Science and Technology
RMUTT Journal
Vol.9 No.1 (2019): 71-79
www.sci.rmutt.ac.th/stj
Online ISSN2229-1547

การประยุกต์ใช้เทคนิคเหมืองข้อความสำหรับการจัดกลุ่มข้อมูล กรณีศึกษามะเร็ง

Application of Text Mining for Data Clustering: A Case Study for Cancer

สุภาพร วีระพันธ์ยานนท์[1*] และ พยุง มีสัจ[1]

[1]ภาควิชาเทคโนโลยีสารสนเทศ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
1518 ถนนประชาราษฎร์ 1 แขวงวงศ์สว่าง เขตบางซื่อ กรุงเทพมหานคร 10800
*E-mail : supapaorn@gmail.com

บทคัดย่อ

งานวิจัยนี้นำเสนอการประยุกต์ใช้เทคนิคเหมืองข้อความ (Text mining) สำหรับการจัดกลุ่มข้อมูล กรณีศึกษามะเร็ง โดยใช้ชุดข้อมูลทดสอบจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวข้องกับมะเร็ง เช่น มะเร็ง การรักษามะเร็ง อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง อาหารเสริมต้านมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น ผู้วิจัยเสนอวิธีการทำเหมืองข้อความโดยการเปรียบเทียบสมการการทำดัชนีเอกสารได้แก่ TFIDF, WTFIDF และ FTFIDF ร่วมกับการทดลองการจัดกลุ่มข้อความแบบ Hierarchical ด้วยวิธี Single link วิธี Average link และวิธี Complete link ผลการทดลองแสดงให้เห็นว่าการใช้สมการ WTFIDF ร่วมกับอัลกอริธึมการจัดกลุ่มแบบ Complete link มีค่าความถูกต้อง (SSE) ในการจัดกลุ่มข้อความดีกว่าเมื่อเปรียบเทียบกับอัลกอริทึมอื่น ๆ

คำสำคัญ: เหมืองข้อความ จัดกลุ่มข้อความ ดัชนีเอกสาร

Abstract

In this research, application of text mining for data clustering in case study for cancer. We used testing data set by searching a definition keyword on website that related to cancer such as cancer, cancer treatment, cancer symptoms, diet for cancer patients, anti-cancer supplements and cancer treatment herb**.** We propose a simple method of text mining by comparing document indexing using TFIDF, WTFIDF and FTFIDF formulas.

Received: June 20, 2018
Revised: January 10, 2019
Accepted: March 05, 2019

The experiment has been done using hierarchical clustering algorithm such as single link, average link and complete link. The results of testing showed that WTFIDF with Complete link algorithm gives the better accuracy for text classification when compared to other algorithms.

Keywords: Text mining, Text classification, Document indexing

1. บทนำ

แนวโน้มการใช้งานอินเทอร์เน็ตผ่านบริการเวิลด์วายเว็บ (World Wide Web) จากอดีตจนถึงปัจจุบันชี้ให้เห็นถึงปริมาณข้อมูลเอกสารอิเล็กทรอนิกส์และเว็บโดเมนที่เพิ่มขึ้นอย่างต่อเนื่องก่อให้เกิดข้อมูลจำนวนขนาดใหญ่ (Big data) อยู่ตามแหล่งต่าง ๆ บนอินเทอร์เน็ต รูปแบบของข้อมูลส่วนใหญ่เป็นข้อความ (Text) มีทั้งรูปแบบกึ่งโครงสร้าง (Semi-structured data) และแบบไม่มีโครงสร้าง (Unstructured data) [1] ซึ่งการค้นหาข้อมูลหรือดาวน์โหลดข้อมูลปริมาณมากเหล่านี้จึงทำได้ยาก แต่หากมีการรวบรวมและจัดหมวดหมู่ข้อมูลจากเว็บไซต์ต่าง ๆ [2] มารวมไว้แหล่งเดียวกันจะทำให้สามารถค้นหาได้สะดวกรวดเร็วและตรงตามความต้องการมากขึ้น

ซึ่งวิธีการรวบรวมและจัดหมวดหมู่ข้อความจากเว็บไซต์ต่าง ๆ สามารถทำได้โดยใช้วิธีการทำเหมือนข้อความ (Text mining) ซึ่งเป็นเครื่องมือการทำเหมืองข้อมูล (Data mining) แบบหนึ่ง [3] นำมาใช้ในการจัดหมวดหมู่ข้อความที่สืบค้นได้จากเครื่องมือเสิรชเอนจิน (Search Engine) [4-5] ดังตัวอย่างงานวิจัย ได้แก่ การจัดกลุ่มเอกสารข้อความภาษาไทยจากข่าวหนังสือพิมพ์ การจัดกลุ่มจากเอกสารเว็บและจัดกลุ่มผลลัพธ์การสืบค้นเว็บภาษาไทย [1] รวมทั้งการจัดหมวดหมู่ของ Search Engine แต่งานวิจัยเหล่านี้ยังขาดการจัดกลุ่มผลลัพธ์จากการสืบค้นข้อมูลภาษาไทยและในกรณีศึกษาโรคมะเร็ง

งานวิจัยฉบับนี้มุ่งเน้นศึกษาการประยุกต์ใช้เทคนิคเหมืองข้อความ (Text mining) สำหรับการจัดกลุ่มข้อมูล กรณีศึกษามะเร็ง โดยใช้ชุดข้อมูลทดสอบจากเว็บไซต์ที่สืบค้นด้วยคำสำคัญที่เกี่ยวข้องกับมะเร็ง เช่น มะเร็ง รักษามะเร็ง อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง อาหารเสริมสำหรับมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น ทั้งนี้กระบวนการทำเหมืองข้อความได้ใช้ข้อมูลที่ผ่านการหาค่าน้ำหนักคำ (Document indexing) จากสมการ 3 สมการ ได้แก่สมการ TFIDF [6] สมการ WTFIDF [7] และสมการ FTFIDF [8] ร่วมกับการจัดกลุ่มข้อความแบบ Hierarchical ด้วยวิธี Single link วิธี Average link และวิธี Complete link เพื่อเปรียบเทียบสมการใดให้ประสิทธิภาพการจัดกลุ่มข้อความวิธีใดดีที่สุด

2. ทฤษฎีที่เกี่ยวข้อง

2.1 โรคมะเร็ง

โรคมะเร็งเป็นปัญหาสาธารณสุขที่สำคัญของประชากรทั่วโลก สมาพันธ์ควบคุมโรคมะเร็งสากล (Union for International Cancer Control: UICC) และองค์การอนามัยโลก (World Health Organization: WHO) จึงกำหนดให้วันที่ 4 กุมภาพันธ์ของทุกปีเป็น "วันมะเร็งโลก" หรือ

"World Cancer Day" เพื่อเผยแพร่องค์ความรู้ให้คนทั่วโลกตระหนักถึงความสำคัญและสาเหตุของโรคมะเร็ง ซึ่งเป็นสาเหตุการเสียชีวิตอันดับต้น ๆ ของประชากรทั่วโลกส่วนสถานการณ์โรคมะเร็งในประเทศไทยจากสถิติพบว่าโรคมะเร็งเป็นสาเหตุการเสียชีวิตอันดับ 1 รองลงมาคืออุบัติเหตุและโรคหัวใจ ซึ่งข้อมูลล่าสุดจากกระทรวงสาธารณสุขยังพบว่าคนไทยเสียชีวิตด้วยโรคมะเร็งประมาณ 60,000 คนต่อปีหรือเฉลี่ยชั่วโมงละเกือบ 7 ราย และจากลักษณะการเกิดของโรคมะเร็งภายในร่างกายสามารถแบ่งออกได้เป็นกลุ่ม ๆ ตามอวัยวะส่วน ต่าง ๆ ของร่างร่างกาย เช่น ศีรษะและลำคอ และสามารถแบ่งเป็นกลุ่มย่อย ๆ คือ มะเร็งที่เกิดอยู่ในอวัยวะนั้น ๆ ปัจจุบันโรคมะเร็งเป็นโรคที่สามารถป้องกัน ถ้าหากผู้ป่วยหรือบุคคลทั่วไปได้รับข้อมูลข่าวสารจากการรวบรวมและการจัดหมวดหมู่ข้อมูลมะเร็งบนเว็บไซต์ เป็นไปอย่างถูกต้องและรวดเร็ว

2.2 การสร้างดัชนีเอกสาร

การสร้างดัชนีเอกสาร (Document Indexing) เป็นขั้นตอนการหาค่าน้ำหนักของคำที่จะใช้เป็นตัวแทนของแต่ละเอกสาร เพื่อเตรียมไปใช้ในการจำแนกข้อความและการสืบค้น [9] การทำดัชนีคำ (Indexing) เป็นศาสตร์หนึ่งของการนำแบบจำลองเวกเตอร์มาประยุกต์ใช้เป็นตัวแทนของเอกสารได้ [15-16] ดังตัวอย่างงานวิจัยของ Nuipian [6] เปรียบเทียบคำสำคัญและวลีสำคัญเพื่อจำแนกเอกสารโดยใช้การเลือกคุณลักษณะจากบทคัดย่องานวิจัยในฐานดิจิทัล ACM โดยใช้เทคนิค Chi-Square, Information Gain, Gain Ratio และความถี่เอกสาร (Document frequency) สกัดคุณลักษณะด้วยเวกเตอร์สเปสโมเดล งานวิจัยของ Ren [7] เสนอวิธีการจำแนกข้อความสำหรับดัชนีการให้น้ำหนักคำด้วยวิธี Class indexing based TF.IDF.ICSδF กับวิธีให้น้ำหนักคำที่แตกต่างกันด้วยการใช้อัลกอริทึมเซนทรอย นาอีฟเบย์ และซัพพอร์ทเวกเตอร์แมทชีนเพื่อแก้ปัญหาการจำแนกเอกสารอัตโนมัติ และ Kwok [8] นำเสนออัลกอริทึม Support Vector Machine ร่วมกับ LSI (Latent Semantic Indexing) ซึ่งจะทำการวิเคราะห์โครงสร้างความสัมพันธ์ของคำที่อยู่ในที่เก็บเอกสาร(Collection) โดยค้นหารูปแบบและแปลงให้อยู่ในเซตของเวกเตอร์เอกสาร เพื่อลดมิติเอกสาร เพื่อให้สามารถปรับการเพิ่มเอกสารลงในคอลเล็กชันแบบไดนามิก

2.3 การวิเคราะห์การจัดกลุ่ม

การวิเคราะห์การจัดกลุ่มคือเทคนิคในการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning technique) ซึ่งมีเป้าหมายเพื่อการจำแนกกลุ่มข้อมูลที่มีคุณลักษณะคล้ายกันอยู่ในนกลุ่มเดียวกัน [10] โดยข้อมูลแต่ละกลุ่มจะถูกเรียกว่า คลัสเตอร์ (Cluster) การวิเคราะห์หรือจำแนกกลุ่มข้อมูลนั้นสามารถแบ่งออกได้เป็น 2 ประเภทได้แก่ วิธีการแบบลำดับชั้น (Hierarchical algorithms) และ วิธีการแบบไม่เป็นลำดับชั้น (Non-hierarchical algorithms) [11]

2.4 การวัดประสิทธิภาพการจัดกลุ่มข้อมูล

วิธีการวัดประสิทธิภาพการจัดกลุ่มข้อมูลการวิจัยครั้งนี้ใช้วิธีการประเมินผลการจัดกลุ่มข้อมูลมะเร็งโดยพิจารณาจากค่า Sum of Square Error (SSE) โดยหากค่าความผิดพลาดมีค่าสูงแสดงว่าข้อมูลภายในกลุ่มมีการกระจายตัวมาก ตรงกันข้ามหากพบว่าความผิดพลาดมีค่าต่ำแสดงว่าข้อมูลที่ถูกจัดกลุ่มนั้นเกาะกลุ่มกันมาก สำหรับฟังก์ชันการคำนวณค่า SSE แสดงดังสมการที่ 1

$$SSE = \sum_{i=1}^{k} \sum_{p \in c_i} d(p, m_i)^2 \qquad (1)$$

โดยที่ $d(p, m_i)$ คือฟังก์ชันคำนวณระยะห่างระหว่างข้อมูล p คือจุดข้อมูลใด ๆ ในกลุ่ม c_i, m_i คือจุดข้อมูลศูนย์กลางประจำกลุ่ม (Centroid of cluster)

3. วิธีดำเนินงานวิจัย

การประยุกต์ใช้เทคนิคเหมืองข้อความสำหรับการจัดกลุ่มข้อมูล กรณีศึกษามะเร็งมีขั้นตอนวิธีการดำเนินงานวิจัยดังนี้

3.1 การเก็บรวบรวมข้อมูล

วิธีการเก็บรวบรวมข้อมูลใช้โปรแกรมคลอเลอร์ (Web Program Crawler) สืบค้นด้วยคำสำคัญที่เกี่ยวข้องกับมะเร็ง เช่น มะเร็ง รักษามะเร็ง อาการของมะเร็ง อาหารสำหรับผู้ป่วยมะเร็ง อาหารเสริมสำหรับมะเร็ง สมุนไพรรักษามะเร็ง เป็นต้น โดยมีจำนวน 640 แถว ซึ่งสามารถแบ่งขอบเขตข้อมูลออกเป็น 7 ส่วน แสดงดังตารางที่ 1 ซึ่งหลักเกณฑ์ที่ใช้กำหนดประเภทเนื้อหาเว็บไซต์ที่น่าเชื่อถือพิจารณาจากเว็บไซต์ประเภทโรงพยาบาลรัฐบาลและเอกชน หน่วยงานด้านสาธารณสุขของภาครัฐ หรือบล็อคของแพทย์ จากนั้นเก็บข้อมูลมะเร็งในฐานข้อมูล SQL

ตารางที่ 1 เขตและรูปแบบข้อมูล

No.	Name	Detail	Type
1	ID	ลำดับ	Text
2	Title	ชื่อหัวข้อเว็บไซต์	Text
3	Content	เนื้อหาในเว็บไซต์	Text
4	CancerDicPlus	ข้อความที่สกัดผ่าน LexTo ร่วมกับ CancerDic+	Text
5	URL	ลิงค์ที่มาของเนื้อหาในเว็บไซต์	Text
6	Label	เนื้อหามะเร็ง เช่น ความรู้ทั่วไป อาการ การรักษา	Text
7	Type Cancer	ชนิดมะเร็ง เช่น โพรงจมูก ท่อน้ำดี ปากมดลูก	Text

3.2 การสกัดข้อความ

ขั้นตอนวิธีการสกัดข้อความเป็นกระบวนการทำการตัดคำ (Word Segmentation) และการกำจัดคำหยุด (Stop Word) เพื่อใช้ในการวิเคราะห์คำออกจากเอกสาร ข่าวสาร ข้อความ และสารสนเทศต่าง ๆ ที่เป็นตัวอักษรเพื่อให้สามารถนำไปทำการจัดกลุ่มข้อมูล (Clustering) สำหรับงานวิจัยนี้ผู้วิจัยได้ใช้โปรแกรม LexTo ซึ่งเป็นลิขสิทธิ์ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติร่วมกับการสกัดข้อความด้วยพจนานุกรมที่ผู้วิจัยสร้างขึ้นชื่อว่า CancerDic+ (URL:http://localhost/cancersearch/cancerdic+)

3.3 การสร้างดัชนีเอกสาร

การสร้างดัชนีเอกสารเป็นขั้นตอนการแปลงเอกสาร ซึ่งเป็นภาษาธรรมชาติให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถประมวลผลได้ ซึ่งจะเป็นการสร้างตัวแทนเนื้อหาเอกสารให้อยู่ในรูปแบบเวกเตอร์ของน้ำหนักคำ (Term Weighting) โดยนิยมใช้รูปแบบของคำเดี่ยว [12-13] ในส่วนของการคำนวณค่าน้ำหนักให้แก่ดัชนีเอกสารสำหรับในงานวิจัยนี้เลือกใช้วิธีการ TFIDF Weighting (Term Frequency-Inverse Document Frequency) โดยเปรียบเทียบกับสมการ 3 สมการเพื่อวัดประสิทธิภาพการให้ค่าน้ำหนักคำ ดังแสดงในสมการที่ 2 สมการที่ 3 และสมการที่ 4

สมการที่ 1 TFIDF [7]

$$TF \times IDF = freq\left(t_j, d\right) \times log\left[\frac{N}{n_j}\right] \quad (2)$$

โดยที่ $freq\left(t_j, d\right)$ คือเอกสารที่มีความถี่คำ j ในเอกสาร d, N คือจำนวนเอกสารทั้งหมดในฐานข้อมูล, n_j คือจำนวนเอกสารที่มีคำ j , d คือเอกสาร

สมการที่ 2 WTFIDF [8]

$$W_{TF.IDF}(ti, dj) = tf(ti, dj)x\left(1 + log\frac{D}{d(ti)}\right) \quad (3)$$

โดยที่ $tf(ti, dj)$ คือค่าความถี่ของคำ i ในเอกสาร j, D คือเอกสารทั้งหมด, $d(ti)$ คือเอกสารที่พบคำ ti, t_i คือคำ i, d_j คือเอกสาร j

สมการที่ 3 FTFIDF [9]

$$\frac{fi(w_j)}{\sqrt{\sum w_j \epsilon D_i fi^2(w_j)}} x\, log(\frac{N}{N(w_j)}) \quad (4)$$

โดยที่ $fi(w_j)$ คือค่าความถี่ของคำ w_j ในเอกสาร D_i, $\sum w_j \epsilon D_i fi^2(w_j)$ คือผลรวมค่าความถี่ของคำ w_j ในเอกสาร D_i ยกกำลัง 2 ของคำ w_j ที่พบในเอกสาร D_iทั้งหมด, N คือเอกสารทั้งหมด, $N(w_j)$ คือเอกสารที่พบคำ w_j

3.4 การทำเหมืองข้อความ (Text Mining)

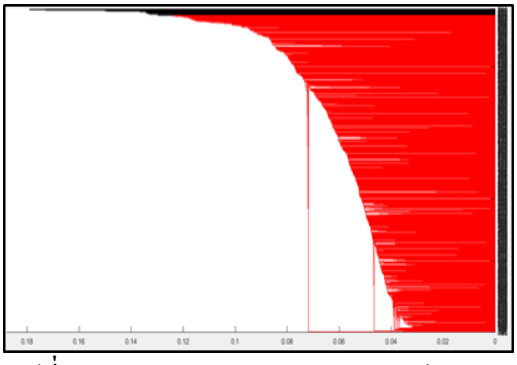
สำหรับขั้นตอนนี้เป็นการจัดกลุ่มข้อความโดยการนำข้อมูลที่ผ่านการหาค่าน้ำหนักคำจากทั้ง 3 สมการ ได้แก่ TFIDF สมการ WTFIDF และสมการ FTFIDF ซึ่งวิธีการจัดกลุ่มผู้วิจัยได้เลือกนำเสนอวิธีการจัดกลุ่มแบบ Hierarchical ด้วยวิธี Single Link วิธี Average Link และวิธี Complete Link เพื่อทำการเปรียบเทียบว่าการหาค่าน้ำหนักคำจากสมการใดให้ประสิทธิภาพการจัดกลุ่มแบบใดดีที่สุด ซึ่งจะทำการทดสอบ การจัดกลุ่มข้อมูลได้ 9 รูปแบบ ดังแสดงในตารางที่ 2

ตารางที่ 2 แสดงวิธีการเปรียบเทียบสมการที่หาค่าน้ำหนักคำกับประสิทธิภาพวิธีการจัดกลุ่ม

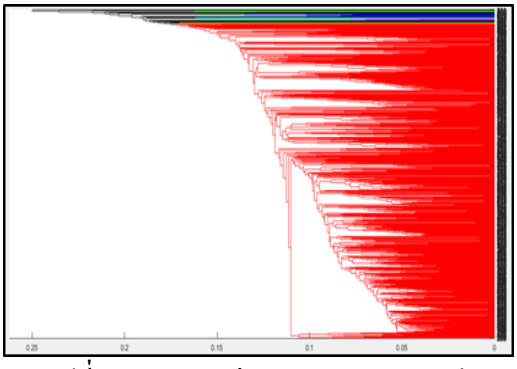
	สมการ	วิธีการจัดกลุ่ม Hierarchical
1	TFIDF	Single Link
2		Average Link
3		Complete Link
4	WTFIDF	Single Link
5		Average Link
6		Complete Link
7	FTFIDF	Single Link
8		Average Link
9		Complete Link

4. ผลการทดลอง

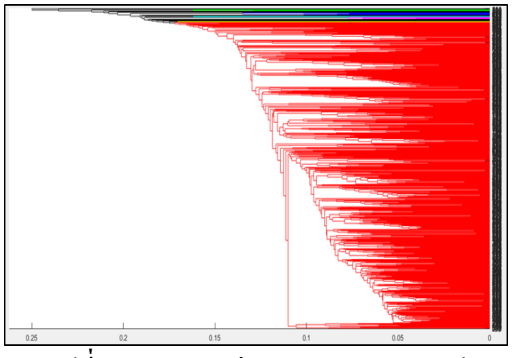
ผลการทำเหมืองข้อความผ่านการหาค่าน้ำหนักคำ (Document indexing) จากสมการ 3 สมการ ได้แก่สมการ TFIDF สมการ WTFIDF และสมการ FTFIDF ร่วมกับการจัดกลุ่มข้อความแบบ Hierarchical ด้วยวิธี Single link วิธี Average link และวิธี Complete link ทำให้ได้ผลลัพธ์เปรียบเทียบการจัดกลุ่มทั้ง 9 รูปแบบ แสดงดังรูปที่ 1 ถึงรูปที่ 9 ตามลำดับ ดังนี้



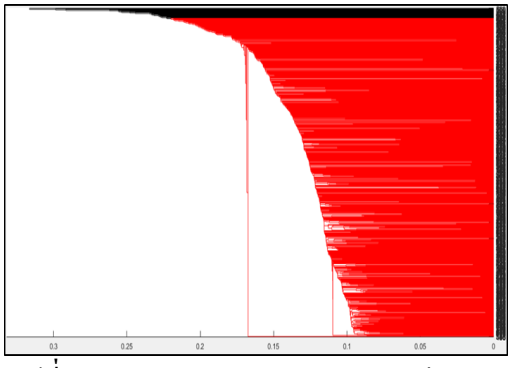
รูปที่ 1 ผลการจัดกลุ่มแบบ Hierarchical วิธี Single Link ด้วยสมการ TFIDF



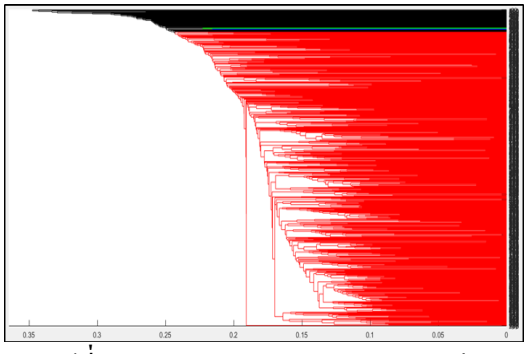
รูปที่ 2 ผลการจัดกลุ่มแบบ Hierarchical วิธี Average Link ด้วยสมการ TFIDF



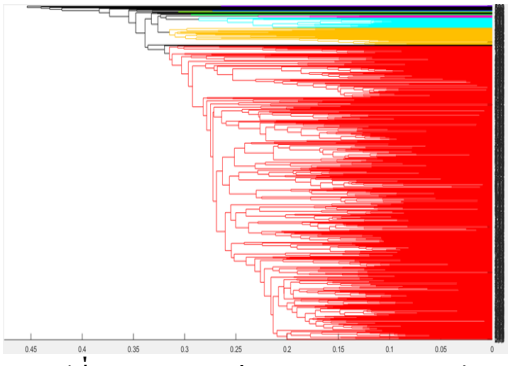
รูปที่ 3 ผลการจัดกลุ่มแบบ Hierarchical วิธี Complete Link ด้วยสมการ TFIDF



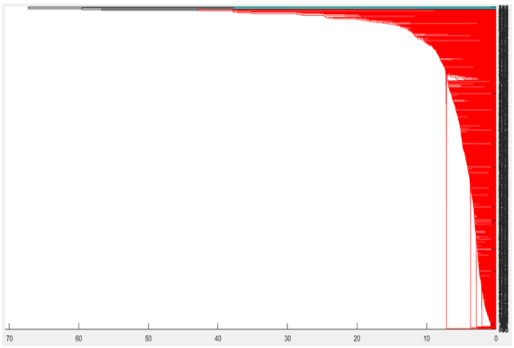
รูปที่ 4 ผลการจัดกลุ่มแบบ Hierarchical วิธี Single Link ด้วยสมการ WTFIDF



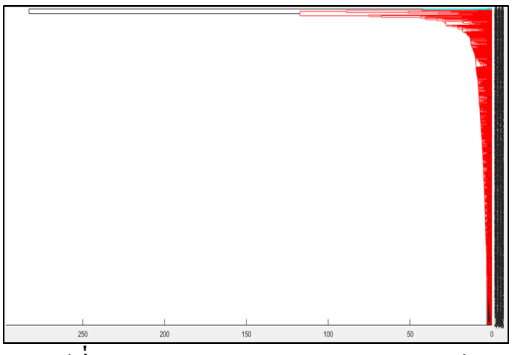
รูปที่ 5 ผลการจัดกลุ่มแบบ Hierarchical วิธี Average Link ด้วยสมการ WTFIDF



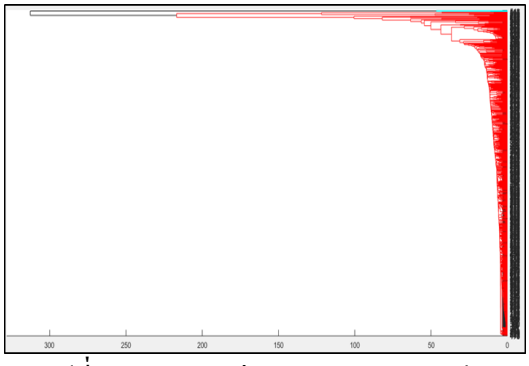
รูปที่ 6 ผลการจัดกลุ่มแบบ Hierarchical วิธี Complete Link ด้วยสมการ WTFIDF



รูปที่ 7 ผลการจัดกลุ่มแบบ Hierarchical วิธี Single Link ด้วยสมการ FFIDF



รูปที่ 8 ผลการจัดกลุ่มแบบ Hierarchical วิธี Average Link ด้วยสมการ FFIDF



รูปที่ 9 ผลการจัดกลุ่มแบบ Hierarchical วิธี Complete Link ด้วยสมการ FFIDF

นอกจากนี้เมื่อทำการเปรียบเทียบอัลกอริทึมจัดกลุ่มข้อมูลมะเร็งโดยพิจารณาจากค่า Sum of Square Error (SSE) ซึ่งหากค่าความผิดพลาดมีค่าสูงแสดงว่าข้อมูลภายในกลุ่มมีการกระจายตัวมาก ตรงกันข้ามหากพบว่าความผิดพลาดมีค่าต่ำแสดงว่าข้อมูลที่ถูกจัดกลุ่มนั้นเกาะกลุ่มกันมาก ผลจากการทดลองแสดงดังตารางที่ 3 พบว่าอัลกอริทึมจัดกลุ่มดีที่สุดคือสมการ WTFIDF-Complete link ให้ค่า SSE ต่ำที่สุดเท่ากับ 40490.54 จำนวนกลุ่มที่ได้เท่ากับ 6 กลุ่ม รองลงมา คือสมการ WTFIDF-Average link มีค่า SSE เท่ากับ 40591.10 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม สมการ TFIDF-Complete Link มีค่า SEE เท่ากับ 40601.72 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม สมการ TFIDF- Average Link มีค่า SSE เท่ากับ 40614.36 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม สมการ FTFIDF- Complete Link มีค่า SSE เท่ากับ 40630.00 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม สมการ TFIDF- Single Link มีค่า SSE เท่ากับ 40667.74 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม สมการ FTFIDF- Single Link มีค่า SSE เท่ากับ 40675.00 สมการ มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม WTFIDF- Single Link มีค่า SEE เท่ากับ 40683.77 มีจำนวนกลุ่มเท่ากับ 5 กลุ่ม และสมการ

FTFIDF- Single Link มีค่า SSE เท่ากับ 40684.00 มีจำนวนกลุ่มเท่ากับ 6 กลุ่ม ตามลำดับ

ตารางที่ 3 แสดงวิธีการเปรียบเทียบสมการที่หาค่าน้ำหนักคำกับประสิทธิภาพวิธีการจัดกลุ่ม

สมการ	วิธีการจัดกลุ่ม	Group	ค่า SSE
TFIDF	Single Link	6	40667.74
	Average Link	6	40614.36
	Complete Link	6	40601.72
WTFIDF	Single Link	5	40683.77
	Average Link	6	40591.10
	Complete Link	6	40490.54
FTFIDF	Single Link	6	40675.00
	Average Link	6	40684.00
	Complete Link	6	40630.00

5. บทสรุป

ผลการวิจัยการประยุกต์ใช้เทคนิคเหมืองข้อความสำหรับการจัดกลุ่มข้อมูล กรณีศึกษามะเร็ง เมื่อพิจารณาจากค่า Sum of Square Error (SSE) พบว่าอัลกอริทึมจัดกลุ่มดีที่สุด คือ การนำข้อมูลที่ผ่านการหาค่าน้ำหนักคำ (Document indexing) จากสมการ WTFIDF ร่วมกับวิธีการจัดกลุ่มแบบ Complete link มีค่า SSE ต่ำที่สุดเท่ากับ 40490.54 จำนวนกลุ่มที่ได้เท่ากับ 6 กลุ่ม ซึ่งแสดงว่าข้อมูลภายในกลุ่มมีการกระจายตัวค่อนข้างมากส่งผลต่อประสิทธิภาพการจัดกลุ่มที่ดีเมื่อเปรียบเทียบกับอัลกอริทึมการจัดกลุ่มวิธีอื่น ๆ

ผลการวิจัยที่ได้ผู้วิจัยสามารถนำอัลกอริทึมการหาค่าน้ำหนักคำจากการสมการ WTFIDF ร่วมกับวิธีการจัดกลุ่มแบบ Complete link นำไปพัฒนาต่อยอดเป็นเว็บท่าเชิงความหมายสำหรับการจัดกลุ่มข้อมูลโดยใช้เทคนิคเหมืองข้อความ กรณีศึกษาโรงมะเร็ง เพื่อใช้ในการรวบรวมและจัดหมวดหมู่ข้อมูลจากเว็บไซต์ต่าง ๆ ที่เกี่ยวกับโรคมะเร็งมารวมไว้แหล่งเดียวกัน จะทำให้ผู้ป่วยโรคมะเร็งหรือบุคคลทั่วไปสามารถค้นหาข้อมูลโรคมะเร็งได้อย่างสะดวกรวดเร็วและตรงตามความต้องการมากขึ้น

6. เอกสารอ้างอิง

[1] ทินกร คุณาสิทธิ์, สิรภัทร เชี่ยวชาญวัฒนา และ คำรณ สุนัติ. “การจัดกลุ่มเอกสารคำอธิบายเว็บภาษาไทยสำหรับผลการสืบค้นด้วยนอนเน็กกาทีฟเมริกซ์แฟกทอไรเซชั่น.” ใน The 4th National conference on Coputer and Information Technology: NCCIT2008. vol. 23–24 พฤษภาคม 2551, no. 4 : 386–391.

[2] Chainapaporn, P. and Netisopakul, P. “Thai Herb Information Extraction from Multiple Websites.” Knowledge and Smart Technology (KST), 2012 4th International Conference. on, Jul. 2012 : 16–23.

[3] Gupta, V. and Lehal, S. G. “A Survey of Text Mining Techniques and Applications.” JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE. Vol.1 No.1 (Aug 2009) : 60-76.

[4] CARPINETO, C. OSI´STANISlAW OSI´, S. ROMANO,G. and WEISS, D. “A Survey of Web Clustering Engines.” ACM Computing Surveys vol. July 2009 : No. 3.

[5] Thanadechteemapat, W. and Chun Che Fung. “Automatic Content Extraction and

Visualization of Thai Websites for Improved Information Representation." IEEE International Conference on Systems, Man, and Cybernetics. vol. 2012, Oct. 2012 : 2230–2234.

[6] Nuipian, V. Meesad, P. and Boonrawd, P. "A Comparison between Keywords and Key-phrases in Text Categorization using Feature Section Technique." 2011 Ninth International Conference on ICT and Knowledge Engineering. (2012) : 156-160.

[7] Ren, F. and Sohrab, G. M. "Class-indexing-based term weighting for automatic text classification." Information Sciences 236 (2013). Vol. 236 No. (July 2013) : 109-125.

[8] Kwok, T-Y. J. "Automated Text Categorization Using Support Vector Machine." In Proceedings of the International Conference on Neural Information Processing (ICONIP). 1998 : 347-351.

[9] Trstenjak, B. Mikac, S. and Donko, D. "KNN with TF-IDF Based Framework for Text Categorization." Proceeding on the 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013. (2013) : 1356-1364.

[10] C.-F. Tsai, C.-T. Tsai, C.-S. Hung, and P.-S. Hwang, "Data Mining Techniques for Identifying Students at Risk of Failing a Computer Proficiency Test Required for Graduation," Australasian Journal of Educational Technology, vol. 27, no. 3, pp. 481–498, 2011.

[11] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques: Concepts and Techniques. Elsevier, 2011.

[12] วราภรณ์ คงสมพงษ์ และธีรพงษ์ สังข์ศรี. "การสกัดเอกสารสนเทศของเอกสารโครงงานนักศึกษาแบบอัตโนมัติบนฐานของกฎ." ใน The 3rd ASEAN Undergraduate Conference in Computing (AUC2) 2015.

[13] Quinlan, J. R. (1986) "Introduction of Decision Trees." In Machin Learning. 81-106