การกรองผลการค้นหาเว็บด้วยความชอบของผู้ใช้และค่าความนิยม
Web Search Filter Using User Preference and Popular Profile

ภชพน ซื่อสัจลือสกุล (Pochpon Suesajlausakul)* และ อรัญญา วลัยรัชต์ (Aranya Walairacht)*

บทคัดย่อ

ในปัจจุบันเครื่องมือค้นหาข้อมูลบนอินเทอร์เน็ต (Search engine) ส่วนใหญ่ เมื่อผู้ใช้ใช้คำค้นหา (Keyword) ที่เหมือนกันมักจะส่งผลการค้นหาให้แก่ผู้ใช้ด้วยผลลัพธ์ที่คล้ายๆ กัน แต่ในความเป็นจริงผู้ใช้แต่ละคนอาจมีความสนใจในเรื่องที่แตกต่างกันออกไป ซึ่งงานวิจัยนี้นำเสนอวิธีการค้นหาที่สามารถแสดงผลลัพธ์ได้ตรงตามความสนใจของผู้ใช้ โดยเก็บประวัติการใช้งานของผู้ใช้ในรูปแบบของข้อมูลเริ่มต้น (Predetermined profile) ข้อมูลผู้ใช้ (User profile) และข้อมูลค่าความนิยมของผู้ใช้ (Popular profile) และใช้ข้อมูลทั้งสามส่วนนี้ในการวิเคราะห์หาความสัมพันธ์ระหว่าง คำค้นหา เอกสาร (เว็บไซต์) และหมวดหมู่ที่เกี่ยวข้อง ทำให้สามารถดึงข้อมูลที่ใกล้เคียงกับความสนใจของผู้ใช้ขึ้นมา ซึ่งผลการทดลองที่ได้แสดงให้เห็นถึงประสิทธิภาพที่เพิ่มขึ้นและได้ผลการค้นหาที่สอดคล้องกับความสนใจของผู้ใช้มากยิ่งขึ้น

คำสำคัญ: เครื่องมือค้นหา ความชอบของผู้ใช้ ข้อมูลเริ่มต้น ข้อมูลผู้ใช้ ข้อมูลค่าความนิยม

Abstract

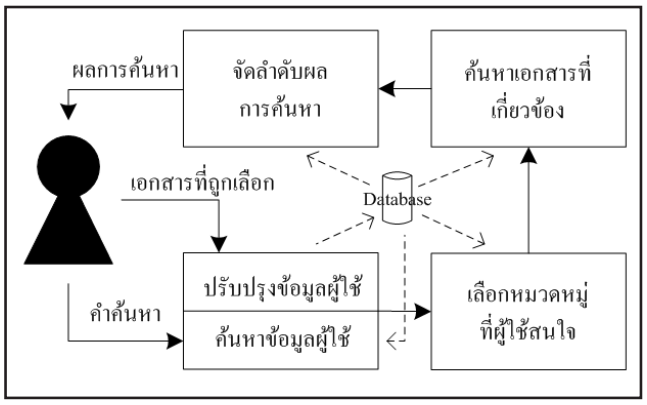
Nowadays many search engines provide the same result to users if they search with the same query. In fact, different users have different purpose. Therefore, this paper presents technique to search and return interesting result for the users. Our system record user's search history to create profile which consists of predetermined profile, user profile and popular profile. We used them to consider the relation between term, document and category. Then we can determine the user preference for each user and return result to user. Experiment result on real click-through show our techniques improve efficiency of search engine and returned better relevance result for users.

Keyword: Search Engines, User Preference, Predetermined Profile, User Profile, Popular Profile.

1. บทนำ

ปัจจุบันจำนวนข้อมูลที่อยู่บนอินเทอร์เน็ตมีจำนวนเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้การค้นหาข้อมูลมีความลำบากมากขึ้น เช่น การค้นหาข้อมูลด้วยคำค้นหาคำเดียวกันแต่ถูกค้นหาโดยผู้ใช้ที่มีความสนใจแตกต่างกัน ผู้ใช้แต่ละคนย่อมต้องการผลการค้นหาที่ตรงตามความสนใจของตนเอง ซึ่งในเครื่องมือค้นหาส่วนใหญ่จะคืนค่าผลการค้นหาให้กับผู้ใช้ด้วยผลลัพธ์ที่เหมือนๆกัน โดยไม่ได้คำนึงถึงความสนใจของผู้ใช้ (User preference) ที่เข้ามาใช้งาน ยกตัวอย่างเช่น การค้นหาคำว่า "แอปเปิล" จะเห็นได้ว่าแอปเปิลอาจประกอบไปด้วย แอปเปิลที่เกี่ยวข้องกับคอมพิวเตอร์ หรือแอปเปิลที่เกี่ยวข้องกับอาหาร ผู้ใช้แต่ละคนย่อมต้องการแอปเปิลที่ต่างกันออกไป ดังนั้นเครื่องมือค้นหาที่ดีควรจะคำนึงถึงความต้องการของผู้ใช้ในส่วนนี้ด้วย ซึ่งในงานวิจัยนี้นำเสนอวิธีที่จะแก้ไขปัญหาดังกล่าวโดยแบ่งการทำงานออกเป็นสองส่วนหลักๆ คือ ส่วนแรกระบบจะค้นหาเซตของหมวดหมู่ที่ผู้ใช้นั้นๆ ให้ความสนใจซึ่งสัมพันธ์กับคำค้นหาที่ป้อนเข้ามา โดยอ้างอิงจากประวัติการค้นหาของผู้ใช้ และในส่วนที่สองจะเป็นการคำนวณค่าน้ำหนักความสนใจของผู้ใช้จากเอกสารที่ผู้ใช้เลือกโดยอยู่ในรูปความสัมพันธ์ระหว่างคำค้นหา, เอกสาร และหมวดหมู่ที่เกี่ยวข้อง และนำค่าน้ำหนักที่คำนวณได้ไปหาเซตของหมวดหมู่ที่ผู้ใช้สนใจในขั้นตอนที่หนึ่ง ซึ่งมีการทำงานดังภาพที่ 1

* สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง



ภาพที่ 1 *ภาพรวมของระบบ*

จากภาพที่ 1 สามารถอธิบายการทำงานได้โดย เริ่มจากผู้ใช้ป้อนคำค้นหาเข้าสู่ระบบ จากนั้นระบบจะนำคำค้นหาที่ป้อนเข้ามาไปหาเซตของหมวดหมู่ที่ผู้ใช้สนใจ และค้นหาเอกสารที่เกี่ยวข้องซึ่งอยู่ในหมวดหมู่นั้นๆ ขึ้นมา ทำการจัดลำดับและส่งเป็นผลการค้นหาให้แก่ผู้ใช้ เมื่อผู้ใช้คลิกเลือกเอกสารใดเอกสารหนึ่ง ระบบจะนำเอกสารนั้นๆ มาคำนวณหาความสัมพันธ์ และนำค่าน้ำหนักที่ได้ไปปรับปรุงข้อมูลผู้ใช้ของผู้ใช้แต่ละคน โดยใช้ฐานข้อมูลเว็บไซต์จาก ODP (Open Directory Project) ซึ่งเป็นฐานข้อมูลที่รวบรวมเว็บไซต์จากทั่วโลกโดยอาสาสมัครที่มีการคัดเลือกเข้ามา ในปัจจุบันมีข้อมูลเว็บไซต์กว่า 5 ล้านเว็บไซต์ แบ่งออกเป็นหมวดหมู่กว่า 1 ล้านหมวดหมู่ และมีอาสาสมัครผู้รวบรวมเว็บไซต์กว่า 90,000 คน

ในส่วนที่ 2 กล่าวถึงงานวิจัยที่เกี่ยวข้อง ส่วนที่ 3 กล่าวถึงวิธีการดำเนินงานวิจัย การวัดผลและประเมินผล ในส่วนที่ 4 เป็นผลการดำเนินงาน และส่วนสุดท้ายส่วนที่ 5 เป็นสรุปผลการดำเนินงาน

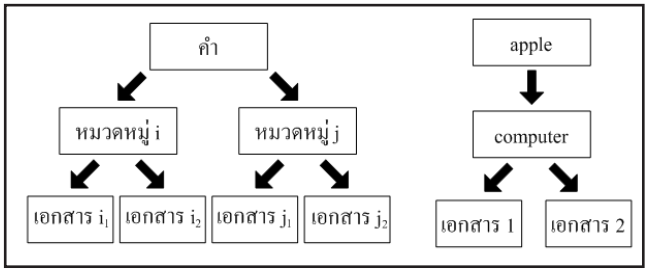
2. งานวิจัยที่เกี่ยวข้อง

ในหลายๆ งานวิจัยก่อนหน้านี้ [1], [2] ได้พัฒนาวิธีการค้นหา และพิจารณาความสัมพันธ์ระหว่างคำค้นหาเข้ากับเซตของหมวดหมู่ หรือหมวดหมู่ที่ได้เลือกเอาไว้ แต่ยังคงส่งผลการค้นหาที่เหมือนๆกันออกมาให้กับผู้ใช้ โดยไม่ได้พิจารณาถึงความสนใจของผู้ใช้ที่ค้นหาคำนั้นๆ ซึ่งอาจกล่าวได้ว่างานวิจัยเหล่านี้ใช้เพียงข้อมูลเริ่มต้นในการแสดงผลการค้นหาให้แก่ผู้ใช้ เท่านั้น ขณะที่ [3] ได้ออกแบบงานวิจัยโดยให้ผู้ใช้งานเลือกหมวดหมู่ที่สนใจก่อนที่จะลงมือค้นหา ซึ่งจะเป็นการสร้างความยากลำบากในการใช้งานและเพิ่มเวลาให้กับผู้ใช้ ส่วน [4] ใช้ความชอบของผู้ใช้ในการค้นหาเอกสารและมีการปรับปรุงคำค้นหาให้กับผู้ใช้ แต่ยังไม่ได้ใช้งานข้อมูลผู้ใช้ในส่วนของการค้นหาข้อมูล โดยจะให้ผู้ใช้ระบุหมวดหมู่ที่ผู้ใช้ต้องการแทน

งานวิจัย [5],[6] เริ่มศึกษาเกี่ยวกับความสนใจของผู้ที่เข้ามาใช้งานโดยใช้ทั้งข้อมูลเริ่มต้น และข้อมูลผู้ใช้ในการแสดงผลการค้นหาออกมา และจัดหมวดหมู่ไว้ให้ผู้ใช้งาน แต่ยังคงจำกัดความลึกของหมวดหมู่ที่ใช้เพียงแค่ 3 ลำดับชั้นเท่านั้น ซึ่งทำให้การแบ่งหมวดหมู่ที่ได้มีความละเอียดลดลง ส่วน [7] ในระหว่างที่มีการใช้งาน เมื่อผู้ใช้คลิกเลือกเอกสารที่สนใจ ระบบจะไม่คำนึงถึงค่าน้ำหนักของการค้นหาในครั้งก่อนหน้าทำให้ระบบไม่มีการเรียนรู้ความสนใจของผู้ใช้งาน และ [8] ได้จัดการการแบ่งหมวดหมู่ของเอกสาร โดยใช้ฐานข้อมูลของ ODP ให้กับผู้ใช้ แต่ยังไม่ได้พิจารณาถึงความสัมพันธ์ระหว่างหมวดหมู่ที่ผู้ใช้เลือกกับหมวดหมู่ที่เคยเลือกไว้ ซึ่งใน [9], [10] จะมีการพิจารณาถึงความสัมพันธ์ดังกล่าว แต่ยังใช้การแสดงผลการค้นหาจากข้อมูลผู้ใช้เพียงอย่างเดียว

3. วิธีการดำเนินงานวิจัย

ในงานวิจัยนี้ใช้การบันทึกประวัติการใช้งานของผู้ใช้เมื่อผู้ใช้คลิกเลือกเอกสาร เพื่อนำไปจัดทำข้อมูลของผู้ใช้ ซึ่งประวัติการใช้งานที่บันทึกไว้จะอยู่ในรูปแบบความสัมพันธ์ของ คำที่ใช้ค้นหา หมวดหมู่ที่เกี่ยวข้อง และเอกสารที่เกี่ยวข้องกับหมวดหมู่นั้นๆ โดยใช้ความสัมพันธ์ทั้ง 3 ส่วนนี้ในการแสดงผลการค้นหาออกมา สามารถอธิบายได้ดังภาพที่ 2



ภาพที่ 2 *โครงสร้างข้อมูล*

จากภาพที่ 2 จะเห็นว่ารากของโหนดคือคำที่ผู้ใช้ต้องการค้นหา ซึ่งในคำหนึ่งๆ นี้ประกอบไปด้วยหมวดหมู่ที่เกี่ยวข้องและในแต่ละหมวดหมู่ก็ประกอบไปด้วยเอกสารต่างๆที่ถูกจัดอยู่ในหมวดหมู่นั้นๆ โดยเก็บข้อมูลทั้งหมด 3 รูปแบบ คือ

ข้อมูลผู้ใช้ ข้อมูลเริ่มต้น และข้อมูลค่าความนิยมของผู้ใช้ สามารถอธิบายได้ดังนี้

3.1 ข้อมูลผู้ใช้ (User Profile)

ข้อมูลผู้ใช้จะแสดงถึงประวัติความสนใจของผู้ใช้แต่ละคน และมีการปรับปรุงข้อมูลทุกๆ ครั้งที่ผู้ใช้เข้ามาใช้งาน โดยจะแสดงเป็นค่าน้ำหนักซึ่งจะบ่งบอกถึงค่าความสนใจของผู้ใช้ที่มีต่อหมวดหมู่ต่างๆ เช่น คำว่า “แอปเปิล” มีค่าน้ำหนักในหมวดหมู่อาหารมากกว่าค่าน้ำหนักในหมวดหมู่คอมพิวเตอร์ แสดงว่าผู้ใช้สนใจหมวดหมู่อาหารมากกว่าคอมพิวเตอร์ โดยการหาค่าน้ำหนักที่กล่าวมานี้ได้จากการคำนวณเมทริกซ์ 3 เมทริกซ์คือ

1) Document-Term Matrix (DTM) เป็นเมทริกซ์ขนาด m x n ซึ่งแสดงความสัมพันธ์ระหว่างเอกสารกับคำ โดยที่ m คือจำนวนเอกสารที่ผู้ใช้เคยเลือกไว้ และ n คือจำนวนคำที่ไม่ซ้ำกันที่ปรากฏอยู่ในเอกสารนั้นๆ โดยที่ถ้าคำและเอกสารมีความสัมพันธ์กันจะมีค่าน้ำหนักมากกว่า 0 และถ้าไม่มีความสัมพันธ์กันจะมีค่าน้ำหนักเท่ากับ 0 ดังตารางที่ 1 โดยค่าน้ำหนักสามารถคำนวณได้จากสมการที่ (3), (4)

ตารางที่ 1 *Document-Term Matrix*

Doc\Term	apple	macbook	bacon	pie
D1	0.10	0.00	0.00	0.00
D2	0.00	0.07	0.00	0.00
D3	0.11	0.08	0.00	0.00
D4	0.07	0.00	0.16	0.00
D5	0.00	0.00	0.00	0.21

Tf-idf (Term Frequency-inverse document frequency) เป็นการพิจารณาค่าน้ำหนักความสำคัญของคำที่ปรากฏอยู่ในเอกสาร หรือในคลังข้อมูล โดยที่ค่าน้ำหนักจะขึ้นอยู่กับจำนวนครั้งที่คำนั้นๆ ปรากฏในเอกสาร และความถี่ของคำที่ปรากฏในเอกสารทั้งหมด ประกอบไปด้วยการคำนวณ 2 ส่วนคือ

- Tf (Term frequency) จะเป็นค่าความถี่ของคำที่ปรากฏในเอกสาร โดยจะคิดเพียงแค่เอกสารเดียวตามสมการ

$$tf_{i,j} = \frac{n_{i,j}}{\sum_{k^n k,j}} \quad (1)$$

โดยที่ $n_{i,j}$ คือจำนวนครั้งที่คำ i ปรากฏในเอกสาร j และตัวหารคือจำนวนคำทั้งหมดที่อยู่ในเอกสาร j

- idf (inverse document frequency) เป็นค่าที่บ่งบอกถึงความสำคัญของคำนั้นๆ กับจำนวนเอกสารทุกๆ เอกสารที่เก็บอยู่ในฐานข้อมูลตามสมการ

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \quad (2)$$

เมื่อ $|D|$ คือจำนวนเอกสารทั้งหมดของผู้ใช้ และ $|\{d : t_i \in d\}|$ คือจำนวนเอกสารที่มีคำ i ปรากฏอยู่ โดยที่

$$(Tf\text{-}idf)_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

LSI (Latent Semantic Indexing) เป็นการพิจารณาค่าน้ำหนักจากความถี่ของคำที่ปรากฏในเอกสารและเอกสารอื่นๆของผู้ใช้ และพิจารณาไปถึงคำที่ใช้ในความหมายที่ใกล้เคียงกัน เช่น drink, drinking, drinker ซึ่งสามารถคำนวณได้จากสมการที่ (4)

$$a_{ij} = g_i \log(tf_{i,j} + 1) \quad (4)$$

โดยที่ g_i จะอธิบายถึงความสัมพันธ์ของความถี่ของคำที่อยู่ในเอกสารที่ผู้ใช้เคยเลือกไว้ ดังสมการ

$$g_i = 1 + \sum_j g_i (p_{ij} \log (p_{ij})(\log(n)) \quad (5)$$

เมื่อ $p_{ij} = tf_{ij} / gf_i$ โดยที่ gf_i คือจำนวนครั้งที่คำ i ปรากฏอยู่ในเอกสารทั้งหมด tf_{ij} คือจำนวนครั้งที่คำ i ปรากฏอยู่ในเอกสาร j และ n คือจำนวนเอกสารทั้งหมด

2) Document-Category Matrix (DCM) เป็นเมทริกซ์แสดงความสัมพันธ์ ระหว่างเอกสารกับหมวดหมู่ที่เกี่ยวข้องโดยแต่ละแถวคือเอกสารที่ผู้ใช้เคยเลือกไว้ และแต่ละคอลัมน์คือหมวดหมู่ที่เกี่ยวข้อง ซึ่งถ้าเอกสารและหมวดหมู่มีความสัมพันธ์กันจะให้ค่าน้ำหนักเท่ากับ 1 และถ้าไม่มีความสัมพันธ์กันจะให้ค่าน้ำหนักเท่ากับ 0 ดังตารางที่ 2

ตารางที่ 2 *Document-Category Matrix*

Doc\Cat	MACINTOSH	MACBOOK	FRUIT	SANDWICH
D1	1.00	0.00	0.00	0.00
D2	0.00	1.00	0.00	0.00
D3	0.00	1.00	0.00	0.00
D4	0.00	0.00	0.00	1.00
D5	0.00	0.00	1.00	0.00

โดยที่หมวดหมู่ MACINTOSH, MACBOOK เป็นหมวดหมู่ย่อยภายใต้หมวดหมู่ COMPUTER และหมวดหมู่ FRUIT,

SANDWICH เป็นหมวดหมู่ย่อยภายใต้หมวดหมู่ COOK

การวัดความเหมือนของหมวดหมู่ (Hierarchical similarity measure)[10] เป็นการวัดความคล้ายคลึงของโหนดแต่ละโหนดที่อยู่ในรูปแบบโครงสร้างต้นไม้ ซึ่งงานวิจัยนี้นำมาใช้วัดค่าความสัมพันธ์ของเอกสารกับหมวดหมู่ที่เกี่ยวข้อง ทำให้ได้ค่าน้ำหนักที่เปลี่ยนไปดังตารางที่ 3 โดยมีวิธีคำนวณดังสมการที่ (6)

$$S(i,j) = e^{-\alpha.l} \cdot \frac{e^{\beta.h} - e^{-\beta.h}}{e^{\beta.h} + e^{-\beta.h}},\ \alpha = 0.2,\ \beta = 0.6 \qquad (6)$$

เมื่อ h หมายถึงความลึกสูงสุดของโหนดที่เชื่อมระหว่างสองโหนดที่ต้องการวัด, l คือจำนวนเส้นเชื่อมระหว่างโหนดที่ต้องการวัด และ (i,j) คือโหนดที่ต้องการวัด โดยที่ค่า α, β คือค่าคงที่ที่เหมาะสมซึ่งได้จากงานวิจัยก่อนหน้านี้ [11]

ตารางที่ 3 *Document-Category Matrix(Similarity Measure)*

Doc\Cat	MACINTOSH	MACBOOK	FRUIT	SANDWICH
D1	1.00	0.54	0.10	0.10
D2	0.54	1.00	0.05	0.05
D3	0.54	1.00	0.05	0.05
D4	0.10	0.05	0.42	1.00
D5	0.10	0.05	1.00	0.42

3) Category-Term Matrix (CTM) สร้างมาจากความสัมพันธ์ของเมทริกซ์ DTM และ DCM เป็นเมทริกซ์ที่แสดงถึงความสัมพันธ์ระหว่างหมวดหมู่และคำที่อยู่ในเอกสารของผู้ใช้ โดยที่แต่ละแถวจะแสดงถึงหมวดหมู่ที่ผู้ใช้สนใจ และแต่ละคอลัมน์จะแสดงถึงคำที่ปรากฏอยู่ในเอกสารดังตารางที่ 4 และ ตารางที่ 5 (ใช้การวัดความเหมือนของหมวดหมู่) โดยสามารถคำนวณค่าน้ำหนักได้ตามสมการที่ (7)

ตารางที่ 4 *Category-Term Matrix*

Cat\Term	apple	macbook	bacon	pie
MACINTOSH	0.11	0.00	0.00	0.00
MACBOOK	0.10	0.21	0.00	0.00
FRUIT	0.06	0.00	0.00	0.20
SANDWICH	0.06	0.00	0.31	0.00

ตารางที่ 5 *Category-Term Matrix (Similarity Measure)*

Cat\Term	apple	macbook	bacon	pie
MACINTOSH	0.20	0.11	0.03	0.02
MACBOOK	0.18	0.21	0.01	0.01
FRUIT	0.12	0.01	0.13	0.20
SANDWICH	0.12	0.01	0.31	0.09

Rocchio-Base Algorithm [3] ส่วนใหญ่จะใช้ในงานเกี่ยวกับการป้อนความเกี่ยวพันธ์ย้อนกลับ (relevance feedback method) ซึ่งในงานวิจัยนี้ใช้สมการของ aRocchio (adaptive Rocchio-Base Algorithm) [5] ในการคำนวณ โดย aRocchio จะมีการเรียนรู้ความสนใจของผู้ใช้ จากการนำค่าน้ำหนักการค้นหาครั้งที่ผ่านมาของผู้ใช้มาคำนวณด้วย ดังสมการ

$$M(i,j)^t = \frac{N_i^{t-1}}{N_i^t} M(i,j)^{t-1} + \frac{1}{N_i^t} \sum_k^m DT(k,j) \times DC(k,i) \quad (7)$$

เมื่อ M^t คือค่าน้ำหนักที่เวลา t (เวลาปัจจุบัน) N_i^t คือจำนวนเอกสารที่เกี่ยวข้องกับหมวดหมู่ i ทั้งหมด, $M(i,j)^{t-1}$ คือค่าน้ำหนักของคำ j ที่สัมพันธ์กับหมวดหมู่ i เมื่อเวลา $t-1$ (เวลาการใช้งานครั้งที่ผ่านมา) m จำนวนคือเอกสารที่อยู่ใน DTM, $DT(k,j)$ คือค่าน้ำหนักของคำ j ที่สัมพันธ์กับเอกสาร k, $DC(k,i)$ คือค่าน้ำหนักของเอกสาร k ที่สัมพันธ์กับหมวดหมู่ i หรืออาจกล่าวได้ว่า $M(i,j)$ ก็คือ ค่าน้ำหนักของคำ j ในทุกๆ เอกสารที่เกี่ยวข้องกับหมวดหมู่ i

3.2 ข้อมูลเริ่มต้น (Predetermined Profile)

ข้อมูลเริ่มต้นคือข้อมูลที่แสดงค่าน้ำหนักของความสัมพันธ์ระหว่าง เอกสาร หมวดหมู่ และคำ ซึ่งประกอบไปด้วยเมทริกซ์ 3 เมทริกซ์เช่นเดียวกับข้อมูลผู้ใช้ โดยจะคำนวณค่าน้ำหนักทั้งหมดในแบบจำลองข้อมูลเริ่มต้นเพียงครั้งเดียวจากฐานข้อมูล ODP ก่อนที่จะเปิดให้ผู้ใช้เข้ามาใช้งาน ซึ่งค่าน้ำหนักในข้อมูลเริ่มต้นนี้จะถูกนำไปใช้คำนวณหาผลการค้นหาให้กับผู้ใช้ทุกๆ คน

3.3 ข้อมูลค่าความนิยมของผู้ใช้ (Popular Profile)

ข้อมูลค่าความนิยมของผู้ใช้ เป็นข้อมูลที่รวบรวมประวัติการใช้งานของผู้ใช้ทั้งหมด ใช้การคำนวณค่าน้ำหนักเช่นเดียว กับข้อมูลผู้ใช้ ซึ่งจะคำนวณค่าน้ำหนักทั้งหมดของผู้ใช้เข้าด้วยกันไม่แยกเป็นรายบุคคล โดยค่าน้ำหนักของข้อมูลค่าความนิยมนี้จะมีค่ามากหรือน้อยขึ้นอยู่กับแนวโน้มการใช้งานของผู้ใช้ทุกๆ คนที่เข้ามาใช้งาน ประกอบด้วยเมทริกซ์ 3 เมทริกซ์เช่นเดียวกัน และนำไปคำนวณผลการค้นหาให้กับผู้ใช้ทุกๆ คน

3.4 การจัดลำดับผลการค้นหา

การจัดลำดับผลการค้นหาเป็นการจัดลำดับจากค่าน้ำหนักที่คำนวณได้จากชุดข้อมูลต่างๆ ซึ่งจะเรียงจากค่าน้ำหนักที่มากที่สุด (ผู้ใช้ให้ความสนใจมาก) ไปยังค่าน้ำหนักที่น้อยที่สุด (ผู้ใช้ให้ความสนใจน้อย) โดยใช้การรวมค่า

น้ำหนักของแต่ละชุดข้อมูลดังสมการที่ (8)

$$AvgProfile = \sum_{i=1}^{n} (Profile_i)ln \quad (8)$$

โดยที่ $Profile_i$ คือค่าน้ำหนักของข้อมูลเริ่มต้น ข้อมูลผู้ใช้ หรือข้อมูลค่าความนิยมที่ต้องการจัดลำดับร่วมกัน และ n คือ จำนวนชุดข้อมูลที่นำมาจัดลำดับร่วมกัน

3.5 การประเมินผล

AvgRank (Average rank) เป็นการประเมินผลการค้นหาว่าระบบมีการแสดงผลที่ตรงกับความสนใจของผู้ใช้หรือไม่ โดยวัดจากลำดับการเลือกเอกสารของผู้ใช้ ถ้าระบบมีการแสดงผลการค้นหาที่ดี เอกสารที่ผู้ใช้สนใจควรจะถูกจัดวางไว้ตำแหน่งต้นๆ ของผลการค้นหา และควรถูกคลิกเลือกโดยผู้ใช้ ซึ่งสามารถวัดผลได้ดังสมการที่ (9)

$$AvgRank(q) = \sum_{p \in S} R(p)/|S|(Profile_i)ln \quad (9)$$

โดยที่ S คือเซตของเอกสารที่ถูกเลือกโดยผู้ใช้สำหรับการค้นหาด้วย q, $R(p)$ คือตำแหน่งของเอกสาร p ที่อยู่ในผลการค้นหา, $|S|$ คือจำนวนเอกสารที่ผู้ใช้เลือก (จำนวนสมาชิกในเซต S) โดยที่ค่า AvgRank ต่ำๆ จะแสดงถึงประสิทธิภาพในการแสดงผลที่ดี

DCG (Discount Cumulated Gain-base measure) [12] เป็นการเปรียบเทียบผลการค้นหา โดยให้ผู้ใช้กรอกค่าคะแนนความสนใจแต่ละเอกสารที่อยู่ในผลการค้นหา และรวมค่าคะแนนความสนใจนั้นเข้าด้วยกัน ดังสมการที่ (10)

$$DCG(i)\begin{cases} G[1] & \text{If i = 1} \\ DCG[i-1]+G[i]/\log[i] & \text{Otherwise} \end{cases} \quad (10)$$

โดยที่ i คือตำแหน่งของเอกสาร, $G[i]$ คือค่าคะแนนความสนใจของผู้ใช้ที่มีต่อเอกสารตำแหน่งที่ i ซึ่งในการทดลองกำหนดค่าคะแนนความสนใจเป็น 3 ระดับคือ $G[i] = 2$ หมายถึงผู้ใช้สนใจมาก $G[i] = 1$ หมายถึงผู้ใช้สนใจ และ $G[i] = 0$ หมายถึงผู้ใช้ไม่สนใจ ซึ่งค่า DCG ที่สูงจะแสดงถึงประสิทธิภาพในการจัดลำดับที่ดี

4. ผลการทดลอง

ในการเก็บผลการทดลองใช้ข้อมูลเว็บไซต์จากฐานข้อมูล ODP ซึ่งปัจจุบันมีจำนวน 5,170,000 เว็บไซต์ แบ่งออกเป็นหมวดหมู่ 1,017,500 หมวดหมู่ โดยให้อาสาสมัครจำนวน 20 คนเข้ามาใช้งาน และบันทึกข้อมูลการค้นหาของผู้ใช้แต่ละคนไว้ แบ่งการทดลองเป็น 3 ชุดด้วยกันคือ

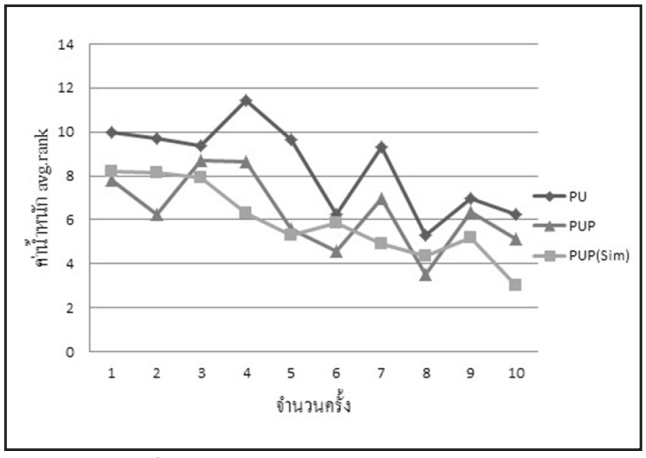
1) การแสดงผลการค้นหาโดยใช้ข้อมูลเริ่มต้น และข้อมูลผู้ใช้ (PU)

2) การแสดงผลการค้นหาโดยใช้ข้อมูลเริ่มต้น ข้อมูลผู้ใช้ และข้อมูลค่าความนิยม (PUP)

3) การแสดงผลการค้นหาโดยใช้ข้อมูลเริ่มต้น ข้อมูลผู้ใช้ ข้อมูลค่าความนิยม และการวัดความเหมือนระหว่างหมวดหมู่ (PUP(Sim)) ซึ่งได้ค่าเฉลี่ยการใช้งานต่อผู้ใช้ดังตารางที่ 6

ตารางที่ 6 *ค่าเฉลี่ยการใช้งานต่อคน*

สถิติ	จำนวน
จำนวนการค้นหา	33.00
จำนวนการคลิก	64.00
จำนวนหมวดหมู่ที่เลือก	11.00
จำนวนคำในเอกสารที่ผู้ใช้สนใจ	513.00
ค่าเฉลี่ยการคลิกต่อหมวดหมู่	5.82
ค่าเฉลี่ยการคลิกต่อการค้นหา	1.94



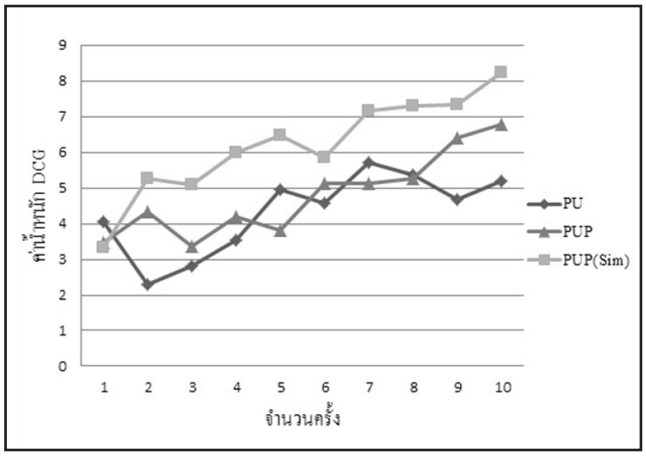
ภาพที่ 3 *ค่าเฉลี่ยลำดับการคลิกเลือกเอกสาร*

การทดลองที่ 1 เป็นการวัดผลค่าเฉลี่ยลำดับการคลิกเลือกเอกสารของผู้ใช้ในการค้นหาจำนวน 10 ครั้ง ซึ่งจากภาพที่ 3 เห็นได้ว่า

• วิธี PU จะให้ค่าเฉลี่ยลำดับการคลิกที่สูงที่สุด (แย่สุด) และเมื่อมีการใช้งานเรื่อยๆกราฟมีแนวโน้มที่ลดลงซึ่งแสดงให้เห็นถึงผลการค้นหาที่ดีขึ้น

• วิธี PUP ให้ค่าเฉลี่ยลำดับการคลิกที่ดีกว่าแบบแรกและกราฟมีลักษณะที่ลดต่ำลงเช่นเดียวกัน

• วิธี PUP(Sim) จะเห็นว่ามีค่าเฉลี่ยลำดับการคลิกดีที่ดีกว่าวิธี PUP อยู่เล็กน้อย แต่กราฟมีอัตราการแกว่งที่น้อยลง เนื่องจากการวัดความเหมือนระหว่างหมวดหมู่ของผู้ใช้จะช่วยให้การเรียงลำดับผลการค้นหาที่ผู้ใช้สนใจมีความละเอียดยิ่งขึ้น



ภาพที่ 4 *ประสิทธิภาพความสนใจของผู้ใช้*

การทดลองที่ 2 เป็นการวัดประสิทธิภาพจากความสนใจของผู้ใช้ โดยให้ผู้ใช้เลือกคะแนนความสนใจในเอกสาร 20 ลำดับแรก ซึ่งจากภาพที่ 4 จะเห็นว่า

• ทั้ง 3 วิธีจะให้กราฟที่สูงขึ้นเรื่อยๆ เมื่อมีการใช้งานแสดงถึงผลการค้นหาที่ดีขึ้น

• การแสดงผลการค้นหาโดยวิธี PU จะให้ประสิทธิภาพที่แย่สุด รองลงมาคือ การแสดงผลการค้นหาโดยวิธี PUP

• การแสดงผลการค้นหาโดยวิธี PUP(Sim) จะให้ประสิทธิภาพที่ดีที่สุด เนื่องจากการวัดความเหมือนทำให้ได้หมวดหมู่ที่ใกล้เคียงกับหมวดหมู่ที่ผู้ใช้สนใจมากยิ่งขึ้น

5. สรุป

งานวิจัยนี้ได้นำเสนอวิธีการแสดงผลการค้นหา โดยอ้างอิงจากความชอบของผู้ใช้ ซึ่งประกอบไปด้วยข้อมูลทั้งหมด 3 รูปแบบคือ ข้อมูลเริ่มต้น, ข้อมูลผู้ใช้ และข้อมูลค่าความนิยมของผู้ใช้ หลังจากเปิดให้ผู้ใช้เข้ามาทดลองใช้งาน พบว่าการแสดงผลการค้นหาโดยวิธี PUP(Sim) ให้ประสิทธิภาพในการค้นหาดีที่สุดเมื่อเปรียบเทียบกับวิธีอื่นๆ ซึ่งการมีข้อมูลค่าความนิยมของผู้ใช้จะช่วยให้สามารถอ้างอิงหมวดหมู่จากผู้ใช้คนอื่นได้ในกรณีที่ประวัติการใช้งานของผู้ใช้ยังมีค่าไม่เพียงพอ และการวัดความเหมือนของหมวดหมู่จะเป็นการกระจายค่าน้ำหนักให้กับหมวดหมู่ที่ใกล้เคียงกันส่งผลให้สามารถแสดงผลการค้นหาเฉพาะหมวดหมู่ที่ผู้ใช้สนใจได้ และในงานวิจัยที่จะดำเนินการต่อ วางแผนที่จะนำไปทดลองกับผู้ใช้งานที่มากขึ้น เปิดให้ใช้งานระยะเวลานานขึ้น และมีการประเมินผลเพิ่มขึ้น ส่วนข้อมูลค่าความนิยมอาจแยกออกมาในลักษณะของเอกสารที่แนะนำ แทนที่จะรวมเข้าไปกับข้อมูลเริ่มต้นและข้อมูลผู้ใช้

6. เอกสารอ้างอิง

[1] R. Dolin, D. Agrawal, A. EI Abbadi, and J. Pearlman. "Using Automated Classification for Summarizing and Selecting Heterogeneous Information Sources." *D-Lib Magazine*, 1998.

[2] C. Yu, W. Meng, W. Wu and K.-L. Liu. "Efficient and Effective Metasearch for Text Databases Incorporating Linkages among Documents." *ACM SIGMOD international conference on Management of data*, New York, NY, USA, pp. 187-198, 21-24 June, 2001.

[3] J. Rocchio, "Relevance Feedback in Information Retrieval." *The Smart Retrieval System: Experiments in Automatic Document Processing*, pp. 313-323, 1971.

[4] E. J. Glover, G. W. Flake, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles, D.Pennock. "Improving Category Specific Web Search by Learning Query Modifications." *IEEE trans. on Applications and the Internet*, pp. 23-32, 2001.

[5] F. Liu, C. Yu and W. Meng. "Personalized Web Search for Improving Retrieval Effectiveness." I*EEE Trans. On Knowledge and Data Engineering*, Vol. 16, No. 1, pp. 28-40, January, 2004.

[6] J. Garofalakis , T. Giannakoudi , and A. Vopi. "Personalized Web Search by Constructing Semantic Clusters of User Profiles." *Knowledge-Based Intelligent Information and Engineering Systems* (*KES08*) Part2, pp. 238-247, 2008.

[7] M.R. Surnalathal, V. Vaidehi, A. Kannan3 and S. Anandhi. "Information Retrieval using Semantic Web Browser–Personalized and Categorical Web Search." *IEEE Trans. on Signal Processing Communications and Networking*

(*ICSCN 2007)*, pp. 238-243, 22-24 February, 2007.

[8] T. Bounoy and A. Walairacht. "User Preference Retrieval using Semantic Categorization for Web Search." *IEEE Trans. Advanced Communication Technology* (*ICACT*), Vol. 2, pp. 1133-1138, 2010.

[9] L. Li , Z.Yang , B. Wang and M. Kitsuregawa. "Dynamic Adaptation Strategies for Long-Term and Short-Term User Profile to Personalize Search." *APWeb/ WAIM'07*, pp. 228-240, 2007.

[10] L. Li, Z.Yang and M. Kitsuregawa. "Using Ontology-Based User Preferences to Aggregate Rank Lists in Web Search." *Pacific-Asia Conference Advances in Knowledge Discovery and Data Mining* (*PAKDD 2008)*, pp. 923-931, 2008.

[11] Y. Li, Z. A. Bandar, and D. McLean. "An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources." *IEEE Trans. Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 871- 882, 2003.

[12] K. Järvelin and J. Kekäläinen. "IR evaluation methods for retrieving highly relevant documents." *ACM SIGIR conference on Research and development in information retrieval*, pp. 41-48, 2000.