Research Article

Received: December 22, 2017; Accepted: February 3, 2018

การประยุกต์ใช้อัลกอริทึมป่าสุ่มและทฤษฎีกราฟสำหรับการวิเคราะห์ข้อความ

Applying Random Forest Algorithm and Graph Theory for Text Analyzing

วัชรีวรรณ จิตต์สกุล* และสุนันทา สดสี

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800

Watchareewan JItsakul* and Sunantha Sodsee

Faculty of Information Technology, King Mongkut's University of Technology North Bangkok,

Wongsawang, Bangsue, Bangkok 10800

บทคัดย่อ

งานวิจัยนี้นำเสนอการประยุกต์ใช้อัลกอริทึมป่าสุ่ม (random forest) และทฤษฎีกราฟสำหรับการวิเคราะห์ข้อความ โดยข้อความที่ใช้ในการทดสอบ 3 ชุด คือ ข้อความการแสดงความคิดเห็นเกี่ยวกับภาพยนตร์ ข้อความการแสดงความคิดเห็นเกี่ยวกับร้านอาหาร และข้อความการแสดงความคิดเห็นเกี่ยวกับสินค้า โดยสกัดคำจากข้อความทดสอบด้วยอัลกอริทึม random forest ต่อมานำข้อความทดสอบไปสร้างความสัมพันธ์ของคำโดยใช้ทฤษฎีกราฟ จากนั้นวัดค่าความเป็นศูนย์กลางเพื่อหาความคำสำคัญของคำ 3 วิธี ดังนี้ วิธีที่ 1 วัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (betweeness centrality, BC) วิธีที่ 2 วัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (closeness centrality, CC) และวิธีที่ 3 ค่าความเป็นศูนย์กลางจากระดับ (degree centrality, DC) เปรียบเทียบคำที่ได้จากอัลกอริทึม Random Forest และทฤษฎีกราฟ ผลลัพธ์แสดงให้เห็นว่างานวิจัยที่นำเสนอสามารถจำแนกคำจากข้อความทดสอบ 3 ชุด ได้ 3 กลุ่ม คือ กลุ่มคำที่ตรงกัน กลุ่มคำที่คล้ายกัน และกลุ่มคำที่ไม่ปรากฏ โดยมีผลลัพธ์ดังนี้ (1) กลุ่มคำที่ตรงกัน มีค่าเฉลี่ย BC, CC และ DC เท่ากับ 94.24010, 2.0369, 23.5736 (2) กลุ่มคำที่คล้ายกัน มีค่าเฉลี่ย BC, CC และ DC เท่ากับ 127.6935, 2.0286, 25.1273 และ (3) กลุ่มคำที่ไม่ปรากฏ มีค่าเฉลี่ย BC, CC และ DC เท่ากับ 38.5155, 2.1053, 18.4643 ตามลำดับ จากผลลัพธ์พบว่าวิธี BC และ DC ให้ผลลัพธ์ของค่าความสำคัญของคำได้เหมาะสมกว่า (มีค่ามากกว่า) วิธี CC ทั้ง 3 กลุ่ม และกลุ่มคำที่คล้ายกันมีค่าเฉลี่ย BC และ DC สูงกว่ากลุ่มคำที่ตรงกัน และกลุ่มคำที่ไม่ปรากฏ

คำสำคัญ : อัลกอริทึมป่าสุ่ม; ทฤษฎีกราฟ; การวิเคราะห์ข้อความ; ความสำคัญของคำ

***ผู้รับผิดชอบบทความ :** Watchareewan.j@it.kmutnb.ac.th

Abstract

This research presents an applying random forest algorithm and graph theory for text analysing. Herein, 3 - benchmark comment datasets collected from “www.imdb.com”, “www.yelp.com”, and “www.amazon.com” given by UCI Machine Learning Repository, are used to evaluate the proposed study. First, the random forest algorithm is applied to extract words from the comment datasets. Secondly, the comment datasets are created relationship between words based on a graph theory. Finally, centrality measures, namely betweeness centrality: BC, closeness centrality: CC, and degree centrality: DC, are used to identify an importance of each word compared between word extracted from the random forest algorithm and graph theory. The results showed that extracted words were classified in three groups based on matched words, similar words, and exception words. From 3-benchmark comment datasets, the matched words contained the average BC, CC and DC as 94.24010, 2.0369, and 23.5736. The similar words showed the average BC, CC and DC as 127.6935, 2.0286, 25.1273, and the exception words presented the average BC, CC and DC as 38.5155, 2.1053, and 18.4643, respectively. From the results, BC and DC, here, were more optimal than CC to text analysing based on centrality of words. Finally, the similar words contained greater average BC and DC than the matched words and exception words.

Keywords: random forest; graph theory; text analysis; word centrality

1. บทนำ

ในปัจจุบันการเข้าถึงหรือค้นหาข้อมูลต่าง ๆ โดยส่วนใหญ่จะกระทำผ่านเครือข่ายอินเทอร์เน็ต [1] ซึ่งเปลี่ยนแปลงจากอดีตที่เน้นการค้นหาข้อมูลจากตำรา เอกสาร หรือสิ่งตีพิมพ์ จึงทำให้ข้อมูลที่เผยแพร่ในเครือข่ายอินเทอร์เน็ตเหล่านี้เป็นเอกสารอิเล็กทรอนิกส์ที่มีลักษณะโครงสร้างไม่แน่นอน (unstructured text) โดยข้อมูลเหล่านี้มีมากถึง 90 % ของข้อมูลทั้งหมด [2] ข้อมูลในเครือข่ายอินเตอร์เน็ตมีจำนวนมหาศาล และมีความซับซ้อนด้านโครงสร้าง ดังนั้นการประมวลผลข้อมูลเพื่อนำมาใช้ประโยชน์ในเรื่องต่าง ๆ เช่น การวิเคราะห์ข้อมูลการให้บริการ การสนับสนุนการตัดสินใจ หรือแนะนำสินค้า โดยมุ่งเน้นความถูกต้องและเวลาในการประมวลผลต่ำนั้นกระทำได้ยาก เพื่อลดข้อจำกัดดังกล่าว มีงานวิจัยจำนวนมากนำเสนออัลกอริทึมเพื่อประยุกต์ใช้ในการวิเคราะห์ข้อมูล เช่น อัลกอริทึมวิธีการจำแนกแบบสัมพันธ์ (associative classification) [3] อัลกอริทึมต้นไม้ตัดสินใจ (decision tree) อัลกอริทึมนาอีฟเบย์ (Naïve Bayes) อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (support vector machine) [4] และอัลกอริทึมเพื่อนบ้านใกล้ที่สุด (k-nearest neighbor) [5] นอกจากนี้มีงานวิจัยที่ประยุกต์ใช้ทฤษฎีกราฟ (graph theory) ในการวิเคราะห์ข้อมูลที่มีคุณลักษณะเป็นข้อความ คำในข้อความจะถูกนำมาสร้างความสัมพันธ์ของคำ และหาความสำคัญของแต่ละคำด้วยวิธีการวัดค่าความเป็นศูนย์กลาง (centrality measure) เช่น วิธี dependency graph [6] วิธี graph structure model [7] degree

centrality, in-degree centrality, out-degree centrality และ closeness centrality [8] ทำให้การประมวลผลข้อความมีประสิทธิภาพมากขึ้น

จากงานวิจัยดังกล่าวข้างต้น งานวิจัยนี้จึงมีแนวความคิดในการวิเคราะห์ข้อมูลเอกสารอิเล็กทรอนิกส์โดยการหาความสำคัญของคำด้วยทฤษฎีกราฟประยุกต์ใช้ร่วมกับอัลกอริทึมสำหรับการจำแนกข้อมูล

โครงสร้างของงานวิจัยนี้ประกอบด้วยทฤษฎีและงานวิจัยที่เกี่ยวข้องนำเสนอในส่วนที่ 2 ส่วนที่ 3 นำเสนอวิธีการดำเนินงาน ส่วนที่ 4 นำเสนอผลการดำเนินงาน และส่วนสุดท้ายนำเสนอการสรุปผลงานวิจัย

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในส่วนของทฤษฎีและงานวิจัยที่เกี่ยวข้องนั้นศึกษาเกี่ยวกับทฤษฎีกราฟ และการวัดค่าความเป็นศูนย์กลาง ดังนี้

2.1 ทฤษฎีกราฟ

กราฟ (graph) เป็นแบบจำลองทางคณิตศาสตร์ คิดค้นโดยนักคณิตศาสตร์ชาวสวิตเซอร์แลนด์ เลออนฮาร์ดออยเลอร์ (Leonhard Euler) [9] กราฟสามารถใช้แทนปัญหาในโลกของความเป็นจริง โดยจำลองปัญหาด้วยแผนภาพที่ประกอบด้วยจุด (point) หรือเรียกว่าโหนด (node) และเส้นที่เชื่อมระหว่างจุด 2 จุด หรือเส้นเชื่อม (edge) ตัวอย่าง เช่น แผนภาพแสดงเส้นทางการบิน แผนภาพแสดงเส้นทางรถไฟฟ้าและวงจรไฟฟ้า อีกทั้งปัจจุบันยังมีการนำไปประยุกต์ใช้ในด้านต่าง ๆ อย่างกว้างขวาง เช่น ปัญหาด้านจิตวิทยาภาษาศาสตร์ เทคโนโลยีคอมพิวเตอร์ และเศษฐศาสตร์ ซึ่งในงานวิจัยฉบับนี้ กราฟถูกนำมาใช้ในการจำลองความสัมพันธ์ของคำในข้อความ โดยกราฟ (G) ประกอบด้วยโหนดสมาชิก (V) และเส้นเชื่อมระหว่างโหนด (E) [9] แสดงดังสมการ

$$G = (V, E)$$

โดยที่ V เป็นเซตจำกัดที่ไม่เป็นเซตว่างของสมาชิกที่เรียกว่าจุดยอด หรือโหนด (node); E เป็นเซตของเส้นเชื่อม (link) ระหว่างโหนด

การวัดค่าความเป็นศูนย์กลาง (centrality measure) หรือความสำคัญของโหนด เป็นการวิเคราะห์การเชื่อมต่อของกราฟ ซึ่งมีการวัดอยู่หลายวิธี ในงานวิจัยนี้ศึกษาการวัด 3 รูปแบบ ดังนี้

2.2.1 วัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (betweenness centrality, BC) เป็นการกำหนดความสำคัญของโหนด โดยพิจารณาว่าโหนดที่มีสถานะเป็นสะพานเพื่อเชื่อมกลุ่มของโหนดต่าง ๆ ที่อยู่ห่างกันให้สามารถเข้าหากัน หรือเป็นตัวกลางในการติดต่อเชื่อมโยงระหว่างสมาชิกอื่น ๆ ดังนี้ [10]

$$BC(i) = \sum_{j \neq i \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$$

โดยที่ $BC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากการคั่นกลางของโหนด i; σ_{jk} คือ จำนวนเส้นเชื่อมโยงที่สั้นที่สุดจากโหนด j ไปยังโหนด k; $\sigma_{jk(i)}$ คือ จำนวนเส้นเชื่อมโยงที่สั้นที่สุดจากโหนด j ไปยังโหนด k ที่ต้องผ่านโหนด i

2.2.2 วัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (closeness centrality, CC) เป็นการกำหนดความสำคัญของโหนดโดยพิจารณาจากความสามารถในการเข้าถึงโหนดอื่น ๆ ด้วยระยะทางที่สั้นที่สุด [11]

$$CC(i) = \frac{1}{\sum_j d(i, j)}$$

โดยที่ $CC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากความใกล้ชิดของโหนด i; $d(i,j)$ คือ จำนวนเส้นเชื่อมโยงในเส้นทางที่สั้นที่สุดจากสมาชิกหนึ่งไปยังอีกสมาชิกหนึ่ง

2.2.3 วัดค่าความเป็นศูนย์กลางจากดีกรี (degree centrality, DC) เป็นการกำหนดความสำคัญของโหนดโดยพิจารณาจากจำนวนเส้นเชื่อมโยงทั้งหมดที่โหนดนั้นๆ เชื่อมต่อกับโหนดอื่น ๆ [11]

$$DC(i) = \frac{\sum_j m_{ij}}{N}$$

โดยที่ $DC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากระดับของโหนด i; m_{ij} = 1 ถ้ามีการเชื่อมต่อระหว่างโหนด; m_{ij} = 0 ถ้าไม่มีการเชื่อมต่อระหว่างกัน; N คือ จำนวนโหนดทั้งหมด

2.2 งานวิจัยที่เกี่ยวข้อง

สำหรับงานวิจัยในปัจจุบันที่เกี่ยวข้องกับการนำทฤษฎีกราฟมาประยุกต์ใช้ ดังงานวิจัยที่เกี่ยวข้องแสดงดังตารางที่ 1

3. วิธีการดำเนินงาน

การวิเคราะห์ความสำคัญของคำโดยใช้ทฤษฎีกราฟ เพื่อนำมาประยุกต์ใช้กับอัลกอริทึมสำหรับการจำแนกข้อความ มีการดำเนินงาน ดังนี้

ขั้นตอนที่ 1 ข้อมูลที่ใช้ในการทดสอบ คือ ข้อมูลจากฐานข้อมูล UCI Machine Learning Repository โดยข้อมูลที่นำมาทดสอบมีจำนวนทั้งหมด 3 ชุด ข้อมูลชุดที่ 1 คือ ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์มาจากเว็บไซต์ www.imdb.com จำนวน 1,000 ระเบียน ข้อมูลชุดที่ 2 ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารมาจากเว็บไซต์ www.yelp.com จำนวน 3,726 ระเบียน และข้อมูลชุดที่ 3 ข้อมูลแสดงความคิดเห็นเกี่ยวกับ สินค้านำมาจากเว็บไซต์ www.amazon.com จำนวน 15,004 ระเบียน ซึ่งข้อมูลทั้ง 3 ชุดนั้น เป็นลักษณะข้อความแสดงความคิดเห็น และมีการจำแนกข้อความออกเป็น 2 คลาส คือ 1 แทนการแสดงความคิดเห็นในเชิงบวก และ 0 แทนการแสดงความคิดเห็นในเชิงบวกเชิงลบการเตรียมข้อมูลกรองเฉพาะข้อความจนเหลือเพียงแต่ข้อความที่นำเสนอความคิดเห็น [17]

ขั้นตอนที่ 2 การสร้างความสัมพันธ์ของคำสร้างความสัมพันธ์ของคำจากข้อมูลประเภทข้อความผ่าน Graph Generator ชื่อว่า textexture (www.textexture.com) โดยสร้างความสัมพันธ์จากการดูคำที่ปรากฏอยู่ในแต่ละข้อความ นำกราฟความสัมพันธ์ของคำมาวิเคราะห์ค่าสมบัติทั่วไป และวัดค่าความเป็นศูนย์กลาง เพื่อหาความสำคัญของคำ

ขั้นตอนที่ 3 การวัดค่าความเป็นศูนย์กลาง เนื่องจากต้องการวิเคราะห์ความสำคัญของคำโดยใช้ทฤษฎีกราฟ ซึ่งต้องใช้การวัดค่าความเป็นศูนย์กลางของทฤษฎีกราฟในการวิเคราะห์หาคำที่สำคัญ ซึ่งในงานวิจัยนี้เลือกวิธีการวัดค่าความเป็นศูนย์กลาง 3 วิธี ดังนี้ วิธีที่ 1 วัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (betweeness centrality, BC) วิธีที่ 2 วัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (closeness centrality, CC) และวิธีที่ 3 ค่าความเป็นศูนย์กลางจากระดับ (degree centrality, DC)

ในส่วนของอัลกอริทึมที่นำมาใช้ในการจำแนกข้อมูลในงานวิจัยนี้ คือ อัลกอริทึม random forest จากการศึกษางานวิจัยเกี่ยวกับการจำแนกข้อมูล พบว่าอัลกอริทึม random forest นั้นมีประสิทธิภาพในการจำแนกข้อมูลที่ดีที่สุด แสดงดังตารางที่ 2 จึงทำให้เลือกอัลกอริทึมดังกล่าวมาสร้างโมเดลเมื่อสร้างโมเดลจะได้คำที่ใช้ในการจำแนกข้อมูลออกมาเพื่อนำไปใช้ในการเปรียบเทียบกับคำที่ได้จากการหาความสำคัญของคำด้วยกราฟ

สำหรับงานวิจัยนี้จะเป็นการวิเคราะห์คำสำคัญที่ได้จากการวัดค่าความเป็นศูนย์กลางโดยทฤษฎีกราฟแล้วนำมาเปรียบเทียบกับคำที่ได้จากอัลกอริทึม random forest เพื่อนำคำที่ได้ไปประยุกต์ใช้ให้การจำแนกมีประสิทธิภาพ โดยมุ่งเน้นพิจารณาคำที่ตรงกัน

ตารางที่ 1 งานวิจัยที่เกี่ยวข้องกับการนำทฤษฎีกราฟมาประยุกต์ใช้

งานวิจัย	ปี	วิธีการ	ผลงานวิจัย
Graph based text representation for document clustering [6]	2015	- TF-IDF - dependency graph	ผลการเปรียบเทียบการจำแนกกลุ่มข่าว จำนวน 20 กลุ่ม โดยแต่ละกลุ่มมีจำนวนเอกสาร 10 เอกสาร โดยเปรียบเทียบการจำแนกโดยใช้วิธี TF-IDF กับวิธี dependency graph พบว่า dependency graph ให้ผลการจำแนกที่ดีกว่า
Text categorization as a graph classification problem [12]	2015	- sub graph - n-gram	เปรียบเทียบการลดคุณลักษณะของข้อมูลโดยวิธี sub graph กับ n-gram โดยทดสอบกับข้อมูล 4 กลุ่ม ดังนี้ WebKB, R8, LingSpam และ Amazon ซึ่งพบว่า วิธี sub graph สามารถลดคุณลักษณะได้ดีกว่า n-gram
Graph-based term weighting for text categorization [8]	2015	- TF - TF-IDF - degree centrality - in-degree centrality - out-degree centrality - closeness centrality	เปรียบเทียบการจำแนกข้อมูล Reuters-21578 และ WebKB พิจารณาในเรื่องน้ำหนักของคำที่เกิดขึ้น โดยเปรียบเทียบกับการคำนวณหาน้ำหนักด้วยวิธีต่าง ๆ พบว่าวิธีการคำนวณหาน้ำหนักแบบวัดค่าความเป็นศูนย์กลางของคำด้วย degree centrality ให้ผลลัพธ์ดีที่สุด
Graph-based techniques for topic classification of tweets in spanish [13]	2013	- PageRank - HITS	จำแนกข้อมูล TASS 2013 ซึ่งเป็นข้อมูลจากทวิตเตอร์ภาษาสเปน ซึ่งมีจำนวนกลุ่มข้อมูลทั้งสิ้น 10 กลุ่ม โดยนำข้อมูลมาสร้างความสัมพันธ์ด้วยกราฟ ในการสร้างความสัมพันธ์ของคำนั้นใช้วิธีการสร้าง 2 วิธี คือ PageRank และ HITS ผลการจำแนกพบว่า สามารถจำแนกกลุ่มได้ความถูกต้องมากกว่า70 %
Graph-based representations for text classification [14]	2011	- TF - TF-IDF - co-occurrence networks - dependency networks	ผลการเปรียบเทียบการจัดหมวดหมู่ของเอกสารข้อมูล 2 ชุด คือ TASA900 และ Reuters พบว่า การสร้างความสัมพันธ์ของคำด้วยกราฟทั้ง 2 วิธี ให้ค่าในการจัดหมวดหมู่ที่สูงกว่าวิธี TF และ TF-IDF
Enhancing text representation for classification tasks with semantic graph structures [7]	2011	- vector space model - graph structure model	เปรียบเทียบการจำแนกกลุ่มข้อความภาษาจีน จาก 15 กลุ่ม กลุ่มละ 20 ข้อความ ผลการทดสอบเพื่อเปรียบเทียบการจำแนกกลุ่มข้อความ พบว่า วิธี graph structure model ให้ประสิทธิภาพที่ดีกว่าวิธี vector space model
Graph-based KNN text classification [15]	2010	- vector space model - graph-based text	เปรียบเทียบการจำแนกข้อความภาษาจีน จำนวน 5 กลุ่ม พิจารณาเรื่องความถูกต้องและเวลาที่ใช้ในการประมวลผล พบว่า graph-based text ให้ค่าความถูกต้องที่สูงกว่า และใช้เวลาน้อยกว่า

ตารางที่ 2 งานวิจัยที่เกี่ยวข้องกับการจำแนกข้อมูล

งานวิจัย	ปี	วิธีการ	ผลงานวิจัย
A comparative study of random forest & K – nearest neighbors on HAR dataset using caret [18]	2017	- random forest - parallel random forest - K-nearest neighbors	เปรียบเทียบการจำแนกข้อมูล HAR (human activity recognition) จากฐานข้อมูล UCI โดยเปรียบเทียบค่าความถูกต้อง ซึ่งพบว่าอัลกอริทึม random forest ให้ค่าสูงที่สุด โดยมีค่าเท่ากับ 0.9313
Comparative study on data mining classification methods for cervical cancer prediction using pap smear results [19]	2016	- Naïve Bayes - support vector machines - random forest	เปรียบเทียบการจำแนกข้อมูลผู้ป่วยโรคมะเร็งปากมดลูกประเทศอินโดนีเซีย โดยทดสอบประสิทธิภาพโดยวัดค่า accuracy, recall, precision และ ROC curve ซึ่งผลการวิจัยพบว่าอัลกอริทึม random forest ให้ผลลัพธ์ดีที่สุด โดยมีค่าเท่ากับ 80.18, 75.96, 80.18 และ 93.39 % ตามลำดับ
Mangrove classification using support vector machines and random forest algorithm: A comparative study [20]	2016	- support vector machines - random forest	จำแนกประเภทของป่าชายเลน โดยใช้ข้อมูล LiDAR และ orthophotographs ซึ่งเป็นข้อมูลชายฝั่งทะเลในประเทศฟิลิปปินส์ โดยเปรียบเทียบประสิทธิภาพ ค่า precision และ recall พบว่าอัลกอริทึม random forest ให้ค่าที่สูงกว่า โดยมีค่าเท่ากับ 100 และ 96.70 % ตามลำดับ

กับคำที่คล้ายกันของผลลัพธ์ที่ได้จากโมเดล และกราฟความสัมพันธ์ของคำ

4. ผลการดำเนินงาน

ในส่วนหัวข้อนี้จะเป็นการนำเสนอในส่วนของผลการสร้างความสัมพันธ์ของคำโดยกราฟ จากนั้นวัดจากค่าความเป็นศูนย์กลางด้วยวิธีต่าง ๆ และเปรียบเทียบกับคำที่ได้จากอัลกอริทึม random forest กับคำที่ได้จากการหาความสำคัญโดยกราฟ มีรายละเอียดดังนี้

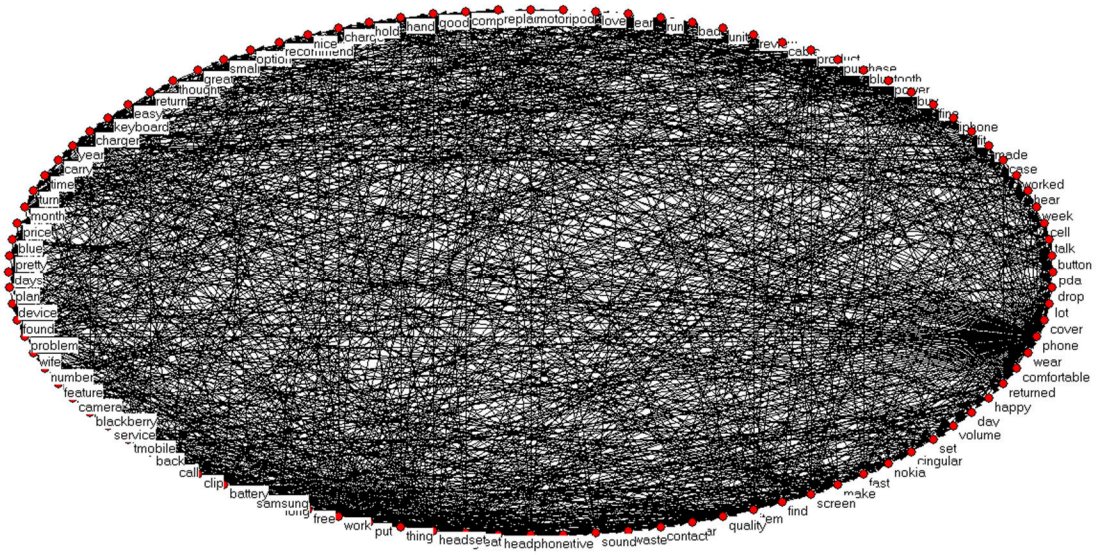
4.1 ผลการสร้างความสัมพันธ์ของคำโดยกราฟ

การสร้างกราฟความสัมพันธ์ของคำของข้อมูลทั้ง 3 ชุด นั้นแทนคำด้วยโหนด และแทนความสัมพันธ์ของคำด้วยลิงค์ โดยกราฟที่ได้อยู่ในรูปแบบของกราฟแบบไม่มีทิศทาง (undirected graph) และความสัมพันธ์ของคำที่เกิดขึ้นจากการปรากฏร่วมของคำต่าง ๆ (co-occurrence) ซึ่งผลการสร้างความสัมพันธ์ของคำโดยกราฟแสดงดังตารางที่ 3

ตารางที่ 3 ผลการสร้างความสัมพันธ์ของคำโดยกราฟของข้อมูลทั้ง 3 ชุดเมื่อกำหนดให้จำนวนคำ (word) สูงที่สุด คือ 100 คำ ผลการวิเคราะห์แสดงให้เห็นว่าความสัมพันธ์ของคำในข้อมูลชุดที่ 3 คำ ในข้อมูลนี้มีการเชื่อมโยงกันจำนวนมากมีความสัมพันธ์กันสูงที่สุด โดยพิจารณาจากจำนวนเส้นเชื่อม (link) ความหนาแน่น (density) และค่าเฉลี่ยดีกรี (average degree) ที่มีค่าสูงกว่าข้อมูลชุดที่ 1 และ 2

ตารางที่ 3 ผลการสร้างความสัมพันธ์ของคำโดยกราฟของข้อมูลทั้ง 3 ชุด

Characteristics	Datasets		
	ข้อมูลชุดที่ 1 (www.imdb.com)	ข้อมูลชุดที่ 2 (www.yelp.com)	ข้อมูลชุดที่ 3 (www.amazon.com)
Network size	100	100	100
Number of links (Edges)	1117	1141	1289
Density	0.1117	0.1141	0.1289
Average degree	22.3400	22.8200	25.7800



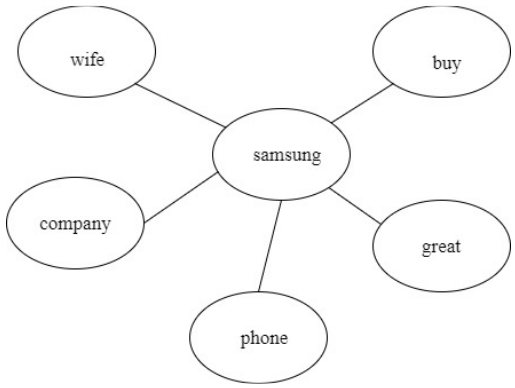
รูปที่ 1 ตัวอย่างกราฟความสัมพันธ์ของข้อมูลชุดที่ 3

ตารางที่ 4 ตัวอย่างคำที่มีความสัมพันธ์กันของข้อมูลชุดที่ 3

คำ	คำที่มีความสัมพันธ์กัน								
	คำที่ 1	คำที่ 2	คำที่ 3	คำที่ 4	คำที่ 5	คำที่ 6	คำที่ 7	คำที่ 8	คำที่ 9
samsung	wife	buy	company	phone	great				
ipod	device	find	fit	thing	fine	great	wear		
pretty	made	cingular	buy	thing	work	case	phone	good	car
price	battery	sound	love	option	motorola	product	fine	headphone	
headset	headset	back	phone	days					

รูปที่ 1 แสดงตัวอย่างกราฟข้อความของข้อมูลชุดที่ 3 กำหนดให้โหนดแสดงถึงคำ และเส้นเชื่อมแสดงถึงความสัมพันธ์ระหว่างคำ เมื่อคำ A ปรากฏแล้ว อยครั้งคำ B จะปรากฏด้วย เส้นเชื่อม (link) ระหว่างคำ A และ B จึงเกิดขึ้น

ตารางที่ 4 เป็นตัวอย่างของคำที่มีความสัมพันธ์กันของข้อมูลชุดที่ 3 จะเห็นได้ว่าแต่ละโหนดหรือคำมีความสัมพันธ์แตกต่างกัน เช่น โหนดหรือคำ "samsung" มีคำที่สัมพันธ์กันอยู่ 5 คำ คือ wife, buy, company, phone และ great ดังกราฟรูปที่ 2



รูปที่ 2 ตัวอย่างกราฟความสัมพันธ์ของ "samsung"

4.2 การวัดค่าความเป็นศูนย์กลาง

ในส่วนนี้ วัดค่าความเป็นศูนย์กลางด้วยวิธีต่าง ๆ และเปรียบเทียบการจำแนกกลุ่มคำกับผลลัพธ์ที่ได้จาก random forest algorithm แสดงให้เห็นว่าข้อมูลชุดทดสอบสามารถจำแนกออกเป็น 3 กลุ่ม คือ (1) กลุ่มคำที่ตรงกัน หมายถึง คำที่ตรงกันตามตัวอักษร คำดังกล่าวจะปรากฏอยู่ในผลลัพธ์การจำแนกจากการวัดค่าความเป็นศูนย์กลางโดยกราฟและ random forest algorithm (2) กลุ่มคำที่คล้ายกัน หมายถึง คำที่มีตัวอักษรไม่ตรงกันทั้งหมด ปรากฏอยู่ในผลลัพธ์การจำแนกจากการวัดค่าความเป็นศูนย์กลางโดยกราฟและ random forest algorithm เช่น “chick” และ “chicken” และ (3) กลุ่มคำที่ไม่ปรากฏ หมายถึง คำที่ปรากฏอยู่ในผลลัพธ์การจำแนกจากการวัดค่าความเป็นศูนย์กลางโดยกราฟ หรือ random forest algorithm อย่างใดอย่างหนึ่งเท่านั้น รายละเอียดแสดงดังตารางที่ 5 ถึง 13

ตารางที่ 5 ตัวอย่างกลุ่มคำที่ตรงกันของข้อมูลชุดที่ 1

No.	Weka	Word	BC	CC	DC
1	acting	actin	215.809	1.7879	47
2	actor	actor	185.394	1.8788	36
3	art	art	13.2589	1.9899	16
4	bad	bad	416.438	1.6869	63
...	...	...	...	...	...
47	year	year	118.623	2.0101	22
Max			1719.99	2.9091	112
Min			1.9368	1.4242	6
Average			112.732	2.0294	24.127

ตารางที่ 5 เป็นผลลัพธ์ของกลุ่มคำที่ตรงกันของข้อมูลชุดที่ 1 พบว่ามีคำที่ตรงกันทั้งหมด 47 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 1719.9943 ค่าต่ำสุดเท่ากับ 1.9368 และค่าเฉลี่ยเท่ากับ 112.7320 closeness centrality ค่าสูงสุดเท่ากับ 2.9091 ค่าต่ำสุดเท่ากับ 1.4242 และค่าเฉลี่ยเท่ากับ 2.0294 degree centrality ค่าสูงสุดเท่ากับ 112 ค่าต่ำสุดเท่ากับ 6 และค่าเฉลี่ยเท่ากับ 24.1277

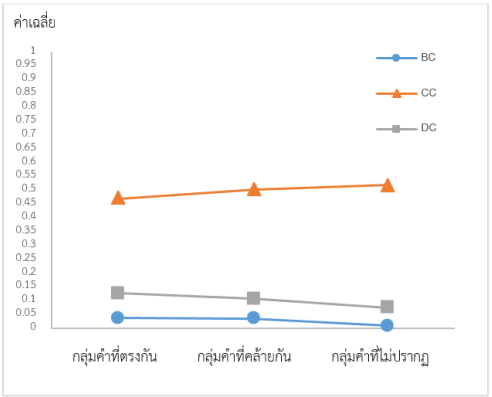
ตารางที่ 6 เป็นผลลัพธ์ของกลุ่มคำที่คล้ายกันของข้อมูลชุดที่ 1 พบว่ามีคำที่คล้ายกันทั้งหมด 47 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 3065.1469 ค่าต่ำสุดเท่ากับ 0.6178 และค่าเฉลี่ยเท่ากับ 106.6432 closeness centrality ค่าสูงสุดเท่ากับ 2.8283 ค่าต่ำสุดเท่ากับ 1.2525 และค่าเฉลี่ย 2.0836 degree centrality ค่าสูงสุดเท่ากับ 148 ค่าต่ำสุดเท่ากับ 6 และค่าเฉลี่ยเท่ากับ 21.2979

ตารางที่ 6 ตัวอย่างกลุ่มคำที่คล้ายกันของข้อมูลชุดที่ 1

No.	Wek	Word	BC	CC	DC
1	abso	absolut	27.959	1.98	17
2	act	action	22.191	2.19	15
3	aw	awful	43.706	2.08	20
4	beau	beautifu	42.324	2.07	15
...	...	...	...	...	...
47	writ	writing	14.233	1.94	18
Max			3065.1	2.82	14
Min			0.6178	1.25	6
Average			106.643	2.083	21.

ตารางที่ 7 ตัวอย่างกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 1

No.	Wek	Word	BC	CC	DC
1	Non	end	21.8970	1.9899	17
2	Non	ending	51.0562	2.1515	13
3	Non	excellent	37.1166	2.1717	16
4	Non	pretty	12.0761	2.2727	13
...	...	...	...	...	...
6	Non	totally	31.5288	2.0707	19
Max			51.0562	2.2727	21
Min			12.0761	1.9899	13
Average			30.0609	2.1094	16.5



รูปที่ 3 ค่าเฉลี่ย BC, CC, DC ของกลุ่มคำที่ตรงกัน กลุ่มคำที่คล้ายกัน และกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 1

ตารางที่ 7 เป็นผลลัพธ์ของกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 1 พบว่ามีคำที่ไม่ปรากฏทั้งหมด 6 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 51.0562 ค่าต่ำสุดเท่ากับ 12.0761 และค่าเฉลี่ยเท่ากับ 30.0609 closeness centrality ค่าสูงสุดเท่ากับ 2.2727 ค่าต่ำสุดเท่ากับ 1.9899 และค่าเฉลี่ยเท่ากับ 2.1094 degree centrality ค่าสูงสุดเท่ากับ 21 ค่าต่ำสุดเท่ากับ 13 และค่าเฉลี่ยเท่ากับ 16.5

รูปที่ 3 จะเห็นว่าค่าเฉลี่ยของ betweenness centrality, degree centrality เป็นไปในทิศทางเดียวกัน นั่นคือ กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมีค่าเฉลี่ยใกล้เคียงกัน ส่วนกลุ่มคำที่ไม่ปรากฏนั้นค่าเฉลี่ยต่างกันและมีค่าต่ำกว่า ซึ่งแตกต่างกับค่าเฉลี่ยของ closeness centrality ที่กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมีค่าใกล้เคียง ส่วนคำที่ไม่ปรากฏมีค่าที่สูงกว่า

ตารางที่ 8 ตัวอย่างกลุ่มคำที่ตรงกันของข้อมูลชุดที่ 2

No.	Weka	Word	BC	CC	DC
1	back	back	235.307	1.848	36
2	bad	bad	43.3654	2.141	19
3	bar	bar	52.0706	2.222	21
4	beer	beer	12.7626	2.484	11
...	...	...	...	...	...
59		worth	87.6642	1.969	27
Max			1253.59	2.535	85
Min			10.4591	1.626	9
Average			98.5703	2.115	21.59

ตารางที่ 8 เป็นผลลัพธ์ของกลุ่มคำที่ตรงกันของข้อมูลชุดที่ 2 พบว่ามีคำที่ตรงกันทั้งหมด 59 คำ

โดย betweeness centrality ค่าสูงสุดเท่ากับ 1253.5934 ค่าต่ำสุด 10.4591 และค่าเฉลี่ยเท่ากับ 98.5703 closeness centrality ค่าสูงสุดเท่ากับ 2.5354 ค่าต่ำสุดเท่ากับ 1.6263 และค่าเฉลี่ยเท่ากับ 2.1154 degree centrality ค่าสูงสุดเท่ากับ 85 ค่าต่ำสุดเท่ากับ 9 และค่าเฉลี่ยเท่ากับ 21.5932

ตารางที่ 9 ตัวอย่างกลุ่มคำที่คล้ายกันของข้อมูลชุดที่ 2

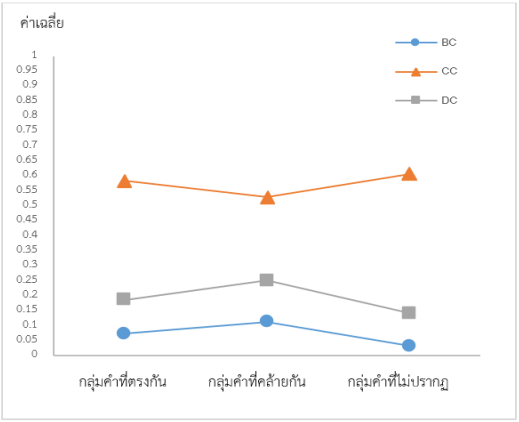
No.	Weka	Word	BC	CC	DC
1	che	cheese	86.991	1.969	25
2	chick	chicken	159.81	2.060	32
3	delici	delicious	37.458	2.151	16
4	din	dine	59.394	2.060	19
...	...	...	...	...	...
34	wait	waiting	28.607	2.292	15
max			986.17	2.505	82
min			5.7306	1.525	7
average			147.41	2.061	26.647

ตารางที่ 10 ตัวอย่างกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 2

No.	Weka	Word	BC	CC	DC
1	Non	busy	60.474	2.010	18
2	Non	friend	39.327	1.979	18
3	Non	friendly	67.087	1.989	25
4	Non	plate	23.587	2.343	13
...	...	...	...	...	...
7	non	sweet	34.567	2.202	16
Max			70.034	2.343	25
Min			23.587	1.979	13
Average			46.545	2.137	18.142

ตารางที่ 9 เป็นผลลัพธ์ของกลุ่มคำที่คล้ายกันของข้อมูลชุดที่ 2 พบว่ามีคำที่คล้ายกันทั้งหมด 34 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 986.1769 ค่าต่ำสุดเท่ากับ 5.7306 และค่าเฉลี่ยเท่ากับ 147.4127 closeness centrality ค่าสูงสุดเท่ากับ 2.5051 ค่าต่ำสุดเท่ากับ1.5253 และค่าเฉลี่ยเท่ากับ 2.0612 degree centrality ค่าสูงสุดเท่ากับ 82 ค่าต่ำสุดเท่ากับ 7 และค่าเฉลี่ยเท่ากับ 26.6471

ตารางที่ 10 เป็นผลลัพธ์ของกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 2 พบว่ามีคำที่ไม่ปรากฏทั้งหมด 7 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 70.0345 ค่าต่ำสุดเท่ากับ 23.5873 และค่าเฉลี่ยเท่ากับ 46.5453 closeness centrality ค่าสูงสุดเท่ากับ 2.3434 ค่าต่ำสุดเท่ากับ 1.9798 และค่าเฉลี่ยเท่ากับ 2.1371 degree centrality ค่าสูงสุดเท่ากับ 25 ค่าต่ำสุดเท่ากับ 13 และค่าเฉลี่ยเท่ากับ 18.1429



รูปที่ 4 ค่าเฉลี่ย BC, CC, DC ของกลุ่มคำที่ตรงกัน กลุ่มคำที่คล้ายกัน และกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 2

รูปที่ 4 จะเห็นว่าค่าเฉลี่ยของ betweeness centrality, degree centrality เป็นไปในทิศทางเดียวกัน นั่นคือ กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมีค่าเฉลี่ยใกล้เคียงกัน ส่วนกลุ่มคำที่ไม่ปรากฏนั้นค่าเฉลี่ยต่างกันและมีค่าต่ำกว่า ซึ่งแตกต่างกับค่าเฉลี่ยของ closeness centrality ที่กลุ่มคำที่ตรงกันกับ

กลุ่มคำที่คล้ายกันมีค่าใกล้เคียง ส่วนคำที่ไม่ปรากฏมีค่าที่สูงกว่า

ตารางที่ 11 เป็นผลลัพธ์ของคำที่ตรงกันของของข้อมูลชุดที่ 3 พบว่ามีคำที่ตรงกันทั้งหมด 46 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 614.2871 ค่าต่ำสุดเท่ากับ 5.5422 และค่าเฉลี่ยเท่ากับ 71.4206 closeness centrality ค่าสูงสุดเท่ากับ 2.7475 ค่าต่ำสุดเท่ากับ 1.5859 และค่าเฉลี่ยเท่ากับ 1.9660 degree centrality ค่าสูงสุดเท่ากับ 84 ค่าต่ำสุดเท่ากับ 9 และค่าเฉลี่ยเท่ากับ 25

ตารางที่ 11 ตัวอย่างคำที่ตรงกันของข้อมูลชุดที่ 3

No.	Weka	Word	BC	CC	DC
1	back	back	70.763	1.949	27
2	bad	bad	44.527	1.899	24
3	bluetoo	bluetoot	152.03	1.818	38
4	button	button	32.227	2.000	22
...	...	...	...	...	...
46	year	year	66.489	1.909	30
Max			614.28	2.747	84
Min			5.5422	1.585	9
Average			71.420	1.966	25

ตารางที่ 12 ตัวอย่างคำที่คล้ายกันของข้อมูลชุดที่ 3

No.	Weka	Word	BC	CC	DC
1	batter	battery	226.833	1.777	48
2	blackb	blackber	23.1413	1.949	20
3	blu	blue	11.7917	2.343	13
4	bougt	bought	146.963	1.808	42
...	...	...	...	...	...
46	work	worked	19.3835	2.010	17
Max			3149.61	2.343	159
Min			8.4985	1.191	11
Average			129.024	1.941	27.43

ตารางที่ 12 เป็นผลลัพธ์ของคำที่คล้ายกันของข้อมูลชุดที่ 3 พบว่ามีคำที่คล้ายกันทั้งหมด 46 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 3149.6141 ค่าต่ำสุดเท่ากับ 8.4985 และค่าเฉลี่ยเท่ากับ 129.0245 closeness centrality ค่าสูงสุดเท่ากับ 2.3434 ค่าต่ำสุดเท่ากับ 1.1919 และค่าเฉลี่ยเท่ากับ 1.9412 degree centrality ค่าสูงสุดเท่ากับ 159 ค่าต่ำสุดเท่ากับ 11 และค่าเฉลี่ยเท่ากับ 27.4348

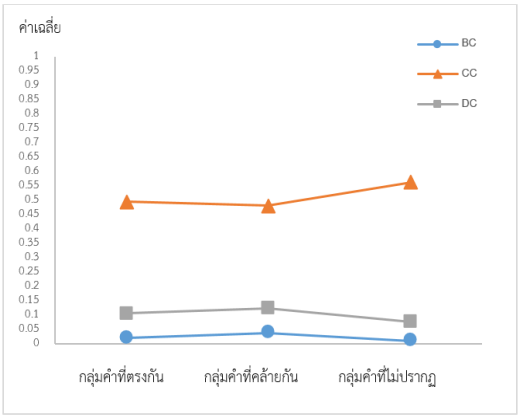
ตารางที่ 13 ตัวอย่างคำที่ไม่มีของข้อมูลชุดที่ 3

No.	Weka	Word	BC	CC	DC
1	None	carry	25.8555	2.343	19
2	None	cingular	17.6171	2.050	16
3	None	iphone	41.0096	1.989	14
4	None	power	9.4071	2.212	13
...	...	...	...	...	...
8	None	great	63.6843	1.979	25
Max			105.054	2.343	38
Min			9.4071	1.868	13
Average			38.9403	2.069	20.75

ตารางที่ 13 เป็นผลลัพธ์ของคำที่ไม่มีของข้อมูลชุดที่ 3 พบว่ามีคำที่ไม่ปรากฏทั้งหมด 8 คำ โดย betweeness centrality ค่าสูงสุดเท่ากับ 105.0547 ค่าต่ำสุดเท่ากับ 9.407124 และค่าเฉลี่ยเท่ากับ 38.94033 closeness centrality ค่าสูงสุดเท่ากับ 2.3434 ค่าต่ำสุดเท่ากับ 1.8686 และค่าเฉลี่ยเท่ากับ 2.0694 degree centrality ค่าสูงสุดเท่ากับ 38 ค่าต่ำสุดเท่ากับ 13 และค่าเฉลี่ยเท่ากับ 20.75

รูปที่ 5 จะเห็นว่าค่าเฉลี่ยของ betweeness centrality, degree centrality เป็นไปในทิศทางเดียวกัน นั่นคือ กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมีค่าเฉลี่ยใกล้เคียงกัน ส่วนกลุ่มคำที่ไม่ปรากฏนั้นค่าเฉลี่ยต่างกันและมีค่าต่ำกว่า ซึ่งแตกต่างกับค่าเฉลี่ย

ของ closeness centrality ที่กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมีค่าใกล้เคียง ส่วนคำที่ไม่ปรากฏมีค่าที่สูงกว่า



รูปที่ 5 ค่าเฉลี่ย BC, CC, DC ของกลุ่มคำที่ตรงกัน กลุ่มคำที่คล้ายกัน และกลุ่มคำที่ไม่ปรากฏของข้อมูลชุดที่ 3

การเปรียบเทียบคำที่ได้จากโมเดลกับคำสำคัญที่ได้จากการวัดค่าความเป็นศูนย์กลางแต่ละวิธีกับข้อมูลทั้ง 3 ชุด โดยเริ่มต้นเปรียบเทียบกับข้อมูลทั้งหมดในแต่ละกลุ่มข้อมูล พบว่าทั้ง 3 ชุดข้อมูล มีลักษณะเดียวกัน คือ พบจำนวนกลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันมากกว่ากลุ่มคำที่ไม่ปรากฏ เมื่อวัดค่าความเป็นศูนย์กลางของข้อมูลทั้ง 3 ชุด พบว่าได้ผลลัพธ์เช่นเดียวกัน นั่นคือในการวัดด้วย betweeness centrality และ degree centrality กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันได้ค่าสูง และค่าของทั้งสองกลุ่มคำมีความใกล้เคียงกัน ส่วนการวัดกลุ่มคำที่ไม่ปรากฏได้ค่าที่ต่ำและค่าต่างจากทั้งสองกลุ่มคำ ซึ่งตรงกันข้ามกับการวัดด้วย closeness centrality ที่ผลลัพธ์ของกลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันถึงแม้ค่าจะใกล้เคียงกันแต่ว่าค่าที่ได้ต่ำ เมื่อเปรียบเทียบกับค่าของกลุ่มคำที่ไม่ปรากฏซึ่งค่าต่างจากทั้งสองกลุ่มคำและค่าที่ได้สูง งานวิจัยนี้ต้องการคำที่ตรงกันกับคำที่คล้ายกันสูง ในการพิจารณาประกอบกับคำในโมเดลเพื่อจำแนกข้อมูลให้มีประสิทธิภาพสูงขึ้น ดังนั้นจากผลลัพธ์ที่เปรียบเทียบพบว่า BC และ DC ให้ผลลัพธ์ของค่าความเป็นศูนย์กลางที่เหมาะสมกว่าวิธี CC ทำให้ในการพิจารณาเลือกวิธีการวัดค่าความเป็นศูนย์กลางในการหาคำสำคัญเพื่อนำไปประยุกต์ใช้กับอัลกอริทึมสำหรับการจำแนก random forest นั้นควรเลือกวิธี BC และ DCเพื่อให้การจำแนกข้อมูลมีประสิทธิภาพสูงขึ้น

5. สรุป

งานวิจัยนี้เป็นการประยุกต์ใช้อัลกอริทึม random forest และทฤษฎีกราฟสำหรับการวิเคราะห์ข้อความ โดยข้อมูลที่นำมาใช้ในการทดสอบเป็นข้อมูลประเภทข้อความจำนวน 3 ชุด ดังนี้ ข้อมูลชุดที่ 1 คือ ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์มาจากเว็บไซต์ www.imdb.com ข้อมูลชุดที่ 2 ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารมาจากเว็บไซต์ www.yelp.com และข้อมูลชุดที่ 3 ข้อมูลแสดงความคิดเห็นเกี่ยวกับสินค้านำมาจากเว็บไซต์ www.amazon.com อัลกอริทึมสำหรับการจำแนก คือ อัลกอริทึม random forest นำข้อความมาสร้างความสัมพันธ์ของคำผ่าน graph generator ชื่อว่า textexture (textexture.com) จากนั้นนำกราฟความสัมพันธ์ของคำมาวิเคราะห์ค่าสมบัติทั่วไป และวัดค่าความเป็นศูนย์กลาง เพื่อหาความสำคัญของคำ 3 วิธี ดังนี้ วิธีที่ 1 วัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (betweeness centrality, BC) วิธีที่ 2 วัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (closeness centrality, CC) และวิธีที่ 3 ค่าความเป็นศูนย์กลางจากระดับ (degree centrality, DC) จากนั้นนำคำที่ได้จากโมเดลที่สร้างจากอัลกอริทึม random forest มาเปรียบเทียบกับคำสำคัญที่ได้จากกราฟ ผลปรากฏว่า สามารถ

จำแนกคำได้ออกเป็น 3 กลุ่ม คือ กลุ่มคำที่ตรงกัน กลุ่มคำที่คล้ายกัน และกลุ่มคำที่ไม่ปรากฏ และเมื่อวัดค่าความเป็นศูนย์กลางของคำทั้ง 3 ชุด ด้วย 3 วิธีนั้น พบว่าผลลัพธ์ของข้อมูลทั้ง 3 ชุด เป็นไปในทิศทางเดียวกัน นั่นคือ ในการวัดด้วย betweeness centrality และ degree centrality กลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกันได้ค่าความเป็นศูนย์กลางของคำสูง และค่าของคำของทั้งสองกลุ่มคำมีความใกล้เคียงกัน ส่วนการวัดกลุ่มคำที่ไม่ปรากฏได้ค่าความเป็นศูนย์กลางของคำที่ต่ำและต่างจากทั้งสองกลุ่มคำ ซึ่งตรงกันข้ามกับการวัดด้วย closeness centrality ที่ผลลัพธ์ของกลุ่มคำที่ตรงกันกับกลุ่มคำที่คล้ายกัน ถึงแม้ค่าความเป็นศูนย์กลางของคำจะใกล้เคียงกันแต่ว่าค่าที่ได้ต่ำ เมื่อเปรียบเทียบกับค่าความเป็นศูนย์กลางของกลุ่มคำที่ไม่ปรากฏ ซึ่งค่าต่างจากทั้งสองกลุ่มคำและค่าที่ได้สูง งานวิจัยนี้ต้องการคำที่ตรงกันกับคำที่คล้ายกันสูงเพื่อไปพิจารณาประกอบกับคำที่ได้จากโมเดลเพื่อจำแนกข้อมูลให้มีประสิทธิภาพสูงขึ้น ดังนั้นจากผลลัพธ์ที่เปรียบเทียบพบว่า BC และ DC ให้ผลลัพธ์ของค่าความเป็นศูนย์กลางที่เหมาะสมกว่าวิธี CC ทำให้ในการพิจารณาเลือกวิธีการวัดค่าความเป็นศูนย์กลางในการหาคำสำคัญเพื่อนำไปประยุกต์ใช้กับอัลกอริทึม สำหรับการจำแนก random forest นั้น ควรเลือกวิธี BC และ DC เพื่อให้การจำแนกข้อมูลมีประสิทธิภาพสูงขึ้น

6. รายการอ้างอิง

[1] Kharea, S.K., Thapab, N. and Sahooc, K.C., 2007, Sahoo internet as a source of information: A survey of Ph.D. scholars, Ann. Library Inf. Stud. 54: 201-206.

[2] Kanimozhi, K.V. and Venkatesan, M., 2015, Unstructured data analysis – A survey, Int. J. Adv. Res. Comp. Commun. Eng. 43: 223-225.

[3] พนิดา ทรงรัมย์, 2559, การจำแนกความคิดเห็นทางการเมืองบนเครือข่ายสังคมออนไลน์ โดยใช้วิธีการจำแนกแบบสัมพันธ์, ว.วิทยาศาสตร์และเทคโนโลยีธัญบุรี 6(1): 83-93.

[4] กานดา แผ่วัฒนากุล, 2555, การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์, วิทยานิพนธ์ปริญญาโท, สถาบันพัฒนบริหารศาสตร์, กรุงเทพฯ, 116 น.

[5] ราชวิทย์ ทิพย์เสนา, ฉัตรเกล้า เจริญผล และแกมกาญจน์ สมประเสริฐศรี, 2557, การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา, ว. วิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม 33(5): 493-502.

[6] Abdulsahib, A.K. and Kamaruddin, S.S., 2015, Graph based text representation for document clustering, J. Theor. Appl. Inf. Technol. 76: 1-13.

[7] Wu, J., Xuan, Z. and Pan, D., 2011, Enhancing text representation for classification tasks with semantic graph structures, Int. J. Innovative Comp. Inf. Cont. 7: 2689-2698.

[8] Malliaros, F.D. and Skianis, K., 2015, Graph-based term weighting for text categorization, pp. 1473-1479, In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM' 15), August 25-28, Paris.

[9] Bondy, J. and Murty, U., 1976, Graph Theory with Applications, The Macmillan Press, Ltd., London.

[10] Hanneman, R.A., 2005, Mark Idle: Introduction to Social Network Methods, University of California, Riverside, California.

[11] Durland, M.M. and Fredericks, K.A. (Eds.), 2006, Social Network Analysis in Program Evaluation, Jossey-Bass, San Francisco.

[12] Kiagias, F.R.E. and Vazirgiannis, M., 2015, Text categorization as a graph classification Problem, pp.1702-1712, In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing.

[13] Cordobés, H., Anta, A.F., Chiroque, L.F., Pérez, F., Redondo, T. and Santos, A., 2013, Graph-based techniques for topic classification of tweets in spanish, Int. J. Artif. Intell. Interact. Multimed. 2(5): 31-37.

[14] Valle, K. and Öztürk, P., 2011, Graph-Based Representations for Text Classification, India-Norway Workshop on Web Concepts and Technologies, Trondheim.

[15] Wang, Z. and Liu, Z., 2016, Graph-based KNN Text Classification, pp. 2363-2366, 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), Yantai.

[16] Liu, J., Wang, J. and Wang, C., 2008, A Text Network Representation Model, pp. 150-154, In Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Washington.

[17] Kotzias, D., Denil, M., De Freitas, N. and Smyth, P., 2015, From Group to Individual Labels using Deep Features, KDD 2015.

[18] Jyothi, K.B., Bindu, K.H. and Suryanarayana, D., 2017, A comparative study of random forest & K – nearest neighbors on HAR dataset using caret, Int. J. Innov. Res. Technol. 3(9): 6-9.

[19] Kurniawati, Y.E., Permanasari, A.E. and Fauziati, S. 2016, Comparative Study on Data Mining Classification Methods for Cervical Cancer Prediction Using Pap Smear Results, 2016 1st International Conference on Biomedical Engineering (IBIOMED), Yogyakarta.

[20] Campomanes, F., Pada, A.V. and Silapan, J., 2016, Mangrove Classification Using Support Vector Machines and Random Forest Algorithm: A Comparative Study, GEOBIA 2016: Solutions and Synergies, Faculty of Geo-Information and Earth Observation (ITC), University of Twente, Overijssel.