การคัดกรองเอกสารที่สืบค้นโดยการแปลงน้ำหนัก-ระยะห่าง

Filtering Search Document using Weight-Distance Transformation

นลินี โสพัศสถิตย์[1] และพีระพนธ์ โสพัศสถิตย์[2]

[1]ภาควิชาวิทยาศาสตร์ประยุกต์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏสวนสุนันทา nalinee.so@ssru.ac.th

[2]ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย peraphon.s@chula.ac.th

บทคัดย่อ

การสืบค้นสารสนเทศจากอินเทอร์เน็ตผ่านระบบสืบค้นมักจะได้ผลลัพธ์เป็นแฟ้มเอกสารจำนวนมาก ซึ่งส่วนใหญ่จะไม่ค่อยตรงกับความต้องการ ปัญหาดังกล่าวเกิดจากการเรียงลำดับคำสำคัญของการสืบค้นให้ถูกต้อง บทความนี้เสนอวิธีคัดกรองแปลงจากเทคนิคต่างๆ ที่ใช้กัน โดยเริ่มจาก ontology tree เป็นพื้นฐานในการเทียบคำสำคัญ เพื่อรวบรวมผลลัพธ์เบื้องต้นที่สืบค้นได้ไปจำแนกเชิงความน่าจะเป็นและจัดกลุ่มตามโดเมนที่เหมาะสม หลังจากนั้นทำการแปลงโดยโยงค่าผลลัพธ์บนปริภูมิเอกสาร ซึ่งอ้างอิงเอกสารอุดมคติที่สร้างจากการใช้น้ำหนัก-ระยะห่างเป็นเกณฑ์ กระบวนการนี้จะดำเนินการซ้ำหลายครั้งจนถึงเพดานที่กำหนด ถือได้ว่าเป็นเทคนิคการกรองเอกสารที่ตรงไปตรงมา ประโยชน์ที่ได้รับโดยตรงคือ จำนวนเอกสารผลลัพธ์ที่ได้ลดลง ผลพลอยได้ที่เป็นประโยชน์ทางอ้อมซึ่งควรดำเนินการเป็นงานวิจัยต่อเนื่องอย่างจริงจังคือ พลังงานที่ใช้ในการประมวลผลและสื่อสารข้อมูลลดลง เหตุผลประการหลังนี้จำเป็นและสอดคล้องกับวัตถุประสงค์การประหยัดพลังงานในปัจจุบัน ความสำเร็จจึงเป็นเพียงปัญหาการถ่ายทอดเทคโนโลยีไปสู่เชิงพาณิชย์อย่างจริงจัง

Abstract

Searching information from the Internet via available search engines is often overwhelmed by myriad of resulting documents that are mostly irrelevant. The problem lies in the use of proper keywords arranged in the right order. This paper proposes an effective filtering approach that exploits various existing techniques through a sequence of transformations. The proposed approach employs ontology tree as a basis for keyword matching, thereby intermediate search results can be stochastically classified and clustered into their respective domains. Subsequent transformations are applied to map the results on to a document space. A user-defined ideal document is utilized to establish measured weights obtained from the Euclidean distance between a given document and the ideal document. This process is repeated until a predetermined threshold is reached, giving rise to a straightforward document filtering technique. The direct benefit from the proposed approach is low number of search documents. This, in turn, entails continual research on reducing energy consumption as the volume of information to be processed and transmitted reduces. This is essential and conforming to today's energy saving directives. It's just a matter of technology transfer to realize the proposed approach in commercial applications.

Keywords: ontology matching tree, document space, document filtering, PLSA domain, search engine.

1. บทนำ

การค้นหาข้อมูลบนอินเทอร์เน็ตมักจะได้ผลลัพธ์ในรูปเอกสารที่มีปริมาณมากเกินความต้องการ แม้ว่าจะมีเครื่องมือที่ช่วยกรองเอกสารที่ไม่ต้องการ แต่ผลลัพธ์ก็ยังมีจำนวนเอกสารมากอยู่ดี อีกทั้งเอกสารผลลัพธ์ที่ได้ก็อาจจะไม่ตรงกับความต้องการทั้งหมด search engine ที่ใช้งานในปัจจุบัน มีเครื่องมือที่พัฒนาตามระเบียบวิธีต่างๆ ช่วยในการสืบค้นและกรองข้อมูล Search Engine Optimization (SEO) เป็นตัวอย่างที่นิยมใช้กันมาก แต่เนื่องจากมีการปรับแต่งเนื้อหาของข้อมูลที่นำเสนอในหน้าเว็ปที่ทำให้ถูกค้นเจอโดย search engine เพื่อยกลำดับความสำคัญของการสืบค้นให้สูงขึ้น ซึ่งเป็นเทคนิคที่ใช้เพื่อวัตถุประสงค์เชิงพาณิชย์โดยตรง จึงไม่ขอกล่าวถึง SEO ในที่นี้ งานวิจัยนี้ เสนอแนวคิดหนึ่งในการกรองข้อมูลตามลำดับความเกี่ยวข้องด้วยระเบียบวิธีที่มีประสิทธิภาพ ทำให้ปริมาณเอกสารผลลัพธ์ลดลง ไม่เสียเวลาในการเสาะหาเอกสารที่ไม่เกี่ยวข้องจำนวนมาก แนวคิดที่จะนำเสนอ เป็นวิธีจัดกลุ่มของคำสำคัญตามลำดับที่จะพบเอกสาร แล้วคัดเลือกเอกสารที่สืบค้นได้ที่ใกล้เคียงกับเอกสารอุดมคติ (ideal document) หรือเอกสารที่ต้องการ

งานวิจัยนี้ไม่ครอบคลุมการคัดกรองเอกสารด้วยการหาความสัมพันธ์ระหว่างความหมาย (semantic) ของคำ [14] ที่ต้องวิเคราะห์ความหมายของเนื้อหาจากเอกสาร โดยสร้างบรรณนิทัศน์ (annotation)

และภววิทยา (ontology) ของคำเหล่านั้น ก่อนจะสรุปผลความเหมือน/คล้ายของเอกสาร งานวิจัยแนวนี้อยู่ในขั้นตอนการสร้างต้นแบบเพื่อหารูปแบบที่เหมาะสม ยังขาดความสมบูรณ์ถึงขั้นที่จะพัฒนาเป็นซอฟต์แวร์เชิงพาณิชย์

บทความนี้นำเสนอสาระงานวิจัยในลำดับต่อไปนี้ หัวข้อที่ 2 จะกล่าวถึงงานวิจัยที่เกี่ยวข้อง หัวข้อที่ 3 เสนอรายละเอียดของระเบียบวิธีที่นำเสนอ โดยมีผลการทดลองสนับสนุนงานวิจัยในหัวข้อที่ 4 หัวข้อสุดท้ายสรุปผลงานวิจัยและแนวทางงานวิจัยที่จะทำต่อในอนาคต

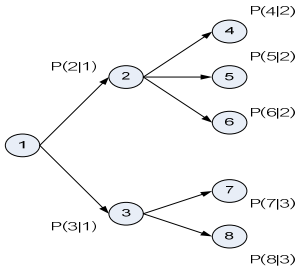
2. งานวิจัยที่เกี่ยวข้อง

Senvar และ Bener [5] กล่าวถึงการใช้ความหมายเพื่อเทียบบริการเว็ปที่ต้องการจากบริการที่มีอยู่ โดยเน้นการกำหนดน้ำหนักของความหมายที่แปลงเป็นระยะห่างตามความผิดเพี้ยนในรูปของต้นไม้ความหมาย เทคนิคดังกล่าวนำมาประยุกต์ในการเรียงลำดับของกลุ่มคำสำคัญ Hofmann [2] นำหลักการของ Probabilistic Latent Semantic Analysis (PLSA) มาวิเคราะห์ความน่าจะเป็นในการเทียบความเหมือนระหว่างสืบค้น Ma, Zhang และ He [3] เสนอวิธีแบ่งกลุ่มตามความหมายเพื่อให้การเทียบมีความถูกต้องมากที่สุดเท่าที่จะมากได้ Balinski และ Danilowicz [1] เสนอวิธีจัดลำดับใหม่โดยใช้ระยะห่างที่แทนความสัมพันธ์ระหว่างเอกสารใกล้เคียงกับเอกสารอุดมคติ โดยแปลงเป็นน้ำหนักเพื่อสะดวกในการเปรียบเทียบความสำคัญ งานวิจัยเกี่ยวกับการคัดกรองเพื่อจัดกลุ่มเอกสารอื่นๆ ยังมีอีกมากมาย เช่น partitioning, hierarchical/flat, non-hierarchical clustering, K-means, Buckshot [13], Fractionation โดยใช้ single word, sentence, snippet เป็นต้น งานวิจัยที่เกี่ยวข้องเหล่านี้สามารถนำมาพัฒนาเป็นระเบียบวิธีการสืบค้นและกรองเอกสาร เพื่อให้ได้เอกสารผลลัพธ์ที่ใกล้กับความต้องการมากที่สุด

3. ระเบียบวิธีที่นำเสนอ

การสืบค้นข้อมูลที่มักจะได้คำตอบเป็นเอกสารในรูปแบบใดรูปแบบหนึ่ง (ที่ถูกควรจะเรียกว่าแฟ้มเอกสารมากกว่า) เช่น เอกสารที่เป็นอักษรล้วน อักษรที่มีรูปภาพประกอบ รูปภาพ ภาพเคลื่อนไหว ฯลฯ การสืบค้นจะเริ่มจากการป้อนคำสำคัญเข้าสู่ระบบสืบค้นหรือ search engine ที่มีอยู่ทั่วไป เมื่อได้ผลลัพธ์ที่เป็นกลุ่มของเอกสาร อาจจะกรองซ้ำอีกครั้งโดยใช้คำสำคัญเพิ่ม เพื่อตัดเอกสารที่ไม่เกี่ยวข้องหรือมีความสำคัญน้อยทิ้งไป ในการสืบค้นข้อมูลเพื่อให้ได้ผลลัพธ์ (เอกสาร) ที่ต้องการนั้น มักจะเปรียบเทียบใน 4 ลักษณะ [5] คือ *exact match, plug-in match, subsume match* และ *fail match* แต่ในงานวิจัยนี้จะพิจารณา partial match แทน fail match เพราะกรณีหลังไม่เกิดประโยชน์ในการทำงาน สำหรับกรณี partial match นั้นยังไม่มีการวิจัยอย่างจริงจัง เนื่องจากผลลัพธ์มักไม่ค่อยมีนัยสำคัญ จึงไม่ได้รับความสนใจจากผู้ใช้ แต่ในความเป็นจริง partial match บ่งบอกถึงหลักการสำคัญประการหนึ่งที่ถูกมองข้าม กล่าวคือ partial match เป็นผลมาจากเซ็ทย่อยของผลลัพธ์ที่มีความถูกต้องเพียงบางส่วนของ input keywords แต่ไม่ถูกต้องทั้งหมด ในภาพรวม หากพิจารณาในมุมมองของการทดสอบการทำงานของ search engine กรณีดังกล่าวเป็นกรณีทดสอบที่ traverse ผ่านเส้นทางทดสอบ (test path) ที่ใช้งานน้อย ซึ่งมักเป็นกรณีที่ไม่ค่อยจะเกิดระหว่างการทำงานของระเบียบวิธีที่ search engine ใช้ พูดง่ายๆ คือมีความน่าจะเป็นที่เส้นทางทดสอบนั้นๆ จะถูกเลือกต่ำ อย่างไรก็ตาม ความครอบคลุมของการทดสอบ (test coverage) จำเป็นต้องรวมกรณีนี้เข้าเป็นส่วนหนึ่งของการวิจัย ดังนั้น การเปรียบเทียบจึงพยายามยึดแนวทาง ontology matching tree [5] ตามลำดับของการสืบค้น พร้อมกับคำนวณความน่าจะเป็นในแต่ละขั้นตอนตามแผนภาพในรูปที่ 1

จากรูปที่ 1 การสืบค้นเริ่มจากบัพ 1 แยกเป็น 2 ทางคือ ตามเส้นทาง 1→2 หรือ 1→3 ผลลัพธ์ที่ได้จะเทียบได้เป็น 1 ใน 4 ลักษณะที่กล่าวข้างต้น เช่น ถ้าเป็น exact match ที่บัพ 2 ความน่าจะเป็น P(2|1) = 1 และ P(3|1) = 0 ทำให้ subtree ใต้บัพ 2 และ 3 หายไปทั้งหมด หากเป็น plug-in, subsume หรือ partial match ที่บัพ 2 และ 3 อาจจะกรองผลลัพธ์ต่อตามเส้นทาง 2→4, 2→5, 2→6 และ 3→7, 3→8 ตามลำดับความน่าจะเป็นของเส้นทางการเลือกแต่ละเส้นสามารถคำนวณได้ตามรูป



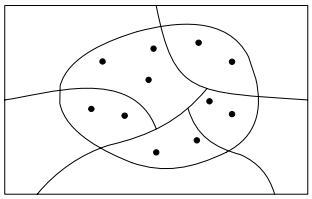
รูปที่ 1 ontology matching tree

จากรูปจะได้ว่าผลลัพธ์ที่บัพใบ P(4|1) = P(2|1) P(4|2) ซึ่งผลรวมของ P(4|1), P(5|1) และ P(6|1) มีค่าเป็น $\sum_{i=4}^{6} P(2|1)P(i|2)$ หรือ P(2|1) $\sum_{i=4}^{6} P(i|2)$ ดังนั้นความน่าจะเป็นรวมที่บัพ 1 สามารถคำนวณได้ตามสมการ

$$P(1) = \sum_{k=2}^{3} P(k|1) = P(2|1) \sum_{i=4}^{6} P(i|2) + P(3|1) \sum_{j=7}^{8} P(j|3) \quad (1)$$

ถ้านำเอกสาร (ผลลัพธ์) ของแต่ละบัพใบข้างต้นไปแทนเป็นจุดในโดเมนความน่าจะเป็นหรือ Probabilistic Latent Semantic Analysis (PLSA) domain ของเอกสารตามวิธีของ [2] ก็สามารถแสดงผลลัพธ์ตามรูปที่ 2 จะเห็นว่าจุดทั้งหมดจะรวมตัวกันภายในกลุ่ม (cluster) ของตนเอง ขึ้นอยู่

กับการพบข้อมูลตามตัวแบบแง่มุม (aspect model) [6] งานวิจัยนี้จะลดทอนความซับซ้อนของการแยกประเภทแง่มุม โดยใช้สัดส่วนของผลลัพธ์จากการสืบค้นด้วย search engine ในการทดลอง เนื่องจากอุปสรรคสำคัญๆ หลายประการ เช่น ปริมาณข้อมูล การแยกประเภท/จัดเก็บ ลิขสิทธิ์ข้อมูล และระบบสืบค้นสมรรถนะสูงที่สามารถประมวลผลข้อมูลปริมาณมากได้อย่างรวดเร็ว เป็นต้น



รูปที่ 2 PLSA domain

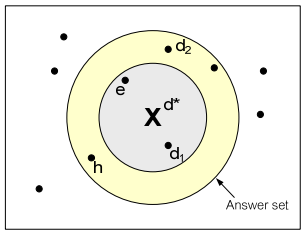
ขั้นตอนถัดไปเป็นการแปลงจุดทั้งหมดในรูปที่ 2 ไปสู่ปริภูมิใหม่ เรียกว่าปริภูมิเอกสาร (document space) โดยมีเอกสารอุดมคติ (ideal document) ซึ่งเป็นเอกสารที่ต้องการมากที่สุดเป็นศูนย์กลาง แล้วกำหนดเนื้อหาสาระของความแตกต่างระหว่างเอกสารที่สืบค้นได้กับเอกสารอุดมคติเป็นน้ำหนักซึ่งแทนด้วยระยะห่างในรูปของ Euclidean distance [1] ตาม triangle inequality ดังนี้

$$| \delta(d_1, d^*) - \delta(e, d_1) | \leq \delta(e, d^*) \leq | \delta(d_1, d^*) + \delta(e, d_1) \qquad (2)$$

สมการ (2) นี้แสดงว่า เอกสารคู่ใดๆ ที่มีน้ำหนักไม่เท่ากัน เอกสารที่มีน้ำหนักมากกว่าจะอยู่ใกล้เอกสารอุดมคติมากกว่า กล่าวอีกนัยหนึ่งคือ มีความคล้ายเอกสารอุดมคติมากกว่า เราสามารถคำนวณน้ำหนักระหว่างคู่เอกสารในระหว่างการแบ่งกลุ่ม โดยประยุกต์จากสมการความคล้าย [3] ที่ใช้เปรียบเทียบ query และ service ในลักษณะเดียวกัน

$$Sim(X, d_1) = | X \cdot d_1 | / (\|X\|^2 \cdot \|d_1\|^2) \qquad (3)$$

ซึ่งทำให้ได้กลุ่มตั้งต้นของเอกสาร เมื่อกำหนด threshold ของระยะห่าง (หรือน้ำหนัก) เพื่อกรองเอกสารที่ไม่ต้องการทิ้งไป ผลที่ได้จะเป็นเซ็ตของคำตอบ (answer set) ในปริภูมิเอกสารที่มีจำนวนเอกสารลดลงเป็นลำดับตามรูปที่ 3



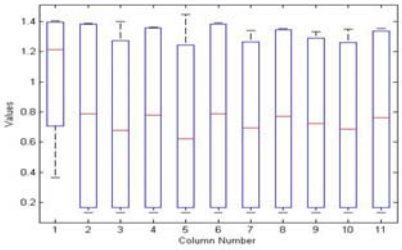
รูปที่ 3 document space

กล่าวโดยสรุป วิธีที่นำเสนอนี้เป็นระเบียบวิธีที่แทนการวัดเชิงคุณภาพ (qualitative measure) ของเอกสารที่ยากต่อการตัดสินใจด้วยการวัดเชิงปริมาณ (quantitative measure) โดยกำหนดเป็นน้ำหนักหรือระยะห่างระหว่างจุด ทำให้เราสามารถบอกได้ว่าเอกสารบางกลุ่มคล้ายกัน เพราะมีน้ำหนักใกล้เคียงกัน หรือระยะใกล้กัน แล้วแบ่งกลุ่มของเอกสาร (document cluster) ตามน้ำหนักที่คำนวณได้ ซึ่งวิธีแบ่งกลุ่มอาจจะต้องจำแนกตามประเภทและชนิดของเอกสารในเบื้องต้น โดยใช้ขั้นตอนวิธี suffix tree clustering ในการแบ่งกลุ่มเอกสาร [11, 12] (เป็นงานวิจัยอีกแนวหนึ่งที่มีบทบาทในการจำแนกเอกสารมาก) สำหรับงานวิจัยนี้ แบ่งกลุ่มด้วยน้ำหนักซึ่งแทนด้วยระยะห่างตามที่กำหนดข้างต้น แล้ววัดผลการทดลองด้วยตัวชี้วัดมาตรฐานของเอกสาร อันได้แก่ entropy, purity, accuracy, recall และ precision รายละเอียดจะกล่าวในส่วนการทดลองต่อไป

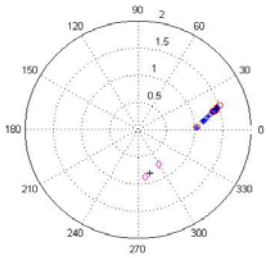
คำถามสำคัญที่ตามมาคือ ทำอย่างไรจึงจะได้เอกสารที่ใกล้เคียงกับเอกสารอุดมคติมากที่สุด พูดง่ายๆ คือมีเนื้อหาสาระใกล้เคียงกับความต้องการของผู้ใช้มากที่สุด นั่นหมายถึงการเลื่อนจุด d_1 หรือ e ให้ใกล้จุด X มากที่สุด โดยประยุกต์ระเบียบวิธี document weight improving (DWI) [1] แล้ววัดระยะห่างหลังการปรับจุด (คำตอบ) แต่ในการทดลองนี้ไม่ได้ใช้วิธี DWI เพราะข้อมูลที่ใช้ได้มาจากอินเทอร์เน็ตด้วยเหตุผล/อุปสรรคที่กล่าวข้างต้น

4. การทดลอง

งานวิจัยนี้ใช้คำสำคัญ 2 ชุดๆ ละ 4 กลุ่ม โดยชุดแรกใช้คำสำคัญสามัญ 3 คำที่รวมเป็นวลีภายใต้บริบทที่เกี่ยวข้องกันคือ computer software process (ศัพท์คอมพิวเตอร์ทั่วไป), travel city temple (ศัพท์ใช้ประจำวัน), music string instrument (เครื่องดนตรีประเภทสาย), religion eastern peace (ศัพท์ทางศาสนา) ส่วนชุดที่สองใช้คำสำคัญเฉพาะสาขา 2 คำที่รวมเป็นวลีในลักษณะเดิมคือ algorithm complexity, design pattern, process interrupt, intrusion cryptography แล้วสลับตำแหน่งของคำสำคัญในทุกๆ permutation เพื่อคำนวณค่าความน่าจะเป็นของแต่ละเงื่อนไขก่อนและหลังเทียบกับความต้องการเอกสารอุดมคติที่สืบค้นจากคำสำคัญทั้ง 4 คำ รูปที่ 4 แสดงผลการทดลองแสดงการกระจายของเอกสารที่สืบค้นได้ รูปที่ 5 แสดงผล plot ในปริภูมิเอกสารโดยมีจุดศูนย์กลางของปริภูมิเอกสารเป็นเอกสารอุดมคติ ส่วนค่าอื่นๆ plot เป็นระยะห่างจากจุดศูนย์กลาง



รูปที่ 4 การกระจายของเอกสารที่สืบค้นได้



รูปที่ 5 ผลการทดลองโดยใช้คำสำคัญทั้งหมด

การประเมินผลการดำเนินงานของระเบียบวิธีที่นำเสนอใช้ตัวชี้วัดต่างๆ [3. 4] ดังนี้ (ในที่นี้กลุ่ม (cluster) ใช้ในความหมายเดียวกับชนิด (category))

- Entropy เป็นการวัดความคงเส้นคงวาของกลุ่ม (cluster consistence) เพื่อเปรียบเทียบความเหมือนของสมาชิกในกลุ่ม

$$E(c_j) = -\sum_{i=1}^{m} \frac{n^i{}_j}{n_j} \bullet \log\left(\frac{n^i{}_j}{n_j}\right)$$

โดย n_j แทนจำนวนเอกสารในกลุ่ม c_j และ n^i_j คือจำนวนเอกสารที่ต้องการ (หรือมีคำสำคัญปรากฏอยู่ในเอกสารนั้น) ในกลุ่ม c_j ของเอกสาร i ส่วน m แทนจำนวนคำสำคัญและ k เป็นกลุ่มทดสอบ

- Purity พิจารณาความถูกต้องของการแบ่งกลุ่ม (classification accuracy) แต่ละ i เป็นสมาชิกของกลุ่ม c_j

$$P(c_j) = \frac{1}{n_j}\sum_{j=1}^{k} \max_i\{n^i{}_j\}$$

- Accuracy วัดความถูกต้องของการจัดกลุ่มเอกสาร

Accuracy = # of identified document / # of document in category

- Recall พิจารณาการเรียกใช้ที่ตรงประเด็น

Recall = relevant document retrieved / relevant document in collection

- Precision วัดสัดส่วนความถูกต้องของเอกสารจากที่มีอยู่ทั้งหมด

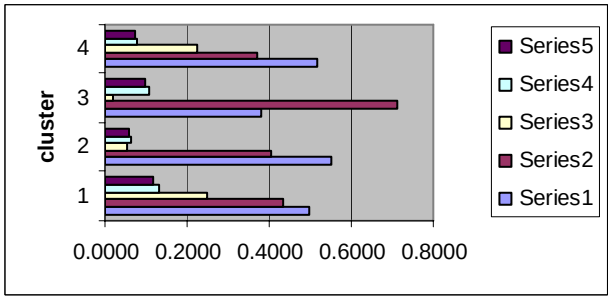
Precision = relevant document retrieved / total document retrieved

จากการทดลองสรุปเป็นตารางดังนี้

cluster	entropy	purity	accuracy	recall	precision
1	0.4997	0.4346	0.2499	0.1337	0.1146
2	0.5534	0.4073	0.0554	0.0637	0.0597
3	0.3786	0.7107	0.0201	0.1056	0.0955
4	0.5172	0.3709	0.2238	0.0786	0.0728

สังเกตุว่าบริบทของความแตกต่างระหว่างคำสำคัญในแต่ละกลุ่มทำให้ค่า entropy และ purity ต่ำ จึงมีระยะห่างจากเอกสารอุดมคติมาก สาเหตุอาจเนื่องมาจากความซับซ้อนของคำที่มีผลข้างเคียงกับบริบท เมื่อมีการสลับตำแหน่งความหมายจึงเปลี่ยนไป ทำให้ค่าอื่นๆ ที่คำนวณได้พลอยต่ำไปด้วย เห็นได้จากกลุ่มที่สามซึ่งคำว่า string มีความหมายอื่นที่แตกต่างจากอีกสองคำที่เหลือ กล่าวคืออาจหมายถึงสายอักขระในบริบทของคอมพิวเตอร์ ไม่จำเป็นต้องหมายถึงเครื่องดนตรีประเภทสายก็ได้ แต่เมื่อรวมกับคำสำคัญที่เหลืออีกสองคำแล้วกลับสอดคล้องตามบริบทเดิม ทำให้การสืบค้นคำว่า string เพียงคำเดียว ให้ผลแตกต่างจากการสืบค้นพร้อมกันทั้งวลี

จะเห็นว่า performance ของระเบียบวิธีที่เสนอก็อยู่ในระดับต่ำเช่นกัน ทั้งนี้เป็นเพราะความสัมพันธ์ของบริบทระหว่างคำสำคัญมีผลโดยตรงต่อเอกสารที่สืบค้น ถือเป็นความซับซ้อนตามกรณี (accidental complexity) ที่มีผลต่อระเบียบวิธีที่เสนอ เป็นการยืนยันผลข้างเคียงของความซับซ้อนระหว่างคำ เป็นผลให้ความถูกต้องของการจัดกลุ่มเอกสารตามระเบียบวิธีต่ำ ซึ่งสอดคล้องกับผลการทดลองในรูปที่ 5 เมื่อเปรียบเทียบเป็นกราฟจะได้ผลดังรูปที่ 6



รูปที่ 6 ผลการทดลองความใกล้เคียงของเอกสารจากการสืบค้น

อย่างไรก็ดี ประโยชน์ส่วนหนึ่งที่ได้รับจากการคัดกรองเอกสารอาจจะนำไปประยุกต์ต่อในการวัดประสิทธิภาพของ web service ที่ให้ผลตอบแทน (กำไร) แก่ผู้ใช้ [10] หรืออาจนำไปประเมินความสามารถของระบบ web service ที่มีอยู่ได้

แนวทางการวิจัยต่อไปคือ สร้างตัวแบบและระเบียบวิธีเพื่อคำนวณหาพลังงานที่ใช้ในการสืบค้นและประมวลผลข้อมูล โดยแนวคิดเบื้องต้นคือการลดขนาดของ ontology matching tree ด้วยการกำหนด threshold ของความน่าจะเป็นเพื่อตัดกิ่งที่มีค่าต่ำกว่า threshold ทิ้ง ขณะเดียวกันก็คำนวณค่าพลังงานในการประมวลผลคำสั่งของการสืบค้นโดยกำหนด weight ใน utility function ที่ derived จาก ontology matching tree ข้างต้นก็จะได้ค่าพลังงานของแต่ละเส้นทางทดสอบ ผลรวมทุกเส้นทางก็คือพลังงานที่ใช้ในการสืบค้นเอกสาร ซึ่งควรจะลดลงจากการสืบค้นโดยไม่มีการคัดกรองก่อน การดำเนินการดังกล่าวถือเป็นประโยชน์ทางอ้อมของการวิจัยการลดปริมาณเอกสารที่ไม่จำเป็น ซึ่งเป็นแนวทางสนับสนุนพลังงานสีเขียวเพื่อลดภาวะโลกร้อนได้มาก

5. บทสรุป

ระเบียบวิธีที่นำเสนอเป็นเพียงแนวทางหนึ่งในการสังเคราะห์กระบวนการที่ตรงไปตรงมา เพื่อลดความซับซ้อนที่ไม่พึงประสงค์ต่างๆ ทั้ง essential และ accidental complexities [7] ที่ฝังตัวในระเบียบวิธี เช่น ความน่าจะเป็นทั้ง 4 กรณีของ matching และการหาความสัมพันธ์ระหว่างบริบทของคำสำคัญ เป็นต้น ซึ่งอาจจะต้องสร้างตัวแบบที่เหมาะสมของ ontology tree ให้ครอบลุมทุกกรณี เพื่อลดจำนวนเอกสารจากการสืบค้นลง

นักวิจัยเริ่มให้ความสนใจมากขึ้นเรื่อยๆ เกี่ยวกับปัญหาของปริมาณเอกสารผลลัพธ์ที่ได้จากการสืบค้นผ่าน search engine เพราะอินเทอร์เน็ตเปรียบเสมือนคลังข้อมูลขนาดใหญ่ที่มีการใช้ web service ในลักษณะ cloud computing หรือ Software as a Service (SaaS) [8] ซึ่งเป็นแนวคิดใหม่ของการประมวลผล หากมีการคิดค่าใช้จ่ายในการสืบค้น (จะด้วยหน่วยวัดใดก็ตาม) รวมทั้งการใช้พลังงานประมวลผลรรชนีของเอกสารนับไม่ถ้วนที่เกี่ยวข้อง ความจำเป็นของการคิดค้นนวัตกรรมใหม่ๆ บนพื้นฐานของเทคโนโลยีที่มีอยู่ [9] ในการพัฒนาระเบียบวิธีการสืบค้นที่มีประสิทธิภาพเพื่อประหยัดพลังงาน ถือเป็นความจำเป็นเร่งด่วน จะเห็นตัวอย่างของ google จากการเริ่มแสดงตัวเลือกของคำคล้ายตามบริบทอักขระขณะเริ่มพิมพ์อักษรตัวแรกๆ แต่ยังไม่ครบถ้วนเป็นคำ ความพยายามดังกล่าวเป็นเพียงตัวอย่างที่ผู้วิจัยจำเป็นต้องเรียนรู้จากระเบียบวิธีที่มีอยู่ (state-of-the-practice) เพื่อสรรสร้างระเบียบวิธีเชิงวิชาการ (state-of-the-art) ที่มีประสิทธิภาพในรูปแบบที่เป็นธรรมชาติ ซึ่งให้ความคุ้นเคยกับผู้ใช้มากที่สุด

เอกสารอ้างอิง

[1] J. Balinski and C. Danilowicz, “Re-ranking Method based on Inter-document Distances”, *Journal of the Information Processing and Management*, Vol. 41, Issue 4, 2005.

[2] T. Hofmann, “Probabilistic Latent Semantic Analysis”, *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*”, Berkeley, California, ACM Press, August 1999, pp. 50-57.

[3] Jiangang Ma, Ynachum Zhang, and Jing He, “Efficiently Finding Web Services Using a Clustering Semantic Approach”, *Proceedings of the 2008 international workshop on Context enabled source and service selection, integration and adaptation: organized with the 17th International World Wide Web Conference (WWW 2008)*, Beijing, China, April 22, 2008.

[4] B. Mandhani, S. Joshi, and K. Kummamuru, “A Matrix Density Based Algorithm to Hierarchically CoCluster Documents and Words”, *Proceedings of the 12th international conference on World Wide Web*, Budapest, Hungary, May 20-24, 2003. pp. 511-518.

[5] Mehmet Senvar and Ayse Bener, “Matchmaking of Semantic Web Services Using Semantic-Distance Information”, *LNCS 4243*, 2006, pp. 177-186.

[6] T. Hofmann, J. Puzicha, and M.I. Jordan, “Unsupervised learning from dyadic data”, *Advances in Neural Information Processing Systems*, vol. 11, MIT Press, 1999.

[7] F.P. Brooks, Jr., “No Silver Bullet: Essence and Accidents of Software Engineering”, *Computer*, vol. 20, no. 4, 1987, pp. 10-19.

[8] H. Erdogmus, “Cloud Computing: Does Nirvana Hide behind the Nebula?”, *IEEE Software*, March/April 2009, pp. 4-6.

[9] G. Booch, “The Resting Place of Innovation”, IEEE Software, March/April 2009, pp. 12-13.

[10] Chawathe, S.S., “Strategic Web-Service Agreements”, International Conference on Web Services (ICWS '06), Sept 18-22, 2006, pp. 119-126.

[11] Jianhua Wang and Ruixu Li, “A New Cluster Merging Algorithm of Suffix Tree Clustering”, 2006, in IFIP International Federation for Information Processing, Volume 228, Intelligent Information Processing III, eds. Z. Shi, Shimohara K., Feng D., (Boston: Springer), pp. 197-203.

[12] Guihong Cao, Dawei Song, and Peter Bruza, “Suffix Tree Clustering on Post-retrieval Documents”, http://www.dstc.edu.au/Research/Projects/Infoeco/publications/tech-report-suffix-tree.pdf. August, 2009.

[13] Douglass R. Cutting, David R. Karger, and Jan O. Pedersen, “Constant interaction-time scatter/gather browsing of very large document collections”, SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, July 1993, pp. 126-134.

[14] Sebastian Schaffert, Francois Bry, Joachim Baumeister, and Malte Kiesel, “Semantic Wikis”, IEEE Software, July/August 2008, pp. 8-11.