การวัดประสิทธิผลสำหรับการค้นหาเอกสาร บนพื้นฐานความหลากหลาย

EVALUATION OF RETRIEVAL EFFECTIVENESS IN DIVERSITY-BASED SEARCH

เอก ตั้งสมบูรณ์

AKE TANGSOMBOON

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

สาขาวิชาเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

พ.ศ. 2558

KMITL-2015-IT-M-001-007

การวัดประสิทธิผลสำหรับการค้นคืนเอกสาร บนพื้นฐานความหลากหลาย

EVALUATION OF RETRIEVAL EFFECTIVENESS IN DIVERSITY -BASED SEARCH

เอก ตั้งสมบูรณ์

AKE TANGSOMBOON

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2558

KMITL-2015-IT-M-001-007

EVALUATION OF RETRIEVAL EFFECTIVENESS IN DIVERSITY-BASED SEARCH

AKE TANGSOMBOON

A THESIS SUBMITTED IN FULFILLMENT OF
THE REQUIREMENT FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2015

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองวิทยานิพนธ์

หัวข้อวิทยานิพนธ์ การวัดประสิทธิผลสำหรับการค้นคืนเอกสารบนพื้นฐานความหลากหลาย

Evaluation of retrieval effectiveness in diversity-based search

นักศึกษา นายเอก ตั้งสมบูรณ์

รหัสประจำตัว 57606156

ปริญญา วิทยาศาสตรมหาบัณฑิต

สาขาวิชา เทคโนโลยีสารสนเทศ

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ธีรพงศ์ ลีลานุภาพ

คณะกรรมการสอบวิทยานิพนธ์	ลายมือชื่อ
ผู้ช่วยศาสตราจารย์ ดร.กันต์พงษ์ วรรัตน์ปัญญา	
ผู้ช่วยศาสตราจารย์ ดร.สุพจน์ นิตย์สุวัฒน์	
ผู้ช่วยศาสตราจารย์ ดร.ธีรพงศ์ ลีลานุภาพ	
รองศาสตราจารย์ ดร.พรฤดี เนติโสภากุล	
ดร.ณฐพล พันธุวงศ์	

วัน / เดือน/ ปี ที่สอบ วันศุกร์ที่ 17 กรกฎาคม 2558 เวลา 09.30 น. เป็นต้นไป

สถานที่สอบ ณ ห้อง 333 ชั้น 3 คณะเทคโนโลยีสารสนเทศ

คณะเทคโนโลยีสารสนเทศรับรองแล้ว

(รองศาสตราจารย์ ดร.จันทร์บูรณ์ สถิตวิริยวงศ์)

คณบดีคณะเทคโนโลยีสารสนเทศ

วันที่ 27 **เดือน** กรกฎาคม **พ.ศ.** 2558

หัวข้อวิทยานิพนธ์	การวัดประสิทธิผลสำหรับการค้นคืนเอกสารบนพื้นฐานความหลากหลาย
นักศึกษา	นายเอก ตั้งสมบูรณ์
รหัสนักศึกษา	57606156
ปริญญา	วิทยาศาสตรมหาบัณฑิต
สาขาวิชา	เทคโนโลยีสารสนเทศ
แขนงวิชา	วิทยาการสารสนเทศ
พ.ศ.	2558
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. ธีรพงศ์ ลีลานุภาพ

บทคัดย่อ

การค้นคืนเอกสารบนพื้นฐานความหลายหลาย (Diversity-based search) เป็นหนึ่งในหัวข้องานวิจัยในสาขาการค้นคืนสารสนเทศ (Information retrieval) โดยการค้นคืนที่เน้นความหลากหลาย มีจุดประสงค์เพื่อพัฒนาเสิร์ชเอนจินที่ตอบสนองความต้องการสารสนเทศของผู้ใช้ที่แตกต่างกัน หรือตอบสนองความต้องการย่อยที่เกี่ยวข้องกับหัวข้อการค้นหา นอกจากนี้ผู้ใช้มักจะใส่คำค้นหาที่กำกวมหรือไม่ระบุเจาะจง ดังนั้นงานวิจัยจำนวนมากจึงมุ่งเน้นที่จะสร้างระบบการค้นคืนสารสนเทศที่ทำให้ผลการค้นหามีความหลากหลาย ซึ่งสามารถตอบสนองความต้องการทั้งหลายภายใต้คำค้นหาหนึ่งๆ โดยการวัดประสิทธิผลของระบบดังกล่าวจำเป็นที่ต้องพัฒนาตัวชี้วัดที่ถูกสร้างมาโดยเฉพาะ ซึ่งมีนักวิจัยจำนวนมากกำลังศึกษา และพัฒนาตัววัดประสิทธิผลสำหรับการค้นคืนเอกสารประเภทความหลากหลายและความซ้ำซ้อน โดยในงานวิทยานิพนธ์นี้ได้บ่งชี้ถึงปัญหาของตัวชี้วัดที่ถูกใช้ในปัจจุบัน และสร้างตัววัดประสิทธิผลเน้นความหลากหลาย โดยตัวชี้วัดใหม่จะถูกประเมินในแง่ของแนวคิดรากฐาน ความสอดคล้องระหว่างตัวชี้วัด ความสอดคล้องระหว่างตัวชี้วัดสำหรับความหลากหลายกับความชอบของผู้ใช้ และความน่าเชื่อถือของตัวชี้วัด โดยผลของการทดสองแสดงให้เห็นว่าตัวชี้วัดใหม่มีความสอดคล้องกับตัวชี้วัดอื่น และมีความสอดคล้องกับผู้ใช้ที่มีความต้องการข้อมูลที่มีหลากหลาย และตัวชี้วัดใหม่มีความน่าเชื่อถือกว่าตัวชี้วัดอื่น

Thesis	Evaluation of retrieval effectiveness in diversity-based search
Student	Mr. Ake Tangsomboon
Student ID.	57606156
Degree	Master of Science
Program	Information Technology
Major	Information Science
Year	2015
Thesis Advisor	Asst.Prof. Dr. Teerapong Leelanupab

ABSTRACT

Diversity-based search is one of the research topics in Information Retrieval (IR). The objective of this topic is to develop search systems, which respond to different user information needs or intents related to a search topic. In addition, a user may often enter an ambiguous or underspecified query. To deal with this uncertainty, much recent research has focused on creating IR systems that diversify search results so as to satisfy the multiple possible information needs underlying the query. To validate these IR systems, many new evaluation measures have been proposed to quantify their effectiveness in terms of diversity and redundancy. In order to evaluate the performance of the systems specifically developed for this problem, several effectiveness measures have been introduced such as Intent recall, a family of intent-aware measures and α-nDCG. In this thesis, we identify the current problems found in existing measures and propose a possible solution to cope with the problems and develop a new diversity measure, called Normalized Coverage Frequency (nCF). In addition, we investigate the intuitiveness of our proposed measure, its correlation with existing measures and user preference, and the reliability of our proposed measure. The experiments show that our new measure correlate well with other exsiting measure and with the assessments of user who is interested in diverse serach result. Also, the experiments show that our new measure is more reliable than other measures.

กิตติกรรมประกาศ

วิทยานิพนธ์นี้ฉบับนี้สำเร็จลุล่วงไปด้วยดี ด้วยความช่วยเหลืองจากผู้ช่วยศาสตราจารย์ ดร. ธีรพงศ์ ลีลานุภาพ อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้ทุ่มเทให้คำปรึกษาและเสียสละเวลาอันมีค่าในการให้คำแนะนำเกี่ยวกับแนวคิดในการทำวิทยานิพนธ์ ตรวจแก้ไขความเรียบร้อย รวมทั้งส่งเสริมงานวิจัยให้ออกสู่สายตาชาวโลก ขอบคุณ นายนนท์ คะนึงสุขเกษม สำหรับข้อแนะนำและความช่วยเหลือในทุก ๆ ด้านในการทำวิจัย ขอขอบคุณ Guido Zuccon ที่แนะนำและให้คำปรึกษางานวิจัย รวมทั้งช่วยเหลือและดูแลอย่างดีในการนำเสนอผลงานที่ประเทศออสเตรเลีย

ขอกราบขอบพระคุณคณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังที่ได้สนับสนุนทุนการศึกษาเล่าเรียนในบัณฑิตศึกษาและทุนสนับสนุนงานวิจัยทำให้กระบวนการงานวิจัยสามารถเป็นไปอย่างราบรื่น

สุดท้ายนี้ขอกราบขอบพระคุณบิดามารดาและพี่น้องที่คอยให้การสนับสนุนด้านการศึกษาและค่อยให้กำลังใจจนการเรียนผ่านพ้นไปได้ด้วยดี ขอกราบขอบพระคุณครูบาอาจารย์ทุกท่านที่ได้มอบความรู้ คุณธรรม และการใช้ชีวิตให้แก่ผู้เขียนตั้งแต่ระดับปริญญาตรีและปริญญาโทจนกระทั่งมีวันนี้ได้

เอก ตั้งสมบูรณ์

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของวิทยานิพนธ์

ระบบค้นหาข้อมูลหรือเสิร์ชเอนจิน (Search engine) ถูกใช้อย่างแพร่หลาย เพื่อสืนค้นข้อมูลหรือเว็บไซต์ ที่ตรงกับความความต้องการของผู้ใช้ (Information need) แต่ผู้ใช้แต่ละคนมีความต้องการข้อมูลไม่เหมือนกัน แต่ใส่คำค้นหา (Query) เดียวกันและมักเป็นคำค้นหาที่สั้นๆ ส่งผลให้คำค้นหานั้นมีความกำกวม (Ambiguous) หรือ เจาะจงไม่เพียงพอ (Underspecified) เช่น คำค้นหา "Panda" มีความกำกวม เพราะสามารถหมายถึง สัตว์เลี้ยงลูกด้วยนม แอนิเมชั่น โปรแกรมแอนตี้ไวรัส เป็นต้น หรือแม้ว่าคำค้นหานั้นไม่มีความกำกวม แต่อาจระบุไม่เจาะจง ทำให้ถูกตีความได้หลายแง่มุม เช่น "Kungfu panda" อาจระบุได้ถึงหลายความแง่มุม เช่น รายละเอียดของหนัง รีวิวหนัง เกม รอบหนัง ดังนั้นระบบค้นหาข้อมูลจำเป็นต้องคืนเว็บไซต์ให้ครอบคลุมทุกความหมายหรือทุกแง่มุม เพื่อตอบสนองความต้องการของผู้ใช้แต่ละคนที่มีความต้องการที่แตกต่างกัน กล่าวคือ การทำให้ผลการค้นหาเกิดความหลากหลาย (Diversity) เพื่อครอบคลุมความต้องการของผู้ใช้ทั้งหมด

นอกจากนี้ความหลากหลายยังสามารถเกิดได้จากกรณีที่ผู้ใช้อาจจะต้องการข้อมูลที่มีความหลากหลาย เช่น ผู้ใช้ที่ต้องการซื้อมือถือใหม่อยากได้ข้อมูลทั้งหมดที่เกี่ยวกับมือถือเช่น ข้อมูลมือถือแต่ละยี่ห้อ รีวิว ราคา สเปค ข่าว เป็นต้น หรือ ผู้ใช้อาจมีความไม่แน่ใจว่าตนเองกำลังจะค้นหาเกี่ยวกับอะไร ดังนั้นผู้ใช้จึงต้องการรู้ข้อมูลโดยกว้างๆ เพื่อให้ผู้ใช้สามารถระบุความต้องการของตนเองได้ชัดเจนมากขึ้น โดยจากปัญหาดังกล่าวทำให้เกิดงานวิจัยขึ้นจำนวนมากในการทำให้ผลลัพธ์มีความหลากหลาย โดยเรียกกลุ่มงานวิจัยนี้ว่าการค้นคืนเอกสารบนพื้นฐานความหลากหลาย (Diversity-based search) ซึ่งหนึ่งในตัวอย่างคือการพิจารณาเอกสารที่จะนำมาถูกจัดอันดับโดยคำนึงถึงเอกสารอื่นที่ถูกจัดอันดับไปแล้ว

นักวิจัยได้พัฒนาตัวชี้วัดประสิทธิผล (Effectiveness measure) ในสาขาการค้นคืนสารสนเทศ (Information Retrieval - IR) เพื่อนำมาใช้ประเมินประสิทธิผลของระบบค้นคืน โดยตัวชี้วัดแต่ละประเภทจะประเมินแตกต่างกันตามประเภทของการค้นหา โดยการค้นคืนแบบเฉพาะกิจ (Ad-hoc retrieval)[1] จะมีตัวชี้วัด เช่น Precision [1], Recall [1] และ F-measure [2] แต่อย่างไรก็ตามตัวชี้วัดเหล่านี้ไม่เหมาะกับการประเมินระบบค้นคืนที่เน้นหลากหลาย ดังนั้นจึงเกิดตัวชี้วัดสำหรับการค้น

---

[1] คำว่าเฉพาะกิจ หมายถึง การค้นหาที่ระบบค้นคืนเอกสารโดยพิจารณาความเกี่ยวข้องของเอกสารให้ตรงกับคำค้นหามากที่สุด

คืนแบบหลากหลายขึ้น โดยเริ่มจากการแยกความต้องการในคำค้นหาหนึ่งๆ ให้เป็นส่วนๆซึ่งเรียกว่าอินเท็น (Intent) หรือความต้องการย่อย อย่างเช่น คำค้นหา "Panda" ซึ่งมีอินเท็นคือ สัตว์เลี้ยงลูกด้วยนม แอนิเมชั่น โปรแกรมแอนตี้ไวรัส เป็นต้น โดยใช้อินเท็นเหล่านี้เพื่อประเมินประสิทธิผลของการค้นคืนแบบหลากหลาย ซึ่งมีวิธีการประเมินความหลากหลายมี 2 ลักษณะคือ *1)* พิจารณาจากจำนวนของอินเท็นที่เกี่ยวข้องในผลการค้นหา โดยเรียกการประเมินนี้ว่า "การประเมินประเภทความหลากหลาย" *2)* การลงโทษระบบเมื่อมีเว็บไซต์ที่มีอินเท็นที่ซ้ำซ้อน หรือเรียกการประเมินแบบนี้ว่า "การประเมินประเภทความซ้ำซ้อน" โดยจากวิธีประเมินที่แตกต่างกันนี้ ทำให้เกิดตัวชี้วัดประเภทความหลากหลาย (Diversity-based measure) และตัวชี้วัดประเภทความซ้ำซ้อน (Redundancy-based measure) ตามลำดับ

1.2 ความมุ่งหมายและวัตถุประสงค์ของการศึกษา

1. เพื่อวิจัยและพัฒนาตัวชี้วัดประเภทความหลากหลายเพื่อประเมินประสิทธิผลของระบบค้นคืนแบบหลากหลาย
2. เพื่อศึกษาและวิเคราะห์ความถูกต้อง ความสอดคล้อง รวมทั้งวิเคราะห์แนวคิดของตัวชี้วัดใหม่ประเภทความหลากหลาย
3. เพื่อวัดความน่าเชื่อถือของในการประเมินของตัวชี้วัดใหม่ เปรียบเทียบกับตัวชี้วัดอื่นๆ

1.3 แนวคิดที่ใช้ในการวิจัย

ตัวชี้วัด Novelty-biased cumulative gain (α-nDCG) [3] และ Intent Mean Reciprocal Rank (ERR-IA) [4] เป็นตัวชี้วัดประเภทความซ้ำซ้อน ซึ่งมีข้อเสียในการประเมินความหลากหลายเพราะเป็นการประเมินที่ให้ความสำคัญกับเอกสารลำดับบนๆ ทำให้ละเลยความหลากหลายในผลการค้นหา ในขณะเดียวกันตัวชี้วัด Intent recall (*I*-rec) [5] เป็นตัวชี้วัดประเภทความหลากหลาย ซึ่งไม่สามารถประเมินระบบค้นคืนได้ทั่วทั้งระบบ เนื่องจากเมื่อระบบคืนเอกสารที่ครอบคลุมอินเท็นทั้งหมดแล้ว *I*-rec ไม่สามารถแยกแยะประสิทธิผลของระบบได้ นอกจากนี้การประเมินประเภทความหลากหลายและการประเมินประเภทความซ้ำซ้อนมีความแตกต่างกัน การที่นำตัวชี้ประเภทความหลากหลายและความซ้ำซ้อนมาประเมินร่วมกันทำให้ผลการประเมินตีความได้ยาก ดังนั้นจึงเกิดเป็นแนวคิดให้ผู้วิจัยพัฒนาตัวชี้วัดตัวใหม่ เพื่อประเมินความหลากหลาย โดยตัวชี้วัดประเภทความหลากหลายนี้มีชื่อว่า Normalized Coverage Frequency (nCF) [6] ซึ่งมีการประยุกต์ใช้แนวคิดจ าก *I*-rec โดย nCF สามารถประเมินระบบค้นคืนได้ทั่วทั้งระบบ และสามารถกำจัดข้อเสียในการประเมินผลของตัวชี้วัดอื่นๆ

1.4 คำถามงานวิจัย (Research questions)

RQ1. ตัวชี้วัดประเภทความซ้ำซ้อน ได้แก่ α-nDCG และ ERR-IA สามารถประเมินความหลากหลายของผลการค้นหาหรือไม่

RQ2. ข้อเสียในการประเมินของตัวชี้วัด I-rec สามารถแก้ไขได้อย่างไร

RQ3. ความน่าเชื่อถือของตัวชี้วัดใหม่เป็นอย่างไร เมื่อเปรียบเทียบกับตัวชี้วัดประเภทความหลากหลายและความซ้ำซ้อนในปัจจุบัน

RQ4. ผลการประเมินของตัวชี้วัดใหม่มีความสอดคล้องกับการประเมินของตัวชี้วัดอื่น และการประเมินของผู้ใช้หรือไม่

1.5 การมีส่วนร่วมต่องานวิจัยในสาขาวิชา (Contribution)

C1. งานวิจัยนี้ได้คิดค้นตัวชี้วัดใหม่ประเภทความหลากหลายที่สามารถกำจัดข้อเสียของตัวชี้วัดประเภทความหลากหลายในปัจจุบัน

C2. ตัวชี้วัดใหม่สามารถประเมินความหลากหลายได้มีประสิทธิผลเมื่อพิจารณาแนวคิดรากฐานของตัวชี้วัดในปัจจุบัน

C3. ทำการทดลองเพื่อตรวจสอบความน่าเชื่อถือของตัวชี้วัดใหม่ และเปรียบเทียบกับตัวชี้วัดอื่นในการประเมินความหลากหลาย

C4. ทำการทดลองกับผู้ใช้ เพื่อตรวจสอบตัวชี้วัดใหม่ว่ามีความสอดคล้องกับตัวชี้วัดอื่นๆ และสอดคล้องกับการประเมินของผู้ใช้จริง

1.6 ขอบเขตการวิจัย

1. การใช้คอลเล็กชั่นทดสอบ (Test collection) ซึ่งประกอบไปด้วย เอกสาร (Document) หัวข้อ (Topic) หรือคำค้นหา (Query) และ การตัดสินความเกี่ยวข้อง (Relevance judgment) หรือเรียกว่า qrel
2. มีการกำหนดตัวแปรต้นซึ่งเป็นปัจจัยตามสมมติฐานที่ต้องการตรวจสอบประสิทธิผลของตัวชี้วัด

1.7 ขั้นตอนการวิจัย

1. ศึกษาตัวชี้วัดประเภทความหลากหลายและความซ้ำซ้อนในปัจจุบัน
2. แสดงตัวอย่างให้ตระหนักถึงประเด็นปัญหาของตัวชี้วัดประเภทความหลากหลายและประเภทความซ้ำซ้อนที่ถูกใช้ในปัจจุบัน ซึ่งละเลยความหลากหลายในผลการค้นหา
3. คิดค้นตัวชี้วัดใหม่ประเภทความหลากหลายที่สามารถแก้ปัญหาของตัวชี้วัดในปัจจุบัน
4. ทำการตรวจสอบความถูกต้องของตัวชี้วัดใหม่จากการวิเคราะห์แนวคิดรากฐาน

5. ทำการทดลองหาความสอดคล้องของตัวชี้วัดใหม่กับตัวชี้วัดที่มีอยู่ในปัจจุบัน รวมทั้งการทดลองหาความสอดคล้องกับความชอบของผู้ใช้
6. ทำการทดลองเพื่อวัดความน่าเชื่อถือของตัวชี้วัดใหม่

1.8 นิยามคำศัพท์

1. ระบบค้นหาข้อมูลหรือระบบค้นคืน หมายถึง โปรแกรมในการค้นหาข้อมูลต่าง ๆ ผ่านระบบเว็บไซต์และเครือข่ายอินเตอร์เน็ต เช่น ระบบค้นหาข้อมูลของ google
2. เอกสาร หมายถึง เว็บไซต์
3. เอกสารที่เกี่ยวข้อง หมายถึง เว็บไซต์ที่เกี่ยวข้องกับสิ่งที่ผู้ใช้กำลังค้นหา
4. อินเท็น (Intent) หรือหัวข้อย่อย (Subtopic) คือ ความหมายหนึ่งหรือแง่มุมหนึ่งในคำค้นหา เช่น ค้นหา "Panda" มีอินเท็นที่หมายถึง สัตว์ และ มีอินเท็นที่หมายถึง โปรแกรมแอนตี้ไวรัส เป็นต้น โดยคำว่า "อินเท็น" จะมองในมุมของผู้ใช้ แต่คำว่า "หัวข้อย่อย" จะมองในมุมของเอกสาร
5. ค้นคืน หมายถึง การที่ระบบคืนหรือแสดงผลการค้นหาให้แก่ผู้ใช้
6. ครอบคลุมอินเท็น หมายถึง ผลการค้นหาประกอบด้วยอินเท็นหนึ่งภายใต้คำค้นหานั้นๆ
7. ผลการค้นหา หมายถึง กลุ่มของเว็บไซต์ที่คืนจากระบบค้นหาข้อมูล
8. คำค้นหา หมายถึง คำที่ผู้ใช้ใส่ในระบบค้นหาข้อมูล เพื่อใช้ในการค้นหาเอกสาร
9. ตัวชี้วัด หมายถึง เครื่องมือที่ใช้ประเมินผลการค้นหา
10. การตัดสินความเกี่ยวข้อง หมายถึง สิ่งที่ระบุว่าแต่ละเว็บไซต์ ประกอบด้วยอินเท็นอะไรบ้าง เช่น เอกสารหนึ่งที่ตรงกับคำค้นหา "Panda" ประกอบด้วยอินเท็น แพนด้าที่เป็นสัตว์ เป็นต้น

1.9 ผลงานที่ได้รับการตีพิมพ์

ผลงานตีพิมพ์ตลอดการทำวิทยานิพนธ์นี้ ทั้งหมด 2 บทความ (สามารถดูได้ที่ภาคผนวก ก)

- *On the Reliability of Diversity and Redundancy-Based Search Metrics*

A. Tangsomboon and T. Leelanupab; in Proceedings of the 7th International Conference on Information Technology and Electronical Engineering, ICITEE 2015, Chiang Mai, Thailand, to appear

- *Evaluating Diversity and Redundancy-Based Search Metrics Independently*

A. Tangsomboon and T. Leelanupab; in Proceedings of the 19th Australasian Document Computing Symposium, ADCS 2014, Melbourne, Australia, (**The Best Student Paper Award**)

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยจะกล่าวถึงทฤษฎีและงานวิจัยที่สำคัญซึ่งเกี่ยวข้องกับตัวชี้วัดในปัจจุบันโดยจุดประสงค์ของเนื้อหาในบทนี้เพื่ออธิบายประเด็นสำคัญซึ่งจำเป็นในการศึกษาและออกแบบงานวิจัยสำหรับตัวชี้วัดใหม่

2.1 ประเภทของการค้นหาในระบบค้นคืน

ประเภทของการค้นหาในการค้นคืนนั้นมีความแตกต่างตามเป้าหมายและความตั้งใจในผู้ใช้แต่ละคน อาทิ ผู้ใช้ต้องการซื้อมือถือใหม่ โดยผู้ค้นหาไม่มีความรู้เกี่ยวกับมือถือเลย ดังนั้นผู้ใช้จึงต้องการรู้ข้อมูลเกี่ยวกับมือถือทั้งหมด ไม่ว่าจะเป็นรีวิว สเปคมือถือ ราคา หรือคำค้นหาที่มีความกำกวมหรือระบุไม่ชัดเจน ทำให้มีหลายความหมายหรือมีหลายแง่มุม ดังนั้นระบบควรค้นคืนเอกสารให้มีความหลากหลายเพื่อตอบสนองให้ครอบคลุมความต้องการของคนทั้งหมด โดยได้จำแนกการค้นหาเป็น 2 ประเภทดังนี้

1. การค้นหาแบบเฉพาะกิจ (Ad-hoc retrieval task)

การค้นคืนแบบเฉพาะกิจเป็นการค้นหาที่พบได้ทั่วไป โดยมีสมมุติฐานว่าผู้ใช้มีความต้องการที่ชัดเจน และแปลงความต้องการนั้นให้อยู่ในรูปของคำค้นหา เพื่อนำไปค้นหาข้อมูลในระบบค้นคืน โดยผู้ใช้ต้องการผลการค้นหาที่ตรงกับคำค้นหานั้นมากที่สุด ซึ่งระบบจะไม่มีการเก็บข้อมูลผู้ใช้เพื่อทำนายความต้องการ หรือปรับปรุงผลการค้นหา กล่าวคือ ระบบจะพิจารณาความเกี่ยวข้องระหว่างเอกสาร และคำค้นหาเท่านั้น โดยความเกี่ยวข้องของเอกสารจะถูกพิจารณาเป็นอิสระกับเอกสารอื่นซึ่งปรากฏอยู่ลำดับก่อนหน้า

2. การค้นหาแบบความหลากหลาย (Diversity retrieval task)

การค้นหาแบบความหลากหลาย เป็นการค้นหาที่ผู้ใช้แต่ละคนมีความต้องการที่ชัดเจนแต่ใส่คำค้นหาเดียวกันที่กำกวมหรือเจาะจงไม่เพียงพอ ซึ่ง Song และคณะ[7] ได้ศึกษาและแบ่งหมวดหมู่ของคำค้นหาเป็น 2 ประเภทคือ คำค้นหากำกวม และคำค้นหาที่เจาะจงไม่เพียงพอ โดยคำค้นหาที่กำกวม ทำให้ตีความได้หลายความหมายโดยเฉพาะคำที่มีความยาว 1 -3 คำ [8] หรือแม้แต่คำค้นหานั้นไม่กำกวม แต่อาจจะถูกพิจารณาว่าเป็นคำค้นหาที่เจาะจงไม่เพียงพอ เพราะระบุแง่มุมไม่ชัดเจน [9], [10] โดยตัวอย่างของคำค้นที่กำกวมเช่น เมื่อผู้ใช้ใส่ค้นหาคำว่า "Panda" ในระบบค้นคืน ระบบค้นคืนก็จะคืนเอกสารที่เกี่ยวกับแพนด้าในหลายความหมายเช่น สัตว์เลี้ยงลูกด้วยนม แอนิ

เมชั่น โปรแกรมแอนตี้ไวรัส เป็นต้น และในกรณีคำค้นหาที่ไม่กำกวม เช่น "Panda animal" ระบบค้นคืนก็จะคืนเอกสารได้หลายแง่มุม เช่น ที่มาของแพนด้า ข่าวเกี่ยวกับแพนด้า เป็นต้น นอกจากนี้ ผู้ใช้อาจมีความต้องการข้อมูลที่หลากหลาย หรือผู้ใช้ที่ไม่แน่ใจว่าตนเองกำลังค้นหาอะไร ผู้ใช้จึงต้องการข้อมูลที่หลากหลายเพื่อเติมเต็มความต้องการให้มีความชัดเจน

ดังนั้นคำค้นหาหนึ่งๆ จะประกอบด้วยหลายความต้องการย่อยหรืออินเท็นซึ่งเป็นความต้องการของผู้ใช้ที่แตกต่างกัน สำหรับนิยามของคำว่า "อินเท็น" [5] ได้นำไปใช้การประชุมเชิงปฏิบัติการเช่น Text REtrieval Conference[2] (TREC) 2009 ถึง 2012 เป็นต้น กล่าวคือ อินเท็นเป็นชิ้นส่วนข้อมูลของคำค้นหาหนึ่งๆ โดยผู้ใช้จะพอใจเมื่อได้รับเอกสารที่เกี่ยวข้อง และ เป็นเอกสารที่ประกอบด้วยอินเท็นที่แตกต่างกันภายใต้หนึ่งคำค้นหาหนึ่งๆ

2.2 ปัจจัยที่ให้เกิดความหลากหลายในการค้นคืน

เพื่อให้เข้าใจถึงปัจจัยต่างๆ ที่ทำให้เกิดความหลากหลายในการค้นคืนมากขึ้น ผู้วิจัยจึงจำแนกปัจจัยของความหลากหลายเป็น 3 ปัจจัยหลักดังนี้

1. คำค้นหาของผู้ใช้มีความกำกวมหรือเจาะจงไม่เพียงพอ

ผู้ใช้แต่ละคน มีความต้องการข้อมูลไม่เหมือนกัน แต่ใส่คำค้นหาเดียวกันที่กำกวม ทำให้ตีความได้หลายความหมาย แม้ว่าผู้ใช้จะมีความต้องการที่ชัดเจน เช่น ผู้ใช้ต้องการหาเกี่ยวกับบริษัท Apple โดยใส่คำค้นหา "Apple" ทำให้ตีความได้หลายความหมายคือ บริษัทด้านเทคโนโลยี หรือผลไม้ นอกจากนี้ความหลากหลายสามารถเกิดจากความสนใจของผู้ใช้ที่แตกต่างกัน เนื่องจากในแต่ละเรื่อง คนแต่ละคนก็มีความรู้หรือความสนใจที่แตกต่างกัน ทำให้คำค้นหาแม้ว่าไม่มีความกำกวม แต่มีความเจาะจงไม่เพียงพอ ทำให้คำค้นหามีหลายแง่มุม เช่น คำค้นหา "Iphone6" ผู้ใช้ที่เป็นช่างซ่อมมือถืออาจจะต้องการวิธีการซ่อม หรือ ผู้ใช้ที่เป็นคนซื้ออาจต้องการข้อมูลในการตัดสินใจซื้อ เป็นต้น

2. ผู้ใช้ต้องการข้อมูลที่หลากหลาย

ผู้ใช้มีความต้องการที่ชัดเจนในการค้นหา และต้องการผลการค้นหาที่แตกต่างกัน เพื่อตอบสนองความต้องการนั้น กล่าวคือผลการค้นหาต้องมีความหลากหลาย เนื่องจากการซ้ำซ้อน

---

[2] การประชุมเชิงปฏิบัติการที่มุ่งเน้นงานในด้านการค้นคืนสารสนเทศ โดยในปัจจุบันมีหลายหมวด เช่น การค้นคืนเว็บ การกระจายปัญหาไปยังกลุ่มค้นเพื่อค้นหาคำตอบ เป็นต้น

ของผลการค้นหาทำให้ผู้ใช้ไม่ได้ประโยชน์ ดังนั้นผู้ใช้จึงต้องการรู้ข้อมูลในหลายแง่มุมของหัวข้อที่กำลังค้นหา เช่น รีวิว ความคิดเห็น เป็นต้น

3. ความไม่สมบูรณ์ของความต้องการผู้ใช้

ผู้ค้นหามีความต้องการที่ไม่ชัดเจน จึงต้องการสำรวจข้อมูลที่มีความหลากหลาย [11] เนื่องจากผู้ใช้ไม่แน่ใจว่าเป้าหมายในการค้นหาของตนเองคืออะไร หรือ ไม่มีความคุ้นเคยกับสิ่งที่ต้องการค้นหา ผู้ใช้จึงต้องการเรียนรู้เพื่อที่จะได้สำรวจ เรียนรู้ และเลือกข้อมูลที่สามารถทำให้ความต้องการที่ไม่แน่นอนมีความชัดเจนมากขึ้น หรือเรียกการค้นหาแบบนี้ว่า การค้นหาแบบสำรวจ (Exploratory search) [12]

2.3 หลักการในสร้างผลการค้นหาให้มีความหลากหลาย

การสร้างผลการค้นหาให้มีความหลากหลายสามารถทำได้ 2 หลักการดังต่อไปนี้

2.3.1. การพิจารณาความเกี่ยวข้องของเอกสารโดยขึ้นอยู่กับเอกสารก่อนหน้า (Inter-dependent Document Relevance Paradigm)

การทำให้ผลการค้นมีความหลากหลายทำได้โดยการพิจารณาจากความสัมพันธ์ระหว่างเอกสาร โดยพิจารณาว่าถ้าเอกสารที่กำลังจะค้นคืนมีความซ้ำซ้อนกับเอกสารที่ค้นคืนไปก่อนหน้า ความเกี่ยวข้องของเอกสารนั้นก็จะลดลง ตัวอย่างวิธีการสร้างผลการค้นหาให้เกิดความหลากหลาย เช่น maximal marginal relevance (MRR) [13] ซึ่งมีพารามิเตอร์เพื่อสมดุลระหว่างความเกี่ยวข้องของเอกสาร กับความซ้ำซ้อนของเอกสาร โดยมีสมการดังต่อไปนี้

$$MMR_{J+1} \equiv \underset{x_i \in I \setminus J}{\operatorname{argmax}}[\lambda S(x_i, q) - (1-\lambda) \max_{x_j \in J} S(x_i, x_j)] \qquad (2.1)$$

โดยเซต I คือกลุ่มของเอกสารที่ตรงกับคำค้นหา และเซต J คือกลุ่มของเอกสารที่คืนไปเรียบร้อยแล้ว โดย $S(x_i,q)$ คือความเกี่ยวข้องของเอกสารที่มีต่อคำค้นหา q และ $S(x_i,x_j)$ คือความซ้ำซ้อนระหว่างเอกสารที่ยังไม่ได้คืนกับเอกสารที่ค้นคืนไปเรียบร้อยแล้ว ดังนั้นระบบจะคืนเอกสารที่ยังไม่ถูกค้นคืนที่มีคะแนนสูงที่สุดจากการนำความเกี่ยวข้องของเอกสารลบด้วยความซ้ำซ้อนในเอกสารที่ถูกคืนไปแล้ว โดยมีพารามิเตอร์ γ คอยถ่วงน้ำหนักระหว่างความเกี่ยวข้องกับความซ้ำซ้อน

2.3.2. การแบ่งแยกเอกสารตามหัวข้อย่อยสำหรับการค้นคืนแบบหลากหลาย(Sub-topic Aware Paradigm for Diversity)

การทำให้ผลการค้นหามีความหลากหลายโดยการจัดกลุ่มเอกสาร (Clustering) ที่มีหัวข้อย่อย (หรือเรียกว่าอินเท็น ในมุมมองของผู้ใช้) เดียวกัน ด้วยวิธีการจัดกลุ่มต่างๆ เช่น K-means

[14], PLSA [15] และ LDA [16] เป็นต้น แล้วจึงเลือกเอกสารจากกลุ่มเอกสาร (Cluster) ที่แตกต่างกัน ทำให้สามารถได้เอกสารที่แตกต่างกันในแต่ละกลุ่ม โดยวิธีนึงในการเลือกเอกสาร คือ การนำกลุ่มของเอกสารมาเรียงตามความเกี่ยวข้องของกลุ่มที่มีต่อคำค้นหา (จากมากไปน้อย) แล้วจึงทำการเลือกเอกสารออกมาจากแต่ละกลุ่มวนไปเรื่อยๆ (Round robin) ดังนั้นทำให้ผลการค้นหามีหลากหลายจากกลุ่มของเอกสารที่แตกต่างกัน

2.4 การประเมินระบบค้นคืน

การประเมินระบบค้นคืนนั้นเป็นส่วนสำคัญเพื่อปรับปรุงระบบค้นคืนให้ดียิ่งขึ้น โดยการประเมินสามารถแบ่งเป็น 2 ลักษณะคือ *1)* การประเมินประสิทธิผล คือ การพิจารณาว่าผลการค้นหาตอบสนองความต้องการของผู้ใช้หรือไม่ *2)* ประสิทธิภาพ คือ ระบบสามารถคืนผลการค้นหาได้เร็วหรือไม่ โดยการทดลองส่วนใหญ่ในด้านการค้นคืนสารสนเทศจะมุ่งเน้นในการประเมินในด้านประสิทธิผลของระบบค้นคืน เพื่อทำให้ระบบค้นคืนสามารถตอบสนองความต้องการของผู้ใช้ได้ดียิ่งขึ้น

การประเมินประสิทธิผลของระบบค้นคืนสามารถวัดได้จากความพึงพอใจของผู้ใช้ที่มีต่อระบบค้นคืน อาทิ ประเมินว่าผู้ใช้พอใจกับผลการค้นหาที่คืนมาจากระบบค้นคืนหรือไม่ ดังนั้นนักวิจัยจึงพัฒนาทฤษฎีของการค้นคืนสารสนเทศหรือโมเดลต่างๆ ซึ่งเปรียบเสมือนตัวแทนผู้ใช้ โดยมีจุดประสงค์เพื่อให้ผู้ใช้พึงพอใจผลการค้นหามากที่สุด [2] โดยการประเมินระบบค้นคืนนั้นขึ้นอยู่กับวิธีการของการทดลอง ที่สามารถสะท้อนถึงความต้องการของผู้ใช้ โดยการประเมินระบบค้นคืนมี 2 แบบคือ

1. การประเมินโดยมีระบบเป็นศูนย์กลาง (System-Oriented Evaluation)

การประเมินโดยมีระบบเป็นศูนย์กลางจะอยู่บนรากฐานบนงานของ Cleverdon และคณะ [17] ซึ่งเป็นคนสร้างคอเลคชั่นทดสอบ (Test collection) สำหรับการประเมินระบบค้นคืนโดยใช้คอมพิวเตอร์ หรือเรียกวิธีประเมินดังกล่าวว่า การประเมินด้วยวิธีแคลนฟิล (Cranfield evaluation paradigm) โดยนักวิจัยได้รวบรวมคอเลคชั่นทดสอบซึ่งประกอบด้วย เอกสาร, คำค้นหา และ การตัดสินความเกี่ยวข้อง โดยการประเมินแบบแคลนฟิลมีหลักการดังต่อไปนี้

- รวบรวมเอกสาร
- สร้างชุดของความต้องการหรือกลุ่มของคำค้นหา
- รวบรวมการตัดสินความเกี่ยวข้องสำหรับคู่เอกสารกับคำค้นหาโดยปกติจะเป็นการประเมินสองทางนั้น คือ เป็นเอกสารที่เกี่ยวข้อง (Relevant) และไม่เกี่ยวข้อง (Irrelevant) กับคำค้นหาหรือไม่ หรือการประเมินความเกี่ยวข้องแบบมีหลายระดับ กล่าวคือ เอกสารมีความเกี่ยวข้องในระดับใดตั้งแต่ 0-4 (0 คือไม่เกี่ยวข้องและ 4 คือเกี่ยวข้องที่สุด) ซึ่งความเกี่ยวข้องเหล่านี้จะนำไปใช้เพื่อประเมินผลการค้นหาที่คืนมาจากระบบค้นคืน

- การประเมินระบบค้นคืนโดยใช้การตัดสินความเกี่ยวข้องผ่านตัวชี้วัดต่างๆ

หลายแคมเปญได้นำวิธีประเมินด้วยวิธีแคลนฟิลในการประเมินระบบค้นคืน อาทิเช่น Text REtrieval Conference (TREC) [18] และ Cross-Language Evaluation Forum (CLEF) [19] โดยวิธีประเมินแบบแคลนฟิลจะเป็นการประเมินที่แม่นยำ แต่การที่จะประเมินให้ครอบคลุมข้อมูลทั้งหมดเป็นไปได้ยาก เนื่องจากการสร้างคอเลคชั่นทดสอบที่ใหญ่ทำได้ยาก โดยเหตุผลหนึ่ง คือการสร้างการตัดสินความเกี่ยวข้องให้สมบูรณ์นั้นเป็นอะไรที่สิ้นเปลือง และใช้จำนวนคนมหาศาล ทำให้ไม่เหมาะกับการสร้างคอเลคชั่นทดสอบขนาดใหญ่ ดังนั้นการประเมินในการประชุมเชิงปฏิบัติการ TREC นั้น การตัดสินความเกี่ยวข้องจะไม่สมบูรณ์ จึงทำให้เกิดวิธีการพูลลิ่ง (Pooling) [20] โดยสร้างการตัดสินความเกี่ยวข้องจาก พูล (Pool) ของเอกสารที่เกิดจากการรวมผลการค้นหาที่เกี่ยวข้องที่ได้จากทีมที่เข้าร่วมใน TREC

2. การประเมินโดยมีผู้ใช้เป็นศูนย์กลาง (User-Centered Evaluation)

ความต้องการนั้นมีการเปลี่ยนแปลงในกระบวนการค้นหา เพราะผู้ค้นหาได้เจอข้อมูลใหม่จากผลการค้นหา ดังนั้นจึงทำให้เกิดการประเมินโดยนำผู้ใช้มาใช้ระบบ และประเมินผลจากพฤติกรรมของผู้ใช้ การโต้ตอบระหว่างผู้ใช้กับระบบ โดยเริ่มตั้งแต่การใส่คำค้นหาของผู้ใช้ การโต้ตอบระหว่างผู้ใช้กับผลการค้นหา พฤติกรรมผู้ใช้ขณะค้นหา เช่น การเคลื่อนของเม้าส์ ตำแหน่งการมอง

การประเมินโดยมีผู้ใช้เป็นศูนย์กลางเป็นการประเมินโดยการนำผู้ใช้มาใช้ระบบ ทำให้เมื่อมีระบบใหม่เกิดขึ้นต้องการคนมาทดลองใช้ระบบใหม่ทุกครั้ง ทำให้เกิดการทดลองซ้ำ ซึ่งทำได้ยากและมีค่าใช้จ่ายที่สูง ทำให้การประเมินในปัจจุบันอาทิ ในการประชุมเชิงปฏิบัติการ TREC จึงใช้การประเมินแบบมีระบบเป็นศูนย์กลาง ซึ่งหนึ่งในขั้นตอนของการประเมินที่มีระบบเป็นศูนย์กลางคือการประเมินระบบด้วยตัวชี้วัด โดยในวิทยานิพนธ์นี้ ผู้วิจัยได้ทำการสร้างตัวชี้วัดประเภทความหลากหลาย เพื่อนำมาใช้ประเมินประสิทธิผลของระบบค้นคืน

2.5 ตัวชี้วัด

เมื่อมีระบบค้นคืนเกิดขึ้น นักวิจัยจึงได้สร้างตัวชี้วัดเพื่อประเมินประสิทธิผลของระบบว่าระบบนั้นเป็นระบบที่ตรงความต้องการผู้ใช้หรือไม่ ส่วนใหญ่มีวิธีการประเมินที่คล้ายกันนั้นคือการพิจารณาจำนวนเอกสารที่เกี่ยวข้อง อาทิ ตัวชี้วัดที่มีชื่อว่า Precision [1] จะประเมินจากอัตราส่วนระหว่าง เอกสารที่เกี่ยวข้องที่ถูกค้นคืน และเอกสารถูกค้นคืนทั้งหมด รวมถึงมีตัวชี้วัดหลายตัวพิจารณาลำดับของเอกสาร กล่าวคือ ประโยชน์ที่ได้จากเอกสารจะถูกชั่งน้ำหนักตามตำแหน่งของ

เอกสารที่ถูกค้นคืน หรือพิจารณาประโยชน์ของเอกสารที่เกี่ยวข้องจากโมเดลอรรถประโยชน์ (Utility model) เช่น โมเดลการเรียกดูเอกสารของผู้ใช้ (User's browsing model)

การสร้างตัวชี้วัดนั้นควรมีสอดคล้องกับประเภทของการค้นคืน เช่น ตัวชี้วัดสำหรับการค้นคืนแบบเฉพาะกิจได้พิจารณา ความเกี่ยวข้องของเอกสารให้ตรงกับคำค้นหาให้มากที่สุด ขณะที่ตัวชี้วัดประเภทความหลากหลายจะไม่เพียงพิจารณาเฉพาะคำค้นหาเท่านั้น แต่จะพิจารณากลุ่มของอินเท็นที่อยู่ภายใต้คำค้นหานั้น เช่น ผู้ค้นหาจะพอใจเมื่อเขาได้รับเอกสารที่ครบทุกอินเท็น ดังนั้นจุดประสงค์ของการค้นหาแต่ละประเภท และโมเดลของผู้ใช้จะสะท้อนถึงตัวชี้วัด ที่ใช้ประเมินระบบค้นคืน โดยสามารถแบ่งตัวชี้วัดเป็น 2 ประเภทได้ดังนี้

1. ตัวชี้วัดสำหรับการค้นคืนแบบเฉพาะกิจ

1.1. Precision

ตัวชี้วัดที่ประเมินผลการค้นหาจากการนับจำนวนเอกสารที่เกี่ยวข้องที่คืนจากระบบมาหารด้วยจำนวนเอกสารที่รับมาทั้งหมด [1]

$$Precision = \frac{\#(relevant\ document\ retrived)}{\#(retrieved\ documents)} \tag{2.2}$$

1.2. Recall

ตัวชี้วัดที่ประเมินผลการค้นหา โดยการนับจำนวนเอกสารที่เกี่ยวข้องที่ถูกค้นคืน หารด้วยจำนวนเอกสารที่เกี่ยวข้องทั้งหมด [1]

$$Recall = \frac{\#(relevant\ documents\ retrieved)}{\#(relevant\ documents)} \tag{2.3}$$

1.3. F-measure

การประเมินโดยใช้ Precision หรือ Recall ไม่สามารถทำให้เห็นมุมมองสมบูรณ์ของประสิทธิผลของระบบค้นคืนได้ โดยระบบในทัศนคตินั้นควรมีความสมดุลระหว่าง Precision และ Recall โดยปรับเปลี่ยนให้ตรงกับจุดประสงค์ของลักษณะงาน โดย F-measure ระบุปัญหาดังกล่าวโดยการสมดุลค่าระหว่าง Precision และ Recall ผ่านพารามิเตอร์ β [2]

$$F_\beta = \frac{(\beta^2 + 1) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \tag{2.4}$$

โดยค่า β ถูกตั้งค่าเป็น 1 ซึ่งทำให้ F เป็นการหาค่าเฉลี่ยแบบฮาโมนิค (Harmonic mean) ระหว่าง Precision และ Recall

1.4. Expected reciprocal rank (ERR)

เป็นตัวชี้วัดที่พิจารณาประโยชน์ ที่ได้จากเอกสารผ่านโมเดลแบบต่อเนื่อง (Cascade model) [4] กล่าวคือ ผู้ใช้จะดูแต่ละเอกสารจากบนลงล่าง โดยประโยชน์ของเอกสารที่กำลังพิจารณาขึ้นอยู่กับความน่าจะเป็นที่ผู้ใช้ไม่พอใจกับเอกสารอันดับก่อนหน้า และมีฟังก์ชันการลดทอน (Discount function) ตามตำแหน่งของเอกสาร รวมทั้งพิจารณาความเกี่ยวข้องแบบมีหลายระดับ (Graded relevance) ในเอกสาร

$$ERR@k = \sum_{r=1}^{k} \prod_{j=1}^{r-1} (1 - P(R_j)) P(R_k) \tag{2.5}$$

โดย $P(R_k)$ คือ ค่าความน่าจะเป็นของเอกสาร ณ ตำแหน่งที่ k และ $\prod_{j=1}^{r-1} (1 - P(R_j))$ คือความน่าจะเป็นที่ผู้ใช้ไม่พอใจในเอกสารก่อนหน้า (ตำแหน่งที่ 1 ถึง k-1) และ $\frac{1}{r}$ เป็นฟังก์ชันการลดทอนโดยการพิจารณาจากตำแหน่งของเอกสาร ซึ่งการคำนวณค่าความน่าจะเป็นสามารถคำนวณได้ดังนี้

$$P(R_i) \approx R(g_i) = \frac{2^g - 1}{2^{g_{\max}}}, g \in \{0, \ldots, 4\} \tag{2.6}$$

โดยที่ g คือระดับความเกี่ยวข้อง ($0 \leq g \leq 4$) ในกรณีที่ใช้เกณฑ์การให้คะแนน 5 ระดับ (แย่, พอใช้, ดี ,ดีมาก, เยี่ยม)

1.5. Normalised Discounted Cumulative Gain (nDCG)

เป็นตัวชี้วัดประเภทความเกี่ยวข้องของเอกสาร โดยพิจารณาความเกี่ยวข้องแบบมีหลายระดับ ซึ่งหมายถึงเอกสารหนึ่งมีความเกี่ยวข้องมากกว่า 2 ระดับ [21] และคำนวณประโยชน์ที่ได้รับจากเอกสารจากพิจารณา จากตำแหน่งของเอกสารที่ถูกค้นคืน กล่าวคือ เอกสารที่มีความเกี่ยวข้องสูง แต่อยู่ในอันดับล่างๆ จะถูกลดตามสัดส่วนของอันดับ โดยเริ่มคำนวณจากสมการ DCG ดังนี้

$$DCG@k = \sum_{r=1}^{k} \frac{J(d_r, q)}{\log_2(1+i)} \tag{2.7}$$

โดย $J(d_i, q)$ เป็นค่าความเกี่ยวข้องของเอกสารที่ i ของคำค้นหา q และ $\log_2(r+1)$ เป็นตัวที่เป็นหารเพื่อลดคะแนนตามอันดับที่เพิ่มขึ้น หลังจากได้ค่า DCG จึงทำการนอร์มัลไลเซชัน

(Normalization) ให้ค่า DCG อยู่ระหว่างค่า 0 กับ 1 เพื่อทำให้สามารถเปรียบเทียบประสิทธิผลของระบบข้ามหัวข้อได้ โดยการสร้างผลการค้นหาที่ทำให้ได้ค่า DCG มากที่สุด ($DCG(k)'$) แล้วจึงนำไปหาร ซึ่งมีสมการดังนี้

$$nDCG@k = \frac{DCG(k)}{DCG(k)'} \tag{2.8}$$

2. ตัวชี้วัดสำหรับการค้นคืนแบบหลากหลาย

ตัวชี้สำหรับการค้นคืนแบบหลากหลาย จะประเมินผลการค้นหาจากการพิจารณาอินเท็นกล่าวคือ แต่ละคำค้นหาจะถูกแบ่งออกเป็นอินเท็น หรือชิ้นส่วนของความต้องการ โดยประสิทธิผลของระบบค้นหาประเภทความหลากหลายจะขึ้นอยู่กับความเกี่ยวข้องของเอกสารที่มีต่ออินเท็น สำหรับตัวชี้วัดสำหรับความหลากหลาย มีการประเมินอยู่ 2 ลักษณะคือ *1)* ประเมินจากความหลากหลายของอินเท็นในเอกสาร *2)* ประเมินด้วยการลงโทษหรือลดคะแนนระบบที่คืนเอกสารที่มีอินเท็นที่ซ้ำซ้อน ดังนั้นจะได้ตัวชี้วัด 2 ประเภทคือ *1)* ตัวชี้วัดประเภทความหลากหลาย *2)* ตัวชี้วัดประเภทความซ้ำซ้อน และ เพิ่มเป็นอีก 1 ประเภทซึ่งไม่มีการจำแนกประเภทที่ชัดเจน เช่น ตัวชี้วัดที่รวมระหว่าง 2 ประเภทเข้าด้วยกัน

2.1 ตัวชี้วัดประเภทความหลากหลาย

2.1.1. Intent recall (*I*-rec)

ตัวชี้วัดที่ประเมินผลการค้นหาโดยดูจากความครอบคลุมในอินเท็นที่มีต่อคำค้นหา [5]

$$I\text{-}rec@k = \frac{\left|\bigcup_{r=1}^{k} intent(d_r)\right|}{|I|} \tag{2.9}$$

โดย $intent(d_r)$ คือจำนวนอินเท็นถึงเอกสารลำดับที่ i และ $|I|$ คือจำนวนอินเท็นทั้งหมดที่เกี่ยวข้องกับคำค้นหา

2.1.2. Intent Mean Reciprocal Rank (*I*-mrr)

Zhai และคณะได้เสนอเป็นตัววัดชี้วัดที่มีลักษณะคล้ายกับ *I*-rec กล่าวคือ *I*-mrr ถูกนิยามว่าเป็นส่วนกลับของตำแหน่งแรกที่ครอบคลุมอินเท็นทั้งหมด

$$I\text{-}mrr@100\% = \frac{1}{\text{first-rank-100\%-coverage}} \tag{2.10}$$

โดยต่อมาได้ถูกปรับเปลี่ยนโดยทำให้สามารถเลือกประเมินบางส่วนของการครอบคลุม เช่น 25% ของอินเท็นทั้งหมด [22], [23]

2.2 ตัวชี้วัดประเภทความซ้ำซ้อน

2.2.1. ตัวชี้วัด Intent-Aware

ตัวชี้วัดตระกูล IA เป็นตัวชี้วัดมีแนวคิดในการประเมินผลการค้นหาโดยใช้แต่ละอินเท็นมาพิจารณา [24] กล่าวคือ นำตัวชี้วัดสำหรับการค้นคืนแบบเฉพาะกิจนำมาพิจารณาแต่ละอินเท็น แล้วจึงคูณกับความน่าจะเป็นในแต่ละอินเท็น เช่น ERR@k และ nDCG@k ก็ถูกเปลี่ยนไปในรูปแบบ IA ดังนี้

$$ERR\text{-}IA@k = \sum_{i=1}^{|I|} P(i \mid q) ERR_i @k \tag{2.11}$$

$$nDCG\text{-}IA@k = \sum_{i=1}^{|I|} P(i \mid q) nDCG_i @k \tag{2.12}$$

โดย $|I|$ คือ จำนวนของ อินเท็นทั้งหมด และ $P(i|q)$ คือความน่าจะเป็นที่ผู้ใช้สนใจ อินเท็น i สำหรับคำค้นหา q และ $ERR\text{-}IA@k$, $nDCG\text{-}IA@k$ คือ ตัวชี้วัดสำหรับการค้นคืนแบบเฉพาะกิจที่พิจารณาอินเท็น i

2.2.2. Novelty-biased cumulative gain (α-nDCG)

ตัวชี้วัดที่ประเมินความซ้ำซ้อน [3] โดยการลงโทษอินเท็น หรือชิ้นส่วนข้อมูลที่ซ้ำซ้อนด้วยฟังก์ชัน $(1-\alpha)^{D_{i,r-1}}$ ซึ่ง $D_{i,r-1}$ คือจำนวนครั้งที่อินเท็น i ปรากฏในเอกสารจนถึงตำแหน่งที่ r-1 ทำให้สามารถลดค่าคะแนนจากอินเท็นที่ซ้ำซ้อนจากเอกสารก่อนตำแหน่งที่ r และ α คือค่าพารามิเตอร์ที่ควบคุมว่าผู้ใช้สามารถทนความซ้ำซ้อนได้เท่าใด (ปกติตั้งค่าเป็น 0.5) และกำหนดให้ $J(d_r,i)$ คือการตัดสินความเกี่ยวข้องว่าเอกสาร d_r เกี่ยวข้องกับอินเท็น i หรือไม่

$$DCNG(k) = \sum_{r=1}^{k} \frac{\sum_{i=1}^{|I|} J(d_r \mid i)(1-\alpha)^{D_{i,r-1}}}{\log_2(r+1)} \tag{2.13}$$

โดย $\log_2(r+1)$ คือฟังก์ชันลดทอน (Discount function) จากตำแหน่งของเอกสารโดย DCNG จะผ่านการนอร์มัลไลเซชั่น โดยสามารถทำได้โดยการหา การจัดอันดับในอุดมคติ (Ideal ranking) ซึ่งทำให้เกิดค่า DCNG มากที่สุดแล้วจึงนำมาหาร ดังสมการต่อไปนี้

$$\alpha\text{-}nDCG@k = \frac{DCNG(k)}{DCNG(k)'} \tag{2.14}$$

โดย DCNG(k)' คือ การจัดอับในอุดมคติ ณ ตำแหน่งที่ k

2.3 ตัวชี้วัดประเภทอื่นๆ

2.3.1. D#-nDCG

ตัวชี้วัด D#-nDCG ถูกสร้างขึ้นเพื่อลดข้อจำกัดของตัวชี้วัด D ซึ่งแนะนำระบบที่ละเลยต่ออินเท็น ที่มีค่าความน่าจะเป็นต่ำ โดยตัวชี้วัด D จะคล้ายกับตัวชี้วัดตะกูล IA แต่จะนำความน่าจะเป็นที่ได้ไปแทนการประเมินความซ้ำซ้อนใน α-nDCG [25] โดยตัวชี้วัด D ในแบบของ D-DCG มีสมการดังต่อไปนี้

$$D\text{-}DCG@k = \sum_{r=1}^{k}\sum_{i}^{|I|}\frac{P(i|q)g_i(r)}{\log_2(r+1)} \tag{2.15}$$

โดย $P(i|q)$ คือค่าความน่าจะเป็นที่ผู้ค้นหาสนใจอินเท็น i ในคำค้นหา q และ $g_i(r)$ คือค่าความเกี่ยวข้องที่ได้จากเอกสาร ณ ตำแหน่ง r สำหรับอินเท็น i และ $\log_2(r+1)$ เป็นฟังก์ชันการลดทอนค่า แล้ว D-DCG จะทำการนอร์มัลไลเซชั่น โดยหารกับค่า D-DCG ที่ได้จากการจัดอันดับที่ทำให้ได้ค่า D-DCG สูงที่สุด

Sakai และคณะ [26] ได้เสนอตัวชี้วัด D# เพื่อการลดข้อจัดกัดของตัวชี้วัด D โดยนำตัวชี้วัดประเภทความหลากหลาย และตัวชี้วัดประเภทความซ้ำซ้อนมารวมเข้ากันแบบเส้นตรง หรือเรียกว่า วิธีการแบบชดเชยกัน ซึ่งจุดประสงค์ของการรวมเพื่อให้สามารประเมินการค้นคืนของเอกสารที่ครอบคลุมอินเท็นต่างๆ ได้เร็วที่สุด พร้อมกับสามารถประเมินการผลการค้นหาของเอกสารที่มีอินเท็นที่เป็นที่นิยม ให้สูงกว่าเอกสารที่มีอินเท็นที่เป็นที่นิยมน้อยกว่า ด้วยเหตุข้างต้นจึงเกิดการรวมกันระหว่าง I-rec และ D-nDCG ดังต่อไปนี้

$$D\#\text{-}nDCG@k = \gamma I\text{-}rec@k + (1-\gamma)D\text{-}nDCG@k \tag{2.16}$$

โดยพารามิเตอร์ γ ถูกใช้ควบคุมการรวมกันระหว่าง I-rec และ D-nDCG โดยถ้า γ = 1 จะหมายถึงการคำนวน I-rec อย่างเดียว

2.6 การประเมินตัวชี้วัด

ผู้วิจัยได้แยกการทดลองเป็น 2 กลุ่มคือ การวัดความสอดคล้อง และการวัดความน่าเชื่อถือ ดังนี้

2.6.1. การวัดความสอดคล้อง

1. ทฤษฎีความสอดคล้องของเคนดัล เทา (Kendal's τ Correlation)

เคนดัลเทา คือตัววัดความสอดคล้องซึ่งวัดพลังความสอดคล้องของอันดับของข้อมูล ค่าคะแนนของเคนดัลเทา (τ) จะอยู่ระหว่าง -1 (อันดับของข้อมูลแรกตรงข้ามกับอันดับที่สอง) ถึง 1 (อันดับของข้อมูลทั้งสองสอดคล้องกันทั้งหมด) โดยค่า τ ที่มากกว่า 0.9 ระบุถึงคู่ของอันดับมีความสอดคล้องกันสูง

2. ความสอดคล้องของตัวชี้วัดกับความชอบของผู้ใช้ (Correlation of evaluation measures with user preferences)

นักวิจัยได้ใช้คอเล็กชั่นทดลองและตัวชี้วัดร่วมกันเพื่อประเมินระบบ ซึ่งการประเมินควรสะท้อนถึงความต้องการของผู้ใช้ในระบบค้นคืนโดยถ้าตัวชี้วัดแนะนำว่าระบบ A ดีกว่าระบบ B แล้วหมายความว่าผู้ใช้จะชอบระบบ A มากกว่าระบบ B ดังนั้นการนำตัวชี้วัดมาเปรียบเทียบกับความชอบผู้ใช้จะสะท้อนให้เห็นถึงประสิทธิผลของตัวชี้วัดว่าสามารถประเมินระบบค้นคืนได้ตรงกับความต้องการผู้ใช้ดังนี้

- ประวัติการใช้งาน (Query log)

Joachims [27] ได้ทำการทดลองโดยแสดงให้ผู้ใช้ดูผลการค้นหาที่แตกต่างกันเพื่อประเมินผู้ใช้ผ่านการวิเคราะห์ประวัติการการใช้ของผู้ค้นหา แม้ว่าจะสามารถวัดความแตกต่างโดยพิจารณาระหว่างปริมาณและอันดับของเอกสารที่เกี่ยวข้อง แต่ Joachim ก็เห็นความแตกต่างเพียงเล็กน้อยในพฤติกรรมการคลิ้กของผู้ค้นหา กล่าวคือ ผู้ใช้แม้จะได้รับผลการค้นหาที่แย่แต่ก็ยังคงเลือกเอกสารที่จัดอันดับต้น ดังนั้นเขาจึงเสนออีกวิธีคือ รวมผลการค้นหาจาก 2 ระบบแล้วสังเกตพฤติกรรมการคลิกเอกสารระหว่าง 2 ผลการค้นหา โดยผลแสดงให้เห็นว่าผู้ใช้มีแนวโน้มคงที่ที่จะคลิกเอกสารของผลการค้นหาที่ดีกว่า โดยวิธีนี้ถูกทำซ้ำในงาน [28] และได้ผลที่คล้ายคลึงกัน

- วัดความชอบผู้ใช้จากการเปรียบเทียบระหว่างผลการค้นหา (Side-by-side evaluation)

Thomas และคณะ [29]ได้เสนอวิธีการใหม่เพื่อสังเกตพฤติกรรมของผู้ใช้โดยแสดง 2 ผลการค้นหาไว้ข้างกันแล้วให้ผู้ใช้พิจารณาว่าผู้ใช้ชอบผลการค้นหาใดมากกว่า โดยใช้ผลการค้นหา 10 อันดับแรกจาก Google เปรียบเทียบระหว่าง ผลการค้นหาลำดับที่ 21-30 โดยผลที่ได้ คือผู้ใช้ชอบผลการค้นหาจาก 10 อันดับแรกมากกว่า

2.6.2. การวัดความน่าเชื่อถือ

1. พลังการแยกแยะ (Discriminative power)

Sakai และคณะ [30] ได้นำเสนอวิธีการเปรียบเทียบระหว่างตัวชี้วัด กล่าวคือ การวัดพลังการแยกแยะซึ่งใช้ความน่าเชื่อถือในแง่ของความเสถียรของตัวชี้วัด ซึ่งวิธีนี้ได้ปรากฏอยู่ในหลายงานวิจัย เช่น [26], [31] โดยการวัดความเสถียรของ Sakai สามารถทำได้โดยการพิจารณาอัตราส่วนของคู่ระบบ ซึ่งเป็นศักยภาพของคู่ระบบที่แตกต่างอย่างมีนัยสำคัญ โดยจะประเมินความเสถียรภาพของตัวชี้วัด เมื่อตัวชี้วัดประเมินคอเลคชั่นทดสอบที่แตกต่างกัน โดยใช้วิธีการบูทสแตร็ป (Bootstrap) เพื่อสุ่มตัวอย่าง แล้วจึงใช้การทดสอบนัยสำคัญทางสถิติสองทาง (Two-tailed statistical significance test) โดยใช้คู่ของระบบที่แตกต่างกัน

2. วิธีการสลับเปลี่ยน (Swap method)

Buckley และคณะ [32] ได้นำเสนอการวัดเสถียรภาพโดยใช้วิธีการสลับเปลี่ยน ซึ่งต่างจากกับวิธีการวัดพลังการแยกแยะโดยจะไม่เกี่ยวข้องกับการทดสอบนัยสำคัญโดยตรง แต่ขึ้นอยู่กับวิธีการลองผิดลองถูก (Heuristic approach) ซึ่งนับความแตกต่างของศักยภาพระหว่าง 2 ระบบ กล่าวคือ จะประมาณโอกาสที่จะได้รับผลการค้นหาที่ขัดแย้งกันจากชุดของหัวข้อที่ต่างกัน เช่น การสุ่มตัวอย่างจากวิธีการบูทสแตร็ป โดยวิธีนี้จะทำการสุ่มเซตของคำค้นหา และ สร้างกราฟระหว่างค่า Minority Rate (MR) และ Proportion of Ties (PT) โดยค่า MR ระบุถึง การขาดความน่าชื่อถือ เมื่อเกิดการเปลี่ยนแปลงในกลุ่มของคำค้นหา และ PT ระบุถึง การขาดพลังในการแยกแยะระบบค้นคืน ดังนั้นการได้ค่า MR และ PT น้อยจะบ่งบอกถึงระบบมีความน่าเชื่อถือมาก

บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้ผู้วิจัยจะกล่าวถึงวิธีดำเนินการวิจัย โดยจุดประสงค์ของเนื้อหา เพื่อชี้ให้เห็นถึงปัญหาของตัวชี้วัดสำหรับความหลากหลายในปัจจุบัน และได้นำเสนอตัวชี้วัดใหม่ประเภทความหลากหลาย โดยจะกล่าวถึงปัญหาของตัวชี้วัดในปัจจุบัน และแสดงตัวอย่างจากระบบจำลองเพื่อทำให้เข้าใจพฤติกรรม ของแต่ละตัวชี้วัดมากยิ่งขึ้น รวมทั้งเสนอตัวชี้วัดประเภทความหลากหลาย และแสดงแนวคิดรากฐานของตัวชี้วัดของตัวชี้วัดใหม่เปรียบเทียบกับตัวชี้วัดอื่น

ในการประชุมเชิงปฏิบัติการ TREC 2009 ถึง 2012 ในหมวดของเว็บ (สามารถดูได้ที่ภาคผนวก ข) และ NTCIR 9 และ 10 ถูกจัดขึ้นเพื่อการแข่งพัฒนาระบบค้นคืน ซึ่งประกอบด้วยการค้นหา 2 ประเภทคือ การค้นหาแบบเฉพาะกิจ และการค้นหาแบบความหลากหลาย โดยแต่ละระบบจะถูกประเมินด้วยตัวชี้วัดต่างๆ โดยในงานวิจัยที่ผ่านมา นักวิจัยได้สร้างตัวชี้วัดขึ้นเพื่อประเมินประสิทธิผลของการค้นหาแบบความหลากหลาย โดยสองตัวชี้วัดที่เป็นที่นิยมในการวัดประสิทธิผลระบบค้นคืน คือ α-nDCG และ ERR-IA โดยทั้งคู่เป็นตัวชี้วัดประเภทความซ้ำซ้อน โดยจะลดทอนความเกี่ยวข้องของเอกสารจากการพิจารณา *1)* ตำแหน่งของเอกสาร *2)* ความซ้ำซ้อนของอินเท็น แต่อย่างไรก็ตาม ตัวชี้วัดเหล่านี้ประเมินผลโดยละเลยการครอบคลุมอินเท็น เนื่องจากการประเมินที่ให้ความสำคัญกับการลำดับของเอกสาร ขณะที่ *I*-rec ซึ่งเป็นตัวชี้วัดประเภทความหลากหลาย แต่มีข้อเสียอยู่หลายประการ เช่น หลังจากระบบครอบคลุมอินเท็นทั้งหมดแล้ว *I*-rec ไม่สามารถประเมินอินเท็นเดิมได้ นอกจากนี้ *I*-rec ยังถูกนำไปรวมกับตัวชี้วัดอื่นๆ เช่น D#-nDCG และ α-nDCG-IA ทำให้ข้อเสียของ *I*-rec สืบทอดไปยังตัวชี้วัดอื่นๆ

3.1 ปัญหางานวิจัย

1. ตัวชี้วัด α-nDCG และ ERR-IA เป็นตัวชี้วัดประเภทความซ้ำซ้อน ซึ่งประเมินผลการค้นหาโดยละเลยการครอบคลุมอินเท็น เนื่องจากตัวชี้วัดให้ความสำคัญกับการคืนเอกสารในลำดับต้นๆ
2. *I*-rec เป็นตัวชี้วัดประเภทความหลากหลาย มีข้อเสียในการแยกแยะประสิทธิผลของระบบ
3. D# เป็นตัวชี้วัดที่ประเมินรวมระหว่างตัวชี้วัดประเภทความหลากหลาย และ ตัวชี้วัดประเภทความซ้ำซ้อนด้วยวิธีแบบเส้นตรง อย่างไรก็ตามการประเมินของ D# ทำให้ตีผลได้ยาก

3.2 สมมติฐานงานวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษา และชี้ให้เห็นถึงปัญหาของตัวชี้วัดในปัจจุบัน รวมถึง สร้างตัวชี้วัดใหม่ที่มีชื่อว่า Normalized Coverage Frequency (nCF) เพื่อสามารถนำไปใช้เพื่อประเมินระบบค้นคืน ในบริบทของความหลากหลาย ดังนั้นจึงเกิดสมมตติฐานดังต่อไปนี้

1. แนวคิดรากฐานของตัวชี้วัดใหม่เปรียบเทียบกับตัวชี้วัดเดิม ตัวชี้วัดใหม่สามารถสะท้อนแนวคิดที่ตรงกับความต้องการในการค้นหาแบบความหลากหลายได้ดีกว่าตัวชี้วัดอื่นๆ
2. การประเมินระบบค้นคืนของตัวชี้วัด nCF มีความสอดคล้องกับการประเมินของตัวชี้วัดสำหรับการค้นคืนแบบหลากหลายอื่น
3. การประเมินระบบค้นคืนของตัวชี้วัด nCF มีความสอดคล้องกับความชอบของผู้ใช้ที่มีความต้องการอินเท็นทั้งหมด
4. การประเมินระบบค้นคืนของตัวชี้วัด nCF และการประเมินระบบค้นคืนร่วมกันระหว่างตัวชี้วัด nCF กับตัวชี้วัดประเภทความซ้ำซ้อน มีความน่าเชื่อถือกว่าตัวชี้วัดอื่น

3.3 การวิเคราะห์ปัญหาของตัวชี้วัดปัจจุบัน

3.3.1. Intent recall

I-rec เป็นตัวชี้วัดประเภทความหลากหลาย แต่ยังเป็นตัวชี้วัดที่หยาบ เนื่องจากผลกระทบจาก 4 ข้อเสียดังต่อไปนี้

1. *I*-rec นั้นไม่พิจารณาถึงตำแหน่งของเอกสารหรือจะกล่าวได้ว่า I-rec ไม่สามารถแยกความแตกต่างประโยชน์ที่ได้จากอินเท็นที่เกี่ยวข้องที่ถูกค้นคืนในตำแหน่ง *k-1* และ *k-2* ได้
2. ทันทีระบบคืนเอกสารที่ประกอบด้วยอินเท็นใดแล้ว *I*-rec จะไม่สามารถแยกความแตกต่างเอกสารอันดับต่อมาที่ประกอบด้วยอินเท็นนั้นได้ อย่างไรก็ตามการได้รับอินเท็นที่ซ้ำซ้อน อาจทำให้ผู้ค้นหาพอใจ แต่ก็ดีกว่าการได้รับอินเท็นที่ไม่เกี่ยวข้อง
3. เมื่อระบบคืนเอกสารครอบคลุมทุกอินเท็นแล้ว *I*-rec ไม่สามารถแยกแยะความแตกต่างของเอกสารที่ค้นคืนต่อมาได้ เนื่องจาก *I*-rec จะสามารถประเมินความหลากหลายของอินเท็นได้จนถึงตำแหน่งที่ครอบคลุมอินเท็นทั้งหมดเท่านั้น ดังนั้นหลังจากครอบคลุมอินเท็นทั้งหมด *I*-rec ไม่สามารถประเมินระบบค้นคืนที่พยายามทำให้ผลการค้นหามีความหลากหลายได้
4. *I*-rec ไม่พิจารณาความน่าจะเป็นของอินเท็นในคำค้นหา เนื่องความความน่าจะเป็นจะสะท้อนถึงปริมาณประชากรที่มีความต้องการในอินเท็น

3.3.2. α-nDCG และ ERR-IA

α-nDCG และ ERR-IA เป็นตัวชี้วัดที่ออกแบบบนพื้นฐานของความซ้ำซ้อน แต่ประเมินผลการค้นหาโดยละเลยการครอบคลุมอินเท็น ซึ่งในปัญหาดังกล่าวจะถูกแสดงให้เห็นในตัวอย่างในหัวข้อถัดไป

3.3.3. D#-nDCG

D#-nDCG เป็นตัวชี้วัดที่รวมตัวชี้วัดประเภทความซ้ำซ้อนและความหลากหลายเข้าด้วยวิธีการแบบชดเชยกัน ทำให้การตีความทำได้ยาก เช่น จะทราบได้อย่างไรว่าผลการค้นหานั้นดีในแง่ของความหลากหลายที่สูง โดยนักวิจัยได้นำเสนอเกณฑ์ทางเลือกอื่นของการประเมินแบบเป็นอิสระหรือเรียกว่า การรวมแบบวิธีการแบบขั้นลำดับ แต่ผู้วิจัยก็ไม่ได้จำกัดการรวมด้วยวิธีการแบบชดเชยกันโดยทั้งคู่จะถูกนำไปใช้ในสำหรับการประเมินร่วมระหว่างตัวชี้วัดใหม่กับตัวชี้วัดประเภทความซ้ำซ้อน

3.4 การวิเคราะห์ปัญหาจากการระบบจำลอง

การวิเคราะห์ตัววัดชี้วัดทำให้เราเข้าใจถึงตัวชี้วัดแต่ละตัวว่าสามารถประเมินความหลากหลายในระบบค้นคืนได้ดีหรือไม่ ดังนั้นผู้วิจัยจึงสร้างระบบจำลองโดยใช้กลุ่มของเอกสารจาก TREC 2012 เพื่อแสดงให้เห็นถึงเหตุการณ์การประเมินระบบค้นคืนที่เป็นไปได้ เพื่อทำให้ง่ายต่อการเข้าใจพฤติกรรมของแต่ละตัวชี้วัด

จากตารางที่ 3.1 แสดง 5 ระบบจำลองโดยแต่ละระบบประกอบด้วย 10 เอกสาร โดยคำค้นหาประกอบด้วย 4 อินเท็น ซึ่งกำหนดการตัดสินความเกี่ยวข้องเป็นแบบความเกี่ยวข้อง 2 ระดับ (Binary relevance) และกำหนดความน่าจะเป็นของแต่ละอินเท็นเท่ากัน คอลัมน์ชื่อว่า intent คือการปรากฏของอินเท็น (1, 2, 3, 4) ในแต่ละอันดับ (#1, #2,…,#10) คอลัมน์ชื่อว่า cn (cumulative nugget) คือ ผลรวมของจำนวนอินเท็นในแต่ละตำแหน่งของเอกสาร ในส่วนตารางด้านบนระบุถึงผลการประเมินระบบจำลอง จากตัวชี้วัด α-nDCG D#-nDCG และ *I*-rec โดยแต่ระบบจะมีการจัดอับดับเช่น #1 (..คะแนน..) หมายถึงมีคะแนนอันดับที่ 1 ของตัวชี้วัดนั้นๆ

จากตารางที่ 3.1 จะเห็นได้ว่าตัวชี้ด α-nDCG และ ERR-IA ประเมิน SyntheticSys1 ให้เป็นระบบที่ดีที่สุดเพราะ มีการคืนเอกสารที่เกี่ยวข้องอยู่อันดับต้นๆ ของผลการค้นหา แต่อย่างไรก็ตามระบบ SyntheticSys1 คืนเอกสารที่ครอบคลุมจำนวนอินเท็นจาก 3 ใน 4 เท่านั้น ดังนั้นจะเห็นได้ว่า α-nDCG และ ERR-IA ไม่สามารถประเมินความหลากหลายได้ เนื่องจากตัวชี้วัดดังกล่าวพิจารณาจำนวนของชิ้นส่วนข้อมูล และ ลดทอนคะแนนจากตำแหน่ง และความซ้ำซ้อนของอินเท็น ขณะที่ SyntheticSys2 ถึง 5 นั้นถูก α-nDCG และ ERR-IA จัดอันดับต่ำกว่าแม้ว่าทั้งสี่ระบบนั้นคืนเอกสารที่ครบอินเท็นทั้งหมด ด้วยจำนวนชิ้นส่วนข้อมูลที่เท่ากัน (cn =12) ยิ่งไปกว่านั้น Systhetic2 3 และ 4

พยายามทำให้ผลการค้นหามีความหลากหลาย โดยการคืนเอกสารที่ครอบคลุมอินเท็นมากกว่า 1 รอบ ดังนั้นแล้ว α-nDCG และ ERR-IA ไม่ควรประเมินความหลากหลาย เนื่องจาก α-nDCG และ ERR-IA ไม่พิจารณาความครอบคลุมของอินเท็น ซึ่งประเด็นดังกล่าวถูกบอกโดยนัยในงานวิจัยของ Gobus และ คณะโดยได้นำเสนอตัวชี้วัดที่ชื่อว่า α#-nDCG-IA เพื่อที่จะลดปัญหาการไม่พิจารณาความครอบคลุมของอินเท็นโดยการรวม α-nDCG และ *I*-rec เข้าด้วยกัน [25] และในงานวิจัยของ Leelanupab และคณะ [33], [34] ได้แสดงตัวอย่างประเด็นที่คล้ายกันของ α-nDCG และ ERR-IA ที่ปรากฏในเหตุการณ์ของการประเมินระบบใน TREC

ตารางที่ 3.1 ระบบจำลอง 5 ระบบที่คืนเอกสารจาก TREC 2012 ในค้นหาที่ 154 โดยถูกประเมินด้วย α-nDCG, ERR-IA, *I*-rec และ D#-nDCG (ตารางด้านบน)

System ranking observed at rank 10, according to:					
α-nDCG	#1 (0.677)	#2 (0.621)	#3 (0.616)	#4 (0.587)	#5 (0.534)
ERR-IA	#1 (0.675)	#2 (0.553)	#3 (0.552)	#4 (0.468)	#5 (0.449)
I-rec	#2 (0.750)	1# (1.000)			
D#-nDCG	#4 (0.555)	#1 (0.656)		#3 (0.642)	#2 (0.644)

	SyntheticSys1		SyntheticSys2		SyntheticSys3		SyntheticSys4		SyntheticSys5	
rank	intent	cn	intent	cn	intent	cn	intent	cn	intent	cn
#1	2 3 4	3	1 3	2	1 3	2	4	1	4	1
#2	2 4	5	4	3	4	3	1 3	3	1 3	3
#3		5		3		3		3		3
#4		5		3		3		3		3
#5	3 4	7	2 3 4	6	2 3 4	6	2	4	2 4	5
#6	4	8	4	7	4	7	3 4	6	3 4	7
#7		8		7		7		6		7
#8		8	1 3 4	10	1 3 4	10	1 2	8	4	8
#9	2 4	10	4	11	4	11	2 3 4	11	3 4	10
#10	2 3	12	3	12	4	12	1	12	3 4	12

สำหรับตัวชี้วัด *I*-rec ประเมินให้ระบบ SyntheticSys1 เป็นระบบที่แย่ที่สุด เนื่องจาก *I*-rec เป็นตัวชี้วัดที่ประเมินแบบตรงไปตรงมา โดยพิจารณาความครอบคลุมของอินเท็น ดังนั้น *I*-rec จึงพิจารณาระบบ SyntheticSys1 แย่ที่สุด เพราะครอบคลุมอินเท็นแค่ 3 ใน 4 ของอินเท็นทั้งหมด อย่างไรก็ตามเมื่อพิจารณา SyntheticSys2 ถึง 5 แล้ว *I*-rec ไม่สามารถแยกความแตกต่างของระบบได้หลังจากระบบครอบคลุมอินเท็นทั้งหมดแล้ว ดังนั้นเมื่อพิจารณาความหลากหลายของอินเท็น

แล้วระบบ SyntheticSys2, 3 และ 5 ควรถูกประเมินให้เป็นระบบที่ดีหรือไม่ ถ้าในกรณีที่อินเท็นที่ 2 ในเอกสารอันดับที่ 5 ไม่สามารถตอบสนองความต้องการของผู้ค้นหาได้ และเหมาะสมหรือไม่ที่ตัวชี้วัดเพิกเฉยการให้คะแนนแก่ระบบ SystheticSys4 ที่พยายามทำให้ผลการค้นหามีความหลากหลาย (ด้วยอินเท็นที่ 2) ดังนั้นการค้นคืนที่ตามมาของอินเท็นที่ 2 โดย SystheticSys4 อาจจะตอบสนองความต้องการของผู้ใช้ได้ดีกว่า เนื่องจากผู้ใช้อาจจะต้องการอินเท็นที่ 2 อีก ถ้าผู้ใช้ได้รับข้อมูลน้อยจากเอกสารที่ประกอบด้วยอินเท็นที่ 2 ที่ผู้ใช้อ่านก่อนหน้า

D#-nDCG ได้สืบทอดข้อเสียของ I-rec โดยแสดงในกรณีที่ระบบ SyntheticSys5 ถูกประเมินให้ดีกว่า SyntheticSys4 แต่ SyntheticSys4 พยายามทำให้ผลการค้นหามีความหลากหลาย หลังจากระบบครอบคลุมทุกอินเท็น ณ ตำแหน่งที่ 5 และ กรณีที่คล้ายกันเกิดขึ้นกับระบบ SyntheticSys2 และ 3 โดยให้คะแนนเท่ากัน (D#-nDCG = 0.656) เนื่องจากหลังจากคืนเอกสารที่ครอบคลุมทุกอินเท็นทั้งหมดแล้ว การประเมินของ D#-nDCG จะขึ้นอยู่กับคะแนนของ D-nDCG เท่านั้น

อย่างไรก็ตามประเมินร่วมกับ D-nDCG ก็มีข้อดี กล่าวคือ สามารถประเมินระบบจากค่าความน่าจะเป็นของอินเท็น การใช้ความเกี่ยวข้องหลาย แต่อย่างไรก็ตามการใช้วิธีรวมกันแบบวิธีการแบบชดเชยกัน (อย่างที่ D#-nDCG ใช้) ทำให้ยากต่อการตีความผล อย่างในกรณีสมมติ ถ้ามีระบบใหม่ชื่อว่า SyntheticSys6 คืนเอกสารที่เกี่ยวข้องทั้งผลการค้นหาโดยแต่ละเอกสารคืนอินเท็น 3 ใน 4 (เช่น 2 3 และ 4) ซึ่งคะแนนของ SyntheticSys6 คือ 0.75 แม้ว่าระบบจะไม่เคยครอบคลุมอินเท็นทั้งหมดเลย และเมื่อนำมาเปรียบเทียบกับระบบ SyntheticSys5 ซึ่งมีการครอบคลุมอินเท็นทั้งหมด 1 ครั้ง ดังนั้นจึงยากต่อการตีความผลที่ได้จาก D#-nDCG ว่าระบบนั้นเกิดจากระหว่างระบบมีความหลากหลายมาก หรือ ระบบมีความซ้ำซ้อนน้อย

3.5 การพัฒนาตัวชี้วัดประเภทความหลากหลาย

ผู้วิจัยได้สร้างตัวชี้วัดใหม่ประเภทความหลากหลายมีชื่อว่า Normalized Coverage Frequency (nCF) โดยใช้แนวคิดของ *I*-rec ซึ่งเป็นแนวคิดในการประเมินระบบโดยพิจารณาจำนวนอินเท็นที่แตกต่างกันในเอกสาร แต่สิ่งที่ต่างออกไปจาก *I*-rec คือ nCF สามารถวัดจำนวนครั้งที่มีการครอบคลุมอินเท็นทั้งหมด นอกจากนี้ ผู้วิจัยได้ตั้งข้อสันนิษฐานเพื่อทำให้ง่ายต่อการพัฒนาตัวชี้วัดดังนี้

1. ในแต่ละอินเท็นจะไม่มีข้อมูลที่ทับซ้อนกัน
2. พิจารณาบางปัจจัย เนื่องจากไม่เหมาะสมเมื่อนำมารวมกับการประเมินความหลากหลายเช่น ค่าความน่าจะเป็นของอินเท็น การประเมินความซ้ำซ้อน เป็นต้น โดยในภายหลังจะถูกพิจารณาโดยตัวชี้วัดอื่นเช่น ERR-IA

โดยผู้วิจัยจะแบ่งนิยามการพัฒนาตัวชี้วัดเป็น 3 ส่วนดังนี้

➢ **นิยามที่ 1 การครอบคลุมอินเท็น (Intent Coverage)**

$$cov(n,k) = \frac{\left|\bigcup_{r=n}^{k} intent(d_r)\right|}{|I|} \tag{3.1}$$

Intent Coverage มีแนวคิดมาจากตัวชี้วัด *I*-rec โดย *cov(n, k)* เป็นฟังก์ชันนับจำนวนอินเท็นที่แตกต่างกันจากตำแหน่ง n ถึง k โดยพิจารณาคำค้นหา (q) ประกอบด้วยอินเท็นจำนวน $|I|$ ซึ่ง I ประกอบด้วยอินเท็น $(i_1, i_2, ..., i_I)$ และเอกสารที่ถูกจัดอันดับ $(d_n, ..., d_k)$ โดย n คือตำแหน่งเริ่มต้น และ k คือตำแหน่งสุดท้าย และฟังก์ชัน $intent(d_r)$ คือ เซตของอินเท็นที่เกี่ยวข้องของเอกสาร d_r

➢ **นิยามที่ 2 จำนวนครั้งที่ครอบคลุมอินเท็น (Coverage Time)**

Coverage time คือ จำนวนครั้งที่มีการครอบคลุมอินเท็นทั้งหมด โดยให้ r คือ ตำแหน่งของเอกสาร d_r และ k คือ ตำแหน่งของผลการค้นหาที่ต้องการประเมิน เพื่อคำนวณหา coverage time ผู้วิจัยเริ่มคำนวณหา interval coverage หรือ *ic(r, k)* โดยเริ่มหาคำนวณหาความครอบคลุมอินเท็นย่อยจากเอกสารที่ถูกจัดอันดับจากตำแหน่งที่ r จนถึงตำแหน่งที่ k โดยการคำนวณ *ic(r, k)* นั้นประกอบด้วย 3 ตัวแปรหลัก ได้แก่ I คือ เซตของอินเท็นทั้งหมด และ C คือ เซตของอินเท็นที่ครอบคลุมใน r และ k คือ ตำแหน่งของเอกสารโดยมีเงื่อนไขดังต่อไปนี้

$$ic(r,k) = \begin{cases} 1, & \text{if } |C \cup intent(d_r)| = |I| \\ \text{cov}(p,k), & \text{if } r = k \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

$$C = \begin{cases} \varnothing, & \text{if } |C \cup intent(d_r)| = |I| \\ C \cup intent(d_r), & \text{if } |C \cup intent(d_r)| \neq |I| \text{ and } r \neq k \end{cases} \tag{3.3}$$

$$p = r+1, \qquad \text{if } |C \cup intent(d_r)| = |I| \tag{3.4}$$

โดย p คือ ตัวเก็บค่าจุดเริ่มต้นของแต่ละรอบกล่าวคือ เมื่อเอกสารครอบคลุมเอกสารทั้งหมดแล้ว ($|C \cup intent(d_r)| = |I|$) ค่า $ic(r,k)$ จะเป็น 1 และจุดเริ่มต้น p จะถูกตั้งค่าใหม่และให้เซต C เป็นเซตว่าง และในกรณีที่เอกสารไม่ครอบคลุมเอกสารทั้งหมด และค่า r ยังไม่ถึงตำแหน่งสุดท้าย ($|C \cup intent(d_r)| \neq |I|$ and $r \neq k$) ค่า $ic(r,k)$ จะเป็น 0 และ จะทำการรวมอินเท็นในเอกสาร

นั้นลงในเซต C และถ้าค่า r ถึงตำแหน่งสุดท้าย ($r = k$) ค่า $ic(r,k)$ จะเข้าฟังชัน $\text{cov}(p,k)$ โดยในครั้งแรกที่ $ic(r,k)$ คำนวน C จะถูกตั้งค่าเป็นเซตว่าง และค่า p ถูกเซตเป็น 1 โดย $ic(r,k)$ จะถูกรวมใน coverage time ซึ่งถูกนิยามดังนี้

$$ct(k) = \sum_{r=1}^{k} ic(r,k) \tag{3.5}$$

โดย *ct(k)* คือ จำนวนครั้ง (รวมถึงสัดส่วนที่เหลือของอินเท็น) ที่ครอบคลุมทุกอินเท็น ณ ตำแหน่ง k

➢ **นิยามที่ 3 ความถี่ของการครอบคลุม (Coverage frequency)**

ความถี่ของการครอบคลุมเป็นอัตราส่วนระหว่างค่า coverage time และระยะเวลาที่ผู้ค้นหาค้นหาเอกสารและหยุด ณ ตำแหน่ง k โดยเมื่อแทนที่ระยะเวลา หรือความพยายามของผู้ค้นด้วยตำแหน่งสุดท้ายที่ผู้ใช้พิจารณา จะได้สมการ CF ดังนี้

$$CF(k) = \frac{ct(k)}{k} \tag{3.6}$$

การประเมินนั้นจะมีการคำนวนเฉพาะตำแหน่ง 5 10 และ 20 ส่วนหนึ่งเนื่องจาก การใช้พูลลิ่งเทคนิค ทำให้ถูกจำกัดจำนวนการตัดสินความเกี่ยวข้องในการวิเคราะห์ที่ตำแหน่งต้นๆเท่านั้น โดยหลังจากมีการคำนวนค่า CF แล้ว ค่า CF จะการนอร์มัลไลเซชั่นด้วยค่าสูงสุด ที่ได้จากผลการค้นหาในอุดมคติของ CF เพื่อเปรียบเทียบเมื่อพิจารณาข้ามแต่ละหัวข้อ

$$nCF@k = \frac{CF(k)}{CF(k)'} = \frac{\sum_{r=1}^{k} ct(r)}{\sum_{r=1}^{k} ct(r)'} \tag{3.7}$$

โดย *CF(k')* คือผลการค้นหาที่ทำให้เกิดคะแนน CF สูงที่สุด ณ ตำแหน่งที่ k ซึ่งเกิดขึ้นด้วยอัลกอลิทึมแบบละโมภ (Greedy algorithm) โดยเลือกเอกสารที่มีจำนวนอินเท็นมากที่สุด แล้วจึงเลือกเอกสารที่มีอินเท็นที่หายไปมากที่สุด ด้วยจำนวนอินเท็นที่น้อยที่สุด จนกว่าเอกสารจะครอบคลุมอินเท็นทั้งหมด และเมื่อครอบคลุมอินเท็นทั้งหมดแล้ว ก็จะเริ่มกระบวนการเดิมอีกครั้ง โดยวิธีดังกล่าวจะทำซ้ำจนถึงตำแหน่งที่ k

3.6 การใช้งานอย่างมีศักยภาพของตัวชี้วัด

การใช้งาน nCF ให้มีศักยภาพมากขึ้นกระทำโดยพิจารณารวมกับตัวชี้วัดอื่น เช่น ERR-IA และ MAP-IA [24] เพื่อให้สามารถวัดปัจจัยอื่นๆในการค้นคืน เช่น การซ้ำซ้อนของเอกสาร การพิจารณาความเกี่ยวข้องหลายระดับ และความน่าจะเป็นของอินเท็น โดย nCF ใช้ ERR-IA รวมเพื่อพิจารณาความซ้ำซ้อนรวมถึงพิจารณาความเกี่ยวข้องหลายระดับ และความน่าจะเป็นซึ่งสามารถรวมกันได้ 2 ลักษณะคือ วิธีการแบบขั้นลำดับ และวิธีการแบบชดเชยกัน

3.6.1. วิธีการแบบขั้นลำดับ

ในบริบทของการค้นคืนสารสนเทศเชิงโต้ตอบ ผู้ใช้เป็นไปได้ที่จะพิจารณาความหลากหลายก่อน และพิจารณาความซ้ำซ้อนตามมา โดยตามข้อแนะนำจาก TREC 2010 ถึง 12 ในหมวดของเว็บได้กล่าวไว้ว่า ระบบค้นคืนควรสร้างการจัดลำดับของเอกสารโดย "คืนเอกสารให้ครอบคลุมพร้อมกับหลีกเลี่ยงความซ้ำซ้อน" ดังนั้น จึงพิจารณาจาก nCF ซึ่งเป็นตัวชี้วัดประเภทความหลากหลายแล้วจึงโดยพิจารณา ERR-IA ซึ่งเป็นตัวชี้วัดประเภทความซ้ำซ้อน กล่าวคือ เมื่อคะแนน nCF ของระบบค้นคืนเท่ากันจึงจะพิจารณา ERR-IA โดยวิธีนี้ถูกตั้งชื่อว่า SW-nCF ⟶ERR-IA

3.6.2. วิธีการแบบชดเชยกัน

วิธีรวมแบบชดเชยกันจะรวมตัวชี้วัดแบบเส้นตรง (Linear combination) ระหว่าง nCF และ ERR-IA ดังสมการต่อไปนี้

$$nCF+ERR\text{-}IA@k = \gamma nCF@k + (1-\gamma)ERR\text{-}IA@k \qquad (3.8)$$

โดยที่ γ คือความสัมพันธ์ระหว่างความหลากหลาย และ ความซ้ำซ้อนโดยจะถูกตั้งเป็น 0.5 ซึ่งเป็นตัวปรับสมดุลระหว่างความหลากหลาย และ ความซ้ำซ้อนโดยวิธีนี้ถูกตั้งชื่อว่า nCF+ERR-IA

3.7 วิเคราะห์แนวคิดรากฐานของตัวชี้วัดของ nCF

จากตารางที่ 3.2 ในส่วนของตารางด้านล่างระบุถึงระบบจำลองที่ถูกประเมินจากตัวชี้วัด nCF ซึ่งประกอบด้วย *1)* nCF *2)* SW-nCF⟶ERR-IA *3)* nCF+ERR-IA โดยมี ct (coverage time) คือจำนวนรอบที่มีการครอบคลุมอินเท็นทั้งหมด

ตัวชี้วัด nCF ประเมินให้ SyntheticSys4 เป็นระบบค้นคืนที่ดีที่สุด เพราะเอกสารที่ค้นคืนครอบคลุมอินเท็นทั้งหมดจำนวน 3 ครั้งด้วยคะนน nCF@10 เท่ากับ 0.3 ในขณะเดียวกัน nCF ประเมินให้ SyntheticSys1 เป็นระบบที่แย่ที่สุด เพราะเอกสารไม่ครอบคลุมอินเท็นทั้งหมด แม้แต่ครั้งเดียว โดยครอบคลุมเฉพาะอินเท็น 3 ใน 4 ของอินเท็นทั้งหมด และ nCF ประเมิน SystheticSys2

และ 3 ให้มีคะแนนเท่ากันคือ nCF@10 เท่ากับ 0.175 เนื่องจากทั้งสองระบบยังคงคืนเอกสารที่หลากหลาย แม้ว่าจะครอบคลุมหัวข้อทั้งหมดแล้ว นอกจากนี้ เพื่อให้สามารถประเมินความซ้ำซ้อนและความน่าจะเป็นของอินเท็นจึงเกิดการรวมกันระหว่าง nCF และ ERR-IA โดย 2 วิธีการคือ วิธีการแบบขั้นลำดับ และวิธีการแบบชดเชยกัน

สำหรับวิธีแบบลำดับขั้น (SW-nCF⟶ERR-IA) แนะนำระบบ SyntheticSys2 ว่าเป็นระบบที่ดีกว่าระบบ SyntheticSys3 เนื่องจาก ระบบแรกมีความซ้ำซ้อนของอินเท็นที่ 3 น้อยกว่า แต่เมื่อพิจารณาด้วยวิธีแบบชดเชยกัน (nCF+ERR-IA) แล้วพบว่าประเมินให้ระบบ SyntheticSys4 เป็นระบบที่ดีที่สุด และตามด้วยระบบ SyntheticSys1 ซึ่งไม่เคยค้นคืนอินเท็นที่ 1 เลย ดังนั้นการรวมกับแบบเส้นตรงนั้นส่งผลให้ยากต่อความเข้าใจระบบค้นคืน เช่น SyntheticSys1 ดีกว่าในด้านการประเมินความหลากหลาย หรือความซ้ำซ้อน เพราะตัวชี้วัด 2 ตัวรวมกัน มาจากลักษณะการประเมินที่แตกต่างกัน

ตารางที่ 3.2 ระบบจำลอง 5 ระบบที่คืนเอกสารจาก TREC 2012 ในค้นหาที่ 154 โดยถูกประเมินด้วย α-nDCG, ERR-IA, I-rec และ D#-nDCG (ตารางด้านบน) และถูกประเมินด้วย nCF, SW-nCF⟶ERR-IA และ nCF+ERR-IA (ตารางด้านล่าง)

System ranking observed at rank 10, according to:					
α-nDCG	#1 (0.677)	#2 (0.621)	#3 (0.616)	#4 (0.587)	#5 (0.534)
ERR-IA	#1 (0.675)	#2 (0.553)	#3 (0.552)	#4 (0.468)	#5 (0.449)
I-rec	#2 (0.750)	1# (1.000)			
D#-nDCG	#4 (0.555)	#1 (0.656)		#3 (0.642)	#2 (0.644)

	SyntheticSys1				SyntheticSys2			SyntheticSys3			SyntheticSys4			SyntheticSys5		
rank	intent	ct	cn		intent	ct	cn	intent	ct	cn	intent	ct	cn	intent	ct	cn
#1	2 3 4	0.75	3		1 3	0.5	2	1 3	0.50	2	4	0.25	1	4	0.25	1
#2	2 4	0.75	5		4	0.75	3	4	0.75	3	1 3	0.75	3	1 3	0.75	3
#3		0.75	5			0.75	3		0.75	3		0.75	3		0.75	3
#4		0.75	5			0.75	3		0.75	3		0.75	3		0.75	3
#5	3 4	0.75	7		2 3 4	1.00	6	2 3 4	1.00	6	2	1.00	4	2 4	1.00	5
#6	4	0.75	8		4	1.25	7	4	1.50	7	3 4	1.50	6	3 4	1.50	7
#7		0.75	8			1.25	7		1.50	7		1.50	6		1.50	7
#8		0.75	8		1 3 4	1.75	10	1 3 4	1.75	10	1 2	2.00	8	4	1.50	8
#9	2 4	0.75	10		4	1.75	11	4	1.75	11	2 3 4	2.75	11	3 4	1.50	10
#10	2 3	0.75	12		3	1.75	12	4	1.75	12	1	3.00	12	3 4	1.50	12

nCF	#4 (0.075)	#2 (0.175)		#1 (0.300)	#3 (0.150)
SW-nCF⟶ ERR-IA	#5 (0.075)	#2 (0.175)⟶(0.553)	#3 (0.175)⟶(0.552)	#1 (0.300)	#4 (0.150)
nCF+ERR-IA	#2 (0.375)	#3 (0.364)	#4 (0.363)	#1 (0.384)	#5 (0.299)
System ranking observed at rank 10, according to:					

บทที่ 4
การทดลองความสอดคล้องของตัวชี้วัด

ในบทนี้ผู้วิจัยจะกล่าวถึงความสอดคล้องในการประเมินระหว่างตัวชี้วัด และความสอดคล้องในการประเมินระหว่างตัวชี้วัดกับความชอบของผู้ใช้ โดยมีสมมติฐานดังนี้

- ผลการประเมินระบบค้นคืนของตัวชี้วัด nCF มีความสอดคล้องกับตัวชี้วัดอื่นๆ
- ผลการประเมินระบบค้นคืนของตัวชี้วัด nCF มีความสอดคล้องกับความชอบผู้ใช้ที่ต้องการอินเท็นทั้งหมด
- ผลการประเมินระบบค้นคืนของตัวชี้วัด α-nDCG มีความสอดคล้องกับความชอบผู้ใช้ที่ต้องการอินเท็นเดียว

4.1 ความสอดคล้องระหว่างตัวชี้วัด

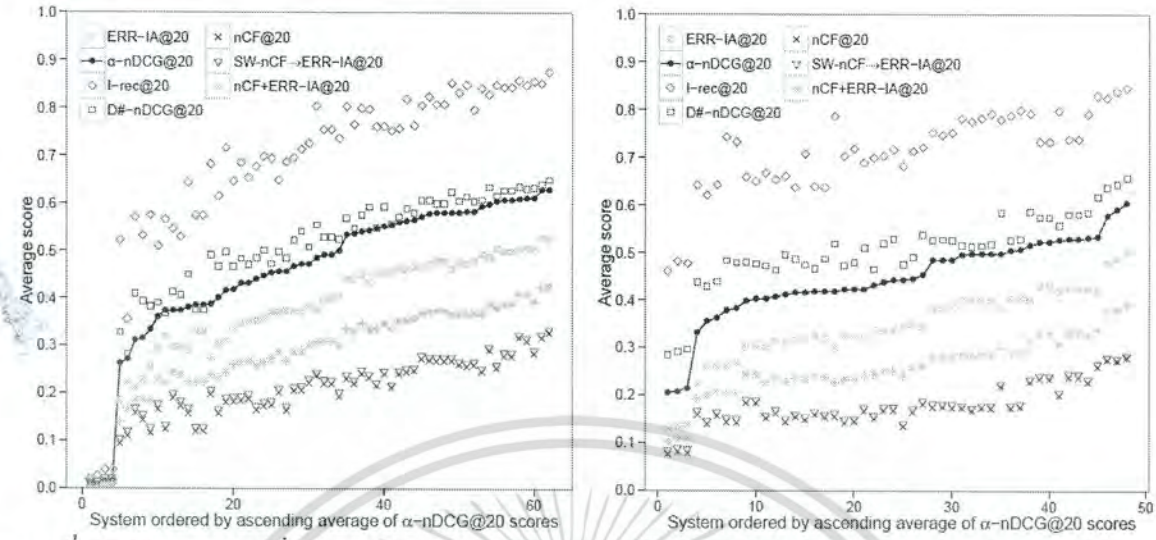
4.1.1. แผนการทดลอง

การทดลองหาความสอดคล้องระหว่างตัวชี้วัดโดยใช้วิธีของเคนดัลเทา ซึ่งเปรียบเทียบตัวชี้วัดใหม่กับตัวชี้วัดเดิมว่ามีการจัดอันดับของการประเมินระบบเหมือนกันแตกต่างกันอย่างไร จากการวิเคราะห์ระบบจาก TREC 2011 และ TREC 2012 ในหมวดของเว็บ [35] โดยได้นำระบบที่ถูกส่งเข้ามาใน TREC ในแต่ละปี จากผู้เข้าร่วมในหมวดเว็บ ซึ่งรวบรวมได้ 62 ระบบจาก 16 กลุ่มใน TREC 2011 และได้ 20 ระบบจาก 8 กลุ่มใน TREC 2012 นอกจากนี้ผู้วิจัยได้ลดจากความเกี่ยวข้องแบบมีหลายระดับเป็นความเกี่ยวข้อง 2 ระดับ และให้ค่าความน่าจะเป็นของแต่ละอินเท็นเท่ากัน เพื่อการประเมินที่เป็นมาตรฐานเดียวกับ TREC

4.1.2. ผลการทดลอง

ผู้วิจัยได้แบ่งผลการทดลองความสอดคล้อง ออกเป็น 2 มิติคือ 1. มิติของศักยภาพของ คือ มิติที่แสดงถึงผลการประเมินของของระบบ ที่ได้จากตัวชี้วัดที่แตกต่างกัน 2. มิติของค่าความสอดคล้องของระบบ คือ มิติที่แสดงถึงค่าความสอดคล้องของการประเมินผลการค้นหา ที่ได้จากการคำนวณค่าความสอดคล้องเคนดัลเทา

1. มิติของศักยภาพของระบบ



ภาพที่ 4.1 ค่าคะแนนเฉลี่ยของระบบของ TREC 2011 (ซ้าย) และ 2012 (ขวา) โดยประเมินจาก nCF, α-nDCG, ERR-IA, *I*-rec และ D#-nDCG ณ ตำแหน่งที่ 20

ภาพที่ 4.1 แสดงถึงศักยภาพของระบบที่ถูกประเมินจาก 7 ตัวชี้วัดโดยแต่ละจุดบนกราฟคือ แต่ละระบบ โดยระบบจะถูกเรียงลำดับตามคะแนนของ α-nDCG ตามแนวแกนนอน และแนวแกนตั้งแสดงถึงคะแนนที่ได้จากตัวชี้วัด และสัญลักษณ์ในกราฟจะบ่งบอกตัวชี้วัดดังต่อไปนี้

ก. สัญลักษณ์รูปเพชรสีแดง คือ *I*-rec

ข. สัญลักษณ์สี่เหลี่ยมจัตุรัสสีม่วงคือ D#-nDCG

ค. สัญลักษณ์ทรงกลมสีดำคือ α-nDCG

ง. สัญลักษณ์สี่เหลี่ยมรวมกับกากบาทสีเขียวคือ ERR-IA

โดยตัววัดดังกล่าวถูกคำนวณจาก ndeval (สามารถดูได้ที่ภาคผนวก ค) ที่ใช้ในหมวดเว็บของ TREC และ สัญลักษณ์ในกราฟของตัวชี้วัดที่ผู้วิจัยสร้างขึ้นมีดังนี้

ก. สัญลักษณ์รูปดาวสีเหลืองคือ nCF+ERR-IA

ข. สัญลักษณ์สามเหลี่ยมสีน้ำตาลคือ SW-nCF ⟶ ERR-IA

ค. สัญลักษณ์กากบาทสีน้ำเงินคือ nCF

การขึ้นหรือลงของแต่ละจุดของ nCF (สีน้ำเงิน) ในกราฟบ่งบอกข้อขัดแย้งกับ α-nDCG และ เมื่อพิจารณาถึงแนวโน้มของกราฟแล้ว การขึ้นและลงของ nCF มีความสอดคล้องกับ D#-nDCG ซึ่งจะถูกยืนยันจากค่าความสอดคล้องในมิติถัดไป

ผู้วิจัยได้วิเคราะห์ถึงความไม่สอดคล้องที่เกิดขึ้นกับตัวชี้วัด nCF โดยการพิจารณาจากผลลัพธ์จากตารางที่ 4.1 โดยแสดงถึงความต่างของตัวชี้วัดในการประเมินความหลากหลายของระบบ เหตุผลหลักของความไม่สอดคล้องเกิดจาก α-nDCG และ ERR-IA ลดประโยชน์ของเอกสาร

ที่เกี่ยวข้องขึ้นอยู่กับตำแหน่ง และความซ้ำซ้อนของอินเท็นในเอกสาร และD#-nDCG ไม่สามารถประเมินความหลากหลายได้ต่อเมื่อครอบคลุมอินเท็นทั้งหมดแล้ว

2. มิติของค่าความสอดคล้องของระบบ

ตารางที่ 4.1 ค่าความสอดคล้อง Kendal's τ ของระบบจาก TREC 2011 และ 2012 โดยถูกประเมิน ณ ตำแหน่งที่ 20 ระหว่าง nCF เปรียบเทียบกับ α-nDCG, ERR-IA, *I*-rec และ D#-nDCG

	TREC	α-nDCG	ERR-IA	I-rec	D#-nDCG
nCF	2011	0.87626	0.862822	0.868421	0.899776
	2012	0.787234	0.766844	0.703901	0.806738

ผู้วิจัยได้ใช้วิธีการวัดที่มีชื่อว่า Kendall's τ เพื่อวัดความสอดคล้องแต่ละตัวชี้วัดโดย Kendall's τ มีค่าอยู่ระหว่าง -1 (การจัดอันดับระหว่าง 2 ตัวชี้วัดกลับกัน) ถึง 1 (การจัดอันดับระหว่าง 2 ตัวชี้วัดเหมือนกัน) โดยงานวิจัยก่อนหน้า [36] ได้แนะนำว่าค่า τ ที่เท่ากับ 0.9 หรือมากกว่าระบุถึงมีความสอดคล้องสูงแต่ถ้าค่า τ น้อยกว่า 0.8 ระบุถึงความแตกต่างของการจัดอันดับ โดยจากตารางที่ 4.1 การวิเคราะห์แสดงถึงความสอดคล้องของ nCF ค่อนข้างสูงกับตัวชี้วัดอื่นโดยเฉพาะอย่างยิ่งตัวชี้วัด D#-nDCG มีค่ามากกว่า 0.8 ในทุกปี

4.2 ความสอดคล้องระหว่าตัวชี้วัดกับความชอบของผู้ใช้

ในการทดลองนี้ประกอบด้วย 4 องค์ประกอบหลักคือ 1. คอเลคชั่นทดสอบ 2. การประเมินความชอบผู้ใช้ 3. ผู้ร่วมการทดลอง 4. การเลือกคู่ระบบเพื่อแสดงให้ผู้ใช้ โดยองค์ประกอบเหล่านี้จะถูกอธิบายในลำดับถัดไป

1. คอเลคชั่นทดสอบ

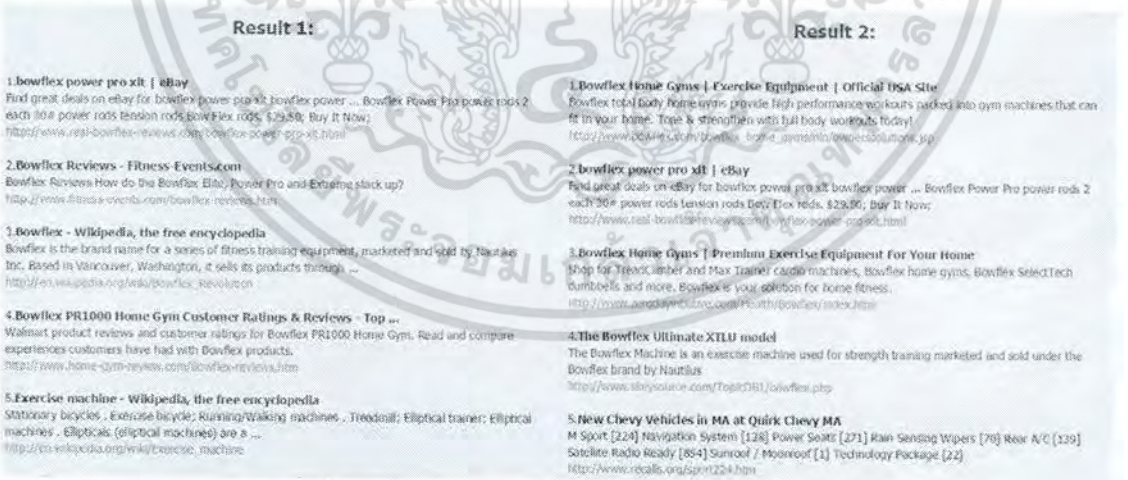
ผู้วิจัยได้ใช้คอเลคชั่นที่มีชื่อว่า Clueweb09 ซึ่งประกอบไปด้วย 50 ล้านเอกสาร และเป็นเอกสารที่ใช้ใน TREC 2012 ในหมวดเว็บ โดยแต่ละอินเท็นจะมี 2 ประเภทเป็น ข้อมูลที่เจาะจง (Navigational) หรือข้อมูลทั่วไป (Informational) [37] และคำค้นหาจะมีประเภทเป็น คำค้นหาที่มีความกำกวม หรือคำค้นหาที่มีหลายแง่มุมหรือเจาะจงไม่เพียงพอ ดังตัวอย่างจากภาพที่ 4.2

```
<topic number="155" type="faceted">
  <query>last supper painting</query>
  <description>
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </description>
  <subtopic number="1" type="nav">
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </subtopic>
  <subtopic number="2" type="nav">
    Are tickets available online to view da Vinci's Last Supper in Milan, Italy?
  </subtopic>
  <subtopic number="3" type="inf">
    What is the significance of da Vinci's interpretation of the Last Supper in
    Catholicism?
  </subtopic>
</topic>
```

ภาพที่ 4.2 ตัวอย่างคำค้นที่ 155 จาก TREC 2012

2. การประเมินความชอบผู้ใช้

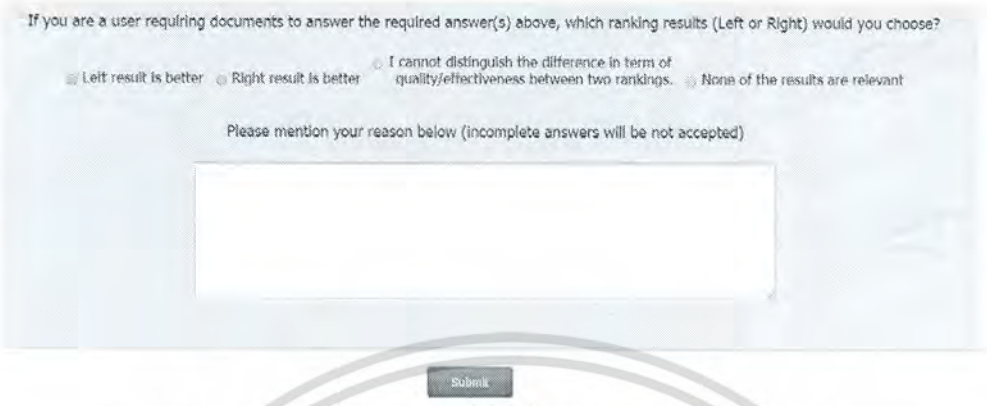
ผู้วิจัยได้เลือกการประเมินความชอบผู้ใช้ระหว่างคู่ระบบที่แตกต่างกัน โดยใช้วิธีเทียบเคียงจาก Thomus [29] โดยแต่ละระบบจะประกอบด้วย 10 เอกสารซึ่งประกอบด้วย ชื่อเรื่อง คำอธิบายเว็บ และยูอาร์แอล (Url) ซึ่งจะถูกแสดงให้กับผู้ทดลอง พร้อมกับคำค้นหาที่ใช้ในการค้นหา รวมทั้งอินเท็นที่ต้องการให้ผู้ใช้หา โดยคำอธิบายของเว็บสร้างขึ้นจาก Bing api หรือใช้ข้อมูลจากแท็ก (Tag) ในเอกสาร ดังในภาพที่ 4.3



ภาพที่ 4.3 คู่ของผลการค้นหาที่แสดงให้ผู้ใช้พิจารณา

ผู้ใช้จะถูกให้เลือกว่าผลการค้นหาใดของ 2 ผลการค้นหาที่ตรงกับความต้องการที่กำหนดให้ดังภาพ แล้วจึงนำผลมาวัดความสอดคล้องระหว่างผู้ใช้และตัวชี้วัด โดยจะมีตัวเลือกให้ผู้ใช้ 4 ตัวเลือกคือ *1)* ผลการค้นหาทางซ้ายดีกว่าทางขวา พอๆกัน *2)* ผลการค้นหาทางขวาดีกว่า

ทางซ้าย *3)* ไม่สามารถแยกความแตกต่างระหว่าง 2 ผลการค้นหา *4)* ทั้ง 2 ผลการค้นหาไม่ดีทั้งคู่ และให้ผู้ใช้ใส่เหตุผลในการเลือก ซึ่งแสดงดังในภาพที่ 4.4



ภาพที่ 4.4 ตัวเลือก 4 ตัวที่ให้ผู้ใช้ใน Mturk เลือกระหว่าง 2 ผลการค้นหา

3. ผู้ร่วมทดลอง

เป้าหมายของการทดลองเพื่อที่จะประเมินความชอบผู้ใช้จากคู่ระบบที่แตกต่างกันโดยใช้ระบบการกระจายปัญหาไปยังกลุ่มค้นเพื่อค้นหาคำตอบ (Crowdsourcing) ของ Mechanical Turk [38] โดยผู้ใช้จาก Mechanical Turk จะตัดสินแต่ละคู่ระบบ รวมทั้งให้ใส่เหตุผลในการเลือก

4. การเลือกคู่ระบบ

การเลือกคู่ระบบเพื่อแสดงให้ผู้ใช้นั้นจะอยู่บนพื้นฐานของระบบค้นคืนที่หลากหลาย และสามารถเปรียบเทียบระหว่างตัวชี้วัดได้ ดังนั้นผู้วิจัยจึงสุ่มระบบมาทั้งหมด 1000 ระบบในแต่ละหัวข้อโดยคัดเลือกระบบที่มีค่า ERR และ α-nDCG มากกว่า 0.7 เพราะ ต้องการระบบที่มีประสิทธิผลสูงแล้ว จึงจับคู่ในทุกระบบเพื่อนำมาคัดเลือก ดังนั้นจะมีทั้งหมด 1000 คูณ 999 คู่ในแต่ละหัวข้อ โดยจะกรองคู่ระบบ ที่ทำให้เกิดความขัดแย้งระหว่าง ตัวชี้วัด nCF เปรียบเทียบกับ ERR-IA และ α-nDCG กล่าวคือ เลือกคู่ผลการค้นหาที่ผลการค้นหาแรกถูกประเมินว่าดีกว่าผลการค้นหาที่สองในตัวชี้วัดแรก แต่ในผลการค้นหาที่สองถูกประเมินดีกว่าผลการค้นหาแรกในตัวชี้วัดที่สอง โดยสามารถเขียนได้ตามสมการในภาพที่ 4.7 หลังจากสร้างคู่ระบบแล้วจึงเลือกคู่มาเพื่อให้ผู้ใช้พิจารณาโดยมีทั้งหมด 10 คู่ระบบ จาก 10 หัวข้อ โดยจะแบ่งการทดลองเป็น 2 ประเภทตามความต้องการผู้ใช้ดังนี้

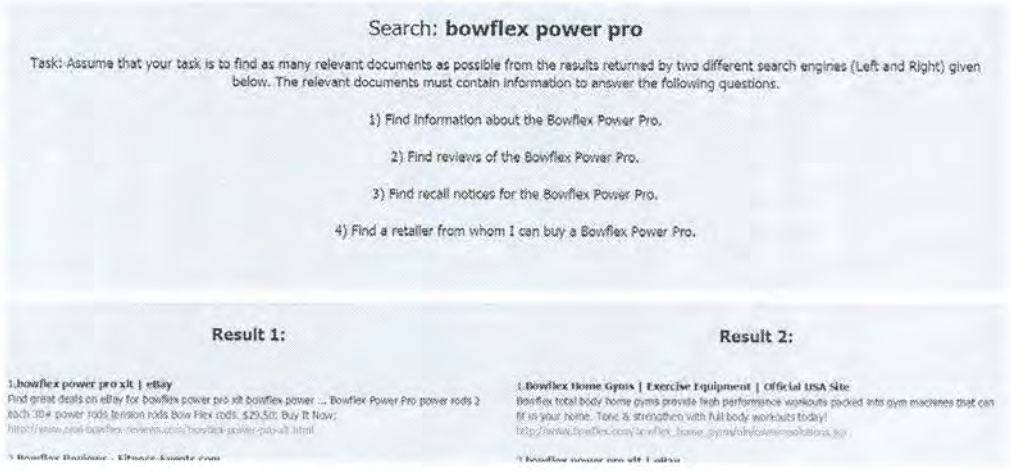
ก. ผู้ใช้ที่มีความต้องการอินเท็นเดียว

ผู้ใช้ที่มีความต้องการเพียงหนึ่งเดียว แต่ใส่คำค้นหาสั้นๆ และกำกวมหรือระบุไม่เจาะจง ทำให้ถูกตีความได้หลายความหมาย หรือมีหลายแง่มุม โดยผู้ใช้ประเภทนี้จะพอใจ เมื่อได้เอกสารที่ตรงกับอินเท็นเพียงหนึ่งเดียวที่ผู้ใช้ต้องการ โดยไม่สนใจข้อมูลในแง่มุมอื่น เช่น คำค้นหา "Panda" ซึ่งอาจตีความหมายได้ว่าเป็น *1)* สัตว์เลี้ยงลูกด้วยนม *2)* โปรแกรมสแกนไวรัส *3)* แอนิเมชั่น แต่ผู้ใช้นั้นสนใจเฉพาะ Panda ที่เป็นสัตว์เท่านั้น โดยให้ผู้ร่วมทดลอง 3 คนประเมินผลการค้นหาบนความต้องการในแต่ละอินเท็น โดยจะแสดงให้ผู้ใช้ดังภาพที่ 4.5

ภาพที่ 4.5 คู่ระบบที่แสดงให้ผู้ใช้ที่ต้องการอินเท็นเดียว

ข. ผู้ใช้ที่มีความต้องการอินเท็นทั้งหมด

ผู้ใช้ที่มีความต้องการข้อมูลที่หลากหลาย หรือผู้ใช้ที่ไม่แน่ใจว่าต้องการอะไรจึงต้องการข้อมูบที่หลากหลาย ดังนั้นผู้ใช้ประเภทนี้จะพอใจเมื่อได้หาที่ข้อมูลที่คลอบคลุมทั้งหมด เช่น คำค้นหา "Panda" ผู้ใช้ต้องการข้อมูลทั้งหมดไม่ว่าจะเป็น สัตว์เลี้ยงลูกด้วยนม โปรแกรมสแกนไวรัส และ แอนิเมชั่น โดยจะให้ผู้ร่วมทดลอง 10 คนประเมินผลการค้นหาบนความต้องการอินเท็นทั้งหมด ซึ่งมีจำนวนน้อยกว่าผู้ร่วมทดลองที่มีความต้องการในอินเท็นเดียว เนื่องจากผู้ทดลองในอินเท็นเดียวมีจำนวนมากกว่า (3 คนต่อ 1 อินเท็น) ดังนั้นเพื่อให้เกิดความสมดุล ผู้วิจัยจึงกำหนดผู้ร่วมการทดลองมีจำนวนที่มากกว่า โดยจะแสดงให้ผู้ใช้ดังภาพที่ 4.6



ภาพที่ 4.6 คู่ระบบที่แสดงให้ผู้ใช้ที่ต้องการอินเท็นทั้งหมด

4.2.1. ผลการทดลอง

ผลการทดลองที่ได้จาก Mechanical Turk เกิดจาก การตัดสินของผู้ทดลอง โดยคัดการตัดสินเฉพาะผู้ร่วมทดลอง ที่มี Approval rate[3] มากกว่า 70 % และให้เหตุผลอย่างมีเหตุมีผล ซึ่งมีการตัดสินจากผู้ใช้ทั้งหมด จำนวน 205 การตัดสิน ซึ่งประกอบด้วย คำตัดสินของผู้ใช้ที่ต้องการอินเท็นเดียวมีทั้งหมด 105 การตัดสิน ในตารางที่ 4.2 และ คำตัดสินของผู้ใช้ที่ต้องการทุกอินเท็นมีทั้งหมด 100 การตัดสิน ในตารางที่ 4.3 โดยตารางที่ 4.2 แสดงให้เห็นว่า เมื่อผู้ใช้มีความต้องการในอินเท็นเดียว ผู้ใช้จะมีความชอบ ERR-IA และ α-nDCG มากกว่า nCF และ รองลงมาคือ *I*-rec ที่ไม่สามารถประเมินประสิทธิผลของระบบค้นคืนได้เลย ในทางกลับกัน ตารางที่ 4.3 แสดงให้เห็นว่า เมื่อผู้ใช้มีความต้องการอินเท็นทั้งหมด ผู้ใช้จะมีความชอบ nCF มากกว่า ERR-IA และ α-nDCG โดย *I*-rec ยังคงไม่สามารถประเมินประสิทธิผลของระบบค้นคืนได้

จากผลการทดลองสรุปได้ว่า *I*-rec นั้น ไม่สามารถประเมินระบบการค้นคืนได้ตรงกับความต้องการของผู้ใช้เลย เนื่องจาก *I*-rec ไม่สามารถแยกแยะ ความแตกต่างของสองระบบได้ ในขณะที่ ERR-IA และ α-nDCG สามารถประเมินระบบได้ตรงกับความต้องการของผู้ใช้ที่ต้องการอินเท็นต์เดียว เนื่องจาก เมื่อประเมินด้วย ERR-IA และ α-nDCG สามารถประเมินระบบที่คืนเอกสารได้เร็วกว่าในอันดับแรก แต่อย่างไรก็ตาม ERR-IA และ α-nDCG ไม่ได้คำนึงถึงความหลากหลายในระบบ ทำให้ผลการประเมินของ nCF สามารถประเมินระบบได้ตรงกับความต้องการของผู้ใช้ที่ต้องการอินเท็นต์ทั้งหมด

---

[3] อัตราส่วนระหว่างจำนวนการตอบรับของการตัดสินและจำนวนการตัดสินทั้งหมดของผู้ใช้นั้นๆ

ตารางที่ 4.2 ตารางแสดงความชอบผู้ใช้ทั้งหมด เมื่อเปรียบเทียบกับตัวชี้วัด nCF α -nDCG ERR-IA และ *I*-rec เมื่อผู้ใช้ต้องการอินเท็นต์เดียว

User	nCF	α-nDCG	ERR-IA	*I*-rec
Agree	45 43%	58 55%	58 55%	0 0%
Rank equal	0 0%	0 0%	0 0%	2 2%
Disagree	60 57%	47 45%	47 45%	103 98%
	105	105	105	105

ตารางที่ 4.3 ตารางแสดงความชอบผู้ใช้ทั้งหมด เมื่อเปรียบเทียบกับตัวชี้วัด nCF α -nDCG ERR-IA และ *I*-rec เมื่อผู้ใช้ต้องการอินเท็นต์ทั้งหมด

User	nCF	α -nDCG	ERR-IA	*I*-rec
Agree	57 57%	30 30%	30 30%	0 0%
Rank equal	0 0%	0 0%	0 0%	13 13%
Disagree	43 43%	70 70%	70 70%	87 87%
	100	100	100	100

Algorithm 1 Generating pairs of systems of each query.

Require: $D_{150} = \{d_{1,150}, d_{2,150}, ..., d_{n,150}\}$, a set of documents in TREC 2012 pool, where $d_{n,q}$ is a document n with respect to query q.

Initialize:
  $outputs = \{\}$
procedure GENERATE PAIRS
  if $|I_{150}| >= 3$ **then**
    for i=1 to 1000 **do**
      create document set S of size $n = 10$ by randomly sampling without replacement from D_{150}.
      if $ERR\text{-}IA(S)$ and $\alpha\text{-}nDCG(S) > 0.7$ **then**
        $setRandoms = setRandoms \cup \{S\}$
      end if
    end for
  end if
  for each system pair $(S_x, S_y) \in setRandoms$ **do**
    if $ERR\text{-}IA(S_x) > ERR\text{-}IA(S_y)$ and $nCF(S_x) < nCF(S_y)$ **then**
      $outputs = outputs \cup \{(S_x, S_y)\}$
    end if
  end for
  return *outputs*, a set of system pairs to be manually selected.
end procedure

ภาพที่ 4.7 วิธีการสร้างคู่ระบบจำลองโดยการสุ่ม 1000 ระบบของแต่ละหัวข้อจาก qrel ของ TREC 2012

บทที่ 5
การทดลองความน่าเชื่อถือของตัวชี้วัด

เป้าหมายของการทดลองนี้เพื่อประเมินความน่าเชื่อถือของ nCF เมื่อเทียบกับตัวชี้วัดอื่นประเภทความหลากหลายและความซ้ำซ้อน ซึ่งผู้วิจัยทำการทดลองบนระบบค้นคืนที่มีศักยภาพสูงในด้านความหลากหลายของผลการค้นหา โดยมีสมมติฐานดังต่อไปนี้

- ตัวชี้วัด nCF ซึ่งเป็นตัวชี้วัดประเภทความหลากหลาย มีความน่าเชื่อถือกว่าตัวชี้วัดอื่นๆ
- เมื่อ nCF ประเมินผลการค้นหาร่วมกับตัวชี้วัดประเภทความซ้ำซ้อน แล้ว nCF มีแนวโน้มที่จะปรับปรุงความน่าเชื่อถือของตัวชี้วัดนั้นกว่าการนำ I-rec ประเมินร่วมกับตัวชี้วัดประเภทความซ้ำซ้อน

5.1 การวัดความน่าเชื่อถือของตัวชี้วัด

5.1.1. แผนการทดลอง

การทดลองหาความสอดคล้องระหว่างตัวชี้วัด เกิดจากการวิเคราะห์จากระบบใน TREC 2011 และ TREC 2012 ในหมวดเว็บ [35] ซึ่งรวบรวมได้ 62 ระบบจาก 16 กลุ่มในปี 2011 และได้ 20 ระบบจาก 8 กลุ่มในปี 2012 นอกจากนี้ผู้วิจัยได้ลดความเกี่ยวข้องแบบมีหลายระดับ เป็นความเกี่ยวข้อง 2 ระดับ และให้ค่าความน่าจะเป็นของแต่ละอินเท็นเท่ากัน โดยนำไปประยุกต์ใช้ในทุกตัวชี้วัด เพื่อการประเมินที่เป็นมาตรฐานเดียวกับ TREC

คอเลกชั่นทดสอบเป็นปัจจัยหนึ่งในการประเมินความหลากหลายและความซ้ำซ้อนให้มีประสิทธิผล ดังนั้นการที่ปรับเปลี่ยนคอเลกชั่นทดสอบนั้นไม่ได้ทำไปเพื่อให้ nCF มีความน่าเชื่อถือที่สูง ยกตัวอย่างเช่น ถ้าเอกสารทั้งหมดในคอเลกชั่นครอบคลุมเพียงแค่ 1 อินเท็น แม้แต่ระบบที่ดีที่สุดก็ไม่สามารถสร้างผลการค้นหาที่มีความหลากหลายได้หรือกล่าวได้ว่า ถ้าระบบไม่สามารถครอบคลุมอินเท็นทั้งหมดได้เลย แล้วตัวชี้วัดประเภทความหลากหลายจะสามารถประเมินผลการค้นหาได้อย่างไร

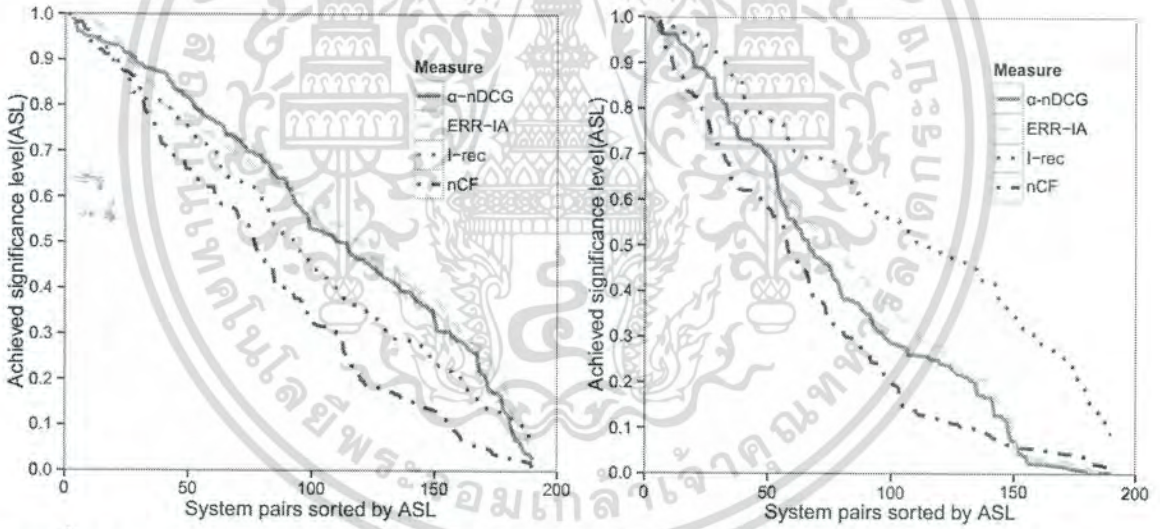
ดังนั้น ผู้วิจัยจึงคัดเลือกระบบที่มีความหลากหลายสำหรับการทดลอง โดยจะมีการเลือกคำค้นหาที่มีอินเท็นมากกว่า 3 แล้วจึงกรองคำค้นหาที่ไม่มีระบบใดเลยสามารถครอบคลุมอินเท็นทั้งหมด ณ ตำแหน่งที่ 10 (I-rec@10=1) หรือ กล่าวคือ คำค้นหานั้นต้องมีอย่างน้อย 1 ระบบที่บรรลุเป้าหมายของการค้นหาประเภทความหลากหลาย ดังนั้นหลังจากผ่านกระบวนการเลือกขั้นต้น ทำให้เหลือคำค้นหาทั้งหมด 40 คำค้นหาใน TREC 2011 และ 43 คำค้นหาใน TREC 2012 แล้วจึงเลือก

20 ระบบที่มีจำนวนคำค้นหาสูงสุด (จำนวนคำค้นหาที่ระบบสามารถทำคะแนน *I*-rec เท่ากับ 1) ดังนั้นการทดลองจะเน้นระบบที่มีศักยภาพสูงและมีผลการค้นหาที่มีความหลากหลายเท่านั้น

5.1.2. ผลการทดลองความน่าเชื่อถือด้วยวิธีการวัดพลังการแยกแยะ

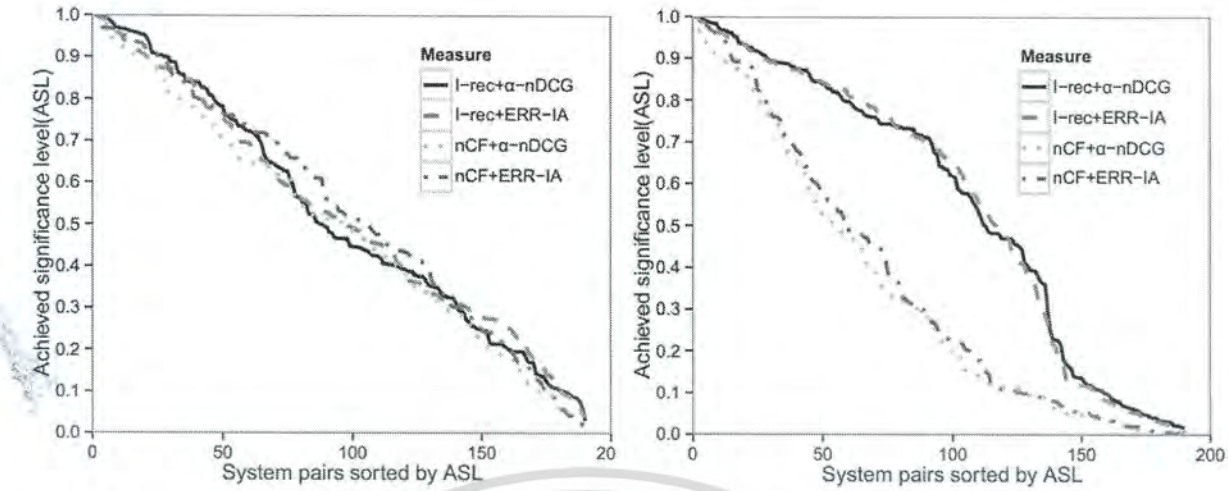
ผู้วิจัยใช้วิธีการวัดพลังการแยกแยะที่เสนอโดย Sakai [14] ซึ่งอยู่บนพื้นฐานของสมมติฐานบูทสแตร็ป (Bootstrap hypothesis) และใช้วิธีการทดสอบนัยสำคัญทางสถิติสองทางจากการสุ่ม 1000 ตัวอย่าง โดยกำหนดระดับนัยสำคัญเป็น 0.05

โดยในการทดลองนี้ ผู้วิจัยได้ทดลอง 2 ตัวชี้วัดประเภทความหลากหลายและ 2 ตัวชี้วัดประเภทความซ้ำซ้อน ดังนี้ *I*-rec, nCF, α-nDCG และ ERR-IA และผู้วิจัยได้ประเมินผลการค้นหาร่วมกันระหว่างตัวชี้วัดประเภทความหลากหลาย และความซ้ำซ้อนแบบเส้นตรง ซึ่งทำให้เกิด 4 ตัวชี้วัดแบบผสมคือ การประเมินร่วมกันระหว่าง *I*-rec กับตัวชี้วัดประเภทความซ้ำซ้อน นั้นคือ *I*-rec+ERR-IA และ *I*-rec+α-nDCG และ การประเมินร่วมกันระหว่าง nCF กับตัวชี้วัดประเภทความซ้ำซ้อน นั้นคือ nCF+ERR-IA และ nCF+α-nDCG



ภาพที่ 5.1 คะแนน ASL ของระบบจาก TREC 2011 (ซ้าย) และ 2012 (ขวา) ที่ถูกประเมินโดย 2 ตัวชี้วัดประเภทความหลากหลาย และ 2 ตัวชี้วัดประเภทความซ้ำซ้อน

ภาพที่ 5.1 และ ภาพที่ 5.2 แสดงกราฟของ Achieved Significance Level (ASL) สำหรับตัวชี้วัดประเภทความหลากหลาย และความซ้ำซ้อนในการประเมินระบบจาก TREC 2011 และ TREC 2012 และแกนในแนวนอนแสดงถึงคู่ระบบ 190 คู่เรียงตามคะแนนของ ASL จากน้อยไปมาก และแกนแนวตั้งแสดงค่าของ ASL โดยถ้ากราฟของตัวชี้วัดใดเข้าใกล้จุดเริ่ม หมายถึงมีความน่าเชื่อถือมาก



ภาพที่ 5.2 คะแนน ASL ของระบบจาก TREC 2011 (ซ้าย) และ 2012 (ขวา) ที่ถูกประเมินโดย 4 ตัวชี้วัดแบบผสม

ภาพที่ 5.1 แสดงให้เห็นว่า ERR-IA และ α-nDCG มีความน่าเชื่อถือน้อยที่สุดใน TREC 2011 และ *I*-rec มีความน่าเชื่อถือน้อยที่สุดใน TREC 2012 แต่สำหรับ nCF มีความน่าเชื่อมากกว่าตัวชี้วัดอื่นทั้ง TREC 2011 และ 12 และ ภาพที่ 5.2 แสดงตัวชี้วัดแบบผสมโดยใน TREC 2011 ไม่ค่อยเห็นความแตกต่างมากนัก แต่ใน TREC 2012 นั้นเมื่อประเมินด้วย nCF ร่วมกับตัวชี้วัดประเภทความซ้ำซ้อนแล้ว ตัวชี้วัดแบบผสมมีความน่าเชื่อถือเพิ่มมากขึ้นกว่าการประเมินร่วมกับตัวชี้วัด *I*-rec

ตารางที่ 5.1 แสดงให้เห็นว่ามีกี่คู่ระบบที่ตรงตามเงื่อนไข ASL น้อยกว่า 0.05 โดยคอลัมน์ที่สองคือ อัตราส่วนของพลังการแยกแยะ และคอลัมน์ที่สามคือความแตกต่างที่คาดการ (Estimated difference) ของระหว่างคู่ระบบ โดยถ้ามีค่ามากกว่าหรือเท่ากับ 0.9 แปลว่า ระบบทั้ง 2 มีความแตกต่างอย่างมีนัยสำคัญ [30] โดยใน TREC 2011 นั้นตัวชี้วัด nCF มีอัตราส่วนของพลังการแยกแยะมากที่สุดเป็น 10% และเมื่อประเมินด้วย nCF ร่วมกับตัวชี้วัดประเภทความซ้ำซ้อน ทำให้ความน่าเชื่อถือเพิ่มมากขึ้น (nCF+ERR-IA เป็น 4.20% และ nCF+α-nDCG เป็น 4.21%) และใน TREC 2012 นั้น α-nDCG มีพลังการแยกแยะมากที่สุดเป็น 38% และเมื่อประเมินด้วยกับ nCF แล้วความน่าเชื่อถือของตัวชี้วัดเพิ่มขึ้นรวมทั้งการประเมินร่วมกับ ERR-IA แต่สำหรับการประเมินด้วย I-rec ร่วมกับ α-nDCG และ ERR-IA ความน่าเชื่อถือลดลงหรือไม่เปลี่ยนแปลง

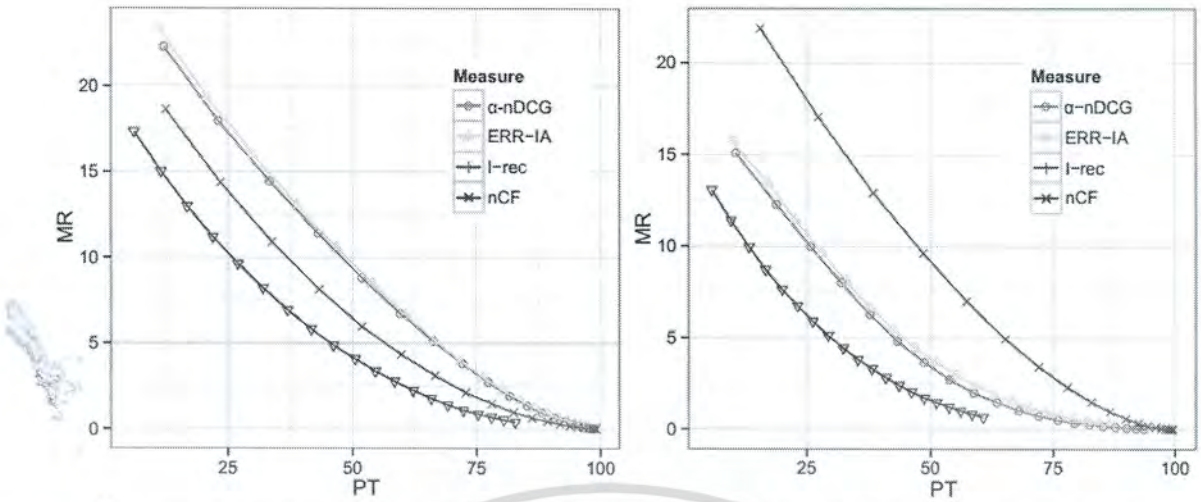
ตารางที่ 5.1 พลังการแยกแยะของตัวชี้วัดที่ระดับนัยสำคัญเท่ากับ 0.05 จาก TREC 2011 และ 2012

TREC 2011	ASL < 0.05	estimated diff.
I-rec	0/190=0%	0.11
nCF	19/190=10%	0.09
α-nDCG	5/190=2.63%	0.08
ERR-IA	1/190=0.52%	0.08
nCF+ERR-IA	8/190=4.2%	0.08
I-rec+ERR-IA	1/190=0.52%	0.08
nCF+α-nDCG	8/190=4.21%	0.08
I-rec-α-nDCG	1/190=0.52%	0.09

TREC 2012	ASL < 0.05	estimated diff.
I-rec	0/190=0%	0.08
nCF	23/190=10.52%	0.08
α-nDCG	38/190=20%	0.09
ERR-IA	18/190=9.47%	0.09
nCF+ERR-IA	36/190=18.94%	0.08
I-rec+ERR-IA	18/190=9.47%	0.09
nCF+α-nDCG	41/190=21.57%	0.09
I-rec-α-nDCG	17/190=8.94%	0.09

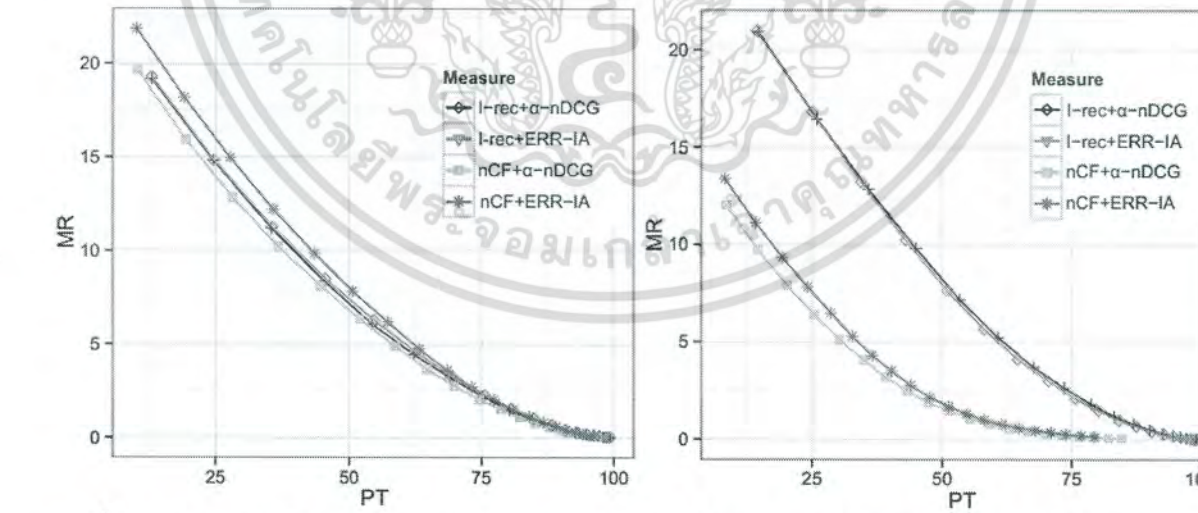
5.1.3. ผลการทดลองความน่าเชื่อถือด้วยวิธีการสลับเปลี่ยน

ผู้วิจัยได้ใช้วิธีการสลับเปลี่ยนกับการสุ่มตัวอย่างด้วยวิธีบูทสแตร็ป และพิจารณาระบบเดียวกันกับการทดลองแบบการวัดพลังการแยกแยะ



ภาพที่ 5.3 กราฟ MR กับ PT ของระบบจาก TREC 2011 (ซ้าย) และ 2012 (ขวา) ที่ถูกประเมินโดย 2 ตัวชี้วัดประเภทความหลากหลายและ 2 ตัวชี้วัดประเภทความซ้ำซ้อน

ภาพที่ 5.3 และ ภาพที่ 5.4 แสดงถึงกราฟระหว่างค่า minority rate (MR) และค่า proportion of tie (PT) ของ TREC 2011 และ 2012 ซึ่งค่า MR แสดงถึงโอกาสของการได้รับข้อสรุปที่ขัดแย้ง (Contradictory conclusion) ในคู่ระบบ และค่า PT แสดงถึงการขาดพลังการแยกแยะ ดังนั้นแล้วค่า MR และ PT น้อยแสดงถึงความน่าเชื่อถือของตัวชี้วัดมีมาก หรือกล่าวคือ ยิ่งกราฟเข้าใจจุดเริ่มต้น ยิ่งมีความน่าเชื่อถือสูง



ภาพที่ 5.4 กราฟ MR กับ PT ของระบบจาก TREC 2011 (ซ้าย) และ 2012 (ขวา) ที่ถูกประเมินโดย 4 ตัวชี้วัดแบบผสม

ภาพที่ 5.3 แสดงกราฟของ 2 ตัวชี้วัดประเภทความหลากหลายและ 2 ตัวชี้วัดประเภทความซ้ำซ้อน โดยรูป ภาพที่ 5.3 ระบุว่า nCF มีความน่าเชื่อถือกว่าตัวชี้วัดอื่นในทุกชุดข้อมูล

นอกจากนี้ภาพที่ 5.4 แสดง 4 ตัวชี้วัดแบบผสมของทั้งสองชุดข้อมูล โดย nCF+α-nDCG มีความน่าเชื่อถือมากที่สุดทั้งสองชุดข้อมูล โดยเฉพาะชุดข้อมูลจาก TREC 2012

5.1.4. สรุปผล

จากผลลัพธ์ที่ได้จากการทดลองความน่าเชื่อถือของตัวชี้วัดต่างๆ โดยผลการทดลองแสดงให้เห็นว่าประเมินของ nCF มีความน่าเชื่อถือในชุดข้อมูลที่แตกต่างกัน รวมทั้งเมื่อประเมิน nCF ร่วมกับตัวชี้วัดประเภทความซ้ำซ้อน ทำให้ตัวชี้วัดดังกล่าวมีความน่าเชื่อถือมากขึ้น ในทางกลับกันเมื่อนำตัวชี้วัด *I*-rec ไปรวม ความน่าเชื่อถือไม่เพิ่มขึ้นมากนักเมื่อเปรียบเทียบแบบยังไม่ได้รวม

บทที่ 6
สรุปผลการวิจัยและข้อเสนอแนะ

6.1 สรุปผลการวิจัย

วิทยานิพนธ์นี้มุ่งเน้นการพัฒนาตัวชี้วัดประเภทความหลากหลายใหม่ที่มีชื่อว่า Normalized Coverage Frequency (nCF) โดยมุ่งเน้นให้ตัวชี้วัดสามารถวัดความหลากหลายในผลการค้นหาได้มีประสิทธิผลมากขึ้นโดยตัวชี้วัดถูกออกแบบเพื่อกำจัดข้อเสียของตัวชี้วัด *I*-rec โดย nCF สามารถวัดได้ตรงตามจุดประสงค์ของความหลากหลาย โดยมีการพิจารณาความหลากหลายจากความถี่ของความครอบคลุมอินเท็น และสามารถประเมินผลการค้นหาได้ทั่วทั้งผลการค้นหา และสามารถนำไปใช้รวมกับตัวชี้วัดประเภทความซ้ำซ้อนด้วยวิธีแบบขั้นลำดับ หรือวิธีแบบชดเชยกัน นอกจากนี้นักวิจัยได้แสดงให้เห็นถึงปัญหาของตัวชี้วัดเดิมทั้งประเภทความซ้ำซ้อนและความหลากหลายโดยตัวชี้วัดประเภทความซ้ำซ้อนไม่ควรนำใช้ในประเมินความหลากหลายในผลการค้นหา ตัวชี้วัดประเภทความหลากหลายเดิมไม่สามารประเมินเอกสารได้ทั่วทั้งผลการค้นหา รวมทั้งการรวมกันระหว่างตัวชี้วัดประเภทความหลากหลายและตัวชี้วัดประเภทความซ้ำซ้อนด้วยวิธีการรวมแบบชดเชยกันทำให้การตีผลทำได้ยาก

6.1.1. การวิเคราะห์แนวคิดรากฐาน

จากการวิเคราะห์แนวคิดรากฐาน nCF สามารถประเมินความหลากหลายในผลการค้นหาได้ดีที่สุดและรองลงมาคือ D#-nDCG แต่สำหรับ α-nDCG และ ERR-IA ซึ่งเป็นตัววัดประเภทความซ้ำซ้อน ไม่สามารถประเมินความหลากหลายในผลการค้นหาได้อย่างมีประสิทธิผล นอกจากนี้ *I*-rec ซึ่งเป็นตัวชี้วัดประเภทความหลากหลายไม่สามารถแยกแยะความแตกต่างของความหลากหลายในผลการค้นหาได้

6.1.2. ความสอดคล้องระหว่างตัวชี้วัด

ความสอดคล้องระหว่างตัวชี้วัด nCF ได้ถูกนำเสนอในสองมิติกล่าวคือ มิติของศักยภาพของการประเมินระบบ และมิติของค่าความสอดคล้อง โดยตัวชี้วัด nCF มีความสอดคล้องกับตัวชี้วัดอื่นๆโดยเฉพาะอย่างยิ่งตัวชี้วัด D#-nDCG ที่มีค่าความสอดคล้องมากกว่า 0.8 ในทุกปี

6.1.3. ความสอดคล้องระหว่างตัวชี้วัดกับความชอบผู้ใช้

การทดลองวัดความสอดคล้องของตัวชี้วัด nCF, *I*-rec, α-nDCG และ ERR-IA กับความชอบผู้ใช้โดยทดลองอยู่บน Mechanical Turk ซึ่งเป็นระบบการกระจายปัญหาไปยังกลุ่มค้นเพื่อค้นหาคำตอบ (Crowdsourcing) ซึ่งผลการทดลองชี้ให้เห็นว่า nCF สามารถประเมินระบบได้

ตรงกับความต้องการของผู้ใช้ที่ต้องการอินเท็นเดียว และ α-nDCG สามารถประเมินระบบได้ตรงกับความต้องการของผู้ใช้ที่ต้องการอินเท็นทั้งหมด ในขณะที่ I-rec ไม่สามารถประเมินระบบได้สอดคล้องกับความต้องการของผู้ใช้

6.1.4. ความน่าชื่อถือของตัวชี้วัด

สำหรับการประเมินความน่าเชื่อถือของ nCF เมื่อเทียบกับตัวชี้วัดในปัจจุบัน ผู้วิจัยได้ใช้วิธีการวัด 2 วิธีคือ วิธีการใช้การวัดพลังการแยกแยะ และทดลองด้วยวิธีการสลับเปลี่ยน โดยผลการทดลองแสดงให้เห็นว่า nCF มีความน่าเชื่อถือว่าตัวชี้วัดอื่นๆรวมทั้งเมื่อมีการรวมระหว่าง nCF กับตัวชี้วัดประเภทความซ้ำซ้อนแล้วความน่าเชื่อถือของตัวชี้วัดประเภทความซ้ำซ้อนเพิ่มขึ้นในทางกลับกันเมื่อนำไปรวมกับ *I*-rec ความน่าเชื่อถือมีความคงที่หรือลดลง

6.2 ข้อเสนอแนะ

ในงานวิจัยนี้ได้ทำการสร้างตัวชี้วัดประเภทความหลากหลายโดยสามารถแก้ปัญหาของตัวชี้วัดปัจจุบัน และทำการทดลองในทางทฤษฎีและทางปฏิบัติจากความชอบของผู้ใช้ โดยปัญหาที่ได้พบในการทำงานวิจัยและแนวคิดที่ใช้ในการปรับปรุงตัวชี้วัดในอนาคตมีดังต่อไปนี้

- ในการคิดค้นตัวชี้วัดนั้นนอกจากปัจจัยต่างๆที่ได้กล่าวในวิทยานิพนธ์นี้ไปแล้วก็ยังมีปัจจัยอื่นๆที่สามารถนำมาพิจารณาต่อยอดให้สามารถประเมินระบบได้ตรงกับความต้องการของผู้ใช้ในปัจจุบัน อาทิ การเกิดขึ้นของข้อมูลในรูปแบบต่างๆ เช่น รูปภาพ วีดิโอ หรือเสียง ทำให้การประเมินแบบเดิมไม่สามารถประเมินได้ตรงตามความต้องการของผู้ใช้ที่เปลี่ยนไป
- การสร้างตัวชี้วัดให้สามารถประเมินได้ทุกแง่มุมเป็นไปได้ยากเนื่องจากการประเมินรวมกันของปัจจัยต่างๆที่แตกต่างกัน อาทิ ความหลากหลายกับความซ้ำซ้อน ทำให้ผลการประเมินตีความได้ยากหรืออาจจะไม่สามารถตีความได้เลย

เอกสารอ้างอิง

[1] A. Kent, M. M. Berry, and J. W. Perry, "Machine Literature Searching II. Problems in Indexing for Machine Searching," *Am. Doc.*, vol. 5, no. 1, pp. 22–25, 1954.

[2] C. J. van Rijsbergen, *Information Retrieval*, 2nd ed. Butterworth, 1979.

[3] C. L. A. Clarke, M. Kolla, G. V Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and Diversity in Information Retrieval Evaluation," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 659–666.

[4] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected Reciprocal Rank for Graded Relevance," in *Proceeding of the 18th ACM conference on Information and knowledge management*, 2009, pp. 621–630.

[5] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003, pp. 10–17.

[6] A. Tangsomboon and T. Leelanupab, "Evaluating Diversity and Redundancy-Based Search Metrics Independently," in *Proceedings of the 2014 Australasian Document Computing Symposium*, 2014, pp. 42:42–42:49.

[7] R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon, "Identification of Ambiguous Queries in Web Search," *Inf. Process. Manag.*, vol. 45, no. 2, pp. 216–229, 2009.

[8] M. Sanderson, "Ambiguous Queries: Test Collections Need More Sense," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 499–506.

เอกสารอ้างอิง (ต่อ)

[9] F. Radlinski and S. Dumais, "Improving Personalized Web Search Using Result Diversification," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 691–692.

[10] R. L. T. Santos, C. Macdonald, and I. Ounis, "Intent-Aware Search Result Diversification," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, 2011, pp. 595–604.

[11] P. Ingwersen, *Information Retrieval Interaction*. Taylor Graham Publishing, 1992.

[12] R. W. White and R. A. Roth, "Exploratory Search: Beyond the Query-Response Paradigm," *Synth. Lect. Inf. Concepts, Retrieval, Serv.*, vol. 1, pp. 1–98, 2009.

[13] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 335–336.

[14] J. B. MacQueen, "Some Methods of Classification and Analysis of Multivariate Observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.

[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

เอกสารอ้างอิง (ต่อ)

[17] C. W. Cleverdon, J. Mills, and M. Keen, "Factors Determining the Performance of Indexing Systems," *Tech. report, ASLIB Cranf. Proj.*, 1966.

[18] E. M. Voorhees and D. K. Harman, *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.

[19] F. C. Gey, N. Kando, and C. Peters, "Cross-Language Information Retrieval: the way ahead," *Inf. Process. Manag.*, vol. 41, no. 3, pp. 415–431, 2005.

[20] K. Spärck-Jones and C. J. van Rijsbergen, "Report on the Need for and Provision of an ``Ideal'' Information Retrieval Test Collection," British Library Research and Development Reports, 1975.

[21] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[22] V. Raghavan, P. Bollmann, and G. S. Jung, "A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance," *ACM Trans. Inf. Syst.*, vol. 7, no. 3, pp. 205–229, 1989.

[23] G. Zuccon and L. Azzopardi, "Using the Quantum Probability Ranking Principle to Rank Interdependent Documents," in *Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval*, 2010, pp. 357–369.

[24] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying Search Results," in *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, 2009, pp. 5–14.

[25] P. B. Golbus, J. A. Aslam, and C. L. Clarke, "Increasing Evaluation Sensitivity to Diversity," *Inf. Retr.*, vol. 16, no. 4, pp. 530–555, Aug. 2013.

เอกสารอ้างอิง (ต่อ)

[26] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Y. Lin, "Simple Evaluation Metrics for Diversified Search Results," in *Proceedings of the 3rd International Workshop on Evaluating Information Access*, 2010, pp. 42–50.

[27] T. Joachims, "Evaluating Retrieval Performance Using Clickthrough Data." 2002.

[28] F. Radlinski, M. Kurup, and T. Joachims, "How Does Clickthrough Data Reflect Retrieval Quality?," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008, pp. 43–52.

[29] P. Thomas and D. Hawking, "Evaluation by Comparing Result Sets in Context," in *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, 2006, pp. 94–101.

[30] T. Sakai, "Evaluating Evaluation Metrics Based on the Bootstrap," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006, pp. 525–532.

[31] C. L. A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A Comparative Analysis of Cascade Measures for Novelty and Diversity," in *Proceedings of the 4th ACM international conference on Web search and data mining*, 2011, pp. 75–84.

[32] C. Buckley and E. M. Voorhees, "Evaluating Evaluation Measure Stability," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 33–40.

[33] T. Leelanupab, G. Zuccon, and J. M. Jose, "A Comprehensive Analysis of Parameter Settings for Novelty-biased Cumulative Gain," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012, pp. 1950–1954.

เอกสารอ้างอิง (ต่อ)

[34] T. Leelanupab, G. Zuccon, and J. M. Jose, "Is Intent-Aware Expected Reciprocal Rank Sufficient to Evaluate Diversity?," in *ECIR{\textquoteright}13*, 2013, pp. 738–742.

[35] C. L. A. Clarke, N. Craswell, I. Soboroff, and G. V Cormack, "Overview of the TREC 2010 Web Track," in *the 19th Text Retrieval Conference, 2010.*, 2010.

[36] C. Buckley and E. M. Voorhees, "Retrieval Evaluation with Incomplete Information," in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 25–32.

[37] A. Broder, "A Taxonomy of Web Search," *ACM SIGIR Forum*, vol. 36, pp. 3–10, 2002.

[38] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for Relevance Evaluation," *SIGIR Forum*, vol. 42, no. 2, pp. 9–15, Nov. 2008.

ภาคผนวก ก.
ผลงานวิจัยที่ได้รับการตีพิมพ์

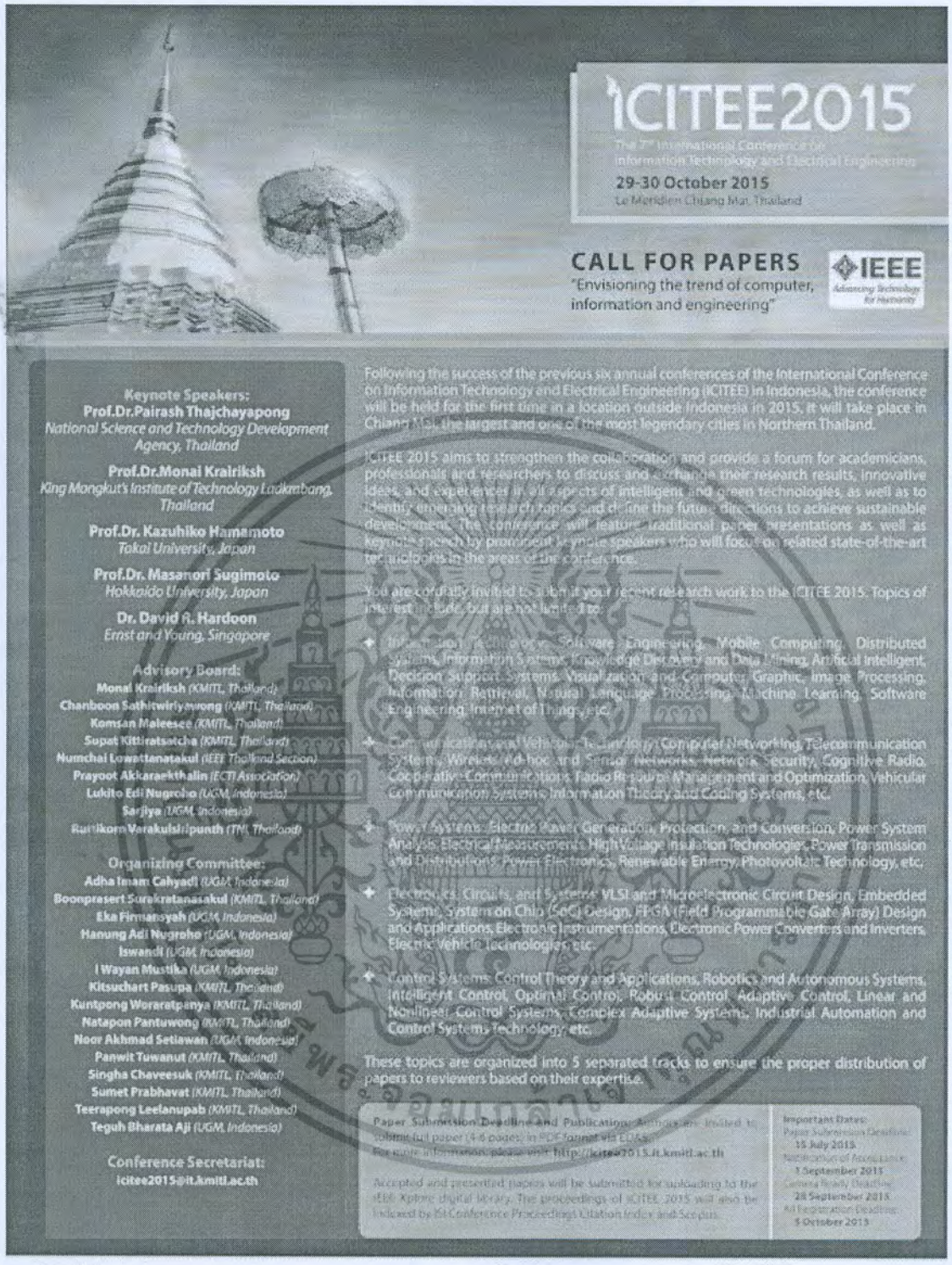
ภาคผนวกนี้ได้นำเสนอบทความฉบับสมบูรณ์ของผลงานวิจัยซึ่งได้รับการตอบรับและตีพิมพ์ลงในการประชุมวิชาการระดับนานาชาติ ซึ่งเป็นส่วนหนึ่งของการทำวิทยานิพนธ์เล่มนี้ โดยมีรายการดังต่อไปนี้

- *On the Reliability of Diversity and Redundancy-Based Search Metrics*

  A. Tangsomboon and T. Leelanupab; in Proceedings of the 7th International Conference on Information Technology and Electronical Engineering, ICITEE 2015, Chiang Mai, Thailand, to appear

- *Evaluating Diversity and Redundancy-Based Search Metrics Independently*

  A. Tangsomboon and T. Leelanupab; in Proceedings of the 19th Australasian Document Computing Symposium, ADCS 2014, Melbourne, Australia, (**The Best Student Paper Award**)



On the Reliability of Diversity and Redundancy-Based Search Metrics

Ake Tangsomboon, Teerapong Leelanupab
Faculty of Information Technology, King Mongkuts Institute of Technology Ladkrabang (KMITL)
Bangkok, Thailand, 10520
Email: a.tangsomboon@gmail.com, teerapong@it.kmitl.ac.th

***Abstract*—Traditional approaches to ranking documents in Information Retrieval (IR) are under the assumption that the representation of information needs is clear and well-defined. This representation, which is usually in the form of a search query, is arguably considered ambiguous or underspecified. To deal with this uncertainty, much recent research has focused on creating IR systems that diversify search results so as to satisfy the multiple possible information needs underlying the query. To validate these IR systems, many new evaluation metrics have been proposed to quantify their effectiveness in terms of diversity and redundancy. Among these, a new diversity-based metric, called normalized Coverage Frequency (nCF), has lately been proposed to quantify diversity in a ranking. When a new metric is proposed, its reliability needs to be validated. This paper conducts an empirical experiment to compares and contrast state-of-the-art diversity and redundancy-based metrics, in term of discriminative power and stability of system rankings. Our experiment shows that the nCF is rated the best among all the studied metrics. Moreover, this finding is confirmed by when nCF is interpolated with other redundancy-based metrics (i.e., ERR-SA and A-nDCG). the nCF is considered more relatively robust than another diversity metric, subtopic-recall.**

I. Introduction

The task of diversity document retrieval stems from the need to handle ambiguous and underspecified queries. These queries are due to the fact that users' information needs are inherently complex and often represented in the form of short text queries. For example, a user who issues the ambiguous query [1], [2] "endorphin" may be interested in either the natural pain medication hormone in animals, motion synthesis software or the Australian or Thai band. Although a query is not ambiguous, it may be underspecified and may not exactly express the users' information need. For instance, another user enters the query "Da Endorphin" (Thai female singer). This user may be searching for either her biographies, her album/single, or news about her activities [3].

In addition, the documents retrieved by a search system may convey more or less redundant information [4]. Typical approaches in Information Retrieval (IR) is on the basis of independent-document ranking such as probability ranking principle, by which documents are retrieved without considering other documents having been earlier ranked [5]–[7]. Unfortunately, these approaches may lead to the cost for users to examine redundant information conveyed by documents in a ranking. To achieve the problems of ambiguity and redundancy in search, numerous alternative methods are introduced for search result diversification, so-called diversity retrieval [8]–[10]. Among these, most of them are inspired by the Maximal Marginal Relevance approach of Carbonell and Goldstein [11]. In diversity retrieval, the ideal document ranking should be both diverse and novel. That is, it should cover as many subtopics[1] as possible and avoids redundant information of documents considered as a whole in a ranking.

To evaluate search system in this task, many evaluation metrics have been proposed in the literature. We previously divided these metrics into two groups according to their measurements, *i*) redundancy and *ii*) diversity. The examples of *redundancy*-based metrics are, Novelty-Biased (Alpha) Normalized Discounted Cumulative Gain (A-nDCG) [3] and Subtopic-Aware Expected Reciprocal Rank[2] (ERR-SA) [12]. In this evaluation context, A-nDCG and ERR-SA are the two most widely used metrics in annual TREC[3] workshop and IR experiments. Although these two metrics were theoretically devised based upon the notion of redundancy, they are in practice used to measure diversity in a document ranking. This issue was discovered and arguably discussed by in our previous works [6], [13]–[15], where we empirically presented a series of comprehensive studies to find out the problems. We also highlighted the drawbacks of the only one previously available native *diversity*-based metric, Subtopic Recall (*S*-recall), and proposed a new metric, called Normalized Coverage Frequency (nCF), to surpass *S*-recall's limitations [15]. Even though our proposed nCF has been well theoretically grounded, further studies are required to examine its concordance with human assessment and its reliability in measuring system performance.

The objective of this paper is to continue our study on the proposed nCF by focusing on the reliability aspect. To this end, we use the discriminative power method proposed by Sakai [16] and the swap method proposed by Voorhees and Buckley [17], with two data sets comprising test collections and submitted runs from TREC 2011-12 Web diversity tracks. We test two metrics from each diversity and redundancy category, i.e., *S*-recall, nCF, A-nDCG and ERR-SA. In addition, we test four hybrid metrics, combining between redundancy and diversity metrics. Our experiment demonstrates that the nCF performs the best among all the single-aspect metrics (i.e., *S*-recall, A-nDCG and ERR-SA) in terms of discriminative power

[1]The term "subtopic" is used to refer to a piece of information in documents that addresses one possible interpretation of an issued ambiguous or underspecified query. It is also known as "intent", "nugget", or "aspect". The difference in this terminology is due to the different viewpoints of how methods or metrics are modeled.

[2]ERR-SA is the most popular metric in the family of intent-aware metrics. Here, we however focus only on ERR-SA as a representative metric and adhere with the term "subtopic" for consistency.

[3]http://trec.nist.gov

and stability. Furthermore, when nCF is interpolated with other redundancy metrics (i.e., ERR-SA and A-nDCG), it is more robust than that combined with another diversity metric, *S*-recall.

The contributions of this paper are as follows: we are the first to investigate the reliability of the new diversity metric, nCF and the effect of integrating diversity metrics (esp. nCF) and novelty metrics in terms of discriminative power and the stability. We carried out an extensive study, comparing nCF with the three state-of-the-art metrics using two standard test collections in IR.

The rest of this paper is structured as follows. The next section surveys related work in state-of-the-art evaluation metrics of diversity retrieval as well as the methods for measuring their reliability. Section III describes the formalization of our nCF in detail. In Section IV, we outline our experiments, including the experimental plan, research questions and results obtained in this study. Finally, we conclude our findings and briefly present our future work in Section V.

II. Related Work

A. Evaluation Metrics in Diversity Retrieval

The basis for relevance judgment in diversity-retrieval is on mutually exclusive subtopics, by which the relevance of a document is assessed according to each piece of information underlying information needs. In other words, the document's relevance is *not* judged with respect to a topic/query, but is judged with respect to each subtopic or a single piece of information containing in a document. Therefore, given a document ranking, the effectiveness of diversified IR systems is mainly dependent on the amount of relevance to every possible subtopic. These subtopics are then employed to evaluate IR systems for their performance in retrieving documents with the diversity and novelty of such subtopics. Whereas diversity is measured by the coverage of unique subtopics, novelty is mostly evaluated by deducting the ranking's utility from the redundant subtopics conveyed by later retrieved documents. A brief explanation of common metrics, categorized into two groups, is given in the following:

1) Diversity Metrics:

*a) **Subtopic recall (S-recall):*** is a set-based metric that refers to the percentage of relevant subtopics covered by documents up to a given rank [18]. At a ranking position r, it is formally defined as the number of returned subtopics divided by the total number of possible subtopics given a query. *S*-recall evaluates the diversity of relevant information in terms of subtopic coverage. However, the major limitations of *S*-recall is that it can only evaluate the ranking effectiveness until the subtopic coverage is achieved. Afterwards, it cannot quantify the degree of diversity.

*b) **Normalized Coverage Frequency (nCF):*** As an extension to *S*-recall, we recently introduced a new set-based metric, called nCF, which can assess diversity throughout a ranking [15]. nCF is a rank-based metric which quantifies the diversity by accumulating the fraction of subtopic coverage up to a given rank r. In other words, it refers to the number of times that subtopic coverage is successfully retrieved by diversified search systems. We will describe our nCF in detail in Section III since nCF is our main focus of this work.

2) Redundancy Metrics:

*a) **A-nDCG:*** A popular redundancy-based metric, Novelty-Biased Normalized Discounted Cumulative Gain (A-nDCG)[4], was introduced by Clark et al. [3]. A-nDCG extends the traditional DCG metric with the main discount function that reduces the original gain of retrieved relevant subtopics if they are redundant to the subtopics ranked before. In other words, each subsequent retrieval of the same subtopic result in a discounting return so as to indicate the diminished value provided by redundant information.

*b) **Subtopic-Aware Expected Reciprocal Rank (ERR-SA):*** is one of the Subtopic-Aware family metrics, based upon a cascade model of expected reciprocal rank (ERR) [12]. According to ERR, the probability that the user is not satisfied with the documents, having been viewed previously, discounts the utility or expected probability of a currently viewing document. Again, the relevance of each document is based on the relevance of documents retrieved in early ranks. To get evaluation scores, ERR-SA individually calculates ERR for each subtopic and averages them with weights that can be defined by the subtopic probability or its importance.

B. Reliability of Evaluation Metrics

1) Discriminative Power: is one of the primary tools in quantifying the reliability of evaluation metrics through sensitivity. It appeared in several IR research works such as [19]–[21]. Sakai's discriminative power measures the sensitivity by determining the percentage of system pairs, performances of which are significantly different [16]. It assesses how sensitive metrics are when evaluating search systems on different test collections. To obtain various collections for testing metrics' ability of system discrimination, this method relies on the technique of bootstrap sampling. It then performs a two-tailed statistical significance test using on different pairs of experimental runs.

In our context, the evaluation of diversity and redundancy in a document ranking necessarily depend on collection over which the search is performed. That is, by altering a test collection, we assume that our nCF does not sacrifice the sensitivity of evaluation to changes in system rankings. For example, If all documents in a collection covers only a single subtopic, even the best search system will not be able to create a diverse ranked list. In other words, if any system cannot achieve a complete coverage of all subtopics in a ranking, how diversity metrics can evaluate the rankings quantitatively. Besides, after all subtopics are successfully covered, it depends on the systems whether they try to further diversify the search result or not.

In this experiment, we aim to isolate diversity from these confounding factors. As such, we first selected only topics/queries that have three or more subtopics and then filtered out topics in which none of the TREC systems can achieve a complete coverage at a ranking position 10 (*S*-recall@10 = 1). In other words, for a selected search topic, there is at least one system that can attain the goal of diversity task. As a result, there are 40 and 43 topics left after a selection criteria for

[4]"A" stands for Alpha, denoting the modified version of "Novelty-Biased" of Discounted Cumulative Gain. It is a tunable parameter, reflecting the probability of user's intolerance to redundant relevant subtopic.

TREC 2011 and 2012, respectively. At last, we chose only the top 20 systems according to the number of topics in which they can perform S-recall@10 = 1. By doing so, our experiment can be focused on the high performing search systems as well as the cases where the different outcomes of measurement from distinct metrics occur.

2) Stability on Swap Method: In order to examine the reliability of the system rank correlation between diversity and redundancy metrics, we employ the stability metrics based on the swap method presented by Buckley et al. [17]. Unlike Sakai's discriminative power, the swap method is not directly related with significance tests. It instead hinges on a heuristic approach that counts the difference in performance between two systems' pairs. More precisely, the swap method estimates what are the chances of obtaining a contradictory result from different topic sets such as bootstrap samples used in the method of discriminative power. If these chances are lower than a specific threshold θ (e.g., $\theta \leq 5\%$), one system is considered better than the other. For a given IR metric and the threshold, this stability method samples n query sets Q' from the original query set Q, and yields a trade-off curve between the Minority Rate (MR) and the Proportion of Ties (PT). The former indicates the lack of stability with respect to changes in the query set, and the latter indicates the lack of discriminative power. We refer interested readers for more detail in the exact stability algorithm we used in [19].

III. Normalized Coverage Frequency

Normalized Coverage Frequency (nCF) is proposed to overcome S-recall's drawbacks by being able to assess diversity all over a document ranking [15]. nCF is modeled based upon a similar concept of S-recall by extending it from the set-based metric to the rank-based one. It explicitly quantifies how many times the percentage of different subtopics are repeatedly covered in a ranking. Formally, let $S\text{-}cov(j,r,q)$ denote the *subtopic coverage* redefined from S-recall. $R_q^{(j,r)}$ is a set of documents retrieved from a ranking position j to r given a query q. The subtopic coverage can be defined by quantifying the number of unique subtopics covered by documents $d_i \in R_q^{(j,r)}$, according to:

$$S\text{-}cov(j,r,q) = \frac{\cup_{d_i \in R_q^{(j,r)}} |S_q \cap S_{d_i}|}{|S_q|} \quad (1)$$

where S_q is the set of relevant subtopics that a query q possibly refers to, and S_{d_i} is the set of relevant subtopics containing in a document d_i retrieved at rank i. In order to enumerate the number of subtopics that can be achieved at a given rank r, the subtopic coverage time $S\text{-}ct(i,r,q)$ is defined as:

$$S\text{-}ct(i,r,q) = \begin{cases} 1, \ S' = \varnothing, \ k = i+1 & \text{if } |S' \cup S_{d_i}| = |S_q| \\ S\text{-}cov(k,r,q) & \text{if } i = r \\ 0, \ S' = S' \cup S_{d_i} & otherwise \end{cases}$$

Given a query q, $S\text{-}ct(i,r,q)$ is defined as the *subtopic coverage time*, starting at rank i until the rank k at which a document ranking is evaluated. In order to count how many times the percentage of different subtopics are repeatedly covered in a ranking, the function $S\text{-}ct(i,r,q)$ is conditioned on the number of total subtopics $|S_q|$, the set of subtopics covered within a interval cycle of subtopic coverage S', and a range of examined ranking positions which will ends the computation when $i = r$. Afterwards, the *subtopic coverage frequency* can be derived:

$$S\text{-}CF(r,q) = \frac{\sum_{i=1}^{r} S\text{-}ct(i,r,q)}{r} \quad (2)$$

where $S\text{-}CF(r,q)$ is the fraction of the coverage time to the duration for which a user reads documents in a ranked list from the top and stop at some rank r. In order to compare a system across a set of queries (due to each topic that has the different number of relevance judgments), the obtained $S\text{-}CF(r,q)$ is normalized by the coverage frequency of a ideal ranking, $S\text{-}CF(r,q)'$. Hence, $nCF(r,q)$ is defined as follows:

$$nCF(r,q) = \frac{S\text{-}CF(r,q)}{S\text{-}CF(r,q)'} \quad (3)$$

where $CF(r,q)'$ is the ideal cover frequency that could be obtained at rank r. In addition, to evaluate other factors, e.g., subtopic redundancy, interval per-subtopic graded relevance and subtopic probability, we can combine our nCF with redundancy-based metrics $Red(r,q)$, such as either ERR-SA or A-nDCG. The hybrid function linearly combines nCF and $Red(r,q)$ as follows:

$$nCF\text{+}Red(r,q) = \lambda nCF(r,q) + (1-\lambda) Red(r,q) \quad (4)$$

where λ is a parameter that trade off between nCF and Red(r,q). It is set 0.5 as a default whereas Red(r,q) can be replaced by either ERR-SA or A-nDCG.

IV. Experiments

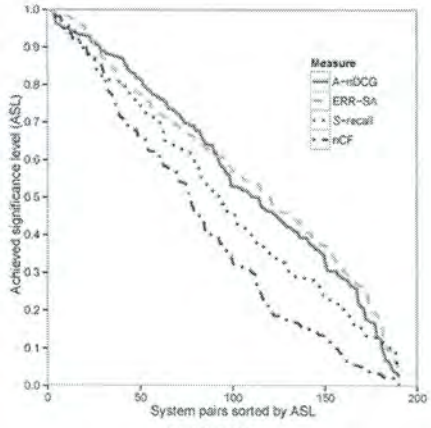
A. Experimental Plan

Our primary goal of this experiment is to evaluate the reliability of the nCF, compared with other state-of-the-art metrics for diversity and redundancy. In particular, we focus on the high performance search systems in result diversification and we investigate the following research questions:

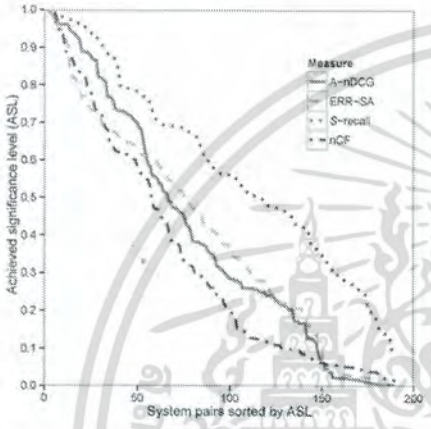
RQ1: Among diversity-based metrics (e.g., S-recall), how reliable is nCF for measuring diversity of document rankings?

RQ2: When nCF is combined with other redundancy-based metrics such as ERR-SA and A-nDCG, does nCF lead to improvements of reliability with respect to those combined with S-recall?

To answer the research questions, we analyzed real document rankings obtained from the TREC 2011-12 Web Diversity tracks [22]. We acquired the experimental runs submitted to TREC each year by Web track participants: a total of 62 systems by 16 groups in 2011 and 20 systems by 8 groups in 2012. We reduced graded relevance to binary relevance and set each intent with an equal probability. These settings was applied for all metrics used in our experiments. For the selection of topics and systems used in this experiment, we have previously described in Section IV-B1.

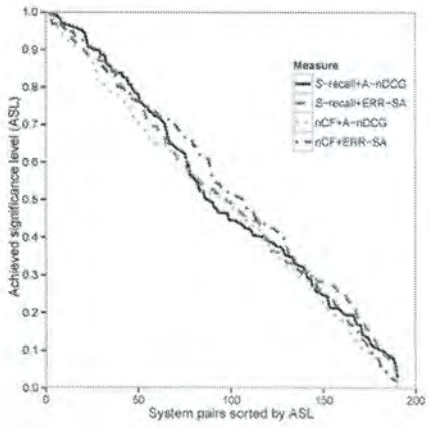


(a) TREC 2011

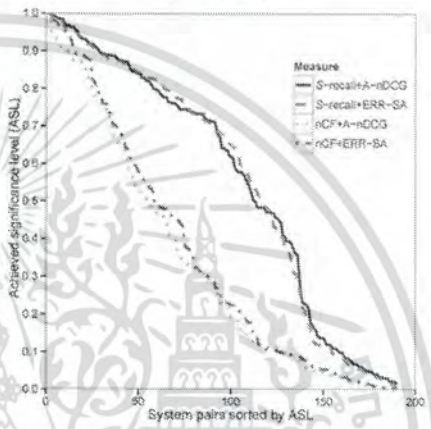


(b) TREC 2012

Fig. 1. ASL score of TREC 2011-12 Web diversity runs evaluated by two redundancy and two diversity metrics at 10.



(a) TREC 2011



(b) TREC 2012

Fig. 2. ASL score of TREC 2011-12 Web diversity runs evaluated by four hybrid metrics at 10.

In this experiment, we used two diversity and two redundancy metrics. These are *S*-recall, nCF, A-nDCG and ERR-SA. Also, we investigated four hybrid metrics, linearly combining redundancy and diversity metrics. By doing this, it produced two hybrid metrics combined with nCF (i.e., nCF+ERR-SA and nCF+A-nDCG) and two hybrid metrics combined with *S*-recall (i.e., *S*-recall+ERR-SA and *S*-recall+A-nDCG).

B. Results and Discussion

1) Discriminative Power: We used the method proposed by Sakai [16] based on the bootstrap hypothesis. All the participating systems submitted to the TREC 2011-12 Web Diversity Tracks were taken into account, but we then selected only the top 20 systems with respect to the number of topics in which they can perform *S*-recall@10 = 1. Hence, we obtained 20*(20-1)/2 = 190 pairs of systems for both TREC 2011 and 2012. For a significance test, we employed a two-tailed paired bootstrap test with 1000 samples and a fixed significant level of 0.05. The bootstrap samples were obtained by sampling queries with replacement.

Fig. 1 and Fig. 2 illustrate the Achieved Significance Level (ASL) curves for diversity and redundancy metrics for TREC 2011 (a) and 2012 (b). The horizontal axis represents the 190 run pairs sorted in a decreasing order of ASL. The vertical axis denotes the ASL. When considering ASL plots, metrics curves of which are closer to the origin are deemed having more reliable than the others.

Among diversity and redundancy metrics (See Fig. 1), the analysis indicates that ERR-SA and A-nDCG are the least reliability in TREC 2011 and *S*-recall is the least reliability in TREC 2012. However, it can be stated that nCF is more reliable than the others. These findings are valid for both datasets. The results in Fig. 2 demonstrate that although the ASL of all hybrid metrics are very similar in TREC 2011, the ASL of only the hybrid metrics using nCF are much more reliable than that of those using *S*-recall in TREC 2012.

Table 1 reports how many pairs of TREC systems fulfill the condition ASL < 0.05. The second column reports the discriminative power. The third one reports the estimated

TABLE I. DISCRIMINATIVE POWER OF METRICS AT SIGNIFICANT LEVEL=0.05 FROM TREC 2011-12.

TREC 2011	ASL < 0.05	estimated diff.
S-recall	0/190=0.00%	0.11
nCF	19/190=10.00%	0.09
A-nDCG	5/190=2.63%	0.08
ERR-SA	1/190=0.52%	0.08
nCF+A-nDCG	8/190=4.21%	0.08
S-recall+A-nDCG	1/190=0.52%	0.09
nCF+ERR-SA	8/190=4.20%	0.08
S-recall+ERR-SA	1/190=0.52%	0.08

TREC 2012	ASL < 0.05	estimated diff.
S-recall	0/190=0.00%	0.08
nCF	23/190=10.52%	0.08
A-nDCG	38/190=20.00%	0.09
ERR-SA	18/190=9.47%	0.09
nCF+A-nDCG	41/190=21.57%	0.09
S-recall-A-nDCG	17/190=8.94%	0.09
nCF+ERR-SA	36/190=18.94%	0.08
S-recall+ERR-SA	18/190=9.47%	0.09

differences required for satisfying the condition ASL < 0.05. If the estimated difference between two systems is 0.09 or larger, then the performances of the two systems are regarded as significantly different [16]. In TREC 2011 dataset, nCF obtained the highest discriminative power ratio (i.e., 10%) at significant level at 0.05. Furthermore, when nCF is interpolated with the redundancy metrics, nCF increases their reliability (i.e., 4.20% for nCF+ERR-SA and 4.21% for nCF+A-nDCG). In TREC 2012 dataset, A-nDCG has the most discriminative power ratio (i.e., 20%). However, when nCF is combined with either A-nDCG or ERR-SA, nCF improved the reliability of its hybrid metric. On the contrary, combining A-nDCG and ERR-SA with *S*-recall leads to their reliability, decreased or unchanged

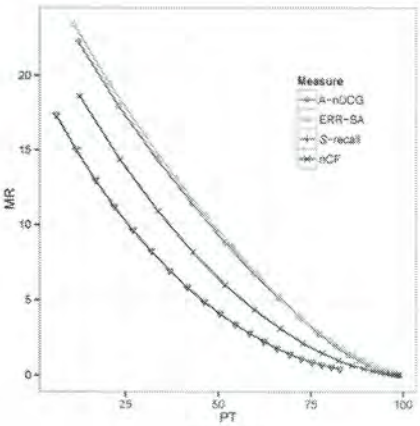
2) Stability Using Swap Method: In this study, we used the swap method with bootstrap samples, and considered the same runs used in previous section to generate systems pairs.

Fig. 3 and Fig. 4 present the plot of the minority rate (MR) against the proportion of ties (PT) for TREC 2011 and 2012, where the fuzziness values were set according to [23]. MR represents the chance of obtaining a contradictory conclusion given a system pair, whereas PT represents the absence of discriminative power. Thus, a reliable metrics is characterized by small values of MR and PT. In other words, the closer a curve is to the origin and the better the associated metric is.

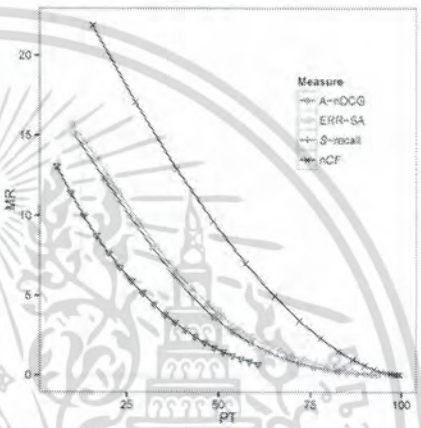
In Fig. 3, MR-PT curves of the two diversity and two redundancy metrics are individually analyzed. The analysis indicates that nCF is more reliable than the other three comparative metrics for both TREC 2011 (a) and 2012 (b) datasets. Similarly, Fig. 4 illustrates the MR-PT curves of four hybrid metrics between diversity and redundancy. As shown in Fig. 4, nCF+A-nDCG is the most reliable among hybrid metrics for both datasets, especially more noticeable in TREC 2012.

V. CONCLUSION AND FUTURE WORK

This paper compared each group of two metrics designed for *diversity* and *redundancy* evaluation, in particular our



(a) TREC 2011

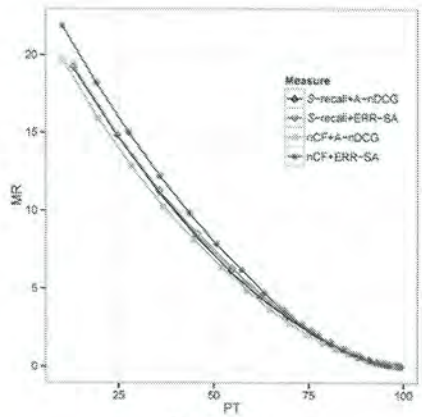


(b) TREC 2012

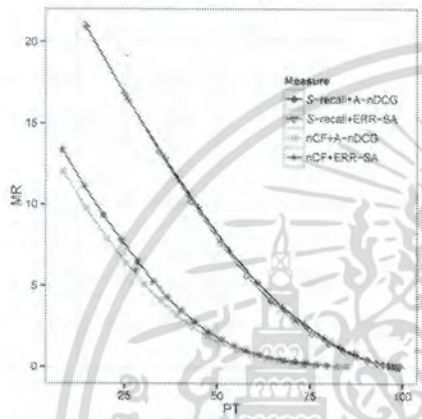
Fig. 3. MR-PT curves of A-nDCG of TREC 2011-12 Web diversity runs evaluated by two redundancy and two diversity metrics at 10.

recently proposed diversity metric, Normalized Coverage Frequency (nCF), in term of reliability and stability of system rankings. In addition to the four investigating metrics, we also test four hybrid metrics, combining nCF or *S*-recall with A-nDCG or ERR-SA. The experiment was carried out using two different standard data sets from TREC 2011-12 Web Diversity track. Our results suggested that our nCF is the best for reliability and stability among redundancy and diversity metrics when considering the high performing systems and the search topics, in which some TREC runs could potentially achieve *S*-recall@10 = 1. Moreover, when redundancy metrics (i.e., ERR-SA and A-nDCG) are interpolated with nCF, their reliability is more relatively improved than that with *S*-recall. In the future, we plan to conduct a user experiment to validate our nCF with user preferences in the diversity context.

Acknowledgments. This research is fully supported by the National Science and Technology Development Agency, Thailand, with a grant awarded to T. Leelanupab (NSTDA funded project, SCH-NR2012-223).



(a) TREC 2011



(b) TREC 2012

Fig. 4. MR-PT curves of TREC 2011-12 Web diversity runs evaluated by four hybrid metrics at 10.

References

[1] S. Cronen-Townsend and W. B. Croft, "Quantifying query ambiguity," in *Proceedings of the Second International Conference on Human Language Technology Research*, ser. HLT '02. San Diego, California: Morgan Kaufmann Publishers Inc., 2002, pp. 104–109.

[2] R. Song, Z. Luo, J.-Y. Nie, Y. Yu, and H.-W. Hon, "Identification of Ambiguous Queries in Web Search," *Information Processing Management*, vol. 45, pp. 216–229, 2009.

[3] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and Diversity in Information Retrieval Evaluation," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '08. Singapore, 2008, pp. 659–666.

[4] Y. Bernstein and J. Zobel, "Redundant Documents and Search Effectiveness," in *Proceedings of the 14th ACM international conference on Information and knowledge management*, ser. CIKM '05. Bremen, Germany, 2005, pp. 736–743.

[5] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims, "Redundancy, Diversity and Interdependent Document Relevance," *ACM SIGIR Forum*, vol. 43, pp. 46–52, 2009.

[6] T. Leelanupab, "A Ranking Framework and Evaluation for Diversity-Based Retrieval," Ph.D. dissertation, University of Glasgow, 2012.

[7] G. Zuccon, "Document Ranking With Quantum Probabilities," Ph.D. dissertation, University of Glasgow, 2012.

[8] H. Chen and D. R. Karger, "Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06. Seattle, Washington, USA, 2006, pp. 429–436.

[9] J. Wang and J. Zhu, "Portfolio Theory of Information Retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '09. Boston, USA, 2009, pp. 115–122.

[10] R. L. T. Santos, C. Macdonald, and I. Ounis, "Search result diversification," *Foundations and Trends® in Information Retrieval*, vol. 9, no. 1, pp. 1–90, 2015.

[11] J. Carbonell and J. Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '98. Melbourne, Australia, 1998, pp. 335–336.

[12] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected Reciprocal Rank for Graded Relevance," in *Proceeding of the 18th ACM conference on Information and knowledge management*, ser. CIKM '09. Hong Kong, China, 2009, pp. 621–630.

[13] T. Leelanupab, G. Zuccon, and J. M. Jose, "A Query-Basis Approach to Parametrizing Novelty-Biased Cumulative Gain," in *Proceedings of the 3nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ser. ICTIR '09. Bertinoro, Italy, 2011.

[14] T. Leelanupab, G. Zuccon, and J. M. Jose, "Is Intent-Aware Expected Reciprocal Rank Sufficient to Evaluate Diversity?" in *Proceedings of the 35th European Conference on Advances in Information Retrieval*, ser. ECIR '13. Moscow, Russia, 2013, pp. 738–742.

[15] A. Tangsomboon and T. Leelanupab, "Evaluating Diversity and Redundancy-Based Search Metrics Independently," in *Proceedings of the 2014 Australasian Document Computing Symposium*, ser. ADCS '14. Melbourne, VIC, Australia, 2014, pp. 42:42–42:49.

[16] T. Sakai, "Evaluating Evaluation Metrics Based on the Bootstrap," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06. Seattle, Washington, USA, 2006, pp. 525–532.

[17] C. Buckley and E. M. Voorhees, "Evaluating Evaluation Measure Stability," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '00. Athens, Greece, 2000, pp. 33–40.

[18] C. X. Zhai, W. W. Cohen, and J. Lafferty, "Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '03. Toronto, Canada, 2003, pp. 10–17.

[19] T. Leelanupab, G. Zuccon, and J. M. Jose, "A Comprehensive Analysis of Parameter Settings for Novelty-biased Cumulative Gain," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, ser. CIKM '12. Maui, Hawaii, USA, 2012, pp. 1950–1954.

[20] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Y. Lin, "Simple Evaluation Metrics for Diversified Search Results," in *Proceedings of the 3rd International Workshop on Evaluating Information Access*, ser. EVIA '10, 2010, pp. 42–50.

[21] C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A Comparative Analysis of Cascade Measures for Novelty and Diversity," in *Proceedings of the 4th ACM international conference on Web search and data mining*, ser. WSDM '11, 2011, pp. 75–84.

[22] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack, "Overview of the TREC 2010 Web Track," in *the 19th Text Retrieval Conference, 2010.*, 2010.

[23] E. M. Voorhees and C. Buckley, "The Effect of Topic Set Size on Retrieval Experiment Error," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '02, Tampere, Finland, 2002, pp. 316–323.

ADCS

Proceedings of the 19th Australasian Document Computing Symposium

Melbourne, Australia November 27-28, 2014

ISBN: 978-1-4503-3000-8

Edited by J. Shane Culpepper, Laurence Park, and Guido Zuccon

bing

RMIT UNIVERSITY

SIGIR
Special Interest Group on Information Retrieval

ICPS

Evaluating Diversity and Redundancy-Based Search Metrics Independently

Ake Tangsomboon, Teerapong Leelanupab
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang (KMITL)
Bangkok, Thailand. 10520
a.tangsomboon@gmail.com, teerapong@it.kmitl.ac.th

ABSTRACT

This paper proposes a new evaluation metric, *normalized Coverage Frequency (nCF)*, which aims to explicitly evaluate the *diversity* of search results, going beyond the drawbacks of previously proposed measures. In fact, two of the most widely adopted metrics for the diversity retrieval task, namely α-nDCG and Intent-Aware Expected Reciprocal Rank (ERR-IA), explicitly evaluate redundancy, but not diversity. While there exists a genuine *diversity*-based metric called Intent Recall (*I*-rec), it has some drawbacks. These drawbacks may be inherited by other derived metrics such as D#-nDCG, which combines *I*-rec with a modified version of nDCG.

The proposed nCF metric assesses how often query-intents are successfully covered throughout a ranked list up to a given rank position. A comprehensive study is conducted using both real and synthetic data to compare nCF with α-nDCG, ERR-IA, *I*-recall and D#-nDCG. Results show that the proposed metric correlates well with the existing ones while it is capable of capturing other factors, e.g., a series of coverage. In addition, we categorize the existing metrics into two distinct groups, i.e., diversity and novelty, based upon their intuitive measurements and suggest that they be used independently according to what they quantify for the ease of performance interpretation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Systems and Software**]: Performance Evaluation

Keywords

Effectiveness Metrics, Diversity, Redundancy

1. INTRODUCTION

The notion of relevance is a vital aspect of many theoretical and practical information retrieval (IR) models. Traditional IR models have mostly focused on satisfying clearly specified user information needs. However, users typically enter short queries to identify their information needs and these queries are often ambiguous and/or unspecified. Consider, for instance, an ambiguous query "Panda" that may refer to a species of animal, the animation movie "Kung fu Panda" or the "Panda security" antivirus software. Without further refinements to these information needs that can disambiguate an actual user intent, a search system could alternatively diversify search results to address all of its possible interpretations (or intents) that may satisfy the information needs [8]. In light of this, search result diversification has gained large attention by, for example, modeling document dependencies in ranking functions [1, 13, 22, 23, 25]. In this work, we are particularly interested in evaluating diversified search results, where each result is assumed to be a single ranked list of documents.

For the evaluation of diversified search systems, many metrics have been devised for measuring diversity and novelty. To this end, the information need underlying a query is decomposed into a set of intents (or subtopics[1]). These intents are employed to evaluate search systems by mostly using two distinct paradigms: either *i*) counting the number of relevant intents covered by documents or *ii*) penalizing the number of redundant (though relevant) intents included by subsequently ranked documents[2]. While the latter aim is basically to measure novelty, it also potentially reduces the amount of relevant information retrieved to a user when it is used as an objective function in a ranking approach.

Within this context, TREC Web Diversity tracks 2009-12 and NTCIR-9&10 INTENT tasks [10] have been created to evaluate IR systems that can retrieve relevant documents while maintaining diversity of intents. In novelty and diversity retrieval, one search topic may have more than one intent and thus the relevance of a document is taken into consideration with respect to intents but not to a topic. Various evaluation metrics have been proposed in the literature. Among these, two *redundancy*-based metrics, i.e., α-nDCG, and Intent-Aware Expected Reciprocal Rank (ERR-IA), are widely used in TREC workshops. They are rank-based measures that discount the usefulness of each relevant document, based upon *i*) its rank position and *ii*) the *redundancy* of

[1]In this work, we adhere to the terms "intent". Whereas intents or query-intents are the terms used to address information needs from a user's point of view, subtopics refer to pieces of information in a document that satisfies user information needs.

[2]In order words, such intents are not previously seen in the ranking.

ADCS '14, November 27 - 28 2014, Melbourne, VIC, Australia

ACM 978-1-4503-3000-8/14/11 ...$15.00
http://dx.doi.org/10.1145/2682862.2682881

relevant intents. Nevertheless, these evaluation metrics are practically employed to measure *diversity* as well [10]. Although studied in previous works [9,11,12,19], it is yet unclear whether α-nDCG and ERR-IA should be used for evaluating diversity. It is our conjecture that this is not a sound practice because novelty and diversity are two distinct features that require different interpretations of search results. Systems considered providing highly novel results do not necessarily mean they also provide highly diverse results. While there exists a genuine and simple *diversity*-based metric, Intent Recall (*I*-rec), it has some drawbacks, such as not addressing relevant intents after complete intent coverage is achieved. These drawbacks may also affect two other recently proposed metrics, D#-nDCG and α#-nDCG-IA [14], which both combine *I*-rec with its modified version of nDCG and α-nDCG respectively.

In this paper, we propose a new evaluation metric, *normalized Coverage Frequency (nCF)*, which aims to specifically evaluate *diversity* within the search results. Two challenges arise due to the complex nature of diversity evaluation. First, different factors can sometimes be contradictory with each other (diversity and novelty) and thus combining those two factors, like D#-nDCG does, may make the task of result interpretation even more complex. Second, diversity is a factor that quantifies the amount of *unique* intents covered by a set of considered documents. Thereby, stripping the consideration of per-intent graded relevance and intent-probability in diversity evaluation is possibly reasonable enough to model an editorial metric, leaving the estimate of those two aspects to, for example, IA-measures. Our intentions of this work are *i*) to explicitly quantify the amount of intent diversity all over a ranked list, but *ii*) not to formulate an universal measure that can take into account all aspects in diversity retrieval.

To achieve this, we devise a new editorial metric that measures how frequent query-intents are successfully covered throughout a ranked list. This can be seen as an extension of the classical set-based metric of *I*-rec to the rank-based metric. In this way, the metric can quantify the utility of top r ranked documents without any condition on the achievement of intent coverage as restricted by *I*-rec. We argue and present empirically that our nCF metric intuitively assesses intent diversity and is thus a more appropriate metric for evaluating IR systems; especially, the systems that aim to diversify search results by returning diverse relevant intents, but not those that return a few redundant relevant intents.

The contributions of the paper are the following: we proposed a novel diversity measure and demonstrated its effectiveness in assessing systems by their diversity. We conducted an extensive study, comparing our nCF with the four state-of-the-art metrics, i.e., α-nDCG, ERR-IA, *I*-recall and D#-nDCG. To our knowledge, this is the first endeavor to revisit and extend an existing *diversity* measure to be rank-based. Moreover, we showed that redundancy-based measures such as α-nDCG and ERR-IA should be used for measuring novelty and interval per-intent graded relevance, but not for intent coverage/diversity.

2. DIVERSIFIED SEARCH EVALUATION

Regarding diversification in search results, several evaluation measures have been proposed in recent years. Zhai *et al.* [24] studied the diversity retrieval task in the context of the TREC Interactive track, and defined straightforward evaluation measures for measuring diversity such as Intent Recall (*I*-rec) and Intent Precision. Similar to the notion of *I*-rec, Intent Mean Reciprocal Rank is defined as the inverse of the first rank at which the specific percentage of intent coverage is achieved (e.g., 50%, 75%, 100%) [8]. With the introduction of α-nDCG, Clarke *et al.* [11] proposed an evaluation strategy that represents multiple intents underlying a query as information nuggets[3]. Also introduced by Clarke *et al.*, NRBP takes ideas from α-nDCG and blend them into Rank-Biased Precision [12]. In α-nDCG and NRBP, document utility is determined by penalizing the number of redundant nuggets covered by later ranked documents.

A family of Intent Aware (IA) measures have been proposed by Agarwal *et al.* [1], in which the likelihood of query-intents is considered via a probability distribution. Similar to IA-measures, Sakai *et al.* [19] introduced an alternative family of "D" measures, which also take into account the probability of intents given a query. The difference between IA and D metrics is the utilization of intent probability in their scoring functions. In IA metrics, the probability of intent is computed after the evaluation is applied on multiple rankings considered separately for each intent, whereas, in D metrics, it is computed for a single ranking considering all intents together in every position. However, both the IA and the D metrics have a limitation that does not enforce a high coverage of multiple query-intents by design. Consequently, some metrics of these two families, e.g., nDCG-IA and D-nDCG, ignore minor intents with a low intent probability and end up with unintentionally rewarding a system that covers only a few yet major intents. Similar cases have been found by Leelanupab *et al.* [15,16], where α-nDCG with α=0.5 and ERR-IA do not measure *diversity* as desired in specific circumstances.

With an aim to overcome the limitation of D-nDCG, a compensatory method that linearly combines and balances a diversity metric (i.e., *I*-rec) with D-nDCG was proposed by Sakai *et al.* [19,20]. The resulting metric is called a "D#" metric and officially used in NTCIR-9&10 INTENT tasks. Recently, Golbus *et al.* [14] proposed another compensatory measure, namely, α#-nDCG-IA, which again linearly interpolates *I*-rec with an IA version of α-nDCG. Chandar *et al.* [6] proposed preference-based search metrics for a diversity task. Their metrics are flexible to various user models and based on explicit preference judgments to interested intents. Only each of these measures used in this study will be explained in more detail below.

All of the existing measures are on the basis of mutually exclusive intents: decomposition of a given query into several separate pieces of information (such as subtopics, nuggets, or aspects) underlying information needs. The effectiveness of diversified search systems is merely dependent on the document's relevance to an intent. Evaluation measures in this context must account for diversity and novelty in a ranking. While diversity is evaluated by the coverage of unique intents, novelty is mostly assessed by penalizing redundancy of relevant information caused by subsequently ranked documents. Hence, to clarify the notion of evaluation metrics, we accordingly divide them into two main categories plus one category that merges both together. A brief description of commonly used metrics is given as follows:

[3]The term "nugget" refers to a piece of information relevant to a user intent.

2.1 Diversity-Based Measure

Intent Recall is defined as a ratio of unique intents to the total number of intents at a given rank [24]. Suppose that a query q consists of the total number of $|I|$ intents. Given a cutoff k, this genuine *diversity* evaluation metric quantifies the degree to which the top k documents in a ranking contain the unique intents of the query q, according to:

$$I\text{-}rec@k = \frac{|\bigcup_{r=1}^{k} intent(d_r)|}{|I|} \quad (1)$$

where d_r is a document at rank r and $intent(d_r)$ is the set of intents to which d_r is relevant.

I-rec is a set-based metric and thus accounts only for binary relevance on its purpose. It explicitly evaluates the diversity of relevant information in terms of intent coverage in a document ranking. Therefore, the greater I-rec is, the higher the number of different intents covered in a ranking. However, when used on its own, I-rec is a rather coarse metric because it is affected by four drawbacks.

1. Similar to typical recall, I-rec is a non-position based metric and thus does not account for the positions at which relevant documents are retrieved. In other words, at a cutoff k, I-rec does not distinguish the utility of retrieving relevant intents at rank k-1 or k-2.
2. Once an intent has already been covered, I-rec does not distinguish between subsequent retrievals of (documents covering) the same relevant intent. Although retrieving a redundant relevant intent may be undesirable for users, it is still considered to be more useful than retrieving a non-relevant intent. We later call this issue the *interval*[4] per-intent relevance.
3. The second issue is a generalization of the third drawback: once *all* intents are covered, I-rec does not further distinguish between retrieving relevant or non-relevant documents. In fact I-rec is only able to address the intent diversity up to the rank at which all relevant intents are retrieved. After complete intent coverage is achieved, I-rec always equals 1 and thus it cannot identify how well a retrieval system further diversifies a search result.
4. I-rec does not account for the probability of different intents given a query. Ideally, this probability should reflect the proportion of the population interested in information needs underlying the query.

Note that the proposed normalized Coverage Frequency (nCF) does not account for the second and forth issues by design. We purposely leave them be handled by existing redundancy-based measures, which will subsequently be combined with our nCF by two possible approaches.

2.2 Redundancy-Based Measure

ERR-IA is a measure based on a cascade model of expected reciprocal rank (ERR) [7]. According to ERR, the utility or expected probability of a currently viewed document is "diminished", depending on the probability that the user is *not* satisfied with the documents having been viewed previously. In other words, the relevance of each document is based upon the relevance of documents ranked above it. The discount function of ERR is thus not only dependent on the rank, but also on the relevance of formerly ranked documents. Although ERR is *not* originally modeled by the notion of redundancy, we categorize ERR-IA as a redundancy-based measure since it discounts the utility of document (and thus its contained intents) based upon that of earlier ranked documents containing the same (or redundant) information.

Let $P(R_r)$ denote the relevance probability of a document at rank r (typically defined as $(2^g - 1)/2^{g_{max}}$, where g is a relevance grade) and let $\prod_{j=1}^{r-1}(1-P(R_j))$ be the probability that the user is not satisfied with documents from ranks j=1 to r−1. Then, ERR-IA is defined as a weighted average of ERR computed individually for each intent:

$$ERR\text{-}IA@k = \sum_{i=1}^{|I|} P(i|q) \sum_{r=1}^{k} \frac{1}{r} \prod_{j=1}^{r-1}(1 - P(R_j))P(R_r) \quad (2)$$

where $P(i|q)$ denotes the probability that the user is interested in intent i for query q and $1/r$ is the utility function based on ranked positions.

α-nDCG Novelty-Biased normalized Discounted Cumulative Gain is a native redundancy-based measure, modeled to evaluate novelty by penalizing *redundant* relevant intents (or information nuggets called by Clarke *et al.*) [11]. The key function of α-nDCG is a discount function $(1-\alpha)^{D_{i,r-1}}$ that diminishes the gain by the amount of redundant intents already covered by documents ranked before r. Let $J(d_r, i)$ is a relevance judgment identifying the extent to which document d_r is relevant to intent i, and $D_{i,r-1}$ is the number of times that intent i appears in documents up to rank $r-1$. α (typically set to a value of 0.5) is a parameter to control how much redundancy is penalized over relevance. In theory, the parameter is used to define the probability of user's intolerance to redundant relevant intent. Before normalization by an ideal gain, *α-DCG* can be defined as:

$$\alpha\text{-}DCG@k = \sum_{r=1}^{k} \frac{\sum_{i=1}^{I} J(d_r|i)(1-\alpha)^{D_{i,r-1}}}{log_2(r+1)} \quad (3)$$

where $log_2(r+1)$ is a common discount function based on document-positions in a ranking. In order to obtain α-nDCG, α-DCG must be normalized for comparing performance across various topics. This is done by finding an "ideal" ranking that maximizes α-DCG. Furthermore, this is usually done by using a greedy algorithm, in which the computation of the ideal ranking is an NP-Complete problem [5].

Whereas ERR-IA and α-nDCG are devised based on redundancy, in practice they are exploited to evaluate diversity as well, such as in TREC Web Diversity tracks 2009-12. In Section 3.1, we will explain in detail the issues with ERR-IA and α-nDCG that ignore diversity in evaluation by using synthetic scenarios of retrieval results.

2.3 Other Related Measure

D#-Measure is devised to compromise the limitation of D-measures, which unexpectedly rewards diversified systems for ignoring minor intents with a low intent probability $P(i|q)$. Similar to IA-measures, the D-measures purposely aimed to evaluate *novelty* by accumulating the expected gain of every covered intent multiplied by its intent probability and then summing them up across all given ranked documents. By

[4]The term "interval" here means a set of documents within a cycle of intent coverage.

doing so, instead of discounting the gain derived from *redundant* relevant intents in the same way as α-nDCG does, D-measure rewards the raw gain to a system that returns each relevant intent [14]. A D-measure version of D-DCG is defined as:

$$D\text{-}DCG@k = \sum_{r=1}^{k} \sum_{i}^{|I|} Pr(i|q) g_i(r) / log_2(r+1) \quad (4)$$

where $P(i|q)$ is probability that the user is interested in intent i for query q, $g_i(r)$ is the gain value of a document at rank r with respect to intent i, and $log_2(r+1)$ is a position-based discount function. To obtain D-nDCG, normalization is performed by dividing D-DCG by that of an ideal ranking.

In order to alleviate the limitation of D-measure, D#-measure is proposed by Sakai *et al.* [19] to merge two properties in a compensatory fashion between diversity and novelty. The purpose of the first property is to consider the retrieval of documents that cover many different intents as early as possible[5] and that of the second property is to account for the ranking of documents relevant to more popular intents higher than documents relevant to less popular intents. Thereby, D#-nDCG is a linear combination of I-rec and D-nDCG, computed as follows:

$$D\#\text{-}nDCG@k = \gamma I\text{-}rec@k + (1-\gamma) D\text{-}nDCG@k \quad (5)$$

where a parameter γ is used for controlling the mixture between I-rec and D-nDCG, with $\gamma = 1$ being equivalent to pure I-rec.

Some questions arise when combining two distinct measurements. First, how do we interpret the scores of such a measure, for example, when an obtained result is high? This is because a document ranking contains either high diversity or high novelty. Second, are diversity and novelty criteria compensatory? That is, can higher novelty compensate for lower diversity or vice versa? We leave these issues as open arguments, where this work preliminarily investigates their effects and introduces an alternative criterion of independent evaluation using a *stepwise* process. Nevertheless, we do not limit the possibility for a compensatory integration of our newly proposed diversity measure and existing redundancy ones.

3. PROPOSED METRIC

3.1 Analysis of Existing Measures

Before going into detail about our proposed measure, we need to understand how well existing measures, i.e., α-nDCG, ERR-IA, I-rec and D#-nDCG, can evaluate diversified rankings. To this end, we generated simulated data of document rankings using the "TREC qrel" from the Web Diversity dataset. By doing so, we can simplify a case to demonstrate possible evaluation scenarios to be easier for understanding of behaviors of each evaluation measure.

Table 1 presents an synthetic example with 10 retrieved documents and 4 corresponding intents of 5 simulated systems. In this example, we assume binary relevance and the intent probability $P(i|q)$ to be equal for all intents. In Table 1, the column "intent" is the intent numbers (i.e., 1, 2, 3, 4) relevant to each document and the column "*ct*" stands for coverage times which is the amount of times that all *unique* intents are covered up to a given rank. The column "*cn*" is cumulative nugget that aggregates the number of ordinary nuggets (i.e., relevant intents). Four rows of the top subtable report the evaluation results of simulated systems measured by α-nDCG, ERR-IA, D#-nDCG and I-rec, respectively. The systems are sorted in the order of α-nDCG.

As suggested by α-nDCG and ERR-IA, "SyntheticSys1" is the best performing system because it retrieves relevant documents with five cumulative nuggets in early ranks, relative to three nuggets performed by the other systems. However, until rank 10, these documents returned by "SyntheticSys1" cover only 3 out of the 4 intents (i.e., i=2, 3 and 4). This result indicates that α-nDCG and ERR-IA do not actually measure diversity of intents. In fact, they account for the number of nuggets diminished by their returned positions and redundancy. In contrast, the simulated runs, identified as "SyntheticSys2-5", are ranked lower than "SyntheticSys1" by α-nDCG and ERR-IA, although they successfully return all four unique intents with the same number of nuggets (12) and the higher number of relevant documents (7). Furthermore, the "SyntheticSys2, 3 & 4" attempt to diversify results by covering more than one round of all intents. This can be seen that α-nDCG and ERR-IA, devised based on redundancy, should be evaluations for *novelty* only, not for both *novelty* and *diversity* as done by previous works. That is, α-nDCG and ERR-IA do not assess the coverage of intents. This issue is later implied by Gobus *et al.* [14], where they introduce α#-nDCG-IA to alleviate this problem of ignoring minor intents by combining their α-nDCG with I-rec. In addition to this synthetic example, our previous works in [15, 16] demonstrates that the similar issue of α-nDCG and ERR-IA occurs in real case scenarios of TREC systems.

Given I-rec, it is suggested that "SyntheticSys1" performs worst in comparison with the other systems. This result is rather straightforward because I-rec basically evaluates the coverage and thus diversity of query-intents, and "SyntheticSys1" covers only some of the intents. However, when considering "SyntheticSys2-5", I-rec cannot distinguish the systems performance after the complete intent coverage is achieved at rank 5. Some questions arise when considering this drawback of I-rec: Should we evaluate "SyntheticSys2/3/5" with high performance if intent 2 covered by a document at rank 5 does not fully satisfy user information needs? Is it appropriate that evaluation metrics just ignore to reward a system, such as "SyntheticSys4", that attempts to further diversify results (with intent 2)? The subsequent retrievals of intent 2 produced by "SyntheticSys4" may be more highly relevant to and thus fulfill the user information need. More precisely, let us consider two possible scenarios regarding what kind of relevant intents users want. Users might regard relevant intents as those having minimum redundancy with previously read documents, or users might even want to continue searching for documents related to an intent previously found novel. This drawback has already been discussed in Section 2.1 as well as the rank-position *not* being addressed by I-rec. These two issues are what we aim to overcome by our proposed measure.

The drawbacks of I-rec are also derived to D#-nDCG. This issue is shown in the case where "SyntheticSys5" is rated better than "SyntheticSys4" although it attempts to keep diversifying search results after completing intent coverage at rank 5. A similar case presents in the results of

[5]Note that I-rec cannot address intent diversity after all intents are fully covered.

Table 1: Five synthetic runs generated from QREL of TREC 2010 Web Diversity Track on query 76, as evaluated by α-nDCG@10 ($\alpha = 0.5$), ERR-IA@10, I-rec@10 and D#-nDCG@10 ($\gamma = 0.5$) (top subtable); and by nCF@10, SW-nCF→ERR-IA@10, and nCF-ERR-IA@10 ($\gamma = 0.5$) (bottom subtable).

System ranking observed at rank 10, according to:				
α-nDCG	#1 (0.737)	#2 (0.695)	#3 (0.690)	#4 (0.657)	#5 (0.597)
ERR-IA	#1 (0.675)	#2 (0.554)	#3 (0.552)	#4 (0.468)	#5 (0.449)
I-rec	#2 (0.750)	1# (1.000)			
D#-nDCG	#4 (0.616)	#1 (0.708)		#3 (0.690)	#2 (0.692)

	SyntheticSys1			SyntheticSys2			SyntheticSys3			SyntheticSys4			SyntheticSys5		
rank	intent	*ct*	*cn*	intent	*ct*	*cn*	intent	*ct*	*cn*	intent	*ct*	*cn*	intent	*ct*	*cn*
#1	2 3 4	0.75	3	1 3	0.5	2	1 3	0.50	2	4	0.25	1	4	0.25	1
#2	2 4	0.75	5	4	0.75	3	4	0.75	3	1 3	0.75	3	1 3	0.75	3
#3		0.75	5		0.75	3		0.75	3		0.75	3		0.75	3
#4		0.75	5		0.75	3		0.75	3		0.75	3		0.75	3
#5	3 4	0.75	7	2 3 4	1.00	6	2 3 4	1.00	6	2	1.00	4	2 4	1.00	5
#6	4	0.75	8	4	1.25	7	4	1.50	7	3 4	1.50	6	3 4	1.50	7
#7		0.75	8		1.25	7		1.50	7		1.50	6		1.50	7
#8		0.75	8	1 3 4	1.75	10	1 3 4	1.75	10	1 2	2.00	8	4	1.50	8
#9	2 4	0.75	10	4	1.75	11	4	1.75	11	2 3 4	2.75	11	3 4	1.50	10
#10	2 3	0.75	12	3	1.75	12	4	1.75	12	1	3.00	12	3 4	1.50	12

nCF	#4 (0.150)	#2 (0.350)		#1 (0.600)	#3 (0.300)
SW-nCF→ERR-IA	#3 (0.150)	#2 (0.350)→(0.554)	#3 (0.350)→(0.552)	#1 (0.600)	#4 (0.300)
nCF+ERR-IA	#4 (0.412)	#2 (0.452)	#3 (0.451)	#1 (0.554)	#5 (0.374)
System ranking observed at rank 10, according to:				

"SyntheticSys2" and "SyntheticSys3", which obtain the same score of 0.708 by D#-nDCG@10. After the success of intent coverage, D#-nDCG relies on the measurement of D-nDCG, simply aggregating the gain of subsequently retrieved intents without a discounting function.

Nevertheless, a combination with D-nDCG also has its own merits to be able to measure the intent probability, graded relevance and *interval* per-intent relevance. The third point is to aggregate the utility of relevant intents subsequently retrieved within a cycle of intent coverage or after any single intent is already covered. However, we believe that using a combination approach in a compensatory fashion like D#-nDCG may cause difficulties for interpreting the search results. In an extreme case, where a diversified system (let us call "SyntheticSys6") returns all relevant documents up to rank 10 and *every* document contains only 3 out of the 4 intents (e.g., intent 2, 3 and 4) (similar to those simulated in Table 1). D#-nDCG@10 rates such a system at 0.875 even though it does not fully cover all intents. In comparison with "SyntheticSys5" that at least completes one time of intent coverage, it is difficult to interpret the result of "SyntheticSys6" obtained by D#-nDCG: it is highly rated due to either high diversity or high novelty.

3.2 Normalized Coverage Frequency

We propose a *diversity* measure using coverage frequency to assess the effectiveness of systems for the diversity and novelty task. Our proposed measure, named as *normalized Coverage Frequency (nCF)*, is formalized based on a similar concept to I-rec that measures diversity through the number of different intents covered by k documents. Major differences from I-rec are that our nCF aims to *i*) explicitly quantify the amount of intent coverage throughout a document ranking and *ii*) extend the set-based metric of I-rec to the rank-based metric. To this aim, we define assumptions that will simplify the development of our metric. First, to be able to identify different intents, we assume that all intents are mutually exclusive and there is no overlapping among intents. Second, we assume that some factors are not suitable to be included together with the estimation of a diversity factor. These include *a*) intent probability and *b*) interval per-intent graded relevance, which will be later accounted for by other existing metrics such as ERR-IA.

DEFINITION 1 (Intent Coverage)
Inspired by the notion of I-rec, intent coverage is the number of unique intents covered as a function of rank. Given a coverage function "cov", the degree to which the number of different intents covered by documents from rank n to k is the estimate of $cov(n,k)$. More precisely, consider a query q with $|I|$ intents $(i_1, i_2, ..., i_{|I|})$ and a range of ranked documents $(d_n, ..., d_k)$. We can define $cov(n,k)$ as:

$$cov(n,k) = \frac{|\bigcup_{r=n}^{k} intent(d_r)|}{|I|} \quad (6)$$

where $intent(d_r)$ is the set of intents to which d_r is relevant. n is a start position and k is a stop position in a ranking.

DEFINITION 2 (Coverage Time)
Coverage time is the amount of time that the intent coverage can be achieved at a given rank. Let r be a ranking position of document d_r, k be a position at which we evaluate a document ranking. The coverage time at rank r is given by $ct(r,k)$. In order to count how many times the percentage of different intents are repeatedly covered in a ranking, coverage time $ct(r,k)$ is conditioned on the set of total intents I, the set of intents covered within a cycle of intent coverage I', and ranking positions r and k. Since we have conditional cases in triplets, we decompose the coverage time as follows:

$$ct(r,k) = \begin{cases} 1 \;\&\; I' = \varnothing \;\&\; p = r+1, & \text{if } |I' \cup intent(d_r)| = |I| \\ cov(p,k), & \text{if } r - k \\ 0 \;\&\; I' \cup = intent(d_r), & otherwise \end{cases}$$

where p is a pointer that indicates a beginning position of a new cycle of intent coverage after all intents are already covered. At the first time when $ct(r,k)$ is executed, $1'$ is set to $\varnothing$ and p is equal to 1. In other words, a coverage time is the total number of times (with the rest percentage of intent coverege) that all intents have been covered by rank k.

DEFINITION 3 (Coverage Frequency)
Finally, the Coverage Frequency is given by the ratio of the coverage time to the duration for which a user scans document down a ranked list and stops at some rank k. By replacing the duration of time or effort a user spends on examining documents by rank, the coverage frequency can be defined as:

$$CF(k) = \frac{\sum_{r=1}^{k} ct(r)}{k} \quad (7)$$

Traditional IR evaluation computes measures only to rank 5, 10, or 20 (partially due to a pooling technique that restricts the number of relevance judgments for analysis at a shallow depth). We must, hence, normalize the measures so that they will average over a set of queries. As a final step in the computation of nCF, we normalize equation 7 cut off at rank k by the ideal cover frequency.

$$nCF@k = \frac{CF(k)}{CF(k)'} = \frac{\sum_{r=1}^{k} ct(r)}{\sum_{r=1}^{k} ct(r)'} \quad (8)$$

where $CF(k)'$ is the ideal cover frequency that could be obtained at rank k. This can be done by, for instance, using a greedy algorithm. We firstly select an available document with the highest $intent(d_r)$, and then continuously select documents that contain at most missing intents yet with minimum $intent(d_r)$. Until complete intent coverage is achieved, this selection process starts again. This iterative method is proceeded until reaching rank k.

3.3 Potential Uses of nCF

With retrieval results of search systems and per-intent relevance judgments, a series of intent coverage and hence *diversity* in a document ranking can be measured by our nCF metric. To evaluate other factors, e.g., intent novelty, interval per-intent graded relevance and intent probability, there are many available redundancy-based measures such as ERR-IA, MAP-IA [1] and IA-version of α-nDCG [14]. For potential uses of nCF, we opt for ERR-IA as a sample metric to measure *novelty* as well as to incorporate *interval per-intent graded relevance* and *intent probability*. Since we can evaluate a ranking's intent diversity and novelty, two integration strategies can be employed: stepwise or compensatory.

1) ***Stepwise Approach***: In an interactive IR context, users could consider diversity first and then novelty in a stepwise fashion, or vice versa. For the purpose of this preliminary investigation, we simply follow the TREC 2009-12 Web Diversity track guidelines which state that IR systems should produce a document ranking that "together provide a complete coverage for a query, while avoiding excessive redundancy." By doing so, we employ a "first rank-by-diversity and then sort-by-novelty" decision process for system evaluation. As a result, the nCF is used as a prior criterion to the ERR-IA in this study (i.e., when nCF scores of systems are equal, we then consider ERR-IA scores). This stepwise measure is named as SW-nCF→ERR-IA.

2) ***Compensatory Approach***: With a compensatory strategy, a linear combination[6] of nCF and ERR-IA can be employed to obtain the relevance score:

$$nCF+ERR\text{-}IA@k = \gamma nCF@k + (1-\gamma) ERR\text{-}IA@k \quad (9)$$

where γ is the relative weight of diversity and novelty. In this paper, we set γ=0.5 by default. We call this metric the compensatory measure with balanced diversity and novelty (nCF+ERR-IA).

4. EXPERIMENT AND VALIDATION

4.1 Plan of Experiments

The scenarios discussed in Section 3.1 had been useful to understand the behaviors of existing measures. Next, we further analyze our nCF aiming to answer the following research question:

- **RQ1**: Is our proposed measure more intuitive than existing measures with respect to the diversity task?
- **RQ2**: How does our measure evaluate performance of retrieval results when considering "real" systems, e.g. TREC systems?
- **RQ3**: Does our proposed measure correlate with other existing measures?

To answer **RQ2** and **RQ3**, we analyze real document rankings obtained from the TREC 2010-12 Web tracks [10]. In our experiments, we acquired the experimental runs submitted to TREC each year by Web track participants. A total of 32 systems were submitted by 12 groups in 2010, 62 systems by 16 groups in 2011 and 20 systems by 8 groups in 2012. We reduce graded relevance to binary relevance[7] and set each intent with an equal probability. These settings are applied for all metrics used in our experiments. For **RQ1**, we simulate document rankings using TREC relevance judgments of topic 76. By doing so, we can simplify possible evaluation scenarios for better understanding of behaviors of comparative measures.

Approaches used in the past to validate newly proposed metrics include evaluating metrics on discriminative power [18]; using a swap method to examine their stability and sensitivity [2]; and comparing them with user preferences or click metrics [17, 21]. Although each of these approaches has its own advantages, we argue that comparison of the existing measures to our measure in terms of intuitiveness by using simulated data is suitable for this work.

4.2 Intuitiveness

We studied intuitiveness of our diversity measure (nCF) with existing measures by using synthetic systems. Let us revisit retrieval scenarios in Table 1. Three rows of the bottom subtable report the evaluation results of synthetic systems assessed by our three corresponding measures, including nCF, SW-nCF→ERR-IA and nCF+ERR-IA.

[6] We use a linear combination because a harmonic mean of two metrics, for instance, will be zero if either of them is zero. Nevertheless, combining two metrics linearly might cause difficulties in result interpretation as it will reward a system even when it satisfies only one of the two combined metrics.

[7] For simplicity we use binary relevance here like TREC evaluations did, although some metrics (e.g., ERR-IA, α-nDCG) can actually handle graded relevance.

(a) TREC 2010 (b) TREC 2011 (c) TREC 2012

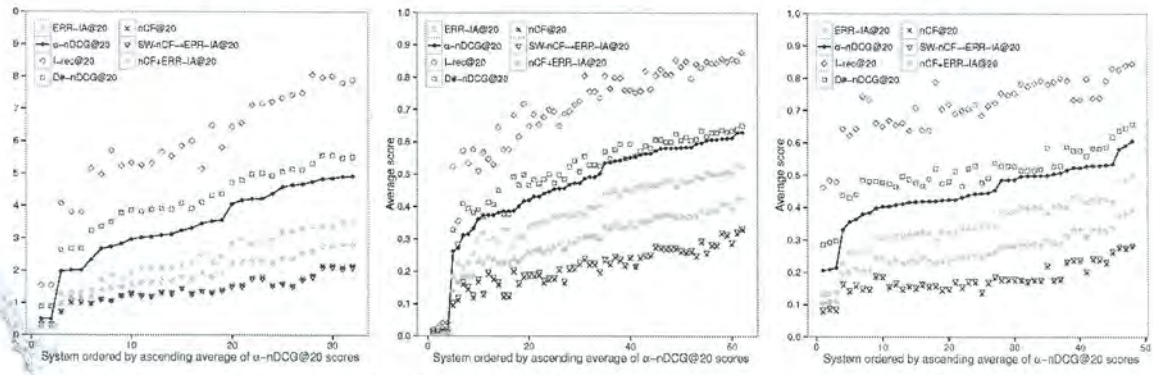


Figure 1: Average performance of TREC 10/11/12 Web diversity runs evaluated by our coverage frequency metric at rank 20 (nCF@20). Compare to α-nDCG, ERR-IA, *I*-rec and D#-nDCG scores

As suggested by nCF, "SyntheticSys4" outperforms other systems because it covers all intents 3 times with nCF@10=0.6. On the other hand, the runs "SyntheticSys1" is assessed as the worst performing system because it does *not* provide complete intent coverage, returning only 3 of the 4 intents. For "SyntheticSys2-3", they obtain the same score of 0.35 by nCF@10 since they further diversify results with additional different intents after complete coverage is achieved. To measure novelty and interval per-intent relevance[8], we integrate our nCF with ERR-IA by using stepwise and compensatory approaches.

Denoted by SW-nCF→ERR-IA, the stepwise approach considers "SyntheticSys2" performing better than "SyntheticSys3" due to the higher score of ERR-IA (0.554 > 0.552). That is, "SyntheticSys2" returns the same number of intents as "SyntheticSys3" does, but with less redundancy of intent 3 relative to 4, as returned at rank 10. Given a compensatory approach (denoted by nCF+ERR-IA), similar results are found, where "SyntheticSys4" performs best followed by "SyntheticSys2" and "SyntheticSys3". Except for the performances of "SyntheticSys1" and "SyntheticSys3", the compensatory approach shows a disagreement with the stepwise. This is due to a linear combination adopted by nCF+ERR-IA. As a result, it is difficult to understand whether a system, such as "SyntheticSys1", peforms better in terms of diversity or novelty since two metrics use different criteria for measurements.

4.3 System Performance

We evaluated all experimental runs submitted to TREC in 2010-12 using our proposed measure with other measures. Figure 1 shows the systems performance with respect to seven studied measures. Each point represents a TREC participant system; they are ordered on the x-axis by α-nDCG. Green squares with cross, black circles, red diamonds and unfilled purple squares represent the values of *ERR-IA*, α-nDCG, *I*-rec and D#-nDCG measures, respectively. All such measures are computed by the ndeval utility used for the Web track[9]. For our measures, blue crosses indicate the *nCF* score, brown triangles are the SW-nCF→ERR-IA, and yellow stars are the nCF+ERR-IA. In these figures we can see that nCF is roughly on the same scale as ERR-IA, though typically 0.1-0.25 lower.

Each increase or drop in the position of blue crosses indicates disagreement with α-nDCG. The increasing trend of the curves with minor fluctuations in Figure 1 indicates that the correlation between nCF and D#-nDCG is high. This is confirmed by system rank correlation between measures, given in the next section.

We analyzed the reason behind disagreements by looking at the results of simulated runs presented in Table 1. We investigated how our proposed measure and the other measures reward diversified search systems. Based on our analysis in Section 4.2, the major reason for disagreements is that α-nDCG and ERR-IA discount the usefulness of each relevant document based upon its rank position and redundancy of relevant intents contained in documents whereas D#-nDCG cannot further evaluate result diversification after all intents are successfully covered.

Table 2: Kendall's τ between rankings of systems submitted to TREC 2010, 2011, 2012 and evaluated at rank 20 with *nCF* against α-nDCG, ERR-IA, *I*-rec and D#-nDCG. All systems are considered.

	TREC	*α-nDCG*	*ERR-IA*	*I-rec*	*D#-nDCG*
nCF	2010	0.810484	0.800403	0.774194	**0.822581**
	2011	0.87626	0.862822	0.868421	**0.899776**
	2012	0.787234	0.766844	0.703901	**0.806738**

4.4 Correlation of System Ranking

We quantify the correlation of measures using Kendall's τ. This is done by ranking the experimental TREC systems under different effectiveness measures. Kendall's τ ranges from -1 (one ranking is the reverse of the other) to 1 (two rankings are the same), with 0 indicating essentially a random reordering. A prior work suggests that a τ value of 0.9 or higher between a pair of ranking indicates high similarity between rankings while a value of 0.8 or lower indicates significant difference [3].

Table 2 shows the Kendall's τ correlation values on our

[8]Note that while this study used binary relevance, ERR-IA, for example, can actually consider graded relevance as well as intent probabilities.

[9]https://github.com/trec-web/trec-web-2013

nCF against α-nDCG, ERR-IA, I-rec and D#-nDCG . The analysis of correlation indicates quite high similarity between systems rankings produced by relative measures. Especially, D#-nDCG has more than 8.0 in every year dataset. However, I-rec is relatively lower than the others.

5. CONCLUSION AND FUTURE WORK

In this work, we proposed a new diversity measure, normalized Coverage Frequency (nCF), which aims to specifically evaluate *diversity* within search results. Our measure is designed to overcome the drawbacks of the current diversity measure, I-rec. The nCF also has several advantages over other existing measures: it explicitly accounts for diversity by the frequency of intent coverage, it can measure the effectiveness of retrieval systems throughout a document ranking up to a given rank position, it does not require any parameter setup, and it can be incorporated with other redundancy-based measures in a stepwise or compensatory fashion. Results from our study showed that our nCF correlates well with existing measures and clearly assesses something different (which is positive for a new measure). Moreover, our measure is more intuitive than I-rec in a diversity aspect.

The clearest direction for future work is to perform a user study, investigating a process of human judgments on system performances in a diversity retrieval task. Particularly, our future study will focus on two key aspects, i.e., diversity and novelty, because these two aspects are fundamentally different but important to a relevance judgment. Different assessment processes, such as compensatory and two-directional stepwise methods, will be defined to determine which method correlates better with human judgments. We plan to start this experiment immediately.

Acknowledgments. This research is fully supported by the National Science and Technology Development Agency, Thailand, with a grant awarded to T. Leelanupab (NSTDA funded project, SCII-NR2012-223).

6. REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *WSDM '09*, pages 5–14, Barcelona, Spain.

[2] C. Buckley and E. M. Voorhees. Evaluating Evaluation Measure Stability. In *SIGIR '00*, pages 33–40, Athens, Greece.

[3] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *SIGIR '04*, pages 25–32, New York, NY, USA. ACM.

[4] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *SIGIR '98*, pages 335–336, Melbourne, Australia.

[5] B. Carterette. An Analysis of NP-Completeness in Novelty and Diversity Ranking. In *ICTIR '09*, pages 89–106, Cambridge, UK.

[6] P. Chandar and B. Carterette. Preference Based Evaluation Measures for Novelty and Diversity. In *SIGIR '13*, pages 413–422, Dublin, Ireland.

[7] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected Reciprocal Rank for Graded Relevance. In *CIKM '09*, pages 621–630, Hong Kong, China.

[8] H. Chen and D. R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In *SIGIR '06*, pages 429–436, Seattle, Washington, USA.

[9] C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *WSDM '11*, pages 75–84.

[10] C. L. Clarke, N. Craswell, I. Soboroff, and G. V. Cormack. Overview of the TREC 2010 Web Track. In *TREC '10*, 2010.

[11] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *SIGIR '08*, pages 659–666, Singapore.

[12] C. L. Clarke, M. Kolla, and O. Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *ICTIR '09*, pages 188–199, Cambridge, UK.

[13] N. Fuhr. A Probability Ranking Principle for Interactive Information Retrieval. *Information Retrieval*, 11:251–265, 2008.

[14] P. B. Golbus, J. A. Aslam, and C. L. Clarke. Increasing evaluation sensitivity to diversity. *Information Retrieval*, 16(4):530–555, Aug. 2013.

[15] T. Leelanupab, G. Zuccon, and J. M. Jose. A Comprehensive Analysis of Parameter Settings for Novelty-Biased Cumulative Gain. In *CIKM '12*, pages 1950–1954, Hawaii, USA.

[16] T. Leelanupab, G. Zuccon, and J. M. Jose. Is Intent-Aware Expected Reciprocal Rank Sufficient to Evaluate Diversity? In *ECIR '13*, pages 738–742, Moscow, Russia, 2013.

[17] F. Radlinski, M. Kurup, and T. Joachims. How Does Clickthrough Data Reflect Retrieval Quality? In *CIKM '08*, pages 43–52, Napa Valley, California, USA, 2008.

[18] T. Sakai. Evaluating Evaluation Metrics Based on the Bootstrap. In *SIGIR '06*, pages 525–532, Seattle, Washington, USA.

[19] T. Sakai, N. Craswell, R. Song, S. Robertson, Z. Dou, and C. Y. Lin. Simple Evaluation Metrics for Diversified Search Results. In *EVIA '10*, pages 42–50.

[20] T. Sakai and R. Song. Evaluating Diversified Search Results Using Per-intent Graded Relevance. In *SIGIR '11*, pages 1043–1052, Beijing, China.

[21] M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do User Preferences and Evaluation Measures Line Up? In *SIGIR '10*, pages 555–562, Geneva, Switzerland, 2010.

[22] R. L. Santos, C. Macdonald, and I. Ounis. Intent-Aware Search Result Diversification. In *SIGIR '11*, pages 595–604, Beijing, China, 2011.

[23] J. Wang and J. Zhu. Portfolio Theory of Information Retrieval. In *SIGIR '09*, pages 115–122, Boston, USA.

[24] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *SIGIR '03*, pages 10–17, Toronto, Canada.

[25] G. Zuccon, L. Azzopardi, and C. J. van Rijsbergen. The Quantum Probability Ranking Principle for Information Retrieval. In *ICTIR '09*, pages 232–240, Cambridge, UK.

ภาคผนวก ข.

ข้อแนะนำของ TREC 2012 ในหมวดเว็บ

TREC 2012 Web Track Guidelines

Charles Clarke, University of Waterloo
Nick Craswell, Microsoft Research
Ellen Voorhees (NIST Contact)

[...]

All judged runs will be fully judged according to both the adhoc and diversity criteria to some minimum depth $k \geq 10$.

If you're planning to participate in the track, you should be on the track mailing list. If you're not on the list, send a mail message to listproc (at) nist (dot) gov such that the body consists of the line "subscribe trec-web *FirstName LastName*".

Timetable

[...]

Overview

Web Tracks at TREC have explored specific aspects of Web retrieval, including named page finding, topic distillation, and traditional adhoc retrieval. Starting in 2009 we introduced a *diversity task* that combines aspects of all these older tasks. The goal of this diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. We continue both the diversity task and a traditional adhoc task for TREC 2012.

The adhoc and diversity tasks share topics, which will be developed with the assistance of information extracted from the the logs of a commercial Web search engine. Topic creation and judging will attempt to reflect a mix of genuine user requirements for the topic. See below for example topics.

Document Collection

The track will again use the ClueWeb09 dataset as its document collection. The full collection consists of roughly 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages. The dataset was crawled from the Web during January and February 2009.

Further information regarding the collection can be found on the associated Website. Since it can take several weeks to obtain the dataset, we urge you to start this process as soon as you can. The collection will be shipped to you on four 1.5TB hard disks at an expected cost of US$790 plus shipping charges.

If you are unable to work with the full dataset, we will accept runs over the smaller ClueWeb09 "Category B" dataset, but we strongly encourage you to use the full "Category A" dataset if you can. The Category B dataset represents a subset of about 50 million English-language pages. The Category B dataset can be ordered through the ClueWeb09 Web. It will be shipped to you on a single 1.0TB hard disk at an expected cost of US$240 plus shipping charges.

Adhoc Task

[...]

Diversity Task

The diversity task is similar to the adhoc retrieval task, but differs in its judging process and evaluation measures. The goal of the diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For this task, the probability of relevance of a document is conditioned on the documents that appear before it in the result list.

For the purposes of the diversity track, each topic will be structured as a representative set of subtopics, each related to a different user need. Example are provided below. Documents will be judged with respect to the subtopics. For each subtopic, NIST assessors will make a binary judgment as to whether or not the

document satisfies the information need associated with the subtopic.

Topics will be fully defined by NIST in advance of topic release, but only the query field will be initially released. Detailed topics will be released only after runs have been submitted. Subtopics will be based on information extracted from the logs of a commercial search engine, and will roughly balanced in terms of popularity. Strange and unusual interpretations and aspects will be avoided as much as possible.

Again this year, the primary evaluation measure for the diversity task will be intent aware expected reciprocal rank (ERR-IA). Developing and validating metrics for diversity tasks continues to be a goal of the track, and we will report a number of other evaluation measures that have been proposed over the past several years. Clarke et al. (WSDM 2011) provides a summary and analysis of many of these evaluation measures, including ERR-IA.

In all other respects, the diversity task is identical to the adhoc task. The same 50 topics will be used. The submission format is the same. The top 10,000 documents should be submitted. You may submit up to three runs, at least one of which will be judged.

Topic Structure

The topic structure will be similar to that used for the TREC 2009 topics. The topics below provide examples.

```
<topic number="6" type="ambiguous">
  <query>kcs</query>
  <description>Find information on the Kansas City
Southern railroad.
  </description>
  <subtopic number="1" type="nav">
    Find the homepage for the Kansas City Southern
railroad.
  </subtopic>
  <subtopic number="2" type="inf">
    I'm looking for a job with the Kansas City
Southern railroad.
  </subtopic>
  <subtopic number="3" type="nav">
```

```
        Find the homepage for Kanawha County Schools in
West Virginia.
      </subtopic>
      <subtopic number="4" type="nav">
        Find the homepage for the Knox County School
system in Tennessee.
      </subtopic>
      <subtopic number="5" type="inf">
        Find information on KCS Energy, Inc., and their
merger with
        Petrohawk Energy Corporation.
      </subtopic>
    </topic>

    <topic number="16" type="faceted">
      <query>arizona game and fish</query>
      <description>I'm looking for information about
fishing and hunting
      in Arizona.
      </description>
      <subtopic number="1" type="nav">
        Take me to the Arizona Game and Fish Department
homepage.
      </subtopic>
      <subtopic number="2" type="inf">
        What are the regulations for hunting and fishing
in Arizona?
      </subtopic>
      <subtopic number="3" type="nav">
        I'm looking for the Arizona Fishing Report site.
      </subtopic>
      <subtopic number="4" type="inf">
        I'd like to find guides and outfitters for
hunting trips in Arizona.
      </subtopic>
    </topic>
```

Initial topic release will include only the *query* field.

As shown in these examples, topics are categorized as either "ambiguous" or "faceted". Ambiguous queries are those that have multiple distinct interpretations. We assume that a user interested in one interpretation would not be interested in the others. On the other hand, facets reflect underspecified queries, with different aspects covered by the subtopics. We assume that a user interested in one aspect may still be interested in others.

Each subtopic is categorized as being either navigational ("nav") or informational ("inf"). A navigational subtopic usually has only a small number of relevant pages (often one). For these subtopics, we assume the user is seeking a page with a specific URL, such as an organization's homepage. On the other hand, an informational query may have a large number of relevant pages. For these subtopics, we assume the user is seeking information without regard to its source, provided that the source is reliable.

For the adhoc task, relevance is judged on the basis of the description field. For the diversity task, a document may not be relevant to any subtopic, even if it is relevant to the overall topic. The set of subtopics is intended to be representative, not exhaustive. We expect each topic to contain 4-10 subtopics.

Submission Format for Adhoc and Diversity Tasks

[...]

ภาคผนวก ค.

การใช้เครื่องมือวัดประสิทธิผล ndeval_nCF

ในวิทยานิพนธ์นี้ ผู้วิจัยได้ทำการสร้างเครื่องมือประเมินประสิทธิผลของระบบค้นคืนสารสนเทศที่เน้นความหลากหลายขึ้นใหม่ ชื่อว่า ndeval_nCF ผู้วิจัยได้เผยแพร่เครื่องมือนี้เพื่อให้นักวิจัยที่สนใจสามารถดาวน์โหลดนำไปใช้งานได้ โดยเครื่องมือ ndeval_nCF นั้นถูกพัฒนาต่อจาก ndeval ดังเดิม ซึ่งเป็นเครื่องมือมาตรฐานสำหรับประเมินประสิทธิผลของระบบค้นค้นสารสนเทศซึ่งเน้นความหลากหลาย โดย ndeval_nCF ซึ่งรวมตัวชี้วัดใหม่ Normalized Coverage Freqnecy (nCF) ที่ผู้วิจัยได้นำเสนอในวิทยานิพนธ์นี้สามารถดาวน์โหลดได้จากลิงค์ข้างล่างนี้

http://www.it.kmitl.ac.th/~teerapong/IT_KMITL/IR_evaluation_tool/ndeval_nCF.c

โปรแกรมถูกพัฒนาโดยภาษาซีพลัสพลัส (C++) และนำไฟล์ดังกล่าวไปคอมไพล์ เพื่อนำไปใช้ในการประเมินผลได้ทันที โดยในการประเมินผลนั้น จำเป็นต้องมี 2 องค์ประกอบหลักคือ 1. ระบบค้นคืน 2. qrel ที่ใช้ประเมินระบบค้นคืนนั้นๆ และ นำข้อมูลดังกล่าวไปประเมินโดยใช้คำสั่งดังต่อไปนี้

ndeval_nCF [options] qrels run

กล่าวคือ ndeval_nCF สามารถระบุตัวเลือก (option) ที่ต้องการได้ และมีการกำหนดไฟล์ qrel และกำหนดไฟล์ของระบบค้นคืน (run) โดยตัวเลือกจะมีดังต่อไปนี้

- -alpha value กำหนดพารามิเตอร์อัลฟาที่ใช้ในตัวชี้วัดเช่นใน α-nDCG เช่น –alpha 0.5
- -beta value กำหนดค่าพารามิเตอร์ของตัวชี้วัด NRBP
- -M depth กำหนดความลึกในการประเมินผลการค้นหา เช่น –M 20 หมายถึง การกำหนดตำแหน่งการประเมิน ณ ตำแหน่งที่ 20

โดยสามารถประเมินด้วยชี้วัดได้ดังต่อไปนี้

- nCF@k
- ERR-IA@k
- nERR-IA@k
- α-DCG@k
- α-nDCG@k
- NRBP
- nNRBP
- MAP-IA

- P-IA@k
- *I*-rec@k

ประวัติผู้เขียน

ผู้เขียน	นายเอก ตั้งสมบูรณ์
วันเดือนปีเกิด	11 มกราคม 2535
สถานที่เกิด	จังหวัด กรุงเทพมหานครฯ
ปริญญา	2557 วิทยาศาสตรบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ (เกียรตินิยมอันดับ2) คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง 2558 วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ แขนงเทคโนโลยีระบบสารสนเทศ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
งานวิจัยที่ตีพิมพ์	*On the Reliability of Diversity and Redundancy-Based Search Metrics* **A. Tangsomboon** and T. Leelanupab; in Proceedings of the 7th International Conference on Information Technology and Electronical Engineering, ICITEE 2015, Chiang Mai, Thailand, to appear *Evaluating Diversity and Redundancy-Based Search Metrics Independently* **A. Tangsomboon** and T. Leelanupab; in Proceedings of the 19th Australasian Document Computing Symposium, ADCS 2014, Melbourne, Australia