การพัฒนาการสืบค้นเอกสารงานวิจัยในเชิงความหมายโดยใช้ออนโทโลยี

นพคุณ บุญสิม

ศูนย์วิจัยและฝึกอบรมปัญญาประดิษฐ์และเทคโนโลยีสารสนเทศ
สาขาวิทยาการคอมพิวเตอร์ วิทยาเขตหนองคาย มหาวิทยาลัยขอนแก่น
Emails: boonsim@nkc.kku.ac.th

บทคัดย่อ

งานวิจัยนี้นำเสนอวิธีการพัฒนาการสืบค้นเอกสารงานวิจัยโดยใช้ เทคนิคการสืบค้นเชิงความหมาย ซึ่งมีการออกแบบออนโทโลยีเข้ามาร่วมด้วย โดยในงานวิจัยประกอบด้วย 2 ขั้นตอนหลัก ขั้นตอนที่ 1 คือการออกแบบการจัดเก็บเอกสาร ขั้นตอนนี้เป็นการหาตัวแทนเอกสาร ซึ่งประกอบด้วยขั้นตอนการสกัดคำสำคัญจากเอกสาร การนับความถี่ของคำ การหาน้ำหนักของคำเพื่อนำมาเป็นตัวแทนเอกสารและมีการสร้างออนโทโลยีสำหรับช่วยในการจัดกลุ่มของเอกสารที่มีความหมายคล้ายคลึงกัน ขั้นตอนที่ 2 คือการสืบค้นเอกสาร มีการออกแบบการสืบค้นข้อมูลเอกสารจากข้อมูลที่จัดเก็บตามโครงสร้างของออนโทโลยีในฐานข้อมูลและมีการใช้ฐานความรู้เวิร์ดเน็ตในการสืบค้น ผลลัพธ์จากการสืบค้นจะมีการจัดลำดับความคล้ายคลึงของข้อมูลด้วย ผลการวิจัยพบว่าวิธีการดังกล่าวมีค่าเฉลี่ยของความถูกต้องสูงกว่าวิธี TF-IDF เท่ากับ 28.32 %

คำสำคัญ-- Document; retrieval; semantic; ontology

1. บทนำ

ปัจจุบันการสืบค้นเอกสารโดยส่วนใหญ่เป็นการใช้คำสำคัญ (keyword) ในการสืบค้น เนื่องจากวิธีดังกล่าวเป็นการเปรียบเทียบทางตรรกะ เอกสารที่จะถูกค้นคืนนั้นจะต้องประกอบด้วยคำสำคัญที่ต้องการสืบค้น ถ้าในเอกสารไม่มีคำสำคัญแต่เป็นเอกสารที่ถูกต้องจะไม่ถูกค้นคืน ตัวอย่างเช่น ถ้าคำสืบค้น มีคำว่า “Car” ถ้าใช้คำสืบค้นแบบคำสำคัญ เอกสารที่มีคำสำคัญคือ “Automobile” “motor” และ “vehicle” จะไม่ถูกค้นคืนแต่เป็นเอกสารที่สืบความหมายถูกต้อง วิธีการดังกล่าวจึงยังไม่เหมาะสมสำหรับการสืบค้น และวิธีการใช้คำสำคัญจะทำการสืบค้นจากชื่อของเอกสาร ชื่อผู้แต่ง หัวข้อและคำสำคัญ ซึ่งข้อมูลดังกล่าวไม่ได้เป็นตัวแทนเอกสารที่แท้จริง จากปัญหาดังกล่าวในงานวิจัยนี้จึงมีแนวความคิดที่จะพัฒนาแนวทางการสืบค้นโดยใช้เทคโนโลยีการสืบค้นในเชิงความหมาย (semantic retrieval) เพื่อให้สามารถค้นคืนเอกสารตามลักษณะในเชิงความหมายได้ บทความนี้ในหัวข้อที่ 2 จะกล่าวถึงงานวิจัยที่เกี่ยวข้องที่ใช้เป็นฐานความรู้ในการพัฒนางานวิจัย หัวข้อที่ 3 เป็นขั้นตอนการดำเนินงานวิจัย หัวข้อที่ 4 แสดงข้อมูลการทดลองและผลการทดลองและหัวข้อที่ 5 นำเสนอบทสรุปของงาน

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การสกัดคำสำคัญ

การสกัดคำสำคัญเป็นงานที่ดำเนินการเพื่อเลือกเอาเฉพาะคำสำคัญที่สื่อความหมายได้โดยใช้รายการคำหยุด (Stop list) [1] รายการคำหยุด เช่น (Do, good, take, a, an, the) เป็นคำที่ไม่สื่อความหมาย การสกัดคำสำคัญเป็นการเลือกคำสำคัญสำหรับการเป็นตัวแทนของเอกสาร

2.2 ความถี่และความถี่ผกผัน

การคำนวณความถี่และความถี่ผกผันเป็นขั้นตอนสำหรับการกำหนดน้ำหนักให้กับคำสำคัญในเอกสารเพื่อใช้เป็นตัวกลุ่มคำตัวแทนของเอกสาร โดยที่ความถี่ (TF, Term Frequency) เป็นการนับความถี่ของความสำคัญในเอกสารเอกสาร ความถี่ผกผัน (IDF, Inverse Document Frequency) [2] เป็นการถ่วงน้ำหนักของเอกสารมีความหมายคือการถ่วงน้ำหนักของความถี่ของคำสำคัญกับเอกสารที่พบคำสำคัญนั้นภายในฐานข้อมูลและน้ำหนักคำสำคัญ (TW, Term Weight) เกิดจากผลคูณของความถี่ของคำสำคัญกับความถี่ผกผันของคำดังกล่าวในฐานข้อมูลดังสมการ (1)

$$TW_{i,d} = TF_{i,d} \times IDF_t$$
$$TF_{i,d} = 1+\log_{10} f_{i,d}$$
$$IDF_t = \log_{10} N/n$$
$$TW_{i,d} = (1+\log_{10} f) \times (\log_{10} N/n) \quad (1)$$

เมื่อ $TW_{i,d}$ คือน้ำหนักของคำสำคัญที่ i ของเอกสาร d
$TF_{i,d}$ คือความถี่ถ่วงน้ำหนักของคำสำคัญที่ i ของเอกสาร d
IDF_t คือความถี่ผกผันของคำสำคัญ t
$f_{i,d}$ คือความถี่ของคำสำคัญที่ i ของเอกสาร d
N คือจำนวนเอกสารทั้งหมดในฐานข้อมูล
n คือจำนวนเอกสารที่ปรากฏคำที่ i

2.3 การหาแกนคำ

การหาแกนคำ (Stemming) เป็นเทคนิคที่ใช้ในการหาคำซึ่งเป็นรากศัพท์ของคำเดียวกันเนื่องจากภาษาอังกฤษนั้นมีการใช้งานคำที่มีความหมายเหมือนกันในหลายลักษณะ โดยการหาแกนคำมีขั้นตอนดังนี้

1) ทำการพิจารณาพยัญชนะและสระภายในคำที่ต้องการหาแกนคำ โดยเมื่อตัดพยัญชนะตัวแรกและตัวสุดท้ายออกไปแล้ว

ถ้าไม่เหลือพยัญชนะและสระจะกำหนดให้ค่า m = 0

ถ้าเหลือพยัญชนะและสระ 1 คู่จะกำหนดให้ค่า m = 1

ถ้าเหลือพยัญชนะและสระ 2 คู่จะกำหนดให้ค่า m = 2

ถ้าเหลือมากกว่านี้สามารถตัดตัวสุดท้ายออกไปได้เลย

ซึ่งค่า m จะเป็นค่าที่ใช้ในการพิจารณาก่อนการตัดตัวอักษรทุกครั้ง

2) การตัดคำให้เพื่อเปลี่ยนจากพหูพจน์เป็นเอกพจน์ เช่น คำที่ลงท้ายด้วย sses ให้ตัดเหลือ ss หรือ ies ให้ตัดเหลือ i หรือลงท้ายด้วย s ตัด s ออกไป ไม่พิจารณาค่า m

3) ตัดคำที่ลงท้ายด้วยคำเหล่านี้ เช่น ICATE ATIVE TION ICAL FUL NESS และ MENT เป็นต้น เพื่อให้เหลือเฉพาะรากศัพท์หรือแกนคำ การตัดคำดังกล่าวจะต้องพิจารณาค่า m ก่อนแล้วจึงทำตัดคำ

การหาแกนคำมีความสำคัญในขั้นตอนการสรุปใจความสำคัญจากเอกสารงานวิจัย ซึ่งวิธีที่ใช้แกนคำจะได้ความถี่ที่แท้จริงของเอกสาร ความถี่ที่ได้จะนำไปคำนวณน้ำหนักของคำในเอกสาร การหาแกนคำนั้นใช้ลำดับขั้นตอนของ พอร์ทเตอร์ (Porter Algorithm) [3]

2.4 ออนโทโลยี

ออนโทโลยี (Ontology) [4] เป็นการนิยามหรือกำหนดรูปแบบโครงสร้างของสิ่งที่สนใจให้มีความหมายตามขอบเขตขององค์ความรู้ซึ่งออนโทโลยีนี้ได้ถูกกล่าวถึงในหลายองค์กรโดยเฉพาะปัญญาประดิษฐ์ ในปัจจุบันได้ถูกนำมาประยุกต์ใช้ในมาตรฐานของการออกแบบจำลองโครงสร้างของ eXtensible Markup Language (XML) และการนิยามรูปแบบแนวคิดของโครงสร้างของฐานข้อมูล เป็นต้น ออนโทโลยีเป็นเทคโนโลยีทางด้านการพัฒนาภาษาเชิงความหมายสมัยใหม่ โดยเป็นภาษาที่เครื่องคอมพิวเตอร์สามารถตีความหมายและทำตามคำสั่งได้ โดยลักษณะของออนโทโลยีคือการนิยามรูปแบบ (model) ภายในขอบเขตขององค์ความรู้เพื่ออธิบายสิ่งที่สนใจ (domain) ให้ได้ใจความและถูกต้องมากที่สุด

2.5 ความคล้ายคลึงของเอกสาร

การคำนวณความคล้ายคลึงของเอกสารหรือข้อมูลเป็นวิธีการสำหรับประเมินผลการเปรียบเทียบข้อมูลของคำสืบค้นกับข้อมูลในฐานข้อมูล หรือข้อมูลกับข้อมูลหรือเอกสารกับเอกสาร โดยใช้สมการ (2) มีการใช้งานในขั้นตอนของการพิจารณาผลลัพธ์ของเอกสาร การจัดกลุ่มเอกสารและการจัดลำดับของผลลัพธ์ของเอกสาร [6]

$$Sim(D_i,D_j) = \sum_{i \in j} wt_k \qquad (2)$$

เมื่อ Wt_k คือน้ำหนักของคำสำคัญ k

$\sum_{i \in j} wt_k$ คือผลรวมของคำสำคัญ i ที่เป็นสมาชิกเอกสาร D_j

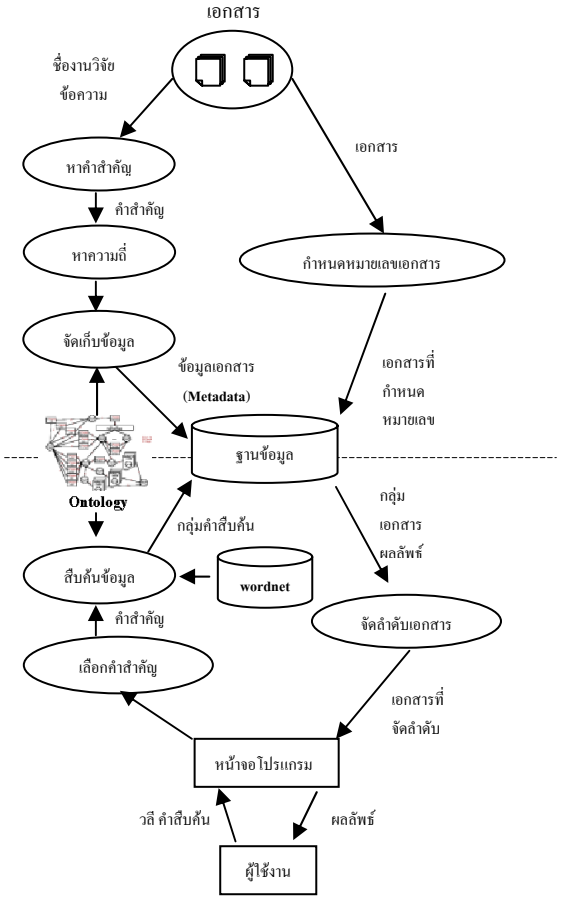
$Sim(D_i,D_j)$ คือผลการคำนวณความคล้ายคลึงระหว่างเอกสาร i กับเอกสาร j

3. วิธีการดำเนินงานวิจัย

3.1 การออกแบบจำลอง

แบบจำลองการสืบค้นเอกสารประกอบด้วยส่วนในการจัดรูปแบบข้อมูลเอกสารและส่วนในการสืบค้นข้อมูลจากรูปที่ 1 ส่วนด้านบนของภาพ เป็นการจัดเก็บข้อมูลจากเอกสาร ประกอบด้วย สกัดคำสำคัญจากข้อความ ซึ่งมาจากเอกสารงานวิจัย ต่อจากนั้นทำการนับความถี่ของคำสำคัญและคำนวณน้ำหนักของคำสำคัญดังกล่าวและจัดเก็บในฐานข้อมูลตามโครงสร้างของออนโทโลยี โดยข้อมูลเอกสารจะประกอบด้วยหมายเลขเอกสารและข้อมูลของเอกสาร (Metadata) ส่วนที่ 2 เป็นส่วนของการสืบค้นข้อมูล เริ่มจากคำหรือวลีสืบค้นของผู้ใช้ที่นำเข้ามาจะนำคำและวลีสืบค้นไปค้นหาในฐานข้อมูลเวิร์ดเน็ตซึ่งจะได้กลุ่มคำสืบค้นและนำกลุ่มคำดังกล่าวทำการสืบค้นข้อมูลในฐานข้อมูลตามออนโทโลยี

ซึ่งผลลัพธ์ ที่ได้จะเป็นกลุ่มของเอกสารผลลัพธ์และทำการจัดลำดับเอกสารที่มีความคล้ายคลึงกับคำสืบค้นมากที่สุดแสดงเป็นลำดับออกมายังผู้ใช้งาน



รูปที่ 1. แบบจำลองการสืบค้นเอกสาร [5]

3.2 การสกัดคำสำคัญจากข้อความเอกสาร

เมื่อนำเข้าเอกสารงานวิจัยสู่ระบบ จะทำการสกัดข้อความ (Text) จากเอกสารและชื่อของงานวิจัยเพื่อใช้ในการทำเป็นตัวแทนของเอกสาร ซึ่งวิธีการสกัดคำสำคัญจากเอกสารคือ ใช้การเปรียบเทียบคำศัพท์ต่าง ๆ ที่พบในเอกสารกับรายการคำหยุดหรือ (Stop list) เพื่อหาคำสำคัญดังรูปที่ 2 ถ้า

พบในรายการคำหยุด จะไม่เลือกเป็นคำสำคัญ เมื่อได้คำสำคัญจากเอกสารแล้วจะเรียกว่าเป็นกลุ่มคำสำคัญของเอกสาร ขั้นต่อมาจะทำการหาแกนคำหรือรากศัพท์ของคำสำคัญทั้งหมดโดยใช้วิธีการของพอร์ทเตอร์ (Porter Algorithm)

word_list : ตาราง

word
amusing
an
ancient
and
angry
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart

appear
appreciate
appropriate
aqua
are
area
areas
aren't
around
as
a's
ashamed
aside
ask
asked
asking

รูปที่ 2. รายการคำหยุด

3.3 การนับความถี่และหาน้ำหนักของคำสำคัญ

จากขั้นตอนที่ผ่านมาได้กลุ่มคำสำคัญจากเอกสาร ขั้นตอนนี้จะเป็นการนับความถี่ของคำสำคัญ เมื่อได้ความถี่ของคำสำคัญแล้วจะนำมาคำนวณน้ำหนักของคำสำคัญตามสมการ (1) จากขั้นตอนนี้จะได้ตัวแทนของเอกสารเป็นกลุ่มของคำสำคัญและน้ำหนักของคำสำคัญและงานวิจัยได้ออกแบบเลือกเอาเฉพาะคำสำคัญที่มีน้ำหนักมากที่สุด 10 อันดับเพื่อเป็นตัวแทนของเอกสารดังรูปที่ 3 ในรูปแสดง doc_id คือหมายเลขของเอกสาร fw คือคำสำคัญที่พบในเอกสาร fn คือความถี่ของคำสำคัญ tf_idf คือน้ำหนักของคำสำคัญของเอกสารในฐานข้อมูล

3.4 การสร้างออนโทโลยี

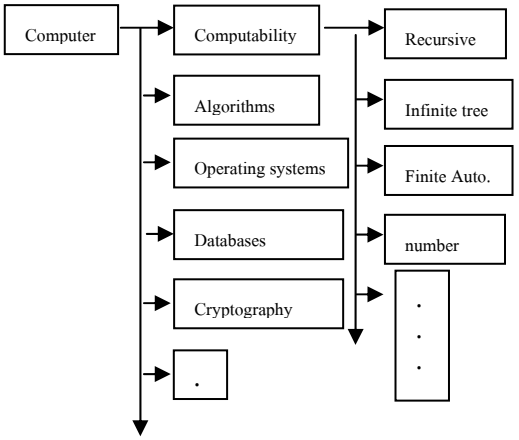
ในขั้นตอนนี้ทำการสร้างออนโทโลยี เพื่อเป็นการนิยามองค์ความรู้ของงานวิจัยเพื่อที่จะเป็นการแบ่งขอบเขต หมวดหมู่หรือประเภทของเรื่องของงานวิจัย โดยในงานวิจัยนี้จะเลือกกรณีศึกษาเป็นงานวิจัยทางด้านวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ดังตัวอย่างรูปที่ 4 แสดงข้อมูลการนิยามออนโทโลยีของคอมพิวเตอร์โดยเริ่มจากคำว่าคอมพิวเตอร์ซึ่งเป็นโหนดแรกของระบบหรือโครงสร้างแบบต้นไม้และในระดับถัดมาก็มีการจำแนกข้อมูลออกมาเป็นหลายโหนด เช่น Computability, Algorithm, Operating System และ Database เป็นต้น

doc_metadata : ตาราง

doc_id	doc_fwn	fw	fn	tf_idf
001	0	algorithm	130	2.5656051468
001	1	filter	92	4.3356683541
001	2	adapt	38	4.4470144224
001	3	subsampl	36	6.3322712581
001	4	comput	28	1.2562080196
001	5	matrix	22	4.6848453616
001	6	transact	18	3.8316403378
001	7	rotat	16	5.459872457
001	8	signal	14	3.0534328338
001	9	updat	14	4.8401035513
002	0	algorithm	84	2.4093392647
002	1	converg	54	5.6045563528
002	2	learn	51	3.3302209916
002	3	rat	33	5.4805161704
002	4	const	32	2.9645706799
002	5	analysi	30	3.0881585346

รูปที่ 3. การจัดเก็บคำสำคัญและน้ำหนักคำสำคัญ

ซึ่งการนิยามหมวดหมู่และจำแนกข้อมูลประเภทของความรู้ทางวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศนั้นทำการอ้างอิงจากการจัดหมวดหมู่ของ The Computer Sciences Accreditation Board (CSAB) [9]



รูปที่ 4. ตัวอย่างการนิยามออนโทโลยี

3.5 การจัดเก็บเอกสารตามโครงสร้างออนโทโลยี

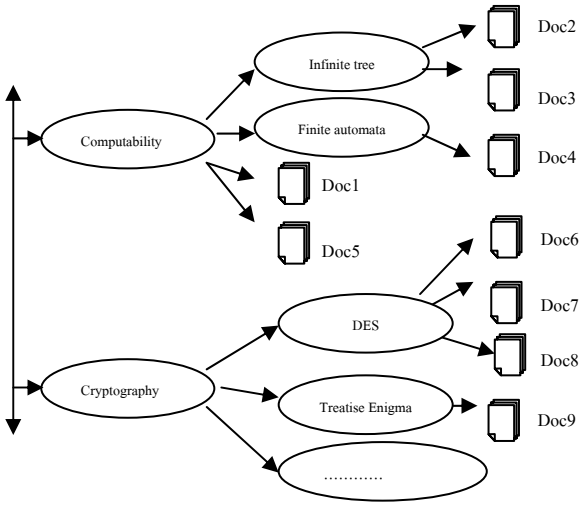
เมื่อได้สร้างออนโทโลยีแล้ว ขั้นต่อมาจะทำการจัดเก็บเอกสารตามโครงสร้างของออนโทโลยี ใช้การคำนวณความคล้ายคลึงของเอกสารกับโหนดของออนโทโลยี ซึ่งมีแนวความคิดคือ จากโหนดของออนโทโลยีที่ประกอบด้วยคำสำคัญตัวอย่างเช่น Cryptographic algorithm จะนำคำสำคัญดังกล่าวมาสืบค้นในตารางจัดเก็บคำสำคัญของเอกสาร ซึ่งแต่ละคำสำคัญของเอกสารจะมีน้ำหนักของคำสำคัญด้วยดังรูปที่ 3 หลังจากนั้นทำการคำนวณความคล้ายคลึงของเอกสารกับออนโทโลยี โดยใช้สมการ (3) และจัดเก็บเอกสารในโหนดที่มีความคล้ายกับเอกสารมากที่สุด ดังรูปที่ 5 และจัดเก็บในฐานข้อมูลดังรูปที่ 6 ถ้ามีผลรวมของการคำนวณความคล้ายคลึงเท่ากันจะทำการพิจารณาโหนดของออนโทโลยีที่อยู่ถัดไป (Super class) จากโหนดดังกล่าวขึ้นไปเรื่อย ๆ จนกว่าจะได้ความคล้ายคลึงมากที่สุด

$$Sim(D_i,O_j) = \sum_{i \in j} wt_k \quad (3)$$

เมื่อ Wt_k คือน้ำหนักของคำสำคัญ k

$\sum_{i \in j} wt_k$ คือผลรวมของคำสำคัญ i ที่พบในออนโทโลยี O_j

$Sim(D_i,O_j)$ คือผลการคำนวณความคล้ายคลึงระหว่างเอกสาร i กับออนโทโลยี O_j



รูปที่ 5. การจัดเก็บเอกสารตามโหนดของออนโทโลยี

doc : ตาราง

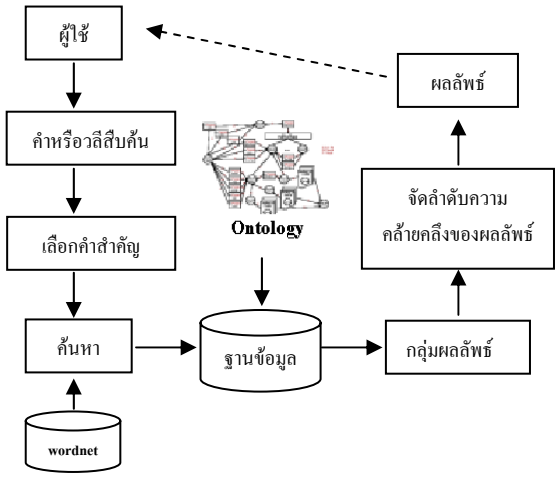
doc_id	doc_name	doc_class	doc_sc
001	Fast Subsampled-Updatin	654	4.4470144224
002	Convergence Analysis of	636	5.5451540811
003	Complementary Two-Way	871	7.9474916183
004	Path Partitions and Forwa	1388	4.3107584852
005	Improving Computational E	485	5.6330201777
006	Empirical Performance Ev	1127	7.5157939635
007	Multichannel Affine and Fa	808	5.7700902291
008	Interaction between Polyn	1033	7.4008205585
009	Analysis, Evaluation,and C	539	7.0566775444
010	Frequency Responses of F	480	4.9345473568
011	Multichannel Recursive-Le	654	4.4270496648
012	Adaptive Scheduling Algor	539	8.0658415513
013	A Precise Termination Cor	488	8.8671867086
014	Two-Dimensional Block A	464	7.6884314335

รูปที่ 6. การจัดเก็บเอกสารตามในฐานข้อมูล

รูปที่ 6 แสดงการจัดเก็บข้อมูลเอกสารในฐานข้อมูลโดย doc_id คือหมายเลขของเอกสาร doc_name คือชื่อของเอกสาร doc_class คือ class หรือโหนดของออนโทโลยีที่เอกสารนี้ถูกจัดเก็บอยู่และสุดท้าย doc_sc คือส่วนแสดงน้ำหนักหรือคะแนนของความคล้ายคลึงของเอกสารกับโหนดของออนโทโลยี

3.6 การสืบค้นข้อมูล

การสืบค้นข้อมูลเริ่มจากคำหรือวลีสืบค้นที่ผู้ใช้ นำเข้าสู่ระบบเพื่อสืบค้นข้อมูล คำหรือวลีดังกล่าวเป็นภาษาธรรมชาติ (Natural Language) ซึ่งการค้นหาด้วยรูปแบบดังกล่าวทำให้ผลลัพธ์จากการสืบค้นไม่ตรงกับความต้องการ จึงทำการออกแบบขั้นตอนการสืบค้นดังขั้นตอนในรูปที่ 7 เริ่มจากคำสืบค้นจากผู้ใช้และพิจารณาเลือกเฉพาะคำสำคัญที่สื่อความหมายได้ โดยใช้วิธีเดียวกันกับขั้นตอนที่ 3.2 คือการสกัดคำสำคัญด้วยรายการคำหยุดและใช้วิธีการหาแกนคำเพื่อหารากศัพท์ของคำสืบค้น จะได้กลุ่มคำสืบค้น Q = {Qi' Qi+1, Qi+2, Qi+3 ...} และจะนำกลุ่มคำดังกล่าวไปสืบค้นคำศัพท์ที่มีความหมายเหมือนกันในฐานข้อมูลเวิร์ดเน็ต หลังจากนั้น ทำการสืบค้นข้อมูลจากฐานข้อมูลโดยใช้การคำนวณความคล้ายคลึงของคำสืบค้นจากสมการ (2) กับข้อมูลงานวิจัยที่จัดเก็บในโครงสร้างของออนโทโลยี



รูปที่ 7. การออกแบบการสืบค้น

เมื่อได้ผลลัพธ์จากการสืบค้น จะนำผลลัพธ์นั้นมาจัดอันดับความคล้ายคลึงของเอกสาร มีแนวความคิดคือ ทำการคำนวณน้ำหนักของคำสืบค้นกับเอกสารผลลัพธ์ที่ประกอบด้วยกลุ่มคำสำคัญและน้ำหนัก โดยใช้สมการ (2) และเมื่อได้น้ำหนักความคล้ายคลึงของเอกสารมาแล้วจะทำการเรียงเอกสารตามน้ำหนักความเหมือนจากมากไปหาน้อย

4. วิธีการทดลองและผลการทดลอง

4.1 วิธีการทดลอง

บทความนี้ใช้เอกสารงานวิจัยงานฐานข้อมูล IEEE จำนวน 900 เอกสาร แบ่งออกเป็น 18 หัวข้อ ได้แก่ algorithm, compiler, computability, data structure, digital, graph, programming language เป็นต้น และสร้างการทดลองออกเป็น 2 แบบ แบบที่ 1 คือการทดลองโดยใช้ Semantic-based (การสืบค้นแบบเชิงความหมายโดยใช้ออนโทโลยี) และการสืบค้นแบบ TF-IDF (การใช้คำสำคัญโดยการนับความถี่ของคำสำคัญ) และกำหนดข้อมูลสืบค้นจำนวน 30 วลีสืบค้นที่ไม่ซ้ำกัน ทำการสืบค้นข้อมูลจำนวน 5 กลุ่มได้แก่ Top-1, Top-5, Top-10, Top-20, Top-30 จากนั้นทำการคำนวณค่าความถูกต้อง Precision และค่าความระลึก Recall จากสมการ (4) และสมการ (5) ตามลำดับ

$$P = \frac{A}{A + B} \quad (4)$$

$$R = \frac{A}{A + C} \quad (5)$$

เมื่อ P คือค่าความถูกต้อง (Precision)

R คือค่าความครบถ้วน (Recall)

A คือจำนวนเอกสารที่ถูกต้องและถูกสืบค้น

B คือจำนวนเอกสารที่ไม่ถูกต้องและถูกสืบค้น

C คือจำนวนเอกสารที่ถูกต้องแต่ไม่ถูกสืบค้น [7]

4.2 ผลการทดลอง

จากการทดลองทำการสืบค้นข้อมูลโดยผู้วิจัยทำการเขียนโปรแกรมด้วย Visual Basic 6 ทำการจัดเก็บข้อมูลทั้งหมดลงในฐานข้อมูล ทำการสืบค้นตามจำนวนคำสืบค้นและบันทึกผลทั้งวิธีการสืบค้นแบบ Semantic-based และการสืบค้นแบบ TF-IDF ตารางที่ 1 แสดงตัวอย่างคำสืบค้น

ตาราง 1. แสดงผลการสืบข้อมูล

ตัวอย่างคำสืบค้น
1. image retrieval techniques
2. Optical character recognition
3. mobile networking model design
4. xml data represent integration
5. web service e-business
6. mobile phone game model
7. algorithm xml mapping
8. visual programming method system
9. multimedia semantic contents
10. semantic annotation RDF model

ตาราง 2. แสดงผลการสืบข้อมูลค่าความถูกต้อง (Precision)

วิธีการสืบค้นแบบ Semantic-based

คำสืบค้น	TOP1	TOP5	TOP10	TOP20	TOP30
1	21.23	16.67	15.93	14.26	12.37
2	13.38	12.24	12.44	12.44	12.44
3	16.16	14.33	13.34	12.69	12.54
4	15.43	17.53	15.07	13.17	11.34
5	39.62	34.11	33.65	30.29	28.26
6	22.2	18.5	14.52	12.22	11.33
7	24.1	19.95	16.22	14.06	11.99
8	29.97	23.56	19.1	16.72	15.29
9	18.02	18.88	15.11	12.44	11
10	30.55	27.49	24.31	21.24	18.98

ตารางที่ 3. แสดงผลการสืบข้อมูลค่าครบถ้วน (Recall)

วิธีการสืบค้นแบบ Semantic-based

คำสืบค้น	TOP1	TOP5	TOP10	TOP20	TOP30
1	3.58	14.9	26.03	44.57	59.92
2	18.92	84.96	100	100	100
3	0.96	4.16	7.53	13.92	19.99
4	1.78	8.1	13.82	24.31	34.72
5	1.02	4.37	8.02	14.32	20.35
6	2.2	8.89	14.51	24.6	33.97
7	1.55	6.19	10.79	18.55	23.35
8	0.86	4.02	7.52	13.68	18.71
9	1.84	7.45	12.99	21.96	29.26
10	1.3	5.49	9.32	16.3	22.6

ตารางที่ 2 และตารางที่ 3 แสดงตัวอย่างผลลัพธ์จากการสืบค้นวิธีการสืบค้นเชิงความหมาย Semantic-based โดยแสดงผลค่าความถูกต้อง (Precision) และความครบถ้วน (Recall) ของผลลัพธ์ Top-1, Top-5, Top-10, Top-20, Top-30 ตามลำดับ

ตารางที่ 4. แสดงผลการสืบข้อมูลค่าครบถ้วน (Recall)

วิธีการสืบค้นแบบคำสำคัญ

คำสืบค้น	TOP1	TOP5	TOP10	TOP20	TOP30
1	14.68	10.71	9.52	8.91	8.59
2	13.38	12.24	12.44	12.44	12.44
3	11.8	11.01	11.07	11.41	9.8
4	15.43	17.53	15.07	13.17	11.34
5	39.62	34.11	33.65	30.29	28.26
6	18	18.57	18.57	18.57	18.57
7	16.79	9.48	6.91	6.16	5.76
8	20.43	15.03	15.09	13.6	12.13
9	14.86	10.71	9.52	8.91	8.59
10	21.29	15.81	14.57	12.78	12.11

ตาราง 5. แสดงผลการสืบข้อมูลความถูกต้อง (Precision)

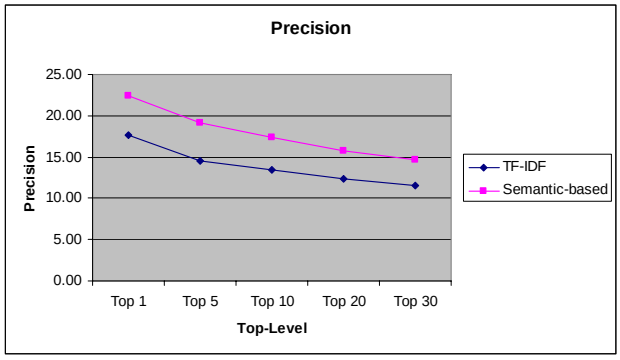
วิธีการสืบค้นแบบคำสำคัญ

คำสืบค้น	TOP1	TOP5	TOP10	TOP20	TOP30
1	14.68	10.71	9.52	8.91	8.59
2	13.38	12.24	12.44	12.44	12.44
3	11.8	11.01	11.07	11.41	9.8
4	15.43	17.53	15.07	13.17	11.34
5	39.62	34.11	33.65	30.29	28.26
6	18	18.57	18.57	18.57	18.57
7	16.79	9.48	6.91	6.16	5.76
8	20.43	15.03	15.09	13.6	12.13
9	14.86	10.71	9.52	8.91	8.59
10	21.29	15.81	14.57	12.78	12.11

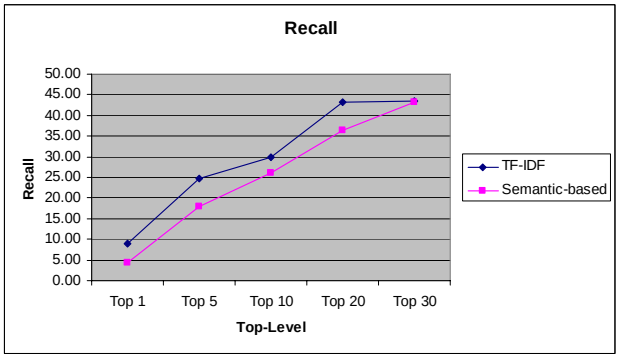
ตารางที่ 4 และตารางที่ 5 แสดงตัวอย่างผลลัพธ์จากการสืบค้นวิธีการสืบค้นแบบคำสำคัญ TF-IDF โดยแสดงผลค่าความถูกต้อง (Precision) และความครบถ้วน (Recall)

ตาราง 6. แสดงค่าเฉลี่ยผลการสืบค้น

	Precision		Recall	
	TF-IDF	Semantic-based	TF-IDF	Semantic-based
Top 1	17.60	22.41	8.94	4.37
Top 5	14.57	19.11	24.63	17.93
Top 10	13.46	17.35	29.83	26.02
Top 20	12.31	15.71	43.12	36.31
Top 30	11.60	14.64	43.44	43.12
Average	13.91	17.84	29.99	25.55



รูปที่ 8. การเปรียบเทียบ Precision ระหว่าง 2 วิธี



รูปที่ 9. การเปรียบเทียบ Recall ระหว่าง 2 วิธี

4.3 สรุปผลการทดลอง

จากการทดลองพบว่า ระดับความแม่นยำของวิธีที่เสนอ (Semantic-based) ให้ผลที่สูงกว่าวิธี TF-IDF โดยเฉลี่ยประมาณ 28.32 % ในขณะเดียวกันที่ระดับการเรียกคืนมีประสิทธิภาพลดลง 14.80 % จากผลการทดลองจะเห็นได้ว่าค่าความแม่นยำของวิธีการสืบค้นเชิงความหมายที่เพิ่มมากขึ้นจากวิธีใช้คำสำคัญ เกิดจากการสืบค้นในแต่ละครั้งทำการค้นคืนข้อมูลจากกลุ่มของเอกสารที่จัดเก็บตามโครงสร้างออนโทโลยีและมีการใช้ฐานข้อมูลเวิร์ดเน็ต [8] ทำให้ได้ผลลัพธ์จากการค้นคืนที่มีเอกสารที่ถูกต้องมากขึ้น

5. บทสรุป

งานวิจัยนี้นำเสนอแนวทางในการพัฒนางานวิจัยเกี่ยวกับการสืบค้นเอกสารในเชิงความหมาย ซึ่งมีขั้นตอนที่สำคัญ 2 ขั้นตอนคือ ขั้นตอนที่ 1 การจัดเก็บเอกสารและขั้นตอนที่ 2 การสืบค้นเอกสาร สำหรับขั้นตอนที่ 1 มีการออกแบบการจัดเก็บเอกสารโดยวิธีการสกัดคำสำคัญจากเอกสาร การหาแกนคำ การหาความถี่ของคำ การหาความถี่ผกผัน การสร้างออนโทโลยีและการจัดกลุ่มเอกสารตามโครงสร้างของออนโทโลยี วิธีการดังกล่าวเป็นขั้นตอนที่สำคัญในการหาตัวแทนเอกสารที่เหมาะสมสำหรับการสืบค้นเอกสาร ขั้นตอนที่ 2 การสืบค้นเอกสารประกอบด้วยขั้นตอนการหาแกนคำ การใช้ฐานข้อมูล เวิร์ดเน็ต และการจัดลำดับของผลลัพธ์การสืบค้นโดยการใช้การคำนวณความคล้ายคลึงของเอกสารกับคำสืบค้น จากการดำเนินงานวิจัยพบว่าวิธีการดังกล่าวให้ผลการสืบค้นที่ดีขึ้นสำหรับการสืบค้นเอกสารงานวิจัยในสาขาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ และงานวิจัยนี้สามารถนำไปประยุกต์ใช้งานกับสาขาอื่น ๆ ได้

6. เอกสารอ้างอิง

[1] R. P. Cook, “Heuristic compression of an English word list: Research Articles”. *Software—Practice & Experience,* Volume 35 Issue 6, 2005.

[2] Y. Rezgui, “Text-based domain ontology building using tf-idf and metric clusters techniques”, *The Knowledge Engineering Review*, Volume 22 Issue 4, Cambridge University Press, 2007

[3] M. F. Porter, “An algorithm for suffix stripping”, *Readings in information retrieval*. San Francisco, CA, USA, 1997.

[4] E. Toppano, V. Roberto, R. Giuffrida, G. B. Buora , “Ontology engineering: reuse and integration”, *International Journal of Metadata, Semantics and Ontologies* , Volume 3 Issue 3 , 2008.

[5] F.B. William, R.B.Yates, *Information retrieval—data structure and algorithm.* Englewood Cliffs, New Jersey, Prentice-Hall, 1992.

[6] J. O'Shea, Z. Bandar, K. Crockett, D. McLean, “A comparative study of two short text semantic similarity measures”, *KES-AMSTA'08: Proceedings of the 2nd KES International conference on Agent and multi-agent systems: technologies and applications.* 2008.

[7] G. Salton, “Automatic text processing the transformation, analysis and retrieval of information by computer”. Addison-Wesley, 1989.

[8] wordnet 3.0. wordnet a lexical database for English language, Retrieved March 01,2010, from http:// wordnet.princeton.edu,2006

[9] Area of computer science, Retrieved March 01, 2010, from http://en.wikipedia.org/wiki/Computer_science