขอบเขตบนซึ่งได้จากการการตัดสินด้วยผู้อ่าน (Judge)

3. TFIDF (Term Frequency Inverse Document Frequency)

Term Frequency Inverse Document Frequency เป็นวิธีการหาค่าน้ำหนักด้วยความถี่ของคำที่ปรากฏในกลุ่มเอกสารเพื่อประเมินค่าความสำคัญของคำนั้น ต่อกลุ่มเอกสาร โดยการหาค่าน้ำหนักจะประกอบด้วยส่วนประกอบสองส่วนดังนี้

- **TF (Term Frequency)** คือ ความถี่ของคำหนึ่งคำ (term) ที่ปรากฏอยู่ในเอกสารฉบับหนึ่ง (Document)

- **IDF (Inverse Document Frequency)** คือ อัตราส่วนกลับของเอกสารทั้งหมด (N) ต่อเอกสารที่คำที่ต้องการประเมินอยู่ (df : document frequency) โดยแสดงถึงค่าความสำคัญของคำต่อเอกสารทั้งหมด ตามสมการดังนี้

$IDF = log\ N/df$

การคำนวณ TFIDF จะคำนวณได้จากค่าของ TF และ IDF ตามสมการดังนี้

$TFIDF = TF * IDF$

4. **ค่าความแม่นยำและค่าการระลึกได้** (Precision and Recall)

เป็นวิธีการวัดผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศ ผลลัพธ์ที่ได้จากการค้นคืนสารสนเทศนั้นจะมีสองแบบ คือ ข้อมูลส่วนที่หาออกมาได้ (Retrieved) และ ข้อมูลส่วนที่เกี่ยวข้องกับสิ่งที่ต้องการ (Relevant) รูปแบบของผลลัพธ์ที่ได้จะสามารถแยกออกได้เป็นสี่กลุ่มดังนี้

4.1 ส่วนที่หาข้อมูลมาได้ และ ข้อมูลส่วนนั้นเกี่ยวข้องกับสิ่งที่ต้องการ

4.2 ส่วนที่หาข้อมูลมาไม่ได้ และข้อมูลส่วนนั้นเกี่ยวข้องกับสิ่งที่ต้องการ

4.3 ส่วนที่หาข้อมูลมาได้ และ ข้อมูลส่วนนั้นไม่เกี่ยวข้องกับสิ่งที่ต้องการ

4.4 ส่วนที่หาข้อมูลมาไม่ได้ และ ข้อมูลส่วนนั้นไม่เกี่ยวกับสิ่งที่ต้องการ

Irrelevant	Retrieved + Irrelevant	Not Retrieved + Irrelevant
Relevant	Retrieved + Relevant	Not Retrieved + Relevant
	Retrieved	Not Retrieved

ภาพ 4 รูปแบบกลุ่มของผลลัพธ์ที่จะได้ทั้งหมดจากการค้นคืนสารสนเทศ

การวัดผลลัพธ์ที่ได้ถูกนำมาคิดด้วยการคำนวณอัตราส่วนสองแบบดังนี้

ค่าความแม่นยำ (Precision) คือ อัตราส่วนระหว่างข้อมูลที่หามาได้และมีความถูกต้องกับ ข้อมูลทั้งหมดที่สามารถหามาได้ นั่นคือ ค่าความแม่นยำ จะบ่งบอกถึงคุณสมบัติถูกต้อง ,ความแม่นยำของข้อมูลที่หาออกมาได้

[Precision = Number of retrieved and relevant / Total number of retrieved]

ค่าการระลึกได้ (Recall) คือ อัตราส่วนระหว่างข้อมูลที่หามาได้แล้วมีความถูกต้อง กับข้อมูลทั้งหมดที่มีความถูกต้องตามความต้องการ นั่นคือค่าการระลึกได้จะบ่งบอกถึงคุณสมบัติความครบถ้วนของข้อมูลที่ถูกต้องที่ถูกหาออกมาได้

[Recall = Number of retrieved and relevant / Total number of relevant]

ในการพิจารณาผลลัพธ์ที่ได้ด้วยค่าความแม่นยำ และ ค่าการระลึกได้นั้น ค่าที่ดีที่สุดที่เป็นไปได้คือ ค่าความแม่นยำเท่ากับค่าการระลึกได้และเท่ากับ 1 นั่นคือค่าผลลัพธ์ที่หาออกมาได้มีความถูกต้องครบถ้วนตามที่ต้องการทั้งหมด ในกรณีที่ผลลัพธ์ที่ได้ไม่เป็นไปตามค่าที่ดีที่สุดนั้น หากค่าความแม่นยำสูงมาก อาจจะทำให้ค่าการระลึกได้ต่ำ ซึ่งจะทำให้ข้อมูลที่ถูกต้องอีกบางส่วนไม่ถูกค้นพบ ในทางกลับกัน หากค่าการระลึกได้สูงมาก อาจจะทำให้ค่าความแม่นยำต่ำ ซึ่งทำให้ข้อมูลที่ได้มานั้นมีข้อมูลที่ไม่เกี่ยวข้อง ไม่ถูกต้องตามที่ต้องการปนมาด้วย

งานวิจัยที่เกี่ยวข้อง

1. การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์

ลักษณะการสื่อสารปากต่อปากนั้นเป็นการสื่อสารจากบุคคลไปสู่บุคคล โดยมีลักษณะการสื่อสารที่ไม่เป็นทางการอย่างต่อเนื่อง[3] มีการกระจายตัวไปได้รวดเร็ว และมีความน่าเชื่อถือสูง เนื่องจากผู้ส่งสารจะมีการถ่ายทอดอารมณ์และความรู้สึกส่วนตัวเข้าไปได้มาก รวมถึงการความสัมพันธ์ระหว่างผู้ส่งสารและผู้รับสารก็เป็นส่วนที่ทำให้ได้รับความน่าเชื่อถือสูงด้วย[4] จึงทำให้การสื่อสารปากต่อปากนั้นมีความสำคัญมากกับการตลาดในการแลกเปลี่ยนความคิดและ

แนะนำสินค้าไปสู่ผู้บริโภคกลุ่มใหม่ๆ [5-6] โดยดั้งเดิมแล้วการสื่อสารปากต่อปากนั้นจะเป็นการสื่อสารด้วยคำพูดเพียงอย่างเดียวเท่านั้น จนกระทั่งในปัจจุบันนี้เทคโนโลยีสารสนเทศมีการพัฒนาขึ้นอย่างมาก การสื่อสารปากต่อปากจึงถูกพัฒนาขึ้นมาอยู่บนระบบสารสนเทศ เป็นการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ (Electronics Word of Mouth) ซึ่งมีอยู่ในหลายช่องทาง เช่น เครือข่ายสังคม (Social Network) , กระดานสนทนา (Web board), จดหมายอิเล็กทรอนิกส์ (Email) เป็นต้น โดยผู้สื่อสารจะสามารถโต้ตอบกันได้ตลอดเวลา รวมถึงข้อมูลสามารถถูกเก็บไว้เพื่อค้นหาในภายหลังได้ ทำให้การเก็บข้อมูลของการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้นสามารถเก็บและนำไปวิเคราะห์ได้สะดวกกว่าแบบดั้งเดิม

2. การค้นพบและวิเคราะห์การสื่อสารที่ต่อเนื่องกันบนระบบเครือข่ายออนไลน์

การค้นพบการสื่อสารที่ต่อเนื่องกันมีงานวิจัยหลายชิ้นที่นำเสนอออกมา กรณีของการสื่อสารผ่านทางทวิตเตอร์ มีการพัฒนาเครื่องมือสรุปรวม (Summize Tool) สำหรับใช้เก็บรวบรวมข้อมูลในไมโครบล็อก (Microblog) ซึ่งใช้ในการรับส่งข้อมูลจากทวิตเตอร์ แล้วนำไปวิเคราะห์ [7] ในการเก็บข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากงานประชุมอภิปรายออนไลน์โดยใช้วิธีการทำเหมืองข้อมูลบนเว็บ (Web Mining) [8]

ในงานวิจัยเรื่องของการติดตามข้อมูลที่ต่อเนื่องกันจากอีเมล [9] ได้เสนอว่า การสรุปใจความสำคัญเป็นแบ่งข้อมูลแต่ละกลุ่มออกจากกันอย่างเป็นอิสระต่อกัน งานวิจัยชิ้นนี้จึงนำเสนอการเปรียบเทียบการสรุปใจความสำคัญของข้อมูลที่ต่อเนื่องกันบนอีเมลสองแบบ คือ การสรุปด้วยเอกสารเดียว (Single-Document Summarization) และการสรุปด้วยเอกสารหลายฉบับ (Muliti-Document Summarization) โดยในกรณีการสรุปใจความสำคัญด้วยเอกสารเดียวจะใช้วิธีการสรุปใจความแบบข้อความเดี่ยว (Individual Message Summarization: IMS) และกรณีสรุปใจความสำคัญด้วยเอกสารหลายฉบับใช้วิธีการสรุปใจความแบบข้อความรวม (Collective Message Summarization: CMS) และผลลัพธ์จากงานวิจัยนั้นสรุปว่าวิธีการสรุปใจความแบบข้อความเดี่ยว (IMS)นั้น สามารถใช้งานกับเอกสารที่มีโครงสร้างแตกต่างกันได้ดี แต่มีข้อเสียคือในการสรุปนั้นมีข้อความที่ไม่เกี่ยวข้องปะปนมาด้วย ส่วนวิธีการสรุปใจความแบบข้อความรวม (CMS) ให้ผลลัพธ์ที่ตรงกันข้าม คือไม่มีปัญหาในเรื่องของการกรองข้อความ แต่ว่ามีปัญหาในเรื่อง

ของโครงสร้างเนื้อหาที่แตกต่างกัน แต่เมื่อเปรียบเทียบทั้งสองวิธีนี้แล้ว วิธีการสรุปใจความแบบข้อความรวม (CMS) ให้ผลลัพธ์ที่ดีกว่าสำหรับการทำงานกับข้อมูลที่ต่อเนื่องกันบนอีเมล

3. การค้นพบข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ด้วยการแบ่งหัวข้อเรื่องและการหาคำสำคัญ

การค้นพบข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากข้อมูลจำนวนมากนั้น มีความสำคัญต่อการพัฒนาผลิตภัณฑ์ใหม่และการวางแผนกลยุทธการตลาดต่างๆ เพราะข้อมูลที่ได้จากการสื่อสารปากต่อปากนั้นเป็นข้อมูลมักจะเป็นข้อมูลจริงจากผู้บริโภคโดยตรง มีศึกษาพบกว่าข้อมูลนั้นมีการแสดงรายละเอียดปลีกย่อยของความต้องการของผู้บริโภคซึ่งเป็นข้อมูลที่เป็นธรรมชาติ โดยไม่มีผลประโยชน์ทางการค้าเข้ามาเกี่ยวข้อง[10] นอกจากนี้ยังสามารถค้นพบง่ายได้ต่างจากการสื่อสารปากต่อปากแบบเดิม ที่จำเป็นต้องมีการสอบถาม หรือสังเกตการณ์ผู้บริโภค ข้อมูลที่ได้เป็นข้อมูลที่อยู่บนเครือข่าย ทำให้สามารถค้นหาได้ในมุมที่กว้างขึ้นและการคงอยู่ของข้อมูลก็มีระยะเวลาที่ยาวนานมากขึ้น

การค้นพบข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากข้อมูลจำนวนมากบนเครือข่ายสังคมนั้นจะใช้สรุปใจความสำคัญ (Text Summarization) และการระบุหัวข้อเรื่องด้วยคำสำคัญ (Keyword)

การสรุปใจความสำคัญนั้นเป็นการสรุปเฉพาะส่วนสำคัญของข้อความทั้งหมดที่มีโดยสรุปออกมามีความยาวไม่เกินครึ่งหนึ่งของข้อความที่นำมาสรุปทั้งหมด [11] ขั้นตอนการสรุปใจความสำคัญนั้นจะแบ่งออกได้เป็น 3 ขั้นตอน คือ การระบุหัวข้อเรื่อง การแปลความหมายของเรื่อง และการสรุปใจความสำคัญ โดยขั้นตอนสำคัญที่จะนำมาใช้ในการค้นพบข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้นคือ การระบุหัวข้อเรื่อง ซึ่งเป็นขั้นตอนที่ใช้กำหนดขอบเขตของข้อมูลในแต่ละหัวข้อออกจากกัน ทำให้สามารถแยกข้อมูลการสื่อสารที่ต่อเนื่องกันออกเป็นหัวข้อย่อยที่เป็นอิสระจากกันได้ งานวิจัยของ [2] ได้เสนอการแบ่งบทความออกเป็นหัวข้อย่อย ด้วยวิธีการตัดแบ่งเนื้อความ (TextTiling) โดยพิจารณาจากการให้คะแนนเชิงภาษาแก่คำในแต่ละช่วงแบ่งข้อความออกจากกันเป็นหัวข้อย่อย การแบ่งหัวข้อย่อยนั้นช่วยให้สามารถจัดโครงสร้างของหัวข้อย่อยได้ชัดเจนยิ่งขึ้น และช่วยให้การค้นคืนสารสนเทศทำได้ง่ายยิ่งขึ้น [12] การแบ่งหัวข้อย่อย

ด้วยวิธีการตัดแบ่งเนื้อความนั้น จะใช้การวัดผลโดยการให้คะแนนค่าความแม่นยำ (Precision) และค่าการระลึกได้ (Recall) โดยเปรียบเทียบกับค่าเฉลี่ยของการตัดสินด้วยมนุษย์

วิธีการตัดแบ่งเนื้อความถูกนำมาเปรียบเทียบกับเทคนิคอื่น โดยมีการเปรียบเทียบระหว่างวิธีการใช้วิธีการตัดแบ่งเนื้อความ กับการใช้ วิธีการ C99 โดยใช้งานกับข้อมูลที่เป็นภาษาอารบิก [13] ซึ่งในการใช้งานวิธีการตัดแบ่งเนื้อความ และ วิธีการ C99 กับภาษาอารบิกนั้นต้องมีการปรับข้อมูลเบื้องต้นให้มีความเหมาะสมกับวิธีการแบ่งหัวข้อจากบทความก่อน โดยลักษณะของภาษาอารบิกนั้น คำที่ใช้จะถูกเขียนต่อกันโดยไม่เว้นวรรค ทำให้จำเป็นต้องมีการตัดคำก่อนจึงจะสามารถนำไปใช้ได้ การทดลองดำเนินการโดยสร้าง ArabTiling ซึ่งเป็นอุปกรณ์ที่ใช้ในการเปรียบเทียบผลลัพธ์ของการทำการตัดข้อความในรูปแบบของวิธีการตัดแบ่งเนื้อความและ วิธีการ C99 ซึ่งผลลัพธ์จากการเปรียบเทียบนั้นแสดงให้เห็นว่า วิธีการตัดแบ่งเนื้อความให้ผลลัพธ์ที่ดีกว่าวิธีการ C99 ทั้งในด้านของความแม่นยำ (Precision) และด้านของการระลึกได้ (Recall) โดยเปรียบเทียบกับมาตรฐานที่ตัดสินโดยมนุษย์ นอกจากนี้ผลการทดลองยังแสดงให้เห็นว่าการแบ่งหัวข้อจากบทความนั้น ช่วยให้ผลลัพธ์ในการทำการค้นคืนสารสนเทศ มีความถูกต้องแม่นยำมากขึ้นอีกด้วย

นอกจากนี้ วิธีการตัดแบ่งเนื้อความถูกนำมาใช้กับภาษาจีน ที่มีลักษณะแตกต่างจากภาษาอังกฤษ โดยงานวิจัยชิ้นนี้นำเสนอการใช้แบ่งหัวข้อจากบทความโดยใช้กับข้อมูลซึ่งเป็นข่าวออกอากาศในภาษาจีน [14] โดยใช้การรู้จำเสียงพูดแล้วทำการแบ่งหัวข้อบทความด้วยวิธี วิธีการตัดแบ่งเนื้อความ รวมกับเทคนิค n-grams ในการปรับค่าให้ได้ความหมายของคำที่ถูกต้อง การทำการแบ่งหัวข้อจากบทความในภาษาจีนนั้น จะมีความแตกต่างจากภาษาอังกฤษ โดยในภาษาอังกฤษนั้นจะลักษณะของคำจะประกอบด้วยตัวอักษรประกอบเป็นคำที่มีการเขียนเว้นวรรคแต่ละคำเอาไว้ แต่ในภาษาจีนตัวอักษรหนึ่งตัวจะแทนค่าเป็นคำหนึ่งคำและเขียนต่อกันไม่มีการเว้นวรรค ทำให้สามารถประยุกต์ใช้งานร่วมกับวิธีการแบ่งหัวข้อบทความด้วย วิธีการตัดแบ่งเนื้อความได้ง่าย แต่ด้วยเหตุผลของลักษณะของภาษาจีนดังกล่าวนี้ ทำให้เกิดปัญหาในการรู้จำเสียงพูด ในกรณีที่การรู้จำเสียงพูดที่กำกวมจะทำให้ได้ตัวอักษรที่ผิดพลาด ทำให้ความหมายของคำเปลี่ยนแปลง รวมถึงกรณีที่เสียงที่รับเข้ามีเสียงเพี้ยนก็จะทำให้ตัวอักษรและความหมายที่ได้เปลี่ยนแปลงเช่นกัน ดังนั้นจึงมีการนำวิธี n-grams เข้ามาแก้ไขปัญหาดังกล่าวก่อนจะนำมาใช้ใน

การแบ่งหัวข้อจากบทความ โดยผลลัพธ์ที่ได้นั้นวัดผลด้วยวิธีการวัดผลแบบ f-measure ได้ผลลัพธ์ที่ดีขึ้น 2.66%

ในส่วนของการทำการแบ่งหัวข้อย่อยจากบทความในภาษาไทยนั้น เนื่องจากภาษาไทยมีลักษณะที่แตกต่างจากภาษาอังกฤษคือ คำในภาษาไทยจะมีการเขียนติดกันโดยไม่มีการเว้นวรรค และไม่สามารถระบุได้ว่าในแต่ละประโยคนั้นเริ่มต้นและจบลงที่ใด ดังนั้นในการแบ่งหัวข้อย่อยจากบทความของภาษาไทยนั้น จำเป็นต้องมีการตัดแบ่งคำให้อยู่ในรูปแบบที่เหมาะสมกับการทำงานก่อน โดยการตัดคำภาษาไทยนั้นมีการนำเสนอ SWATH [15] เป็นเครื่องมือที่ใช้ SWATH จะตัดคำโดยใช้หลักการของคุณลักษณะของคำเป็นหลักในการตัดสินการตัดแต่ละคำ

การระบุหัวข้อเรื่องด้วยคำสำคัญ (Keyword) เป็นการหากลุ่มคำสำคัญจากเอกสารที่มีการแบ่งหัวข้อเรื่องเอาไว้แล้ว โดยการหาคำสำคัญจากเอกสารสามารถหาได้ด้วยการนำค่า TFIDFมาใช้ให้น้ำหนักความสำคัญของคำในเอกสาร เพื่อให้ได้มาซึ่งคำสำคัญในแต่ละเอกสาร [16] โดยในงานวิจัยชิ้นนี้นำเสนอการใช้เทคนิค word-occurrence เข้ามาปรับปรุงการหาคำสำคัญด้วยวิธีพิจารณาค่าน้ำหนักจากค่า TFIDF ให้มีความแม่นยำมากยิ่งขึ้น

TFIDF ถูกนำมาใช้ในการหาคำสำคัญในภาษาอื่นๆอาทิ การนำค่า TFIDF มาใช้ในการให้น้ำหนักความสำคัญของคำสำคัญในข่าวภาษาเกาหลี [17] โดยจะทำการแบ่งขั้นตอนการค้นหาคำสำคัญเป็นสองขั้นตอนคือ ค้นหากลุ่มคำที่มีความสำคัญสูงมาจากเอกสารทั้งหมด แล้วกำจัดคำที่ไม่เกี่ยวข้องหรือไม่มีความหมายตรงตามที่ต้องการไปด้วยการจัดลำดับ ผลลัพธ์ที่ได้มานั้นจะเป็นกลุ่มคำสำคัญที่มีความน่าเชื่อถือสูงขึ้น เนื่องจากผ่านการกรองส่วนที่ไม่เกี่ยวข้องออกไปแล้ว

การใช้ค่าน้ำหนักความสำคัญแบบ TFIDF หาคำสำคัญจากเอกสารประเภทข่าวในภาษาจีน [18] โดยนำมาปรับปรุงให้เหมาะกับภาษาจีน ซึ่งลักษณะของภาษาจีนนั้นเป็นภาษาที่คำและความหมายขึ้นอยู่กับตัวอักษรหนึ่งตัว แตกต่างจากภาษาอังกฤษที่คำเกิดจากการประสมตัวอักษรหลายๆตัวเข้าด้วยกัน

บทที่ 3
วิธีดำเนินการวิจัย

ขอบเขตของการดำเนินการวิจัยนั้น จะใช้ข้อมูลจากเว็บบอร์ด www.pantip.com/cafe/food มาเป็นข้อมูลสำหรับทำงานวิจัย โดยรูปแบบของเว็บไซต์นั้นจะเป็นกระดานสนทนา ที่มีการสนทนาในหัวข้อเรื่องของอาหารในด้านต่างๆ ข้อมูลที่ได้มาเหล่านี้จะถูกนำไปแบ่งหัวข้อย่อยแต่ละกลุ่ม แล้วจึงนำไปค้นคืนสารสนเทศเพื่อหาคำสำคัญในแต่ละหัวข้อ เพื่อให้คำสำคัญมีความสำคัญต่างกันในแต่ละหัวข้อของการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ที่แบ่งออกมาได้

ข้อมูลที่นำเข้ามาใช้ในการคำนวณหากลุ่มคำสำคัญนั้นจะแบ่งออกเป็นสองแบบ คือ ข้อมูลที่เก็บมาได้ทั้งหมด และคำสำคัญที่ต้องการนำไปใช้ในการค้นหากลุ่มคำสำคัญที่เกี่ยวข้อง เมื่อนำข้อมูลทั้งสองแบบมาคำนวณแล้ว ผลลัพธ์ที่ได้ออกมาจะเป็นกลุ่มคำสำคัญที่เกี่ยวข้องกับคำสำคัญที่นำเข้ามาคำนวณในขั้นตอนแรกที่อยู่ในขอบเขตของสายโยงใยเดียวกัน มีการเรียงลำดับตามค่าน้ำหนักของคำจากมากไปหาน้อย เพื่อแสดงให้เห็นลำดับความสำคัญของคำสำคัญแต่ละคำต่อคำสำคัญที่นำเข้ามาคำนวณ โดยขั้นตอนการดำเนินงานวิจัยจะแบ่งเป็นขั้นตอนการทำงานได้ดังนี้

3.1 **การเตรียมข้อมูลก่อนทำการประมวลผล** (Preprocessing)

ข้อมูลที่ทำการเก็บมาจากเว็บไซต์สองเว็บไซต์ที่เลือกไว้นั้น จะถูกนำมาเก็บไว้ในเฉพาะส่วนที่เป็นเนื้อหาสาระสำคัญในกระทู้เท่านั้น นั่นคือรูปแบบของข้อมูลจะมีเพียงหัวข้อของกระทู้และเนื้อความที่ปรากฏในกระทู้ ไม่รวมส่วนของรูปภาพ หมายเลขกระทู้ และหมายเลขลำดับคำตอบในกระทู้ และรายละเอียดย่อยอื่นๆ ลักษณะรูปแบบของข้อมูลจะต่อเนื่องกันไป โดยมีการขึ้นบรรทัดใหม่เมื่อจบกระทู้ ข้อความที่ได้นั้นจะอยู่ในรูปแบบภาษาไทยเป็นเกือบทั้งหมด ดังนั้นก่อนที่จะนำไปดำเนินการในขั้นต่อไปนั้น จะต้องมีการปรับรูปแบบของข้อความในภาษาไทยเพื่อให้เหมาะสมกับการใช้งานก่อน เนื่องจากว่าโดยพื้นฐานแล้ว วิธีการตัดแบ่งเนื้อความ ถูกสร้างมาสำหรับการทำงานกับภาษาอังกฤษ ดังนั้นจึงจำเป็นต้องมีการปรับรูปแบบของภาษาไทยให้มีลักษณะคล้ายภาษาอังกฤษ เพื่อให้การประมวลผลสามารถทำได้ถูกต้อง โดยส่วนหลักที่จำเป็นต้องเปลี่ยนนั่นคือ ภาษาไทยมีลักษณะเป็นภาษาที่เขียนคำต่อเนื่องติดกันไปทั้งประโยค

และไม่มีการแสดงจุดเริ่มต้นและการจบประโยคที่ชัดเจน ซึ่งแตกต่างจากภาษาอังกฤษที่มีการเว้นวรรคช่องว่างของคำแต่ละคำ และมีการใช้ตัวอักษรพิมพ์ใหญ่ในการขึ้นต้นประโยค และใช้สัญลักษณ์ fullstop (.) ในการแสดงการจบประโยค ดังนั้นการปรับรูปแบบของภาษาไทยนี้จะนำ SWATH เข้ามาใช้ โดย SWATH เป็นเครื่องมือที่ใช้ในการตัดคำภาษาไทย โดยอ้างอิงหลักการและคุณลักษณะของคำในภาษาไทยมาใช้ในการตัดคำ ผลลัพธ์ที่ได้จากการเตรียมข้อมูลนั้น จะมีการแสดงออกมาเป็นข้อมูลภาษาไทยที่รวบรวมเนื้อความทุกกระทู้เรียงต่อกันอยู่ในเอกสารเดียวกัน และคำแต่ละคำจะถูกแบ่งคำอย่างชัดเจนด้วยการเว้นวรรค ตรงตามลักษณะที่คล้ายคลึงกับภาษาอังกฤษ

น้ำจิ้มข้าวมันไก่แยกขั้น เสียหรือเปล่าครับ
ไปกินข้าวมันไก่ร้านดัง น้ำจิ้มเค้าใส่โหลไว้ให้ตักเอง ถ้าไปกินช่วงบ่ายๆ เย็นๆ น้ำจิ้มจะแยกตัว ส่วนที่เป็นเนื้อเค้า
เจี๊ยวสีน้ำตาลอ่อนจะอยู่ข้างล่าง ส่วนที่เป็นน้ำสีเข้มจะอยู่บน
ถ้าไปกินตอนกลางวันจะไม่แยกตัว
ไปกินช่วงตอนน้ำจิ้มแยกตัว ทีไร ท้องเสียทุกทีเลย
เป็นที่น้ำจิ้มหรือเปล่าครับ
segment_break
โปรตีนเกษตร !!!!!!!!!!
โปรตีนเกษตรประมาณ 1 ทัพพี นี่กี่กรัมหรอคะ
แบบที่ผัดพริกแกงอ่าค่ะ ^^
พอดีจะไปคำนวนแคลอรี่ :) **[1]**

น้ำจิ้ม ข้าวมัน ไก่ แยก ขั้น เสีย หรือ เปล่า ครับ
ไป กิน ข้าวมัน ไก่ ร้าน ดัง น้ำจิ้ม เค้า ใส่ โหล ไว้ ให้ ตัก เอง ถ้า ไป กิน ช่วง บ่ายๆ เย็นๆ น้ำจิ้ม จะ แยก ตัว
ส่วน ที่ เป็น เนื้อ เค้า เจี๊ ยว สี น้ำตาล อ่อน จะ อยู่ ข้าง ล่าง ส่วน ที่ เป็นน้ำ สี เข้ม จะ อยู่ บน
ถ้า ไป กิน ตอน กลางวัน จะ ไม่ แยก ตัว
ไป กิน ช่วง ตอน น้ำจิ้ม แยก ตัว ที ไร ท้องเสีย ทุกที เลย
เป็น ที่ น้ำจิ้ม หรือ เปล่า ครับ
segment _break
โป ร ตีน เกษตร !!!!!!!!!!
โป ร ตีน เกษตร ประมาณ 1 ทัพพี นี่ กี่ กรัม หรอ คะ
แบบ ที่ ผัด พริก แกง อ่า ค่ะ ^^
พอดี จะ ไป คำ นว น แค ลอรี่ :) **[2]**

ภาพ 5 [1] ตัวอย่างของข้อมูลที่เก็บมาจากเว็บไซต์โดยอยู่ในรูปแบบของข้อความเท่านั้น (text)
[2] ตัวอย่างของข้อมูลที่ผ่านการเตรียมข้อมูลให้อยู่ในรูปแบบที่พร้อมประมวลผลด้วยวิธีการตัดแบ่งเนื้อความ

3.2 **การแบ่งสายโยงใย**(Thread) **ด้วยวิธีการตัดแบ่งเนื้อความ** (Text Tiling Algorithm)

การแบ่งสายโยงใย (Thread) คือการนำข้อมูลจากกระทู้ทั้งหมดที่เก็บได้ในขั้นตอนแรก มาทำการแบ่งหัวข้อย่อย โดยแยกข้อมูลที่การสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ ในแต่ละหัวข้อที่แตกต่างกันออกจากกัน จะดำเนินการด้วยการนำวิธีการตัดแบ่งเนื้อความมาใช้ในการแบ่งการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ออกจากกัน วิธีการตัดแบ่งเนื้อความจะคำนวณให้คะแนนเชิงภาษาแก่คำที่มีความถี่ต่อเนื่องกันในแต่ละช่วง และแบ่งขอบเขตของหัวข้อย่อยแต่ละหัวข้อ เมื่อคะแนนเชิงภาษาที่จุดนั้นมีการลดลงต่ำมากเมื่อเปรียบเทียบกับจุดข้างเคียง วิธีการนี้จะช่วยให้สามารถค้นพบจุดเริ่มต้นและจุดสิ้นสุดของการสื่อสารที่ต่อเนื่องกันเป็นหัวข้อเดียวกันออกจากกันให้อยู่ในรูปแบบของสายโยงใยได้

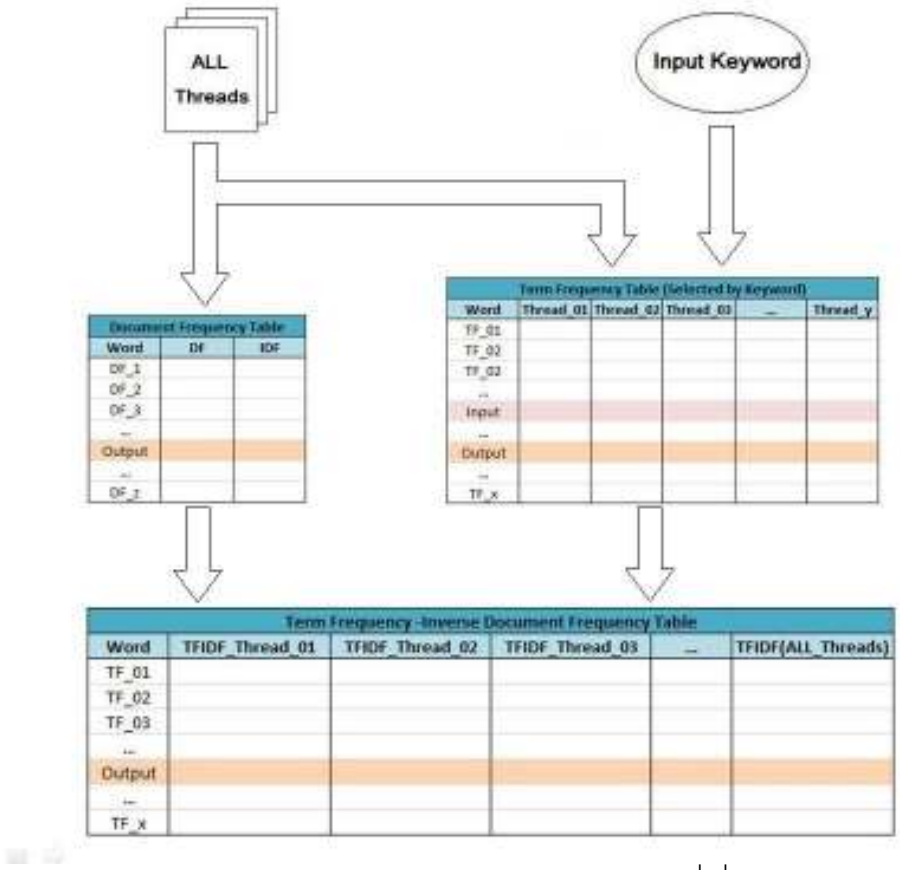
ในการคำนวณการแบ่งสายโยงใยนั้นจะแบ่งข้อมูลที่ใช้เป็นสองส่วนคือ กลุ่มข้อมูลฝึกสอน (Training Set) และ กลุ่มข้อมูลสำหรับทดสอบ (Test Set) โดยกลุ่มข้อมูลฝึกสอนจะใช้ในการสอนและปรับโครงสร้างและค่าของตัวแปรต่างๆในการคำนวณการแบ่งสายโยงใย ส่วนกลุ่มข้อมูลทดสอบนั้นจะใช้ในการทดสอบผลที่ได้หลังจากมีการฝึกสอนเรียบร้อยแล้ว ในด้านของการวัดผลลัพธ์ความถูกต้องนั้น มีการวัดผลเปรียบเทียบกันระหว่างการแบ่งสายโยงใยโดยใช้มนุษย์เข้ามาเป็นผู้อ่านและตัดสิน(Reader Judge)[2,12,13]และการตัดสินด้วยเครื่องมือที่ทำงานด้วยวิธีการตัดแบ่งเนื้อความ โดยใช้ผู้อ่านจำนวนสามคนขึ้นไปในการอ่านและตัดสิน และสรุปผลลัพธ์ตามจำนวนเสียงข้างมาก ผลลัพธ์ของการวัดผลด้วยวิธีการตัดแบ่งเนื้อความ เมื่อเปรียบเทียบกับผู้อ่าน จะแสดงค่าความแม่นยำ (Precision) และ ค่าการระลึกได้ (Recall) ค่าความแม่นยำ จะแสดงอัตราส่วนระหว่างข้อมูลที่หามาได้และมีความถูกต้อง กับ ข้อมูลทั้งหมดที่สามารถหามาได้ ซึ่งแสดงถึงคุณสมบัติความแม่นยำของผลลัพธ์ที่ได้ ส่วนค่าการระลึกได้แสดงอัตราส่วนระหว่างข้อมูลที่หามาได้แล้วมีความถูกต้อง กับ ข้อมูลทั้งหมดที่มีความถูกต้องตามความต้องการ ซึ่งแสดงถึงคุณสมบัติความครบถ้วนของผลลัพธ์ที่ถูกต้องทั้งหมดที่หามาได้ ในกรณีของการแบ่งสายโยงใยของการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้น จะให้ความสำคัญกับการวัดผลค่าความแม่นยำมากกว่าค่าการระลึกได้ เนื่องจากข้อมูลการแบ่งสายโยงใยนั้น ไม่สามารถระบุจำนวนสายโยงใยที่ครบถ้วนชัดเจนได้จริง แม้ว่าจะเป็นการตัดสินด้วยผู้อ่าน จำนวนสายโยงใยที่ได้ยังมีความแตกต่างกัน

3.3 การค้นหาคำสำคัญจากสายโยงใยโดยการให้น้ำหนักความสำคัญด้วย MT-TFIDF

การค้นหาคำสำคัญจากสายโยงใยต่างๆที่แบ่งออกมาแล้วนั้น เป็นการหากลุ่มคำสำคัญที่เกี่ยวข้องกับเรื่องที่ต้องการ โดยจะข้อมูลที่มีการแบ่งสายโยงใยออกเป็นหัวข้อย่อยที่แตกต่างกันเอาไว้นั้น จะได้กลุ่มคำสำคัญที่มีลักษณะต่างกันตามแต่ละเรื่อง ผลลัพธ์ที่ได้จึงเป็นกลุ่มของคำสำคัญที่เกี่ยวข้องอยู่ในขอบเขตของสายโยงใยที่อยู่ในเรื่องเดียวกับคำสำคัญที่ผู้ใช้ระบุเอาไว้ สามารถบ่งบอกว่าแต่ละสายโยงใยนั้นเป็นเรื่องราวเกี่ยวกับอะไร โดยอ้างอิงจากกลุ่มคำสำคัญที่ได้

ในการค้นหากลุ่มคำสำคัญผู้วิจัยได้ปรับปรุงค่า TFIDF เดิมเพื่อให้เหมาะสมกับการนำมาใช้ในข้อความที่มีลักษณะเป็นการพูดคุยตอบโต้กัน ซึ่งโดยปกติแล้ว มักจะพูดถึงเรื่องเดียวกัน และใช้คำซ้ำกัน ซึ่งค่า TFIDF แบบทั่วไปนั้น สามารถคำนวณได้จากความถี่ในการปรากฏของคำนั้นกับอัตราส่วนกลับของการปรากฏของเอกสารที่มีคำนั้นปรากฏอยู่ภายใน การเลือกกลุ่มคำสำคัญออกมาเป็นผลลัพธ์จะคัดจากกลุ่มเอกสารที่มีคำสำคัญที่ผู้ใช้ระบุออกมา

การปรับปรุงค่า TFIDF เพื่อให้สามารถหาคำสำคัญจากสายโยงใยนั้น จะมีการคำนวณที่แตกต่างจากการคำนวณสำหรับข้อมูลจำพวกบทความทั่วไป เนื่องจากผู้ใช้ต้องมีการระบุคำสำคัญในส่วนที่ต้องการเพื่อระบุหาสายโยงใยและคำสำคัญที่เกี่ยวข้องกับคำสำคัญที่ระบุมา ดังนั้นการคำนวณ จะดำเนินการโดยใช้ค่าความถี่ของเอกสารและค่าส่วนกลับความถี่ของเอกสารตามการคำนวณปกติ แต่ส่วนของการคำนวณค่าความถี่ของคำจะแตกต่างออกไป โดยพิจารณาเฉพาะจากสายโยงใยที่มีคำสำคัญตามที่ระบุไว้เท่านั้น ทำให้ผลลัพธ์ที่ได้อยู่ในขอบเขตที่เกี่ยวข้องกับคำสำคัญตามที่ต้องการ



ภาพ 6 ระบบการคำนวณค่า TFIDF สาหรับการหาคำสำคัญที่เกี่ยวข้องจากการใช้คำสำคัญ

การคำนวณค่าความสำคัญของคำ สำหรับข้อมูลที่เป็นการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้น จะต้องมีคำนวณการให้น้ำหนักของคำที่แตกต่างจากการคำนวณด้วยค่า TFIDF ตามปกติ เนื่องจากการคำนวณด้วยค่า TFIDF เพียงอย่างเดียวนั้น มีการให้ความสำคัญกับความถี่ของคำสูง แต่ว่าในการสกัดคำสำคัญจากข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์นั้น จะมุ่งเน้นให้ความสำคัญกับคำที่มีการถูกพูดถึงเป็นจำนวนมากในหลากหลายสายโยงใยมากกว่าการคำนึงถึงความถี่ของคำเพียงอย่างเดียว ดังนั้นการที่คำหนึ่งคำมีความถี่สูงเพียงอย่างเดียวจึงไม่ได้บ่งบอกว่าคำนั้นจะมีความสำคัญสูง ลักษณะของคำสำคัญจากข้อมูลการสื่อสารแบบปากต่อปากที่ตามเป้าหมายของงานวิจัยจะมีลักษณะดังต่อไปนี้

3.3.1 คำที่ปรากฏเป็นจำนวนมากในสายโยงใยที่เกี่ยวข้องกับคำสำคัญที่สนใจ ควรจะมีความสำคัญในลำดับสูง เนื่องจากเป็นคำที่เกี่ยวข้องกับคำสำคัญที่นำเข้ามาพิจารณาโดยตรงคำที่ปรากฏเป็นจำนวนมากในสายโยงใยที่ไม่เกี่ยวข้องกับคำสำคัญที่สนใจ ควรจะมีความสำคัญในลำดับต่ำ เนื่องจากเป็นคำที่ไม่เกี่ยวข้องกับคำสำคัญที่นำเข้ามาพิจารณา อยู่นอกขอบเขตที่สนใจ

3.3.2 คำที่ปรากฏเป็นจำนวนมากทั้งในสายโยงใยที่เกี่ยวข้องและไม่เกี่ยวข้องกับคำสำคัญที่สนใจ ควรจะมีความสำคัญอยู่ในระดับต่ำ เนื่องจากเป็นคำที่ถูกพบโดยทั่วไป ไม่ถือว่าสำคัญต่อคำสำคัญที่นำเข้ามาเป็นพิเศษ

3.3.3 คำที่ปรากฏในสายโยงใยเดียวที่เกี่ยวข้องกับคำสำคัญที่สนใจ ควรจะมีค่าน้ำหนักความสำคัญต่ำกว่า คำที่ปรากฏในจำนวนหลายสายโยงใยที่เกี่ยวข้องกับคำสำคัญที่สนใจ

จากลักษณะดังกล่าว จะต้องมีการปรับค่าน้ำหนักการคำนวณ TFIDF (Modified for Thread-TFIDF, MT-TFIDF) ให้มีความสัมพันธ์กับเรื่องของสายโยงใยที่สนใจและสายโยงใยที่ไม่สนใจรวมและปรับความสำคัญของความถี่ของคำที่ปรากฏในสายโยงใยให้สูงขึ้น โดยมีสมการดังนี้

$$Score_{word} = \left(\frac{\tau}{\tau + \tau'} \right) \left(\frac{\tau}{\mathrm{T}} - \frac{\tau'}{\mathrm{T}'} \right) TFIDF$$

เมื่อ τ เป็นจำนวนของสายโยงใยที่สนใจที่มีคำทีต้องการคำนวณปรากฏ

T เป็นจำนวนของสายโยงใยที่สนใจทั้งหมด

τ' เป็นจำนวนของสายโยงใยที่ไม่สนใจที่มีคำที่ต้องการคำนวณปรากฏ

T ' เป็นจำนวนของสายโยงใยทั้งหมดที่ไม่ถูกสนใจ

TFIDF เป็นค่า TFIDF มาตรฐาน

	มีคำที่ต้องการอยู่ในสายโยงใย	ไม่มีคำที่ต้องการในสายโยงใย
จำนวนสายโยงใยที่สนใจ (T)	τ	X
จำนวนสายโยงใยที่ไม่สนใจ (T')	τ'	Y

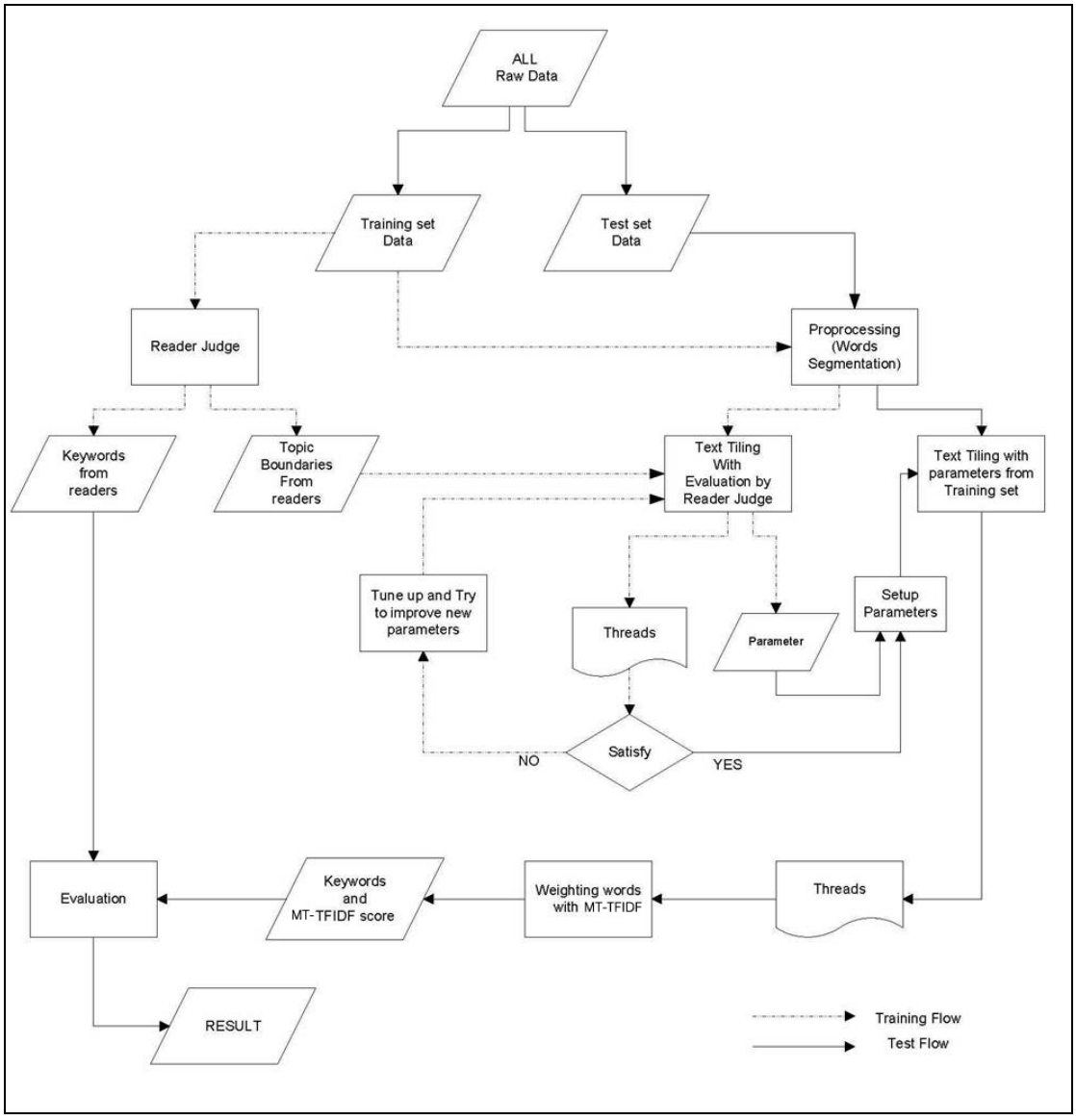
ตาราง 2 การแจกแจงความสัมพันธ์ของตัวแปรที่นำมาคำนวณค่าน้ำหนักคำสำคัญแบบ MT-TFIDF

ผลลัพธ์ที่ได้จากคำนวณน้ำหนักความสำคัญของคำด้วยการใช้ค่า TFIDF และ MT-TFIDF นั้น จะถูกกรองคำที่มีลักษณะไม่เป็นคำที่น่าสนใจออกไป โดยลักษณะที่ไม่เป็นคำและไม่น่าสนใจนั้นเป็นไปตามข้อจำกัดดังนี้

- คำในภาษาไทย จะไม่มีตัวอักษรตัวเดียวโดดๆ เช่น ก ย ร เ โ เป็นต้น

- อักขระเดี่ยวๆไม่สามารถประกอบเป็นคำได้ในภาษาไทย เช่น . / = - * เป็นต้น
- ตัวเลขจะถูกตัดออกไป เนื่องจากไม่สามารถบ่งชี้เป็นคำสำคัญได้ เมื่ออยู่เพียงจำนวนเดียว

จากนั้นจัดลำดับใหม่เรียงตามค่า TFIDF และ MT-TFIDF ที่คำนวณได้จากมากไปหาน้อย ทั้งสองแบบตามลำดับ เพื่อให้สามารถมองเห็นลำดับของความสำคัญของคำสำคัญได้อย่างชัดเจน และสามารถนำมาเปรียบเทียบความถูกต้องของผลลัพธ์จากการให้น้ำหนักในแต่ละแบบได้



ภาพ 7 แผนภาพการทำงานทั้งหมดของระบบการสกัดคำสำคัญจากการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์

บทที่ 4

การทดลองและผลการทดลอง

การทดลองจะถูกแบ่งเป็นสองส่วน คือ ส่วนของการเก็บข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากเว็บไซต์ต้นทางและประมวลผลการแบ่งสายโยงใยออกจากกัน และส่วนของการนำข้อมูลในสายโยงใยมาคำนวณคะแนนอันดับคำสำคัญที่เกี่ยวข้องจากคำสำคัญที่ต้องการ โดยมีรายละเอียดในแต่ละขั้นตอน และรูปแบบของข้อมูลดังต่อไปนี้

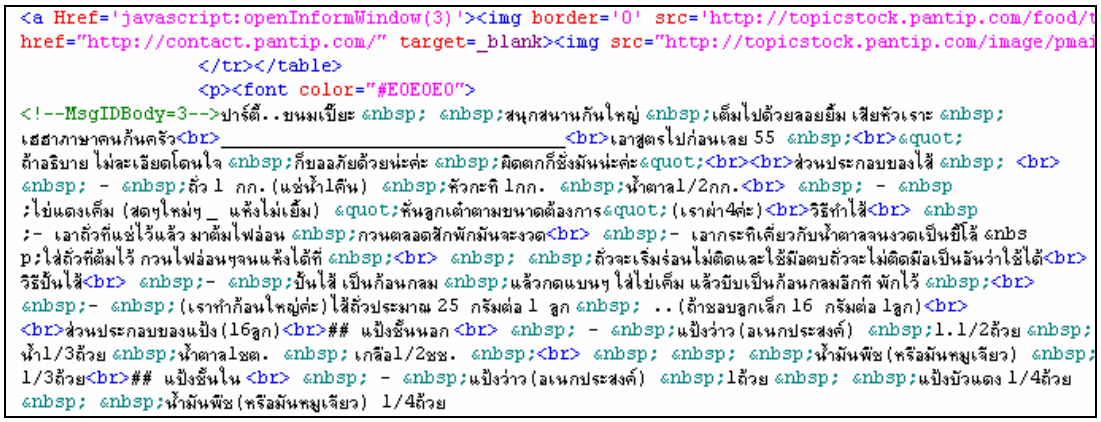
4.1 ข้อมูลที่ใช้ในการทดลอง

ผลลัพธ์ของงานวิจัยนี้ทดสอบโดยนำข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์จากเว็บบอร์ดพันทิพ http://www.pantip.com/cafe/food โดยใช้ข้อมูลนับย้อนหลังเป็นเวลาหนึ่งปี ตั้งแต่เดือนมกราคม ปี พ.ศ. 2554 ถึง เดือนมกราคม ปีพ.ศ. 2555 ข้อมูลทั้งหมดที่ถูกนำมาใช้ในการทดลองนั้นมีจำนวนทั้งสิ้น 17,738 กระทู้ ซึ่งเป็นข้อมูลที่เกี่ยวข้องอยู่ในขอบเขตเรื่องของอาหารทั้งหมด กระทู้ที่เก็บข้อมูลมาจะถูกแบ่งออกเป็นสองกลุ่ม คือ กระทู้ในเว็บบอร์ดปัจจุบัน และกระทู้ในคลังกระทู้ เนื่องจากเว็บไซต์ www.pantip.com/cafe/food นั้น มีการเก็บข้อมูลย้อนหลังเป็นเวลาประมาณหนึ่งเดือนเท่านั้น หลังจากนั้นกระทู้ที่มีอายุเกินหนึ่งเดือนจะถูกคัดเลือกเก็บเอาไว้ เฉพาะกระทู้ที่มีการโหวตให้เก็บเข้าคลังกระทู้ โดยถูกเก็บอยู่ที่ http://topicstock.pantip.com/food/topicstock/ ข้อมูลที่เก็บมานั้นเป็นกระทู้ปัจจุบันจำนวน 10721 กระทู้ ซึ่งเป็นข้อมูลที่เกิดขึ้นในระยะเวลาช่วงเดือนมกราคมของปี พ.ศ.2555 และกระทู้จากคลังกระทู้จำนวน 7017 กระทู้ ซึ่งเป็นข้อมูลในระยะเวลาช่วงเดือนมกราคมของปี พ.ศ. 2554 ถึง ธันวาคม ปี พ.ศ.2555

4.2 รูปแบบและวิธีการเก็บข้อมูลและเตรียมข้อมูลสำหรับการทดลอง

การเก็บข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์บนเว็บบอร์ดนั้น จะทำการเก็บข้อมูลโดยดาวน์โหลดข้อมูลทั้งหมดเข้ามาเก็บไว้ แล้วจึงนำมาประมวลผลเพื่อหาคำสำคัญต่อไป ในการดาวน์โหลดข้อมูลนั้น จะดำเนินการโดยใช้เครื่องมือ HTTrack Website Copier มีการกำหนดคุณสมบัติของการดาวน์โหลดข้อมูลเอาไว้ โดยเลือกเก็บเฉพาะข้อมูลที่มีลักษณะ

อยู่ในรูปแบบของ HTML และอยู่ในขอบเขตของ www.pantip.com/cafe/food และ http://topicstock.pantip.com/food/topicstock/ เท่านั้น ไม่มีการเก็บข้อมูลของภาพและข้อมูลไฟล์แนบอื่นๆที่อยู่ภายในขอบเขตดังกล่าว



ภาพ 8 รูปแบบของข้อมูลที่ได้จากการเก็บข้อมูลด้วยเครื่องมือ HTTrack Website Copier

```
httrack -%L D:/Combination/1.RunHTTrack/WinHTTrack/Link_list.txt
+http://topicstock.pantip.com/food/topicstock/2011*
+http://topicstock.pantip.com/food/topicstock/2012*
-O C:/PantipStock_2011-12/ -pl -%I -D
```

ภาพ 9 รูปแบบคำสั่งที่ใช้ในการเก็บข้อมูลจากคลังกระทู้ในช่วงปี พ.ศ. 2554 -2555 โดยเลือกเก็บเฉพาะส่วนที่เป็นข้อความในรูปแบบของ HTML เท่านั้น

จากนั้น ข้อมูลที่ได้มาจะถูกนำมาปรับเปลี่ยนรูปแบบ จากข้อมูลที่อยู่ในรูปแบบของ HTML ให้อยู่ในรูปแบบของข้อความที่ไม่มีตัวกำกับแบบ HTML (HTML tag) และตัดส่วนของชื่อของผู้ออกความเห็น และลำดับของความเห็นออก ผลลัพธ์ที่ได้จะเหลือเพียง ชื่อกระทู้ เนื้อหากระทู้ และ ความเห็นต่างๆ ที่เรียงกันออกมาอยู่ในรูปแบบของบทความ และมีการคั่นเนื้อหาของกระทู้และความเห็นแต่ละความเห็นออกจากกัน โดยใช้คำว่า segment_break

```
==== ++ Beef Wellington ทำทานเองที่บ้านไม่ยาก ++ ====
segment_break
ขอฝากตัวกับเพื่อนๆ ห้องก้นครัวครับ นี่เป็นกระทู้แรกของผมในห้องนี้

ก็เนื่องมาจากคุณ Diana Grace Monroe เธอเห็นกระทู้ของคุณ AuntPenguin ก็นึกอยากทาน เลยชวนลองทำบ้าง

ผมเองก็เป็นแฟนรายการของ Chef Gordon Ramsay
ทั้ง Hell's Kitchen และ Master Chefอยู่แล้ว

พอดูคลิปวิธีการก็นึกว่าคงไม่ยากมากนัก อุปกรณ์และเครื่องปรุงก็น่าจะหาไม่ยาก เลยบอกว่าตกลงวันหยุดนี้ลองทำกันเลย


segment_break
พอตัดสินใจว่าจะลงมือแล้ว ก็เลยไปหาวัตถุดิบ ที่ Villa สุขุมวิท 33
ก็ได้ครบทุกอย่างที่ต้องการ
segment_break
ก่อนอื่นก็เปิดคลิปดูอีกรอบ จดโน๊ตขั้นตอนแล้วก็เริ่มลงมือ

คุณกอร์ดอนเขาก็สาธิตตามสเต็ปเชฟขั้นเทพ แต่ไม่บอกปริมาณเครื่องปรุง เราก็เลย กะๆ เอาด้วยสายตา

เริ่มแรกก็เตรียมเนื้อ (ผมใช้สันในประมาณ 4 ขีด) และปรุงรสด้วยเกลือ พริกไทย
segment_break
```

ภาพ 10 รูปแบบของข้อมูลที่ถูกตัดส่วนกำกับ HTML ออกไป และแบ่งคั่นกระทู้และความเห็นด้วยการใช้คำว่า segment_break

ข้อมูลที่จะนำไปใช้ในการประมวลผลด้วยวิธีการตัดแบ่งเนื้อความนั้น หากเป็นข้อมูลที่เป็นภาษาไทยจะต้องมีการปรับข้อมูลให้อยู่ในรูปแบบที่เหมาะสมเสียก่อน โดยจะต้องตัดแบ่งคำในประโยคที่แต่ละคำเขียนติดกันในภาษาไทย ให้ออกมาเป็นคำที่เว้นวรรคออกจากกันในแต่ละคำอย่างชัดเจน ในการดำเนินการตัดคำในภาษาไทยนั้น จะนำเครื่องมือ SWATH มาใช้ช่วยในการตัดคำภาษาไทยออกมาเป็นคำๆ โดย SWATH จะใช้หลักการเชิงภาษาในการตัดคำ ข้อมูลที่ได้หลังจากที่ถูกตัดคำเรียบร้อยแล้ว จะเป็นข้อมูลที่ถูกเตรียมสำหรับการประมวลผลขั้นตอนตัดแบ่งเนื้อความต่อไปเรียบร้อยแล้ว

```
= = = =  + +  Beef  Wellington  ทำ ทาน เอง ที่ บ้าน ไม่ ยาก  + +  = = = =
segment_break
ขอ ฝาก ตัว กับ เพื่อนๆ  ห้อง ก้นครัว ครับ  นี่ เป็น กระทู้ แรก ของ ผม ใน ห้อง นี้
ก็ เนื่อง มา จาก คุณ  Diana  Grace  Monroe  เธอ เห็น กระทู้ ของ คุณ  AuntPenguin
ผม เอง ก็ เป็น แฟ น รา ย การ ของ  Chef  Gordon  Ramsay
ทั้ง  Hell ' s  Kitchen  และ  Master  Chef อยู่ แล้ว
พอดู ค ลิ ป วิธี การ ก็ นึก ว่า คง ไม่ ยาก มาก นัก  อุปกรณ์ และ เครื่องปรุง ก็ น่า จะ หาไม่ ยาก  เลย บอก

segment_break
พอ ตัดสินใจ ว่า จะ ลงมือ แล้ว  ก็ เลย ไป หา วัตถุดิบ  ที่  Villa  สุขุม วิ ท  33
ก็ได้ ครบ ทุก อย่าง ที่ ต้องการ
segment_break
```

ภาพ 11 รูปแบบของข้อมูลที่ถูกตัดคำภาษาไทย และสามารถนำไปใช้ในการประมวลผลด้วยวิธีการตัดแบ่งเนื้อความได้

4.3 การดำเนินการตัดแบ่งเนื้อความเพื่อแบ่งสายโยงใยออกจากกัน

ในขั้นตอนของการแบ่งสายโยงใยนั้น จะนำข้อมูลที่ผ่านการเตรียมข้อมูลเรียบร้อยแล้วมาใช้ตัดแบ่งเนื้อความ โดยมีการแบ่งข้อมูลออกเป็นสองชุด คือข้อมูลฝึกสอน และข้อมูลทดสอบ เพื่อใช้ในการทดสอบหาค่าตัวแปรที่เหมาะสมกับการแบ่งสายโยงใยข้อมูลมากที่สุด โดยข้อมูลฝึกสอนนี้จะทำการเปรียบเทียบความถูกต้องของข้อมูลกับ เสียงส่วนใหญ่ของการตัดสินของผู้อ่าน (Reader Judge) จำนวน 3 คน ข้อมูลชุดนี้จะเป็นข้อมูลกระทู้จำนวน 100 กระทู้จากกระดานสนทนาที่เป็นกระทู้ปัจจุบัน โดยผลลัพธ์ที่ได้ออกมานั้นการตัดสินของผู้อ่านนั้นได้แบ่งสายโยงใยออกเป็น 102 สายโยงใย ข้อมูลฝึกสอนจะถูกนำมาแบ่งสายโยงใยด้วยการปรับค่าตัวแปรต่างๆ เพื่อให้สามารถแบ่งสายโยงใยได้ใกล้เคียงกับเสียงส่วนใหญ่ของการตัดสินของผู้อ่าน โดยจะปรับค่าตัวแปรโดยใช้วิธี F-measure มาใช้ในการคำนวณว่าค่าตัวแปรใด ให้ผลลัพธ์ที่ดีที่สุด ซึ่งในกรณีนี้ค่าที่ดีที่สุดคือการปรับค่าตัวแปรซึ่งกำหนดให้ค่าจำนวนหน่วยของคำต่อหนึ่งประโยคเทียมเท่ากับห้า (token = 5) , จำนวนประโยคเทียมต่อหนึ่งบล็อกเท่ากับสิบห้า (tile = 15) และ จำนวนอันดับสายโยงใยสูงสุดที่จะทำการตัดเท่ากับ150 (s = 150)

S=150	Tile	2		5		10		15		20	
Token		Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
2		50	34	49	33	53	36	35	39	50	34
5		51	35	51	35	52	36	34	39	50	34
10		45	39	41	36	42	39	32	40	34	37
15		48	33	**54**	**37**	39	39	26	38	19	36
20		30	40	29	42	25	40	25	41	20	41
50		30	40	11	42	5	32	7	50	0	0
100		8	56	3	38	0	0	0	0	0	100

ตาราง 3 ผลลัพธ์จากการทดสอบหาค่าตัวแปรที่ดีที่สุดสำหรับข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ จำนวน 100 กระทู้ จากกระทู้ปัจจุบันบน www.pantip.com/cafe/food

จากนั้นทำการนำข้อความทั้งหมดเข้าไปทำการตัดแบ่งข้อมูลที่ต่อเนื่องกันออกจากกันเป็นสายโยงใยด้วยตัวแปรที่เหมาะสมที่สุดจากการทดลอง ผลลัพธ์ของการตัดแบ่งเนื้อความจะแสดงออกมาในรูปแบบของจุดแบ่งขอบเขตบนเนื้อหาที่นำมาคำนวน จากนั้นจึงนำขอบเขตนั้นไปพิจารณาตัดแบ่งเนื้อความที่มีออกเป็นสายโยงใยที่แยกจากกันอย่างชัดเจน โดยผลลัพธ์ที่ได้นั้น ถูกแบ่งออกเป็น 25322 สายโยงใยที่เป็นหัวข้อที่แตกต่างกัน ข้อมูลที่ได้จะถูกนำมาแบ่งออกเป็นเอกสารคนละชุดกันอย่างสิ้นเชิง

4306	1	0	0	0.0	0
4307	1	0	0	0.0	1
4308	1	0	0	0.4	0
4309	1	1	1	1.0	1
4310	1	0	0	0.0	0
4311	1	0	0	0.0	0
4312	1	0	0	0.0	0
4313	1	0	0	0.0	1
4314	1	0	0	0.4	0
4315	1	1	1	1.0	1
4316	1	0	0	0.0	0
4317	1	0	0	0.0	0

ภาพ 12 ผลการแบ่งสายโยงใย แสดงผลโดยบอกรายละเอียดที่จำเป็นต่อการตัดแบ่งดังนี้
คอลัมน์แรก หมายถึง ลำดับที่ของส่วนที่คั่นไว้ด้วยคำว่า segment_break
คอลัมน์สุดท้าย หมายถึง จุดนี้คือจุดแบ่งสายโยงใยหรือไม่ หากเป็น 1 คือ จุดแบ่งสายโยงใย

4.4 การคำนวณหาค่าน้ำหนักของคำ และการจัดลำดับน้ำหนักของคำสำคัญ

เมื่อได้สายโยงใยที่มีการแบ่งออกจากกันแล้ว จะเป็นขั้นตอนการคำนวณหาค่ากลุ่มคำสำคัญที่เกี่ยวข้องกับคำสำคัญที่ต้องการ โดยทำการนับจำนวนคำที่มีการปรากฏในแต่ละสายโยงใย ทำให้สามารถนำเอาจำนวนคำที่ได้ไปใช้ในการคำนวณต่อไปได้

ร้าน 10
อาหาร 3
นาน 1
ไม่ 8
ว่า 6

ภาพ 13 รูปแบบของคำที่ถูกนับในแต่ละสายโยงใย จะแสดงด้วย คำ และ จำนวนของคำนั้นที่ปรากฏอยู่ภายในสายโยงใยนั้นๆ

ข้อมูลของคำที่ถูกนับและแบ่งตามสายโยงใยทั้งหมดที่มี จะถูกนำมาคำนวณและสร้างตารางค่าส่วนกลับของความถี่เอกสาร (Inverse Document Frequency : IDF) สำหรับคำทั้งหมด คำนวณโดยการนำคำแต่ละคำที่มีทั้งหมดมานับว่า คำแต่ละหนึ่งคำนั้น มีการปรากฏในสายโยงใยทั้งหมดกี่สายโยงใย ผลลัพธ์ที่นับได้คือค่าของ ความถี่เอกสาร (Document

Frequency : DF) จากนั้นจึงนำค่าความถี่เอกสารที่ได้มาคำนวณหาค่าส่วนกลับของความถี่เอกสาร ตามสมการ IDF = log N/DF

Word	DF	IDF
ฟาร์ม	8	2.395545
ใต้	86	1.364136
พลาด	83	1.379557
ข้างๆ	41	1.685851
บริเวณ	30	1.821514
ก้อน	96	1.316364
ไขมัน	35	1.754567
มอง	108	1.265211
ขูด	23	1.936907
จำนวน	29	1.836237

ตาราง 4 ตัวอย่างตารางผลการคำนวณค่าส่วนกลับของความถี่เอกสาร (IDF)

การคำนวณหาค่าความถี่ของคำ (Term Frequency : TF) ที่จะนำไปใช้ในการคำนวณค่าน้ำหนักแบบ TFIDF นั้น จะทำการคำนวณเฉพาะขอบเขตของจำนวนคำที่มีความเกี่ยวข้องกับคำที่เป็นคำสำคัญที่รับข้อมูลเข้ามาตามที่ต้องการเพื่อหาคำสำคัญที่เกี่ยวข้องเท่านั้น นั่นคือ จะทำการคำนวณเฉพาะ คำที่ปรากฏอยู่ในสายโยงใยเดียวกันกับคำสำคัญที่ต้องการหาคำที่มีความเกี่ยวข้องเท่านั้น และคำนวณเฉพาะจำนวนคำที่อยู่ในสายโยงใยเดียวกันเท่านั้น

การทดลองดำเนินการโดยให้ผู้ใช้งานซึ่งเป็นทีมงานฝ่ายการตลาดของบริษัท CPRAM เข้ามาช่วยในการระบุคำสำคัญที่ต้องการนำมาใช้ค้นหากลุ่มคำสำคัญที่เกี่ยวข้อง โดยตัวอย่างการทดลองในตาราง 5 จะนำเสนอตัวอย่างคำสำคัญ 3 คำ ซึ่งเป็นคำสำคัญที่ได้มาจากการเก็บข้อมูลความต้องการของผู้ใช้งาน โดยมีความแตกต่างในด้านของขอบเขตความหมายของคำอย่างชัดเจน เพื่อจะทำการค้นหากลุ่มคำสำคัญที่เกี่ยวข้องเป็นสามกลุ่มตามคำสำคัญแต่ละคำ ได้แก่ ติ่มซำ, กับแกล้ม และ ยำลูกชิ้น

เมื่อมีข้อมูลส่วนของ ความถี่ของคำและส่วนกลับความถี่ของเอกสารสำหรับในแต่ละคำแล้ว คำนวณหาค่าน้ำหนัก TFIDF โดยการนำค่าของความถี่ของคำและส่วนกลับความถี่ของเอกสารสำหรับในแต่ละคำ มาคูณกัน ผลลัพธ์ที่ได้จะแสดงเป็นค่าน้ำหนักแบบ TFIDF ของคำในแต่ละคำที่แตกต่างกันออกไป โดยผลลัพธ์ที่ได้นั้นพิจารณาจากการให้ผู้ใช้งานเลือกคำสำคัญจำนวน 20 คำ มาจากคำสำคัญจำนวน 200 อันดับแรกของแต่ละกลุ่มของคำสำคัญทั้งสามคำ ดังนี้

ต้มซำ	กับแกล้ม	ยำลูกชิ้น
ร้าน	กับแกล้ม	หมู
ต้มซำ	อร่อย	ผัด
อาหาร	ไก่	อาหาร
ทอด	ชอบ	อร่อย
ราคา	ปลา	ลูกชิ้น
น้ำ	ทอด	ยำ
หมู	เนื้อ	ทอด
ตอน	พริก	ปลา
เนื้อ	รส	เนื้อ
ลอง	ยำ	แม่
ไก่	หอม	ผัก
ปลา	หมู	ชอบ
ข้าว	รสชาติ	ต้ม
ขนม	ใหม่	พริก
ผัด	หลาย	ไก่
จีน	ต้ม	กุ้ง
รส	เผ็ด	ก๋วยเตี๋ยว
เปิด	เหล้า	รส
ซา	อ้วน	เยอะ
แถว	กับข้าว	หลาย

ตาราง 5 ผลลัพธ์ของคำสำคัญ 20 อันดับแรกที่ผู้ใช้งานพิจารณาว่ามีความสำคัญเหมาะสมสามารถนำไปใช้ในการพัฒนาผลิตภัณฑ์ใหม่ในอุตสาหกรรมอาหารได้ โดยเลือกขึ้นมาจากคำสำคัญที่จัดอันดับด้วยวิธีการให้น้ำหนักด้วยค่า TFIDF จำนวน 200 อันดับแรก

ต้มซ่า : 244 สายโยงใย			กับแกล้ม : 63 สายโยงใย			ยำลูกชิ้น : 76 สายโยงใย		
Ranking	Word	TFIDF	Ranking	Word	TFIDF	Ranking	Word	TFIDF
1	ครับ	1497.526	1	ครับ	325.3397	1	ไม่	794.4042
2	ร้าน	1359.32	2	ค่ะ	318.3881	2	ร้าน	729.9408
3	ไม่	1169.707	3	ไม่	281.6822	3	ก็	709.3223
4	ที่	1138.169	4	ก็	277.4028	4	ครับ	673.0109
5	ไป	1107.488	5	ไป	267.1961	5	ค่ะ	659.234
6	ค่ะ	1087.352	6	ร้าน	263.5416	6	ที่	652.6436
7	ก็	1000.177	7	มา	257.7477	7	จะ	604.5852
8	มา	968.4722	8	จะ	228.6247	8	ไป	601.1912
9	ว่า	907.5321	9	เลย	225.7713	9	มา	572.3307
10	คุณ	893.6926	10	กิน	223.5627	10	ว่า	550.346
11	จะ	881.3057	11	เรา	214.9458	11	เรา	548.7233
12	ต้มซ่า	873.9791	12	คุณ	213.2319	12	ได้	509.1664
13	เลย	863.0899	13	แล้ว	212.3704	13	แล้ว	501.4302
14	เรา	849.2979	14	ว่า	211.1189	14	แต่	492.9188
15	มี	845.6554	15	เป็น	198.245	15	กิน	491.1124
16	นี้	835.1796	16	ได้	196.2592	16	คะ	476.4977
17	ได้	833.4961	17	มี	192.244	17	เลย	476.4674
18	กิน	808.9978	18	คน	191.0129	18	มี	474.9771
19	นะ	788.9243	19	แต่	190.9717	19	ใส่	470.4988
20	คะ	787.4559	20	นี้	186.8838	20	ทำ	465.5601
21	แต่	770.9454	21	ทำ	185.6378	21	หมู	464.993
22	แล้ว	760.6663	22	คะ	181.0874	22	นี้	455.9556
23	อร่อย	749.103	23	กับแกล้ม	179.6869	23	ผัด	453.8203
24	ทาน	710.6108	24	นะ	178.818	24	เป็น	437.1974
25	เป็น	691.6186	25	กัน	176.6979	25	ผม	418.7515

ตาราง 6 ผลลัพธ์ของการสกัดคำสำคัญด้วยวิธีการให้น้ำหนักด้วยค่า TFIDF ซึ่งแสดงตัวอย่างในรูปอันดับของคำที่สกัดออกมาได้ 25 อันดับแรก เรียงลำดับความสำคัญด้วยคะแนน TFIDF จากมากไปหาน้อยส่วนที่อยู่ในช่องสีทึบคือคำที่เป็นคำสำคัญตามที่ผู้ใช้งานพิจารณาแล้วว่าเป็นคำที่มีความสำคัญและสามารถนำไปใช้ในการพัฒนาผลิตภัณฑ์ใหม่ในอุตสาหกรรมอาหารสร้างสรรค์ได้

นำมาคำนวณค่าน้ำหนักคำสำคัญอีกครั้ง ด้วยการให้น้ำหนักแบบ MT-TFIDF ซึ่งเป็นการปรับค่าน้ำหนักของคำสำคัญเพื่อให้ได้คำสำคัญตามที่ต้องการมากที่สุด จากสมการดังนี้

$$Score_{word} = \left(\frac{\tau}{\tau + \tau'}\right)\left(\frac{\tau}{\mathrm{T}} - \frac{\tau'}{\mathrm{T}'}\right)TFIDF$$

โดยผลลัพธ์ที่ได้นั้นจะแสดงตัวอย่าง 25 คำแรกที่มีความสำคัญสูงที่สุดโดยพิจารณาจากค่า MT-TFIDF ตามตารางที่ 7

ต้มซำ : 244 สายโยงใย			กับแกล้ม : 63 สายโยงใย			ยำลูกชิ้น : 76 สายโยงใย		
Ranking	Word	MT-TFIDF	Ranking	Word	MT-TFIDF	Ranking	Word	MT-TFIDF
1	ต้มซำ	878.2739	1	กับแกล้ม	179.6869	1	ลูกชิ้น	50.15949
2	เข่ง	16.2394	2	ทอด	0.751547	2	ยำ	31.69739
3	ครับ	14.26757	3	จาน	0.705549	3	หมู	6.439246
4	ร้าน	14.16532	4	กิน	0.652069	4	ผัด	6.02641
5	อร่อย	10.1578	5	ไก่	0.641581	5	ก๋วยเตี๋ยว	5.999136
6	อาหาร	9.955177	6	ร้าน	0.625645	6	ทอด	4.791488
7	โรงแรม	9.574578	7	ปลา	0.580571	7	ใส่	4.705734
8	จีบ	9.505251	8	บ้าน	0.573714	8	ร้าน	4.690338
9	ทาน	9.110746	9	เอา	0.551674	9	ปลา	4.058167
10	ทอด	8.998207	10	พริก	0.548151	10	กุ้ง	3.782892
11	ราคา	8.998064	11	ครับ	0.482226	11	ผัก	3.737811
12	ที่	8.76823	12	เป็น	0.477482	12	น้ำ	3.660471
13	ผม	8.032635	13	กับ	0.463938	13	ต้ม	3.397649
14	ไป	7.235316	14	เหล้า	0.460348	14	เนื้อ	3.341362
15	ว่า	7.079677	15	รส	0.460144	15	ข้าว	3.319405
16	ไม่	7.004963	16	ก็	0.457497	16	เรา	3.207655
17	จีน	6.974909	17	ข้าว	0.457034	17	สั่ง	3.194713
18	เก๋า	6.905024	18	ยำ	0.441344	18	ครับ	3.064339
19	ก็	6.686571	19	ชอบ	0.438571	19	ผม	3.026719
20	นี่	6.443296	20	ไว้	0.428297	20	จาน	2.994319
21	เรา	6.368894	21	เรา	0.422477	21	พริก	2.946819
22	คุณ	6.281527	22	ไม่	0.418113	22	ไก่	2.903081
23	กับ	6.246458	23	คุณ	0.404281	23	เส้น	2.888479
24	บุ	6.056355	24	อร่อย	0.395941	24	ขาย	2.834558
25	มา	6.052022	25	ผม	0.389811	25	เอา	2.807169

ตาราง 7 ผลลัพธ์ของการสกัดคำสำคัญด้วยวิธีการให้น้ำหนักด้วยค่า TFIDF ซึ่งแสดงตัวอย่างในรูปอันดับของคำที่สกัดออกมาได้ 25 อันดับแรก เรียงลำดับความสำคัญด้วยคะแนน TFIDF จากมากไปหาน้อยส่วนที่อยู่ในช่องสีทึบคือคำที่เป็นคำสำคัญตามที่ผู้ใช้งานพิจารณาแล้วว่าเป็นคำที่มีความสำคัญและสามารถนำไปใช้ในการพัฒนาผลิตภัณฑ์ใหม่ในอุตสาหกรรมอาหารสร้างสรรค์ได้

ตัวอย่างผลลัพธ์ที่ได้จากการสกัดคำสำคัญด้วยการใช้ค่า TFIDF และ MT-TFIDF นั้นนำมาเปรียบเทียบแยกตามคำสำคัญที่นำมาใช้ทดลอง โดยแสดงตามตารางที่ 8, 9 และ 10 ดังนี้

ติ่มซำ : 244 สายโยงใย				
Ranking	Word	TFIDF	Word	MT-TFIDF
1	ครับ	1497.526	ติ่มซำ	878.2739
2	ร้าน	1359.32	เข่ง	16.2394
3	ไม่	1169.707	ครับ	14.26757
4	ที่	1138.169	ร้าน	14.16532
5	ไป	1107.488	อร่อย	10.1578
6	ค่ะ	1087.352	อาหาร	9.955177
7	ก็	1000.177	โรงแรม	9.574578
8	มา	968.4722	จีบ	9.505251
9	ว่า	907.5321	ทาน	9.110746
10	คุณ	893.6926	ทอด	8.998207
11	จะ	881.3057	ราคา	8.998064
12	ติ่มซำ	873.9791	ที่	8.76823
13	เลย	863.0899	ผม	8.032635
14	เรา	849.2979	ไป	7.235316
15	มี	845.6554	ว่า	7.079677
16	นี้	835.1796	ไม่	7.004963
17	ได้	833.4961	จีน	6.974909
18	กิน	808.9978	เก๋า	6.905024
19	นะ	788.9243	ก็	6.686571
20	คะ	787.4559	นี้	6.443296
21	แต่	770.9454	เรา	6.368894
22	แล้ว	760.6663	คุณ	6.281527
23	อร่อย	749.103	กับ	6.246458
24	ทาน	710.6108	บุ	6.056355
25	เป็น	691.6186	มา	6.052022

ตาราง 8 ผลลัพธ์ของการจัดอันดับคำสำคัญ โดยนำเสนอ 25 อันดับแรกของ คำสำคัญที่มีความสัมพันธ์กับคำว่า “ติ่มซำ คำที่อยู่ในช่องสีทึบคือคำที่เป็นคำสำคัญที่ทางผู้เชี่ยวชาญเลือกขึ้นมา 20 อันดับแรกซึ่งเป็นคำชุดเดียวกันทั้งวิธีการคำนวณด้วยTFIDF และ วิธีการคำนวณด้วย MT-TFIDF โดยผลที่ได้นั้นแสดงให้เห็นว่าคำสำคัญที่ผู้ใช้งานได้เลือกไว้ในการทดลองด้วยการคำนวณด้วย TFIDF นั้น มีการปรากฏในอันดับต้นๆมากขึ้นในการทดลองด้วยการคำนวณการสกัดคำสำคัญด้วย MT-TFIDF

กับแกล้ม : 63 สายโยงใย				
Ranking	Word	TFIDF	Word	MT-TFIDF
1	ครับ	325.3397	กับแกล้ม	179.6869
2	ค่ะ	318.3881	ทอด	0.751547
3	ไม่	281.6822	จาน	0.705549
4	ก็	277.4028	กิน	0.652069
5	ไป	267.1961	ไก่	0.641581
6	ร้าน	263.5416	ร้าน	0.625645
7	มา	257.7477	ปลา	0.580571
8	จะ	228.6247	บ้าน	0.573714
9	เลย	225.7713	เอา	0.551674
10	กิน	223.5627	พริก	0.548151
11	เรา	214.9458	ครับ	0.482226
12	คุณ	213.2319	เป็น	0.477482
13	แล้ว	212.3704	กับ	0.463938
14	ว่า	211.1189	เหล้า	0.460348
15	เป็น	198.245	รส	0.460144
16	ได้	196.2592	ก็	0.457497
17	มี	192.244	ข้าว	0.457034
18	คน	191.0129	ยำ	0.441344
19	แต่	190.9717	ชอบ	0.438571
20	นี้	186.8838	ไว้	0.428297
21	ทำ	185.6378	เรา	0.422477
22	คะ	181.0874	ไม่	0.418113
23	กับแกล้ม	179.6869	คุณ	0.404281
24	นะ	178.818	อร่อย	0.395941
25	กัน	176.6979	ผม	0.389811

ตาราง 9 ผลลัพธ์ของการจัดอันดับคำสำคัญ โดยนำเสนอ 25 อันดับแรกของ คำสำคัญที่มีความสัมพันธ์กับคำว่า “กับแกล้ม” คำที่อยู่ในช่องสีทึบคือคำที่เป็นคำสำคัญที่ทางผู้เชี่ยวชาญเลือกขึ้นมา 20 อันดับแรกซึ่งเป็นคำชุดเดียวกันทั้งวิธีการคำนวณด้วยTFIDF และ วิธีการคำนวณด้วย MT-TFIDF โดยผลที่ได้นั้นแสดงให้เห็นว่าคำสำคัญที่ผู้ใช้งานได้เลือกไว้ในการทดลองด้วย การคำนวณด้วย TFIDF นั้น มีการปรากฏในอันดับต้นๆมากขึ้นในการทดลองด้วยการคำนวณการสกัดคำสำคัญด้วย MT-TFIDF

ยำลูกชิ้น : 76 สายโยงใย				
Ranking	Word	TFIDF	Word	MT-TFIDF
1	ไม่	794.4042	ลูกชิ้น	50.15949
2	ร้าน	729.9408	ยำ	31.69739
3	ก็	709.3223	หมู	6.439246
4	ครับ	673.0109	ผัด	6.02641
5	ค่ะ	659.234	ก๋วยเตี๋ยว	5.999136
6	ที่	652.6436	ทอด	4.791488
7	จะ	604.5852	ใส่	4.705734
8	ไป	601.1912	ร้าน	4.690338
9	มา	572.3307	ปลา	4.058167
10	ว่า	550.346	กุ้ง	3.782892
11	เรา	548.7233	ผัก	3.737811
12	ได้	509.1664	น้ำ	3.660471
13	แล้ว	501.4302	ต้ม	3.397649
14	แต่	492.9188	เนื้อ	3.341362
15	กิน	491.1124	ข้าว	3.319405
16	คะ	476.4977	เรา	3.207655
17	เลย	476.4674	สั่ง	3.194713
18	มี	474.9771	ครับ	3.064339
19	ใส่	470.4988	ผม	3.026719
20	ทำ	465.5601	จาน	2.994319
21	หมู	464.993	พริก	2.946819
22	นี้	455.9556	ไก่	2.903081
23	ผัด	453.8203	เส้น	2.888479
24	เป็น	437.1974	ขาย	2.834558
25	ผม	418.7515	เอา	2.807169

ตาราง 10 ผลลัพธ์ของการจัดอันดับคำสำคัญ โดยนำเสนอ 25 อันดับแรกของ คำสำคัญที่มีความสัมพันธ์กับคำว่า “ยำลูกชิ้น” คำที่อยู่ในช่องสีทึบคือคำที่เป็นคำสำคัญที่ทางผู้เชี่ยวชาญเลือกขึ้นมา 20 อันดับแรกซึ่งเป็นคำชุดเดียวกันทั้งวิธีการคำนวณด้วยTFIDF และ วิธีการคำนวณด้วย MT-TFIDF โดยผลที่ได้นั้นแสดงให้เห็นว่าคำสำคัญที่ผู้ใช้งานได้เลือกไว้ในการทดลองด้วย การคำนวณด้วย TFIDF นั้น มีการปรากฏในอันดับต้นๆมากขึ้นในการทดลองด้วยการคำนวณการสกัดคำสำคัญด้วย MT-TFIDF

4.5 การวัดผลการทดลอง

จากผลการทดลองที่ได้นั้น เมื่อมีการจัดอันดับของคำสำคัญที่สกัดออกมาได้แล้ว การสกัดคำสำคัญโดยวิธีการประยุกต์การให้น้ำหนักด้วยค่า MT-TFIDF นั้น ได้ผลลัพธ์ที่ดีว่า การสกัดคำสำคัญโดยคำนวณจากการให้น้ำหนักด้วยค่า TFIDF เพียงอย่างเดียว

โดยการวัดผลนั้น ดำเนินการโดยใช้การวัดผลทางสถิติแบบ T-Test โดยมีการกำหนดสมมติฐานดังต่อไปนี้

H0 = วิธีการให้น้ำหนักความสำคัญของคำแบบ TFIDF ไม่แตกต่างกันหรือแย่กว่า วิธีการให้น้ำหนักความสำคัญของคำแบบ MT-TFIDF

H1 = วิธีการให้น้ำหนักความสำคัญแบบ MT-TFIDF ให้ผลลัพธ์ที่ดีกว่า วิธีการให้น้ำหนักความสำคัญแบบ TFIDF

โดยสำหรับคำว่า"ติ่มซำ" ที่นำเข้ามาใช้ในการคำนวณนั้น ผลการทดสอบด้วย การวัดผลทางสถิติแบบ T-Test โดยเปรียบเทียบผลลัพธ์ของการจัดอันดับด้วยวิธีการให้น้ำหนักความสำคัญของคำแบบ TFIDF และ MT-TFIDFนั้น ได้ค่า P-value เท่ากับ 0.09 สามารถสรุปได้ว่า ในกรณีของการทดลองด้วยข้อมูลคำว่า "ติ่มซำ"นั้น ผลลัพธ์ที่ได้จากการให้ค่าน้ำหนักความสำคัญแบบ MT-TFIDF ดีกว่า TFIDF ที่ความเชื่อมั่น 90%

และสำหรับคำว่า “กับแกล้ม” และ คำว่า “ยำลูกชิ้น” นั้น ผลการทดสอบด้วย การวัดผลทางสถิติแบบ T-Test โดยเปรียบเทียบผลลัพธ์ของการจัดอันดับด้วยวิธีการให้น้ำหนักความสำคัญของคำแบบ TFIDF และ MT-TFIDF ได้ค่า P-value เท่ากับ 0.0000086 และ 0.000019 ตามลำดับ ทำให้สามารถสรุปได้ว่า ในกรณีของการทดลองด้วยข้อมูลคำว่า "กับแกล้ม" และ “ยำลูกชิ้น”นั้น ผลลัพธ์ที่ได้จากการให้ค่าน้ำหนักความสำคัญแบบ MT-TFIDF ดีกว่า TFIDF ที่ความเชื่อมั่น 99%

นอกจากตัวอย่างขั้นต้นแล้ว ได้มีการทดลองกับคำสำคัญอื่นๆ คือ คำว่า ข้าวปั้น สปาเก็ตตี้ สเต็ก ทอดมัน สลัด โดยผลลัพธ์ที่ได้จากคำว่า “ข้าวปั้น”, “สเต็ก”และ “ทอดมัน” นั้น ทดสอบด้วย การวัดผลทางสถิติแบบ T-Test โดยเปรียบเทียบผลลัพธ์ของการจัดอันดับด้วยวิธีการให้น้ำหนักความสำคัญของคำแบบ TFIDF และ MT-TFIDF ได้ค่า P-value เท่ากับ 0.072 , 0.073 และ 0.087 ตามลำดับ ทำให้สามารถสรุปได้ว่า ในกรณีนี้ ผลลัพธ์ที่ได้จากการให้ค่าน้ำหนัก

ความสำคัญแบบ MT-TFIDF ดีกว่า TFIDF ที่ความเชื่อมั่น 90% ส่วนคำว่า สปาเก็ตตี้ และ สลัด ทดสอบด้วย การวัดผลทางสถิติแบบ T-Test ได้ค่า P-value เท่ากับ 0.028 และ 0.003 ตามลำดับ สรุปได้ว่า ในกรณีนี้ ผลลัพธ์ที่ได้จากการให้ค่าน้ำหนักความสำคัญแบบ MT-TFIDF ดีกว่า TFIDF ที่ความเชื่อมั่น 95%

ซึ่งจากการทดลองจากตัวอย่างทั้งหมดดังกล่าว สรุปได้ว่าคำสำคัญตามตัวอย่างทั้งหมดนั้น แสดงผลลัพธ์ว่าการให้ค่าน้ำหนักความสำคัญแบบ MT-TFIDF ดีกว่า TFIDF ที่ความเชื่อมั่นมากกว่า 90% ในทุกๆคำสำคัญที่นำมาใช้ทดลอง

บทที่ 5

สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

งานวิจัยชิ้นนี้นำเสนอการสกัดกลุ่มคำสำคัญในหัวข้อที่ต้องการจากข้อมูลการสื่อสารปากต่อปากบนเว็บบอร์ด โดยการระบุคำสำคัญที่ต้องการค้นหากลุ่มคำสำคัญที่มีความสัมพันธ์กันเพื่อให้สามารถนำกลุ่มคำสำคัญมาใช้ประโยชน์ในการตลาดได้ โดยใช้วิธีการแบ่งเนื้อหา มาแบ่งข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ออกเป็นสายโยงใยที่อิสระต่อกัน เพื่อให้น้ำหนักของคำสำคัญที่ได้มีลักษณะเพิ่มขึ้นตามการปรากฏในจำนวนสายโยงใยที่หลากหลายมากกว่าการที่มีคำนั้นปรากฏซ้ำๆในสายโยงใยเดียว และเพื่อสร้างโครงสร้างของข้อมูลการสื่อสารปากต่อปากแบบอิเล็กทรอนิกส์ ให้สามารถค้นคืนสารสนเทศได้ดีขึ้น และใช้การให้น้ำหนักความคำสำคัญของคำ MT-TFIDF มาใช้ในการสกัดและจัดอันดับความสำคัญของคำสำคัญขึ้นมา โดยผลลัพธ์ที่ได้นั้นแสดงให้เห็นว่าการใช้วิธีการแบ่งเนื้อหาและการให้น้ำหนักคำสำคัญด้วยวิธีการที่นำเสนอนั้น ได้ผลลัพธ์ที่ดีกว่า การให้น้ำหนักคำสำคัญด้วยค่า TFIDF ตามปกติ โดยการวัดผลด้วยวิธีวัดผลสถิติแบบ T-Test ที่ความเชื่อมั่นตั้งแต่ 90% ขึ้นไป

ผลลัพธ์ที่ได้จากงานวิจัยนั้นจะถูกแสดงเป็นหน่วยย่อยที่เล็กที่สุดของคำ ทำให้การค้นหาคำที่ยาวๆหรือคำที่เป็นคำประสมมีปัญหาเกิดขึ้น เพราะว่าระบบการตัดคำจะตัดออกเป็นคำย่อยๆออกจากกัน เช่นคำว่า ขนมจีน จะกลายเป็นคำว่า ขนม และ จีน เป็นต้น รวมถึงการตัดคำในบางกรณีที่ตัดออกมาได้ไม่สมบูรณ์ ทำให้ผลลัพธ์ที่ได้คลาดเคลื่อนไป นอกจากนี้ลักษณะของข้อมูลที่ไม่เป็นทางการยังส่งผลให้การตัดคำเป็นไปได้ยากเช่นกัน เพราะคำศัพท์ที่ใช้อาจสะกดไม่ตรงตามพจนานุกรม ทำให้ระบบตัดคำไม่สามารถตัดคำถูกต้องได้ ดังนั้นการพัฒนาระบบตัดคำที่ดีขึ้นมีความเป็นไปได้ที่จะช่วยให้ผลลัพธ์ของงานวิจัยชิ้นนี้ได้ผลลัพธ์ที่น่าดียิ่งขึ้น

รายการอ้างอิง

[1] Wu Tong, Electronics word of mouth in online social networks Communication Systems, Networks and Applications (ICCSNA), 2010 Second International Conference, pp. 249 – 253 Hong Kong, 2010

[2] Hearst, Marti A. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics 23,1 (1997) :33-64

[3] Williams, M. , Word of mouth: A definition of communication. Elmhurst College ,2007

[4] Steffes, E. M., and Burgee, L. E. . Social ties and online word of mouth. Internet Research 19,1 (2009) :42-59

[5] De Bryun, A., and Lilien, G. L. . A multi-stage model of word of mouth influence through viral marketing.Intern. J. Research in Marketing 25 (2008) :151-163

[6] Alexandru, B. The art and science of word of mouth and electronic word of mouth. Fascicle of Management and Technological Engineering 9,19 (2010) :7-16

[7] Bernard, J. J. and Mini, Z., Twitter power: tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology 60,11 (2007) : 2169-2188.

[8] Wong, P., Su, Y., Shih, M., and Luo, S., Analysis of online word of mouth in online forums regarding notebook computers. Journal of Convergence Information Technology 5,5 , (2010) :118-124

[9] Zajic, D. M., Dorr, B. J., and Lin, J., Single-document and multi-document summarization techniques for email threads using sentence compression. Information Processing and Management 44,4 (2008) :1600-1610

[10] Park, C., and Lee, T. M. . Information Direction, Website Reputation and eWOM Effect., Journal of Business Research, 62,1 (2009) :61-67

[11] Hovy, H. E. Automated text summarization. In R. Mitkov (ed), The Oxford Handbook of Computational Linguistics , pp. 583–598. Oxford: Oxford University Press. 2005

[12] Hearst, Marti A., Plaunt, and Christian. Subtopic structuring for full-length document

access. Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Infofmation Retrieval (1993) :59-68

[13] Harrag F., Hamdi-Cherif A., and Al-Salman A.S.. Comparative study of topic segmentation algorithms based on lexical cohesion: Experimental results on Arabic language Arabian Journal for Science and Engineering 35,2C (2010) :183-202

[14] Xie L., Zeng J., Feng W., Multi-scale text Tiling for automatic story segmentation in Chinese broadcast news. Computer Science 4993 (2008) :345-355

[15] Paisarn Charoenpornsawat, SWATH: Thai Word Segmentation[Online].1999. Available from : http://www.cs.cmu.edu/ ~paisarn/software.html [2011,July]

[16] Wartena C., Brussee R., and Slakhorst W.. Keyword extraction using word co-occurrence Proceedings - 21st International Workshop on Database and Expert Systems Applications, pp.54-58. Bilbao, 2010

[17] Lee S.,and Kim H.-J. , News keyword extraction for topic tracking Proceedings - 4th International Conference on Networked Computing and Advanced Information Management, pp. 554-559. Gyeongju, 2008

[18] Li, J., Fan, Q., Zhang, K., Keyword extraction based on tf/idf for Chinese news document Journal of Natural Sciences 12,5 (2007) : 917-921

ประวัติผู้เขียนวิทยานิพนธ์

นายเอกภูมิ ภูมิพันธุ์ เกิดวันที่ 1 ธันวาคม พ.ศ. 2528 จังหวัดนครราชสีมา สำเร็จการศึกษาระดับปริญญาบัณฑิตจากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ศรีราชา มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตศรีราชา เมื่อปีการศึกษา 2550