วารสารแม่โจ้เทคโนโลยีสารสนเทศและนวัตกรรม มหาวิทยาลัยแม่โจ้ ปีที่ 3 ฉบับที่ 1 มกราคม – มิถุนายน 2560

การเพิ่มประสิทธิภาพของการค้นคืนข้อมูลโดยอาศัยการขยายคำสอบถามแบบจัดกลุ่ม

Improvement of Retrieval Efficiency using Clustered Base Query Expansion

รัฐสิทธิ์ สุขะหุต และ คมสันต์ น้อยเทพ

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

Rattasit Sukhahuta and Khomsan Noithep

Department of Computer Science, Faculty of Science,

Chiangmai University

Abstract

There has been an increasing numbers of electronic document distributed online in nowadays. Most of the data available is in a semi-structure and unstructured format. To make use of these data, we need to understand the underlying structure at some level in order for data becomes useful. This can be done using text pre-processing process and data structure analysis with natural language processing process. In information retrieval, interested documents can be retrieved using indexing technique consisting of a list of sorted terms that link to a list of documents. Each document is associated with the frequency count indicating the number of times that a term appear in the document. The users can then search for the interested documents by specify one or more keywords called 'terms query'. These keywords can be either a single word or multiple words together. The key issue in identifying search terms is that most users may not be able to identify the search terms correctly. The problems occur when user do not know what terms are indexed. Therefore, the interested documents can not be found. This paper proposes a technique that expand the query terms based on the term similarity and terms with similar meaning within the same domain. By measuring the value of recall and precision, this technique has proven that this technique resulted in increased document retrieval performance.

Keywords: *Query Expansion, Keyword Search*, *WordNet*

บทคัดย่อ

ปัจจุบันการเผยแพร่เอกสารในรูปของเอกสารอิเล็กทรอนิกส์มีปริมาณที่เพิ่มมากขึ้น และเป็นข้อมูลที่จัดอยู่ในประเภทของข้อมูลที่ไม่มีโครงสร้างตายตัว การค้นคืนข้อมูลจากเอกสารจำเป็นต้องทำความเข้าใจกับรูปแบบโครงสร้างข้อมูลที่จัดเก็บ และต้องมีการเตรียมข้อมูลให้เหมาะสมก่อนการค้นคืน หลักการทำงานของระบบการค้นคืนเอกสารจะต้องอาศัยคำค้นคืน ซึ่งเป็นคำสำคัญที่ผู้ใช้งานสนใจหรือต้องการสืบค้นโดยการระบุคำที่ใช้ในการค้นคืนนั้น อาจจะเป็นคำเดี่ยวหรือคำหลายคำมาประกอบกัน ปัญหาสำคัญของการระบุคำค้นคืนคือผู้ใช้งานส่วนใหญ่อาจจะไม่สามารถระบุคำค้นคืนได้อย่างถูกต้องหรือคำที่ต้องการไม่ได้ถูกจัดให้เป็นดัชนีสำหรับค้นคืน ทำให้ไม่พบเอกสารที่ต้องการ จากการทดลองพบว่าแนวคิดการขยายคำสืบค้นโดยอาศัยหลักการจัดกลุ่มคำจากคำนิยามศัพท์จากเวิร์ดเน็ต จากการวัดผลค่าระลึกและค่าความแม่นยำพบว่าเทคนิควิธีนี้ทำให้ประสิทธิภาพการค้นคืนเอกสารเพิ่มมากขึ้น

คำสำคัญ: การขยายคำสืบค้น, สืบค้นคำสำคัญ, เวิร์ดเน็ต

1. บทนำ (Introduction)

การจัดกลุ่มเอกสารเป็นเทคนิคที่ใช้สำหรับการจำแนกเอกสารตามคุณสมบัติของเอกสารที่กำหนด โดยให้เอกสารที่มีคุณสมบัติที่มีความคล้ายคลึงกันถูกจัดให้อยู่ภายในกลุ่มเดียวกัน เทคนิคการจัดกลุ่มสามารถนำไปใช้กับงานค้นคืนเอกสารเพื่อเพิ่มประสิทธิภาพ ความรวดเร็วในการค้นคืนเอกสาร การจัดกลุ่มสามารถทำได้โดยการแบ่งออกเป็นกลุ่มย่อยๆ เมื่อผู้ใช้มีการระบุคำสำหรับการค้นคืนเพื่อให้มีการจัดกลุ่ม และลดขนาดของจำนวนกลุ่มเป้าหมายที่ต้องการค้นคืนแทนการค้นหาจากเอกสารทั้งหมด การกำหนดคำสำคัญของคำที่ต้องการสืบค้นผู้ใช้จำเป็นต้องเข้าใจถึงเนื้อหาและคำสำคัญที่เป็นตัวแทนของเอกสารที่สนใจ เนื่องจากการประมวลผลข้อมูลต้นทางอยู่ในรูปของภาษาธรรมชาติ แนวคิดของการประมวลผลด้วยภาษาธรรมชาติจึงถึงนำมาใช้เพื่อวิเคราะห์โครงสร้างข้อมูล โดยการศึกษาการสกัดคุณสมบัติของคำสำคัญพบว่ามีการเลือกด้วยคำเดี่ยว พยางค์ วลี หรือกลุ่มคำที่อยู่ในรูปของประโยค ที่อยู่ในระบบการสืบค้นนั้นจะอาศัยการทำดัชนี และการสกัดคุณลักษณะที่สำคัญของเอกสารที่มีการรวบรวมมาจากอินเตอร์เน็ตหรือแหล่งที่เก็บข้อมูลต่างๆ ซึ่งกลุ่มคำเหล่านี้เรียกว่าถุงคำ ที่ถูกนำมาจัดให้อยู่ในรูปของเวคเตอร์ โดยมีการแทนค่าด้วยคุณลักษณะของค่าความถี่ (Frequency) และความถี่ผกผัน (Inverse Document Frequency) ที่ปรากฏและค่าน้ำหนักตามโดเมนที่กำหนด นอกจากนี้การนำแนวคิดของการประมวลผลภาษาธรรมชาติมาใช้เพื่อประมวลผลคำ อาทิการตัดคำ การลดรูปของคำ (Word Inflection) เพื่อให้คำอยู่ในรูปของรากศัพท์ กรณีที่มีการผันคำให้อยู่ในรูปต่างๆ ยังอาศัยการรวมคำเดี่ยวให้เป็นกลุ่มคำโดยอาศัย N-Gram ยังสามารถช่วยแก้ปัญหาในเรื่องของตำแหน่งการเกิดของคำอีกด้วย พิลาวัณย์ พลับรู้การ และกฤษณะ ไวยมัย ได้ศึกษาค้นคว้าเพื่อวัดความคล้ายคลึงของคำเพื่อนำมาจัดกลุ่มโดยอาศัย

คำนิยามศัพท์จากพจนานุกรมเวิร์ดเน็ตเพื่อนำเอาคำที่อยู่ในกลุ่มเดียวกันนี้มาทำการขยายคำสอบถามเพื่อเป็นการเพิ่มประสิทธิภาพของการค้นคืนเอกสารในระบบการสืบค้นต่อไป

2. เอกสารงานวิจัยที่ผ่านมา (Research Background)

จากงานวิจัยที่ผ่านมาพบว่าการค้นคืนเอกสารโดยวิธีการจัดกลุ่มได้มีบทบาทสำคัญเช่น นิเวศ จิระวิชิตชัย (2556) การจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน การปรับค่าพารามิเตอร์เคอร์เนลฟังก์ชั่นแบบต่างๆ โดยทดสอบประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยกับอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) และเนอีฟเบย์ (Naïve-Bayes) โดยใช้วิธีการลดคุณลักษณะร่วมกับอัลกอริทึมเครื่องจักรการเรียนรู้ จากการทดลองพบว่าการลดคุณลักษณะ ด้วยวิธี Information Gain เพื่อลดมิติของข้อมูล แล้วส่งเข้าเครื่องจักรการเรียนรู้และวัดประสิทธิภาพจากค่า F-Measurement สูงสุด สามารถสรุปได้ว่า อัลกอริทึม SVM เคอร์เนลฟังก์ชั่นแบบ Linear และ SVM เคอร์เนลฟังก์ชันแบบ Polynomial Degree = 3 ให้ประสิทธิภาพการจัดหมวดหมู่โดยเฉลี่ยออกมาดีที่สุดคือ 95.1% รองลงมาเป็นอัลกอริทึม SVM เคอร์เนลฟังก์ชั่นแบบ Radial Basis Function (RBF) gamma 0.8 และ 1.0 ให้ประสิทธิภาพการจัดหมวดหมู่ 94.9% อัลกอริทึม Naïve Bays ให้ประสิทธิภาพการจัดหมวดหมู่ 88.7% อัลกอริทึม C4.5 ให้ประสิทธิภาพการจัดหมวดหมู่ 79.9% ตามลำดับ ทั้งนี้จากผลงานวิจัยของ นิเวศ จิระวิชิตชัย และคณะ (2551) ได้ทำการวิจัยการจัดหมวดหมู่เอกสารโดยอาศัยอัลกอริทึม Support Vector Machine มีพฤติกรรมที่จะแยกแยะข้อมูล โดยใช้สมการระนาบหลายมิติโดยจะพยายามหาจุดข้อมูลที่ทำให้ได้สมการระนาบหลายมิติที่ใช้แบ่งแยกดีที่สุด (Optimal Hyperplane) ความถูกต้องที่สุด โดยพิจารณาจากระยะห่าง (Margin) ระหว่างคลาส ซึ่งเส้นระนาบที่ดีที่สุดนี้จะสามารถจำแนกกลุ่มเอกสารออกมาได้อย่างมีประสิทธิภาพ ผลจากการทดลองพบว่าสามารถลดขนาดคุณลักษณะและทดสอบด้วยอัลกอริทึม Support Vector Machine จากกลุ่มตัวอย่าง พบว่าสามารถลดคุณลักษณะลงได้มากถึง 91.37% โดยการลดลงของคุณลักษณะดังกล่าวไม่ส่งผลให้ประสิทธิภาพในการจัดหมวดหมู่เอกสารลดลงแต่อย่างใด แต่สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก จากผลการทดลองนี้สามารถนำไปประยุกต์ใช้ประโยชน์ในการสร้างระบบจัดหมวดหมู่เอกสารอัตโนมัติ และสามารถนำมาประยุกต์ใช้กับงานด้านอื่นๆ เช่น การคัดกรองเอกสาร (Document Filtering) การจัดทำดัชนีอัตโนมัติเพื่อใช้ในการค้นคืนเอกสาร (Automatic Indexing for IR System) การจัดหมวดหมู่ของเว็บเพจ (Web Page Classification) เป็นต้น ในขณะที่ ชุลีรัตน์ จรัสกุลชัยและคณะ (2556) ได้ศึกษาถึงแนวทางการจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย งานวิจัยฉบับนี้ได้ศึกษาวิจัยขั้นตอนวิธีการจัดกลุ่มเอกสารทั้งแบบขั้นตอนวิธี การจัดกลุ่มแบบ Complete link สำหรับการกลุ่มแบบลำดับชั้น และ Single pass สำหรับการจัดกลุ่มแบบไม่เป็นลำดับชั้น โดยประยุกต์งานขั้นตอนวิธีดังกล่าวกับข้อความข่าวภาษาไทย นอกจากนี้

งานวิจัยดังกล่าวยังได้ประยุกต์หลักการประมวลผลแบบขนาน เพื่อแก้ปัญหาในการคำนวณค่าความเหมือนของเอกสาร ผลงานวิจัยเบื้องต้นสรุปได้ว่าขั้นตอนวิธีในการตัดคำไม่มีผลต่อการจัดกลุ่ม และขั้นตอนวิธีในการจัดกลุ่มทั้งสองแบบไม่ได้ให้ผลที่แตกต่างกันอย่างชัดเจน

จิราภรณ์ ถมแก้ว และศรัณย์ อินทโกสุม (2555) ได้นำเสนอแนวคิดของการจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ งานวิจัยชิ้นนี้นำเสนอการทดสอบสมมติฐานที่ว่าการจำแนกข้อมูลจากงานวิจัยมักจะพิจารณาคุณลักษณะทั้งหมดของข้อมูล อย่างไรก็ตามคุณลักษณะบางประการมีความสำคัญน้อยซึ่งเมื่อนำมาร่วมคำนวณด้วยแล้วอาจเป็นสาเหตุทำให้ความแม่นยำในการจำแนกข้อมูลลดลง โดยในการทดลองประยุกต์ใช้อัลกอริธึมแบบตะกละ (Greedy algorithm) เพื่อคัดเลือกคุณลักษณะที่สำคัญของข้อมูล ร่วมกับการจำแนกข้อมูล ผลการทดลองพบว่า การใช้วิธีคัดเลือกคุณลักษณะด้วยกรีดดีอัลกอริธึมร่วมกับการจำแนกข้อมูลด้วย RBF สามารถเพิ่มประสิทธิภาพในการจำแนกข้อมูลให้สูงขึ้นและใช้เวลาประมวลผลลดลงเมื่อเปรียบเทียบกับการจำแนกข้อมูลโดยไม่มีการคัดเลือกคุณลักษณะ

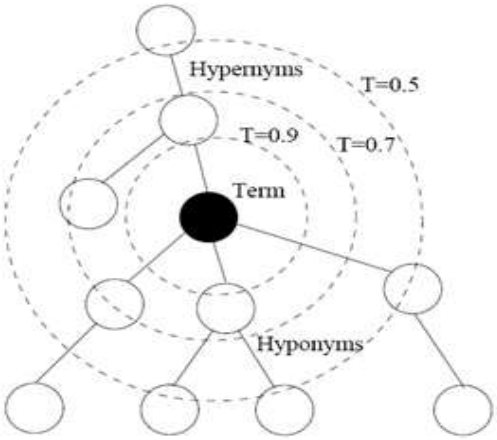
นอกจากนี้การหาค่าความสำคัญของเอกสารซึ่งจะดูที่น้ำหนักของคำที่สนใจโดยใช้ทฤษฎี tf*idf และการหาค่าความคล้ายเอกสารนั้นจะใช้ ทฤษฎี Vector Space Model (VSM) ซึ่งก็คือ Cosine Similarity จะได้ผลดีกว่า Inner Product และต้องเป็นคำคำเดียวกันถ้าเอกสารเป็นคำคนละคำกันแต่มีความหมายเหมือนกันก็จะไม่สามารถทำได้ ซึ่ง SSRM เป็นวิธีที่จะมาใช้แก้ปัญหานี้โดยมีวิธีการดังนี้

(1) Term Re-Weighting การหาค่าน้ำหนักใหม่ของคำ ซึ่งน้ำหนักใหม่ของคำแต่ละคำจะแทนค่าด้วย qi ของแต่ละการสืบค้น i จะปรับโดยดูความสัมพันธ์กับคำที่มีความหมายคล้ายกับคำ j ในเวคเตอร์เดียวกัน ดังสมการ (2.1)

$$q_i = q_i + \sum_{sim(i,j) \geq t}^{j \neq i} q_i sim(i,j) \qquad (2.1)$$

โดยที่ t คือค่าเทรชโฮลด์ (Threshold) ที่ผู้ใช้กำหนดขึ้น (ในที่นี้ t = 0.8) สูตรนี้ใช้เฉพาะคำที่มีคำที่คล้ายกันกับคำที่สืบค้น

(2) เทอมเอ็กซ์แพนชัน การขยายคำศัพท์ ข้อแรกเลือกคำพ้อง หลักจากนั้นเลือกคำใน Hyponyms และ Hypernyms ของคำค้น



ภาพที่ 1 โครงสร้างต้นไม้ของเวิร์ดเน็ต

โดยแต่ละคำนั้นจะสืบค้นจาก โครงสร้างต้นไม้ของเวิร์ดเน็ต (WordNet tree) จากภาพที่ 1 ซึ่งจะมีโครงสร้างของคำคำนั้น คำที่มีค่าเทรชโฮลด์ มากกว่า 0.9 จะถูกนำมาเพิ่มในการค้นคืน คำที่นำมาเพิ่มอาจจะอยู่สูงกว่า หรือต่ำกว่า มากกว่า 1 ขั้นของคำคำนั้นก็ได้ ดังสมการ (2.2)

$$q_i = \begin{cases} \sum_{sim(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j sim(i,j), & i \text{ is a new term} \\ q_i + \sum_{sim(i,j)\geq T}^{i\neq j} \frac{1}{n} q_j sim(i,j), & i \text{ had weight } q_i \end{cases} \quad (2.2)$$

โดยที่จำนวน n คือจำนวนของ Hyponym ของแต่ละคำ j และสำหรับ Hypernym n จะมีค่าเท่ากับ 1 คำที่อยู่ในคำสืบค้นอยู่แล้วอาจจะกลายเป็นคำใหม่สำหรับคำอื่น และคำหนึ่งคำ อาจจะถูกเพิ่มมากกว่าหนึ่งครั้งได้

(3) การหาความคล้ายของเอกสาร (Document Similarity) ใช้สูตรดังนี้

$$sim(q,d) = \frac{\sum_i \sum_j q_i d_i sim(i,j)}{\sum_i \sum_j q_i d_i} \quad (2.3)$$

จากสมการ (2.3) โดยที่ i และ j คือคำที่สนใจและคำในเอกสารตามลำดับ คำที่สนใจจะถูกคำนวณน้ำหนักใหม่ และถูกขยายคำ โดยที่คำในเอกสารจะไม่ต้องทำอะไรนอกจากหาน้ำหนักโดยใช้ สูตร tf*idf เท่านั้น ผลการสืบค้นจะมีค่าระหว่าง 0 กับ 1

(4) การหาค่า tf*idf คือการหาค่าความถี่ของเทอมในเอกสารและความถี่ของเอกสารที่มีเทอมนั้นอยู่ โดยแบ่งสามารถคำนวนหาได้จากสูตรดังต่อไปนี้

การหาค่า ความถี่ของโทเค็น (Token) ที่ปรากฏในเอกสาร หรือก็คือความถี่ของเทอมนั้น (Term weight : Term frequency) ดังสมการ (2.4)

$$F_{ij} = Frequency\ of\ term\ i\ in\ document\ j \tag{2.4}$$

การหาค่าความถี่ของเทอม (tf : Term frequency) ได้จากสมการ (2.5)

$$TF_{ij} = \frac{F_{ij}}{\max\{F_{ij}\}} \tag{2.5}$$

การหาค่าน้ำหนักของเอกสาร (Term weight : inverse document frequency) ที่มีเทอมปรากฏอยู่ในเอกสาร โดยการหาค่าความถี่ของเอกสาร (Document frequency) จากสมการ (2.6)

$$Df_i = \frac{The\ document\ frequency\ of\ t}{Thw\ number\ of\ documents\ that\ contain\ t} \tag{2.6}$$

การหาค่าความถี่เอกสารผกผัน (Inverse Document Frequency) หาได้จากการนำค่า Df ที่คำนวนได้มาคำนวนจากสมการ (2.7) ต่อไปนี้

$$\boldsymbol{IDF_i = \log_{2\frac{n}{Df_i}}} \tag{2.7}$$

การหาค่าน้ำหนักความสัมพันธ์ของเทอมกับเอกสาร (TF-IDF weighting) ตามสมการ (2.8)

$$W_{ij} = TF_{ij} * IDF_{ij} \tag{2.8}$$

วงกต ศรีอุไร และคณะ (2552) ได้นำเสนองานวิจัยเกี่ยวกับการเตรียมฟีเจอร์บนพื้นฐานแบบจำลองหัวข้อสำหรับการจำแนกหมวดหมู่ของเอกสาร งานวิจัยชิ้นนี้กล่าวว่าโดยทั่วไปการจำแนกหมวดหมู่ของเอกสารจะใช้การแทนเอกสารด้วยวิธี Bag of Words (BOW) ซึ่งเป็นวิธีที่ง่ายแต่เป็นวิธีที่ไม่ได้ให้ความสำคัญกับคำที่มีความหมายเหมือนกัน ดังนั้นเมื่อนำข้อมูลไปใช้ในการจำแนกหมวดหมู่จึง

ส่งผลต่อประสิทธิภาพในการจำแนกหมวดหมู่และการค้นคืนเอกสาร งานวิจัยนี้มีวัตถุประสงค์เพื่อปรับปรุงการจำแนกหมวดหมู่ของเอกสารโดยนำเสนอการแทนเอกสารด้วยวิธีสร้างแบบจำลองหัวข้อให้กับเอกสาร

เทคนิคการจัดกลุ่มข้อมูล (Data Clustering) สามารถแบ่งออกได้เป็น 2 ประเภทใหญ่ได้แก่การแบ่งแบบตัดส่วน (Partitioning) โดยแบ่งข้อมูลออกเป็นกลุ่มต่างๆตามจำนวนกลุ่มที่กำหนด และการแบ่งแบบลำดับชั้น (Hierarchical) เป็นลักษณะของการแบ่งเป็นกลุ่มย่อยที่ถูกแบ่งไว้ก่อนหน้านั้นซ้ำๆหลายครั้ง โดยการแบ่งแบบลำดับชั้นนั้นสามารถแบ่งได้ 2 วิธีคือแบบบนลงล่าง หรือล่างขึ้นบน ปัจจุบันการจัดกลุ่มข้อมูลมีอยู่หลายเทคนิคด้วยกัน อาทิ Exclusive Clustering เป็นการแบ่งกลุ่มข้อมูลที่มีลักษณะเหมือนกันมาไว้ในกลุ่มเดียวกัน Overlapping Clustering เป็นการแบ่งกลุ่มข้อมูลให้เป็นเซ็ตย่อยๆซึ่งผลลัพธ์ที่ได้มากกว่าหนึ่งคลัสเตอร์ก็ได้และมีจำนวนสมาชิกภายในเซ็ตต่างๆที่มีค่าแตกต่างกัน, Hierarchical Clustering ซึ่งเป็นการรวมเอาคุณสมบัติของ 2 เทคนิคแรกมาไว้ด้วยกัน และ Probabilistic Clustering ซึ่งเป็นการแบ่งกลุ่มโดยวิธีทางสถิติ การแบ่งกลุ่มข้อมูลเป็นเทคนิควิธีการวิเคราะห์เซ็ตของข้อมูลที่ถูกจัดให้อยู่ในรูปของเวกเตอร์คำเพื่อนำมาพิจารณาความคล้ายจากคุณสมบัติความเหมือน (Similarity) หรือระยะห่าง (Proximity) โดยคำนวณจากระยะห่างระหว่างเวกเตอร์ของเอกสาร เทคนิคประกอบด้วยยูเครเดียน (Euclidean) แบบแมนฮัตตัน (Manhattan) และการเชบิเชฟ (Chebychev) ซึ่งก็จะได้ผลของการจัดกลุ่มที่แตกต่างกันไป นอกจากนี้การแบ่งกลุ่มยังสามารถแบ่งออกเป็นแบบการเรียนรู้แบบมีผู้สอน (Supervised Learning) และไม่มีผู้สอน (Un-Supervised Learning) ได้แก่ K-Means, Hierarchical และ Self-organizing การแบ่งกลุ่มเอกสารในลักษณะนี้การให้คะแนนความคล้ายคลึงระหว่างเอกสารจะกำหนดเป็นแบบไบนารี 0 หรือ 1 หมายถึงเอกสารที่มีคำสำคัญปรากฏอยู่ในทั้ง 2 เอกสาร ซึ่งในบางครั้งเอกสารที่ประกอบด้วยคำที่มีความหมายเหมือนกันแต่เขียนต่างกันก็จะไม่ได้ถูกจัดให้อยู่กลุ่มเดียวกันก็เป็นไปได้ ซึ่งเป็นข้อจำกัดของเทคนิคการจัดกลุ่มแบบนี้

ปัจจุบันพบว่าการจัดกลุ่มได้นำเอาวิธีการฟัซซี (Fuzzy) อาทิ Sequence hierarchical Clustering, Hard C-Mean Clustering ถูกนำมาใช้เพื่อพิจารณาความแปรปรวนในคุณสมบัติของคำสำคัญที่เป็นตัวแทนของเอกสาร ตัวอย่างเช่นการใช้ฟัซซี ซี-มีนส์ (Fuzzy C-Means: FCM) เป็นการแบ่งกลุ่มซ็อตของภาพเคลื่อนไหว เพื่อการแบ่งข้อมูล การแบ่งกลุ่มด้วย ฟัซซี ซี-มีนส์ (Fuzzy C-Means (FCM) Clustering) ศักดิ์ชัย ศรีมากรณ์ (2551) ได้นำเสนองานวิจัยเรื่องการแบ่งกลุ่มซ็อตภาพเคลื่อนไหวโดยใช้วิธีการแบ่งจำนวนกลุ่มโดยไม่ต้องรู้จำนวนกลุ่ม กล่าวถึงวิธีการแบ่งกลุ่มข้อมูลนั้นสามารถทำได้หลายวิธี เช่น Sequential Clustering Hierarchical Clustering, Hard C– Means Clustering ฯลฯ เพื่อช่วยในการแบ่งบริเวณต่างๆ ให้แยกออกจากกันได้ชัดเจนยิ่งขึ้น Fuzzy C– Means Clustering ก็เป็นวิธีการแบ่งกลุ่มข้อมูลวิธีการหนึ่งที่ได้รับความนิยมมาก ขั้นตอนการแบ่งกลุ่มแบบฟัซซี ซี-มีนส์

ประกอบด้วยการกำหนดจำนวนกลุ่มเริ่มต้นในชุดข้อมูล การกำหนดระดับค่าความเป็นสมาชิกในกลุ่ม ข้อมูลทุกตัวจะถูกนำมาคำนวณค่าความเป็นสมาชิกในการหาตัวแทนกลุ่มแต่ละกลุ่มที่มีลักษณะสำคัญของกลุ่มครบถ้วน จากนั้นจะเข้าสู่ขั้นตอนการจัดข้อมูลเข้ากลุ่ม แล้วค่อยมาปรับค่าความเป็นสมาชิกของสมาชิกทุกตัวในแต่ละกลุ่มและการปรับค่าตัวแทน จนกระทั่งตัวแทนปรับค่าศูนย์กลางของกลุ่มได้และค่าดัชนีที่วัดความผิดพลาดของการแบ่งการกลุ่มมีค่าน้อยลงจนถึงจุดที่แสดงว่าการแบ่งกลุ่มข้อมูลน่าจะถูกต้อง สำหรับการจำแนกข้อมูลภาพออกเป็นกลุ่มๆ ข้อมูลที่กล่าวถึงในที่นี้คือค่าที่ใช้แทนแต่ละพิกเซล) จากหลักการในการแบ่งกลุ่มข้อมูลโดยวิธี (ของเฟรมภาพ FCM โดยพิจารณาชุดข้อมูล $\mathbf{X}=\{\underline{\mathbf{x}}_1,\underline{\mathbf{x}}_2,\mathbf{K},\underline{\mathbf{x}}_n\}$ โดยที่ $\underline{\mathbf{x}}_k$ เป็นเวกเตอร์ใน d มิติ ถ้าเราต้องการแบ่งข้อมูลออกเป็น c กลุ่ม และมี ฟัซซี ซูโดพาร์ทิชัน $P=\{A_1,A_2,\mathbf{K},A_c\}$ โดยที่ $A_i(\underline{\mathbf{x}}_k)$ คือ Membership Grades ของ $\underline{\mathbf{x}}_k$ ทั้งหมดที่มีต่อ Cluster i สามารถคำนวณจุดศูนย์กลางของทุกๆ กลุ่ม ได้จาก

$$\underline{\mathbf{v}}_i=\frac{\sum_{k=1}^{n}\left[A_i(\underline{\mathbf{x}}_k)\right]^m\underline{\mathbf{x}}_k}{\sum_{k=1}^{n}\left[A_i(\underline{\mathbf{x}}_k)\right]^m} \quad ; \quad i=1,2,..,c \tag{1}$$

เมื่อ $m>1$ เป็นจำนวนจริงที่ควบคุมผลของค่าความเป็นสมาชิก (Membership Grade) ที่มีต่อการแบ่งกลุ่มข้อมูล ตามนิยามดรรชนีสมรรถนะ (Performance Index; $J_m(P)$) ของ P โดย

$$J_m(P)=\sum_{k=1}^{n}\sum_{i=1}^{c}\left[A_i(\underline{\mathbf{x}}_k)\right]^m\|\underline{\mathbf{x}}_k-\underline{\mathbf{v}}_i\|^2 \tag{2}$$

เป้าหมายในการทำการแบ่งกลุ่ม (Clustering) คือการหา P ที่ทำให้ $J_m(P)$ มีค่าน้อยที่สุด (Minimize $J_m(P)$) นั่นคือ ถ้าค่า $J_m(P)$ มีค่าน้อยก็หมายถึงความไม่คล้ายคลึงกัน (Dissimilarity)

การหาค่าความคล้ายของวูและพาเมอร์นั้นเป็นวิธีการหนึ่งในการหาค่าความคล้ายของเอกสารถือเป็นวิธีการประเภทการหาค่าความคล้ายแบบเอดจ์เคาทิงเมธอด (Edge Counting Methods) มีลักษณะการหาความคล้ายโดยอาศัยระยะห่างของโหนด (Path) ที่เชื่อมต่อกันแต่ละคำ และตำแหน่งในกลุ่มของคำนั้น ๆ ซึ่งจะนิยมนำมาใช้หาค่าความคล้ายของคำหรือเอกสารที่มีลักษณะคล้ายคลึงกันหรือมาจากฐานข้อมูลเดียวกัน (Single) การคำนวนหาค่าความคล้ายของวูและพาเมอร์มีลักษณะสำคัญคืออาศัยความสัมพันธ์แบบแนวดิ่ง (VRs) และความสัมพันธ์แบบแนวระนาบ (HRs) ของกลุ่มคำพ้อง ซึ่งระยะห่างของโหนดและความลึกระหว่างคำทั้งหมดจะถูกกำหนดไว้แล้วในฐานข้อมูลเวิร์ดเน็ต เมื่อเราเรียกใช้คำสั่งเพื่อหาค่าความคล้ายของคำค่าความสัมพันธ์ดังที่กล่าวมาข้างต้นจะถูกนำมาคิดคำนวนแล้วแสดงผลออกมาตามสมการของวูและพาเมอร์ ดังสมการ (3.1)

$$sim_{(x,y)} = Max\left[\frac{2*depth(LCS(x,y))}{length(x,y)+2*depth(LCS(x,y))}\right] \tag{4.1}$$

จากสมการเป็นการหาค่าความคล้ายด้วยวิธีการของวูและพาเมอร์ระหว่างค่าของกลุ่มชื่อพ้อง x และ y โดยค่า depth คือค่าความลึกของโหนดกลุ่มคำพ้องซึ่งมีการกำหนดค่าไว้แล้วในฐานข้อมูลเวิร์ดเน็ต The Lowest Common Subsumer) เป็นโหนดที่อยู่ต่ำที่สุดที่เป็นโหนดเชื่อมระหว่างสองโหนดที่ต้องการหาค่าความคล้ายซึ่งในที่นี้คือค่า x และ y อีกค่าหนึ่งที่สำคัญในสมการวูและพาเมอร์คือค่า length เป็นค่าระยะห่างระหว่างโหนดสองโหนดโดยจะนับเป็นจำนวนของโหนดที่อยู่ระหว่างโหนด x และโหนด y ซึ่งค่า length ถูกกำหนดไว้แล้ว ผลลัพธ์จากการหาค่าความคล้ายจะมีค่าอยู่ในช่วง $0 \geq simx,y \geq 1$ หากผลลัพธ์มีค่ามากแสดงว่ากลุ่มคำพ้องทั้งสองมีความคล้ายมากด้วยเช่นกัน

3. วิธีดำเนินการ (Methods)

การขยายคำค้นคืนเป็นการประมวลโดยอาศัยหลักการหาค่าน้ำหนักของคำในเทอม และนำเทอมที่ได้ไปทำการขยายซึ่งจะอาศัยฐานข้อมูลเวิร์ดเน็ต โดยฐานข้อมูลเวิร์ดเน็ตนั้นมีลักษณะดังที่กล่าวมาแล้ว ส่วนที่เราจะนำมาใช้ในการขยายเทอมของเรานั้นเป็นกลุ่มของคำที่มีความหมายคล้ายคลึงกัน เรียกว่า “synsets” หมายถึงกลุ่มคำที่มีความสัมพันธ์กันในเชิงความหมายของแต่ละคำ ที่มีการจัดเก็บในรูปของออนโทโลยีในพจนานุกรมเวิร์ดเน็ต ตัวอย่างการแสดงข้อมูล wn.synsets(“word”) โดยอาศัยเครื่องมือ NLTK จะได้ผลลัพธ์ synsets ทั้งหมดของ word ยกตัวอย่างเช่นคำว่า dog ดังภาพที่ 2

```
>>> wn.synsets("dog")
[Synset('dog.n.01'), Synset('frump.n.01'), Synset('dog.n.03'), Synset('cad.n.01'), Synset('frank.n.02'),
Synset('pawl.n.01'), Synset('andiron.n.01'), Synset('chase.v.01')]
>>>
```

ภาพที่ 2 ค่า synsets ของคำว่า Dog

จะเห็นได้ว่าคำว่า dog นั้นมี synsets อยู่หลายกลุ่มและหลายชนิดคำโดยจะแบ่งออก 3 ส่วนตามรูปแบบดังนี้คือ synset(‘word.pos.nn’) word คือกลุ่มคำที่เป็น synsets กับคำว่า “dog” pos คือชนิดของคำว่า “dog” ซึ่งมีทั้ง noun และ verb (nn เป็น part-of-speech หมายถึงคำนาม) คือ ตัวเลขที่บอก ลำดับที่ของกลุ่ม นอกจากนี้เรายังสามารถนำค่าต่างๆ ของ synsets ไปประยุกต์ใช้ได้หลายประเภททั้งการหาความสัมพันธ์ของคำประเภทต่างๆ เพื่อหาค่าความคล้ายของคำหรือเอกสาร การแสดงบริบทของคำ การแสดงประโยคตัวอย่างของการใช้คำซึ่งเป็นคุณสมบัติของ synsets ของคำศัพท์ที่จัดเก็บในเวิร์ดเน็ต

```
>>> dog.hypernyms()
[Synset('domestic_animal.n.01'), Synset('canine.n.02')]
>>> dog.hyponyms()
[Synset('puppy.n.01'), Synset('great_pyrenees.n.01'), Synset('basenji.n.01'), Synset('newfoundlan
d.n.01'), Synset('lapdog.n.01'), Synset('poodle.n.01'), Synset('leonberg.n.01'), Synset('toy_dog.
n.01'), Synset('spitz.n.01'), Synset('pooch.n.01'), Synset('cur.n.01'), Synset('mexican_hairless.
n.01'), Synset('hunting_dog.n.01'), Synset('working_dog.n.01'), Synset('dalmatian.n.02'), Synset(
'pug.n.01'), Synset('corgi.n.01'), Synset('griffon.n.02')]
>>> dog.member_holonyms()
[Synset('pack.n.06'), Synset('canis.n.01')]
>>> dog.member_meronyms()
[]
```

ภาพที่ 3 ค่าคำที่มีความสัมพันธ์กับคีย์เวิร์ดในรูปแบบต่างๆ

synsets นั้นก็เป็นกลุ่มของคำที่มีความหมายใกล้เคียงกันหรือเหมือนกันในรูปแบบต่างๆ จากรูปตัวอย่างที่ 3.2 เป็นการเขียนคำสั่งเพื่อให้โปรแกรมแสดงคำที่มีความสัมพันธ์แบบ hypernyms hyponyms holonyms และ meronyms กับคีย์เวิร์ด “dog” หากเราต้องการดูตัวอย่างประโยคของคีย์เวิร์ดและบริบทของ “dog” ก็สามารถทำได้ดังภาพที่ 3

```
>>> wn.synset('dog.n.01').definition
'a member of the genus Canis (probably descended from the common wolf) that has been
domesticated by man since prehistoric times; occurs in many breeds'
>>> wn.synset('dog.n.01').examples
['the dog barked all night']
>>>
```

ภาพที่ 4 นิยามศัพท์ของคำที่เป็นคีย์เวิร์ด

สำหรับในแต่ละ synsets ของคำทุกๆ คำจะสามารถนำมาหาความคล้ายกันของคำได้ โดยวิธีการหาค่าความคล้ายกันนั้นแบ่งออกเป็น 4 วิธีใหญ่ ๆ คือ

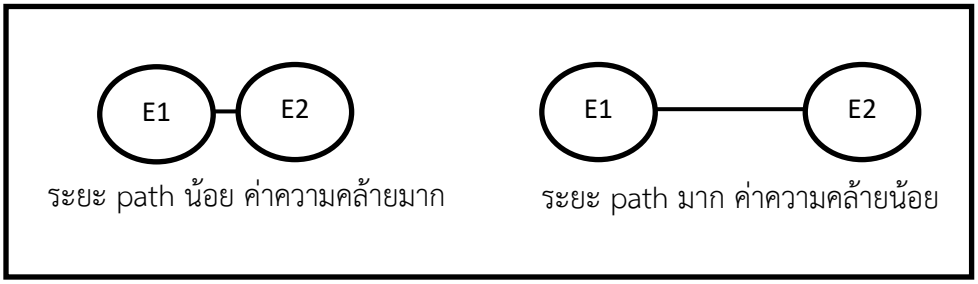
1. Edge Counting Methods วัดค่าความคล้ายกันของคำจากความยาวของ path ที่เชื่อมต่อแต่ละคำ จากคำหนึ่งไปยังอีกคำหนึ่ง
2. Information Content Methods การวัดค่าเนื้อหาของคำโดยใช้ความเป็นไปได้ที่จะเกิดในเอกสาร
3. Feature Based Method วัดค่าความคล้ายกันของคำสองคำจากคุณสมบัติของคำสองคำ
4. Hybrid Method เป็นการรวมวิธีการหาค่าความคล้ายกันของคำจากสามวิธีก่อนหน้านี้ทั้งหมดมารวมไว้ใช้ในวิธีเดียว

โดยทั่วไปแล้วการหาค่าแบบวิธีที่ (1) และวิธีที่ (2) นั้นจะนิยมใช้เปรียบเทียบจากฐานข้อมูลเดียวกัน ส่วนวิธีที่ (3) และวิธีที่ (4) จะใช้เปรียบเทียบจากฐานข้อมูลคนละฐาน ใน NLTK นี้จะใช้วิธีการหาค่าความคล้ายแบบวิธีที่ (1) และวิธีที่ (2) เท่านั้น และวิธีการหาค่าความคล้ายที่ผู้ศึกษาเลือกนำมาใช้ในโปรแกรมคือการหาค่าความคล้ายของ Wu-Palmer Similarity เพราะค่าที่ได้จากการหาค่านั้นจะถูกปรับค่าให้เหมาะสมมาแล้ว คือมีค่าตั้งแต่ 0-1 นอกจากนี้ยังง่ายต่อการใช้งาน มีเพียงแค่ synsets สองค่าก็สามารถนำมาหาค่าความคล้ายกันได้ซึ่งการหาค่าความคล้ายของ Wu-Palmer Similarity นั้น เป็นการหาค่าแบบ Edge Counting Methods ดังภาพที่ 5

```
>>> dog = wn.synset('dog.n.01')
>>> cat = wn.synset('cat.n.01')
>>> dog.wup_similarity(cat)
0.8571428571428571
>>> cat.wup_similarity(dog)
0.8571428571428571
>>>
```

ภาพที่5 การหาค่าความคล้ายของคำด้วย Wu-Palmer Similarity

จากภาพที่ 5 จะเห็นว่าอันดับแรกเราใช้วิธีการหา synsets ทั้งหมดของ “dog” และ “cat” ก่อนแล้วจึงเลือกว่าต้องการหาค่าความคล้ายกันของระหว่างกลุ่มไหนจากทั้งสองคำเมื่อเลือกได้แล้วก็นำมาหาค่าโดยการพิมพ์คำสั่ง dog.wup_similarity(cat) จะเป็นกลุ่มคำไหนไว้ที่หน้าคำสั่งก็ได้แล้วให้อีกคำอยู่ในวงเล็บท้ายคำสั่ง จะเห็นว่าถึงแม้ว่าเราจะวางคำสั่งสลับที่กันก็ตามค่าที่ได้ก็ไม่ต่างกัน จากตัวอย่างลองสลับระหว่าง synsets ‘dog.n.01’ กับ ‘cat.n.01’ ค่าที่ได้คือ 0.8571428571428571 เท่ากัน จะเห็นได้ว่าทั้งสอง synsets นี้มีค่าความคล้ายที่ได้ใกล้เคียง 1 มากเท่าใด นั่นหมายความว่าทั้งสอง synsets มีความคล้ายกันมากจากการหาค่าระยะ path ดังภาพที่ 6



ภาพที่ 6 อธิบายความสัมพันธ์ของค่าความคล้ายกับระยะห่างของ path

ในการทดลองการขยายคำค้นคืนโดยอาศัยเทคนิควิธีการจัดกลุ่มคำ โดยคลังคำที่นำมาใช้ในการทดลองนี้มาจากพจนานุกรม WordNet และเครื่องมือ NLTK สำหรับการประมวลผลภาษาธรรมชาติ และผลลัพธ์สุดท้ายจะนำเสนอข้อมูลผลการจัดกลุ่มในรูปของภาพแผนภูมิ ในการทดลองนี้ คำค้นคืนที่ผู้ใช้ระบุในคิวรีจะถูกนำมากำหนดเป็นกลุ่มคำเริ่มต้นที่จะนำมาขยายเป็นคำค้นคืนชุดใหม่ และนำมาค้นคืนใหม่อีกครั้ง ซึ่งคำที่จะขยายนั้นจะมาจากคำที่ปรากฏอยู่ในกลุ่ม Synset ที่จัดเก็บภายใน WordNet ซึ่งกลุ่มคำเหล่านี้ได้มีการจัดลำดับตามค่า Synset การขยายคำค้นคืนโดยอาศัยเทคนิคการจัดกลุ่มนี้คำเพื่อให้คำที่ต้องการขยายออกไปนั้นมีความหมายที่ใกล้เคียงกับคำค้นคืนเดิมที่ผู้ใช้ระบุ โดยการจัดกลุ่มจะอาศัยคำนิยามศัพท์ เป็นคุณสมบัติหลักจากคลังคำที่ปรากฏ ซึ่งในงานวิจัยนี้ได้เลือกใช้พจนานุกรมเวิร์ดเน็ต จากค่า Synset การจัดรูปแบบของคำนิยามศัพท์ให้อยู่ในรูปของเวกเตอร์ สำหรับการหากลุ่มคำที่มีคุณสมบัติเหมือนหรือคล้ายคลึงกันในการจัดกลุ่มคำค้นคืน สำหรับตัวอย่างคำนิยามศัพท์ ของคำว่า ant bat และ cat ที่เป็นคำนาม ต่อไปนี้

wn.synset(“ant.n.01”).definition() – ‘social insect living in organized colonies; characteristically the males and fertile queen have wings during breeding season; wingless sterile females are the workers’
wn.synset(“bat.n.01”).definition() – ‘nocturnal mouselike mammal with forelimbs modified to form membranous wings and anatomical adaptations for echolocation by which they navigate’
wn.synset(“cat.n.01”).definition() – ‘feline mammal usually having thick soft fur and no ability to roar: domestic cats; wildcats’

ภาพที่ 7 คำนิยามของคำศัพท์จาก WordNet

คำนิยามศัพท์จะถูกนำมาจัดให้อยู่ในรูปของเวกเตอร์และจะถูกประมวลผลด้วยเทคนิคการประมวลผลภาษาธรรมชาติ การเตรียมเอกสารก่อนการประมวลผลประกอบด้วย

1) การประมวลผลคำและการทำรากศัพท์ ประกอบด้วยขั้นตอนการขจัดอักขระพิเศษ ตัวอย่างเช่น punctuations = !()-[]{};:'"\,<>./?@#$%^&*_~ สำหรับเครื่องหมายอักขระพิเศษที่พบจะถูกกำจัดออกจากประโยคนิยามศัพท์เพื่อลดค่ารบกวนที่อาจเกิดขึ้น และการคัดกรองคำที่ไม่สื่อความหมายออกจากเอกสาร คำที่พบบ่อย (Stopword) เพื่อเป็นการลดความถี่ของคำหยุด และลดความคลาดเคลื่อนในการคำนวณ

2) การแปลงข้อความให้อยู่ในรูปของเวกเตอร์และการลดรูปคำ เป็นขั้นตอนในการจัดรูปแบบคำนิยามของคำให้อยู่ในรูปของเวกเตอร์ โดยขั้นตอนนี้จะมีการลดรูปของคำให้อยู่ของรากศัพท์ ยกตัวอย่าง เช่น การตัดคำต่อท้าย (suffix) ออก การตัด –s ออกจากคำพหูพจน์ เป็นต้น ซึ่งจะช่วยลดความหลากหลายของคำ เช่นการเปลี่ยนรูปคำตามเพศ และกาลเวลา โดยอาศัยเทคนิคการทำ Porter Stemming

3) การจัดกลุ่มเอกสาร เป็นการหาค่าความคล้ายคลึงของคำที่ปรากฏในแต่ละเวกเตอร์คำนิยามศัพท์เพื่อนำเอาคำมาจัดกลุ่มตามที่กำหนด โดยคำที่มีความคล้ายคลึงกันจากนิยามศัพท์จะถูกนำมาจัดให้อยู่ในกลุ่มเดียวกัน ซึ่งเราจะใช้คำที่พบในแต่ละกลุ่มคำศัพท์นี้เพื่อขยายคำค้นคืนจากคิวรีเริ่มต้น

```
Threshold = 0.5, Input = [bird]


All synsets of [bird] are  [Synset('bird.n.01'), Synset('bird.n.02'), Synset('dame.n.01'),
Synset('boo.n.01'), Synset('shuttlecock.n.01')]


Similarity between  Synset('bird.n.01')  and  Synset('bird.n.01')  is  1.0
Similarity between  Synset('bird.n.01')  and  Synset('bird.n.02')  is  0.235294117647
Similarity between  Synset('bird.n.01')  and  Synset('dame.n.01')  is  0.571428571429
Similarity between  Synset('bird.n.01')  and  Synset('boo.n.01')  is  0.117647058824
Similarity between  Synset('bird.n.01')  and  Synset('shuttlecock.n.01')  is  0.4

 Synsets have similarity more than threshold are :
       [Synset('bird.n.01'), Synset('dame.n.01')]


Lemma names of  Synset('bird.n.01')  =  ['bird']

Lemma names of  Synset('dame.n.01')  =  ['dame', 'doll', 'wench', 'skirt', 'chick', 'bird']

The result from expandsion are  ['bird', 'dame', 'doll', 'wench', 'skirt', 'chick', 'bird']
```

ภาพที่ 8 การจัดกลุ่มคำค้นคืนสำหรับคำว่า ‘dame’ และ ‘bird’

จากภาพที่ 8 แสดงการจัดกลุ่มเอกสารในตอนเริ่มต้นของการทดลองเรากำหนดค่าเทรชโฮลด์ ไว้ที่ 0.5 ซึ่งผลลัพธ์ที่ได้พบว่าจะมีคำที่ถูกจัดให้อยู่ในกลุ่มเดียวกันทั้งคำที่มีความหมายเหมือนกันหรืออาจจะต่างกันบ้าง แต่เมื่อกำหนดค่าที่สูงขึ้นจำนวนคำที่ปรากฏในกลุ่มจะมีจำนวนที่น้อยลงไปซึ่งจะเหลือเพียงคำที่มีความหมายคล้ายกันมากยิ่งขึ้น ตัวอย่างการขยายคำค้นคืนสำหรับคำว่า ‘dame’ ในรอบที่ 1 จะได้กลุ่มคำ

['dame', 'doll', 'wench', 'skirt', 'chick', 'bird']

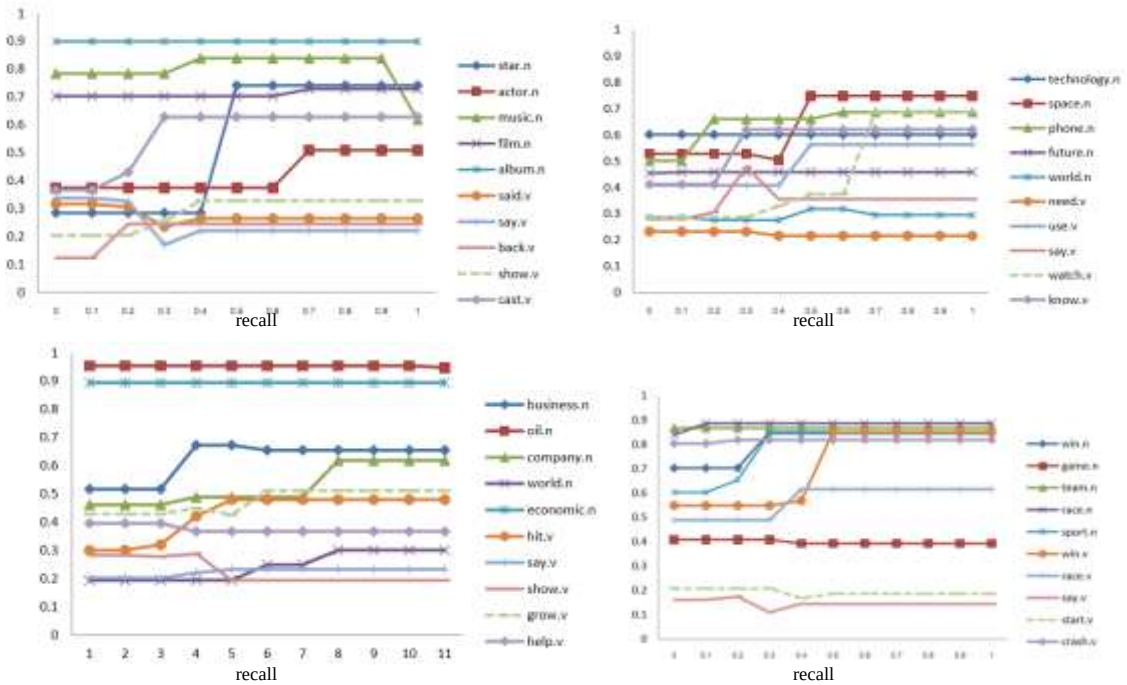
และในรอบที่สองของการขยายคำค้นคืนจากคำว่า ‘bird’ ก็จะได้กลุ่มคำต่อไปนี้

]'bird', 'dame', 'doll', 'wench', 'skirt', 'chick', 'bird']

ด้วยค่าเทรชโฮลด์ เท่ากับ 0.5 จากข้อมูลนำเข้า คำว่า bird คือ จากที่กล่าวมาเป็นเพียงการยกตัวอย่างการทำการขยายคำจากข้อมูลเพียงคำเดียวจริงๆ แล้วข้อมูลนำเข้าอาจจะมีค่ามากกว่าหนึ่งคำจากกลุ่มคำเดียวกัน จากกลุ่มคำหลายๆ กลุ่ม หรือแม้แต่กลุ่มคำเดียวกันสามารถมีชนิดของคำ (part-of-speech) ที่แตกต่างกันได้

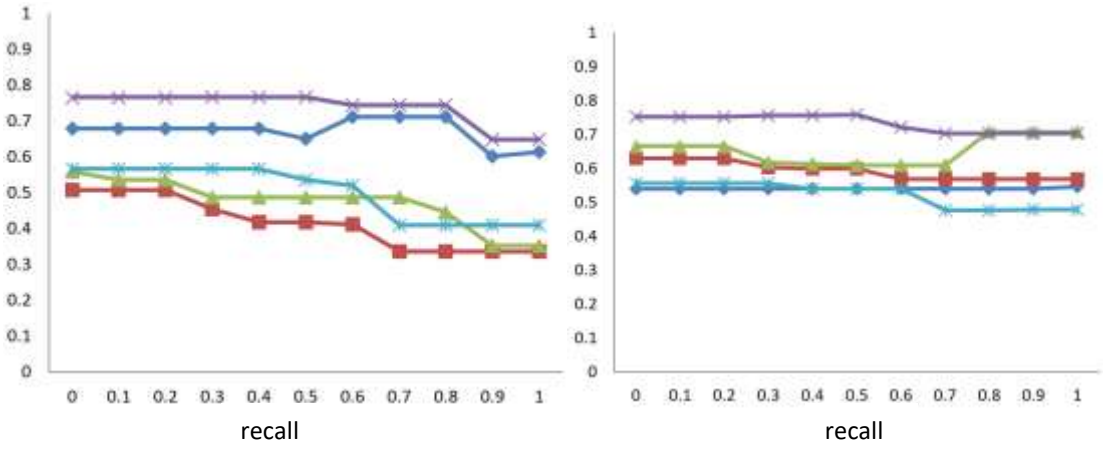
4. ผลการทดลอง (Experiment Results)

จากการทดลองการจัดกลุ่มคำจากนิยามศัพท์เพื่อนำมาใช้สำหรับการขยายคำค้นคืน โดยในงานวิจัยนี้ได้มีคัดเลือกคำจาก 4 โดเมนประกอบด้วย Entertainment, Technology, Business และ Sport โดยนำเอาคำที่พบในแต่ละโดเมนมาทำการหาค่านิยามศัพท์เพื่อนำมาจัดกลุ่ม การวัดผลและประเมินนั้นจะคิดจากค่า Precision ค่า Recall และ ค่า F-measure จากผลการทดลองในรูปที่ 4.1 ที่แสดงความสัมพันธ์ระหว่างค่าเทรชโฮลด์ และค่าเฉลี่ยความแม่นยำจากผลของการทดลองโดยใช้คำค้นสองกลุ่มคือกลุ่มคำนามและคำกริยา ใช้ค่าเทรชโฮลด์ 0.0 - 1.0 ซึ่งค่าเทรชโฮลด์ที่มากขึ้นให้ผลที่ต่างกันออกไปหลายลักษณะ เมื่อนำผลที่ได้มาทำการวิเคราะห์แล้วจะทำให้ทราบว่าค่าความแม่นยำของกลุ่มคำนามจะมีค่าสูงกว่าค่าความแม่นยำของกลุ่มคำกริยา โดยอ้างอิงจากค่าเฉลี่ยความแม่นยำ (Mean Average Precision)



ภาพที่ 9 ผลการประเมินค่าระลึกและค่าความแม่นยำ จาก 4 โดเมน

เนื่องจากกลุ่มคำนามนั้นจะมีความหลากหลายและเฉพาะเจาะจงกว่ากลุ่มคำกริยา ทำให้ผลการค้นหาเอกสารตรงตามคำตอบของการค้นหามากกว่าและค่าความแม่นยำจึงสูงกว่า อีกประการหนึ่งคือคำกริยาบางคำสื่อความหมายได้หลากหลายจึงถูกนำไปใช้ในหลายโดเมน ทำให้วัดค่าความแม่นยำได้ผลที่ไม่สูงมากนัก จากผลที่กล่าวมาข้างต้นนั้นตรงตามสมมุติฐานที่การค้นหาส่วนใหญ่จะใช้คำนามมากกว่าคำกริยา สำหรับกลุ่มคำที่มีค่าแม่นยำในระดับ 1 นั้นจะถือว่าเป็นคำที่มีความคล้ายคลึงกันที่สามารถนำไปใช้ในการขยายคำค้นคืนได้มากที่สุดและหลังจากนั้นจะใช้ค่าที่รองลงมา ซึ่งจากผลการขยายคำเมื่อเปรียบเทียบกับค่า Synset ที่แนะนำโดยเวิร์ดเน็ตพบว่าจะมีความแตกต่างกันบ้างแต่ก็ให้ผลไปในทิศทางเดียวกัน และลักษณะสำคัญที่สังเกตได้จากการทดลองอีกอย่างหนึ่งได้แก่การขยายคำโดยอาศัยค่าความคล้ายจากฐานข้อมูลเวิร์ดเน็ตนั้น บางคำนั้นจะให้ค่าความคล้ายของคำใกล้เคียงหรือเกือบจะเป็น 0 จากการคำนวณ แต่การหาค่าความคล้ายจากนิยามศัพท์นั้น จะยังมีค่าอยู่ทำให้ขอบเขตของการขยายคำกว้างกว่า ช่วยให้สามารถเพิ่มคำในกลุ่มคำที่เราทำการขยายได้มากขึ้นตามไปด้วย เมื่อได้ค่าเทรชโฮลด์จากการทดลองข้างต้นการทดลองต่อไปจะนำค่าที่ได้ไปวัดผลการค้นหาโดยการปรับค่าน้ำหนักตั้งแต่ 0.0 - 1.0 แล้วคำนวนหาค่า Average Precision และ Mean Average Precision ซึ่งจะวัดผลจากสองชุดข้อมูล โดยชุดแรกใช้เฉพาะกลุ่มคำชนิดคำนาม และชุดที่สองใช้กลุ่มคำค้นที่มีทั้งชนิดคำนามและคำกริยาเพื่อนำผลมาวิเคราะห์ความสัมพันธ์ระหว่างค่าน้ำหนักและค่าความแม่นยำ ในการทดลองทำให้เราทราบว่าค่าเฉลี่ยของค่าเฉลี่ยความแม่นยำของการทดลองครั้งที่สองมีค่าสูงขึ้นและได้ค่าน้ำหนัก = 0.2 ที่ค่าเฉลี่ยของค่าเฉลี่ยความแม่นยำสูงที่สุด



ภาพที่ 10 ค่าเฉลี่ยของค่าเฉลี่ยความแม่นยำจากกลุ่มคำนามและคำกริยา

5. สรุปผล และอภิปรายผล (conclusion)

การวัดค่าความคล้ายของคำโดยวิธีการจัดกลุ่มคำนั้นอาศัยความถี่ของการเกิดของคำเป็นหลัก ดังนั้นค่าความคล้ายระหว่างกลุ่มคำค้นคืนจะมีค่าน้อยเนื่องจากคำนิยามศัพท์ที่ได้จากเวิร์ดเน็ตนั้น คำนิยามศัพท์บางคำมีจำนวนคำที่แตกต่างกัน อีกทั้งการวัดความคล้ายโดยอาศัยความถี่นั้นพบว่ามีบางคำที่ซ้ำกันหรือไม่ก็ตามจะทำให้ค่าความคล้ายในเอกสารสูงขึ้นตามไปด้วย ดังนั้นบางเอกสารที่มีคำในกลุ่มคำค้นน้อยกว่าอาจจะมีค่าความคล้ายสูงกว่าเอกสารที่มีคำในกลุ่มคำค้นมากกว่าได้ และการวัดค่าความคล้ายของเอกสารจะอาศัยความถี่ของคำไม่ได้อ้างอิงเชิงความหมาย ดังนั้นเอกสารที่มีโครงสร้างของคำที่เหมือนกันมากจะให้ค่าความคล้ายกันของเอกสารที่มากตามไปด้วย

สำหรับค่าตัวแปรที่นำมาใช้ในโปรแกรมเพื่อใช้ในการขยายคำค้นคืนโดยวิธีการจัดกลุ่มคำจะไม่ตายตัวขึ้นอยู่กับประเภทและลักษณะของคำ บางครั้งอาจจะต้องมีการปรับค่าตัวแปรนั้นๆ ทุกครั้งเมื่อใช้โปรแกรมกับโดเมนของข้อมูลที่ต่างชนิดกัน ประสิทธิภาพของโปรแกรมด้านความเร็ว จะแปรผกผันกับจำนวนของคำทั้งหมดในเอกสารเพราะหากจำนวนคำที่เพิ่มมากขึ้นโปรแกรมจะทำให้ต้องอ่านข้อมูล แปลงข้อมูล นับค่าน้ำหนัก และคำนวนหาผลลัพธ์ต่างๆ มากขึ้นตามไปด้วยและอาจจะพบข้อจำกัดในเรื่องของโปรแกรมจะประมวลคำค้นโดยไม่พึ่งบริบทของคำค้น คำที่มีลักษณะพิเศษเชิงความหมายจะถูกประมวลผลในลักษณะเหมือนคำค้นทั่วไปโดยไม่คำนึงถึงการนำเอาบริบทมาใช้ในการค้นคืน ทำให้สรุปได้ว่าค่าน้ำหนักที่เหมาะสมนั้นมีผลต่อความแม่นยำของการค้นหา และจากการทดลองครั้งที่สองทำให้ได้ค่าน้ำหนักที่ค่าเฉลี่ยความแม่นยำสูงสุดคือ 0.2 ค่าน้ำหนักที่มากเกินไปจะทำให้กลุ่มคำค้นกระจายกันหลายกลุ่มกว่าที่ควรจะเป็น คำบางคำที่จัดกลุ่มรวมกันได้ถูกแยกออกจากกันทำให้ลำดับในการค้นหาถูกแทรกด้วยผลการค้นหาที่ไม่ถูกต้อง ค่าน้ำหนักที่น้อยจนเกินไปทำให้คำค้นทุกคำถูกรวมเป็นกลุ่มเดียวกัน คำบางคำสื่อความหมายได้หลายลักษณะและเมื่อรวมกันทำให้ผลการค้นหาต้องอาศัยเพียงค่าความคล้ายจากความถี่ของเอกสารเท่านั้น และการหาค่าน้ำหนักที่ดีที่สุดจะช่วยเพิ่มประสิทธิภาพในการค้นหาให้แม่นยำมากขึ้น

6. กิตติกรรมประกาศ หรือคำขอบคุณ (acknowledgement)

ขอขอบคุณคณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ที่ให้การสนับสนุนงบประมาณวิจัยจากงบประมาณเงินรายได้ ประจำปี 2558 ประเภทโครงการวิจัยพื้นฐานสาขาวิทยาศาสตร์ ประเภทนักวิจัยรุ่นกลาง

7. เอกสารอ้างอิง (references)

ชุลีรัตน์ จรัสกุลชัย, เจษฎา กันทะเสนา, สถาพร คิ้วสุวรรณสุข. (2556). **การจัดกลุ่มเอกสารสำหรับข้อความภาษาไทย.** รายงานวิจัย, ห้องปฏิบัติการงานวิจัยสารสนเทศอัจฉริยะและฐานข้อมูล ภาควิชาวิทยาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์มหาวิทยาลัยเกษตรศาสตร์.

นิเวศ จิระวิชิตชัย. (2556). **แบบจำลองการจำแนกเอกสารภาษาไทยอัตโนมัติ.** วารสารวิชาการเทคโนโลยีอุตสาหกรรม ปี ที่ 9 ฉบับที่ 1 มกราคม – เมษายน 2556.

ศักดิ์ชัย ศรีมากรณ์. (2551). **การแบ่งกลุ่มช็อตภาพเคลื่อนไหวโดยใช้วิธีการแบ่งจำนวนกลุ่มโดยไม่ต้องรู้จำนวนกลุ่ม.** วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่.

นิเวศ จิระวิชิตชัย, ปริญญา สงวนสัตย์ และพยุง มีสัจ. (2553). **การจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน : Automatic Thai Document Categorization with Support Vector Machines.** The 6TH National Conference on Computing and Information Technology, NCCIT2010-105, 2553.

จิราภรณ์ ถมแก้ว และศรัณย์ อินทโกสุม. (2555). **การจำแนกข้อมูลโดยการคัดเลือกคุณลักษณะที่สำคัญ.** สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพมหานคร, การประชุมวิชาการเสนอผลงานวิจัยระดับบัณฑิตศึกษาแห่งชาติครั้งที่ 23.

วงกต ศรีอุไร, พยุง มีสัจ และชูชาติ หฤไชยะศักดิ์. (2552). **การเตรียมฟีเจอร์บนพื้นฐานแบบจำลองหัวข้อสำหรับการจำแนกหมวดหมู่ของเอกสาร.** The 5th National Conference on Computing and Information Technology, NCCIT 2552.