ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต : กรณีศึกษา มหาวิทยาลัยราชภัฏธนบุรี

พิศิษฐ์ บวรเลิศสุธี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมเว็บ วิทยาลัยครีเอทีฟดีไซน์ แอนด์
เอ็นเตอร์เทนเมนต์เทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิตย์
พ.ศ. 2560

A Model for Ranking Search Results of Documents on the Intranet : Case Study of Dhonburi Rajaphat University

Pisit Bowornlertsutee

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science Program in Web Engineering

College of Creative Design and Entertainment Technology

Dhurakij Pundit University

2017

หัวข้อวิทยานิพนธ์	ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืน ของเอกสารบนอินทราเน็ต กรณีศึกษา : มหาวิทยาลัยราชภัฏธนบุรี
ชื่อผู้เขียน	พิศิษฐ์ บวรเลิศสุธี
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร.วรสิทธิ์ ชูชัยวัฒนา
สาขาวิชา	วิศวกรรมเว็บ
ปีการศึกษา	2559

บทคัดย่อ

วิทยานิพนธ์นี้นำเสนอตัวแบบการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต ระหว่างเทคนิค Query Dependent Ranking คือ เทคนิคการเปรียบเทียบคำค้นและดัชนีของเอกสาร และเทคนิค Query Independent Ranking คือ เทคนิคที่มีการนำคุณภาพและส่วนประกอบอื่นๆ ของเอกสารเข้ามามีส่วนในการเรียงลำดับผลลัพธ์การค้นคืน โดยการนำความเชื่อมโยงของเอกสารภายในเครือข่ายเป็นส่วนประกอบในการเรียงลำดับผลลัพธ์การค้นคืน ซึ่งเป็นการใช้ Similarity Feature ประกอบด้วย ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) และความเชื่อมโยงของเอกสารในเครือข่าย (Location) มาใช้ในการสร้างดัชนี (Index) จากผลการทดลองเบื้องต้นด้วยผู้ทดสอบจำนวน 35 ผู้ทดสอบ และคำค้นทั้งหมดจำนวน 105 คำสืบค้น พบว่า การเรียงลำดับโดยใช้เทคนิค Query Independent Ranking ผสมผสานกับความเชื่อมโยงของเอกสารภายในเครือข่ายให้ผลลัพธ์การค้นคืนเอกสาร 20 อันดับแรกดีกว่าการเรียงลำดับผลลัพธ์การค้นคืนโดยใช้เทคนิค Query Dependent Ranking เพียงอย่างเดียว จากผลการทดลองสรุปได้ว่า ผู้ใช้ให้ความสำคัญกับผลลัพธ์การค้นคืนของเอกสารที่มีความเกี่ยวข้องกับหน่วยงานที่สังกัดมากกว่าเอกสารทั่วๆ ไป

Thesis Title	A Model for Ranking Search Result of Documents on the Intranet Case Study of : DhonburiRajaphat University.
Author	Pisit Bowornlertsutee
ThesisAdvisor	Assistant Professor Dr.Worasit Choochaiwattana
Academic Program	Web Engineering
Academic Year	2016

ABSTRACT

This thesis presents a model for re-ranking of document retrieval on an intranet. There is a combination of query dependent ranking techniques is a technique for re-ranking model search results by comparing keywords and indexes of documents and the query independent ranking technique, it brings quality and other components. The document takes part in the re-ranking of retrieval results. which links the documents within the network. This will use a similarity feature, including title, detail, department name, and Location to index. There are 35 testers and 105 keywords searched. re-ranking using the query independent ranking technique combined with the linking of documents within the location. The top 20 search results are better than queries using only the query dependent ranking technique. It can be concludes that users location more importance on the search results of documents that are closer to oneself and more relevant to the agencies than the general documents.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยความอนุเคราะห์และการสนับสนุนตลอดการดำเนินการวิจัยจาก ผู้ช่วยศาสตราจารย์ ดร.วรสิทธิ์ ชูชัยวัฒนา อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่กรุณาให้คำแนะนำ ความรู้ ความคิดเห็นต่างๆ อันเป็นประโยชน์ในการทำวิจัย

ขอกราบขอบพระคุณคณาจารย์สาขาวิศวกรรมเว็บ วิทยาลัยครีเอทีฟดีไซน์ แอนด์ เอ็นเตอร์เทนเมนต์เทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิตย์ทุกท่าน ที่กรุณาถ่ายทอดความรู้อันเป็นประโยชน์ตลอดการศึกษา

ขอกราบขอบพระคุณ คุณดุลยารัตน์ ขาววิเศษ และบุคลากรของสาขาวิชาทุกท่านที่ช่วยให้คำแนะนำในการติดต่อประสานงานและจัดทำเล่มวิทยานิพนธ์จนสำเร็จลุล่วงไปด้วยดี

ขอกราบขอบพระคุณคณาจารย์ บุคลากร และนักศึกษา มหาวิทยาลัยราชภัฏธนบุรี ที่กรุณาเสียสละเวลาอันมีค่าในการช่วยทดสอบและประเมินระบบ DRU INTRANET SEARCH

ขอขอบคุณเพื่อนๆ พี่ๆ ทุกคนที่คอยให้ความช่วยเหลือ เอื้อเฟื้อด้านต่างๆ รวมถึงกำลังใจที่คอยแบ่งปันให้กันตลอดเวลา

สุดท้ายขอขอบคุณกำลังใจจากครอบครัว ซึ่งเป็นพลังอันสำคัญและยิ่งใหญ่ที่คอยผลักดันให้การทำวิทยานิพนธ์ครั้งนี้สำเร็จลุล่วงไปได้ด้วยดี

พิศิษฐ์ บวรเลิศสุธี

สารบัญ

สารบัญ (ต่อ)

บทที่ **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของงาน

ในปัจจุบันเทคโนโลยีสารสนเทศและอินเทอร์เน็ตถูกพัฒนาไปอย่างรวดเร็ว ทำให้ปริมาณสารสนเทศต่างๆ ถูกเผยแพร่บนระบบอินเทอร์เน็ต (WWW) และบนระบบอินทราเน็ต (Intranet) ภายในองค์กรต่างๆ อย่างมากมายมหาศาล ซึ่งข้อมูลส่วนใหญ่เป็นข้อมูลสารสนเทศที่มีความสำคัญในด้านต่าง ๆ เช่น ข่าวสาร การศึกษา การวิจัย เป็นต้น การพัฒนาระบบสืบค้นภายในองค์กรเป็นที่ต้องการมากยิ่งขึ้น ทั้งนี้เพื่ออำนวยความสะดวกในการสืบค้นข้อมูลที่เป็นประโยชน์สำหรับใช้ในการปฏิบัติงานต่างๆ และผลจากการที่ข้อมูลมีปริมาณเพิ่มมากขึ้นอยู่ตลอดเวลาจึงส่งผลให้การสืบค้นข้อมูลเกิดปัญหาและใช้เวลาในการคัดกรองข้อมูลที่ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น และผู้ใช้ส่วนใหญ่ขาดความรู้ความเข้าใจเกี่ยวกับการใช้คำค้น (Query) ที่เหมาะสมสำหรับการสืบค้น ซึ่งมีผลทำให้ระบบไม่สามารถค้นคืนข้อมูลที่ตรงกับความต้องการของผู้ใช้ได้อย่างแท้จริง เนื่องจากระบบจะแสดงผลลัพธ์การค้นคืน (Results) เฉพาะเอกสารที่ตรงกับคำค้นเท่านั้น (ศิรัตน์ ศิรนานนท์ : 2549,น.1) และอีกปัญหาที่พบบ่อยในระบบสืบค้นข้อมูลภายในมหาวิทยาลัย คือ เอกสารที่มีประกาศภายในมหาวิทยาลัย เอกสารจะมีการเปลี่ยนแปลงหลายครั้งและประชาสัมพันธ์บนเว็บไซต์ทั้งหมด เมื่อทำการสืบค้นผลลัพธ์ของเอกสารที่ได้จะไม่เรียงลำดับผลลัพธ์การค้นคืนตามความต้องการของผู้ใช้ ซึ่งในบางครั้งทำให้ผู้ใช้เกิดความสับสนในการเลือกดูผลลัพธ์เพื่อนำไปใช้ประโยชน์ ปัญหาต่อมาที่พบได้บ่อยในการพัฒนาระบบสืบค้นเกิดจากการสร้างดัชนีที่ไม่มีคุณภาพ ขาดเทคนิคการวิเคราะห์องค์ประกอบที่เกี่ยวข้องกับสถาปัตยกรรมและโครงการสร้างของเว็บเพจ (Web Page) และส่วนประกอบอื่นๆ ที่ปรากฏอยู่นอกเหนือเอกสารมาเป็นส่วนประกอบในการสร้างดัชนี (วรสิทธิ์ ชูชัยวัฒนา : 2555) ซึ่งปัญหาดังกล่าวก็ส่งผลให้การเรียงลำดับผลลัพธ์การค้นคืนไม่ตรงกับความต้องการของผู้ใช้เช่นเดียวกัน

การพัฒนาระบบสืบค้น (Search Engine) ภายในองค์กรส่วนใหญ่นิยมพัฒนาโดยใช้เทคนิคการทำ Full Text Search หรือ Full Text Indexing คือการสืบค้นจากคำที่มีทั้งหมดในเอกสาร โดยจะนำคำค้น (Query) ไปเปรียบเทียบกับเอกสารทั้งหมดที่มีอยู่ในฐานข้อมูล (Database) ซึ่งเป็นที่นิยมและมีการใช้งานในฐานข้อมูลบรรณานุกรมออนไลน์มาตั้งแต่ปี ค.ศ.1990 เช่นเว็บสืบค้น

ข้อมูล AltaVista ใช้เทคนิคการสืบค้นข้อมูลแบบ Full Text Search โดยการสร้างดัชนีจากส่วนหนึ่งของเอกสารบนหน้าเว็บไซต์ที่มีอยู่ทั้งหมดในฐานข้อมูล เมื่อผู้ใช้ทำการสืบค้นระบบก็จะทำการนำคำค้นไปเปรียบเทียบกับคำทั้งหมดที่มีอยู่ในคลังเอกสาร (Document Corpus) และแสดงผลลัพธ์การค้นคืนออกมา ซึ่งเกิดปัญหา คือ ผู้ใช้จะเสียเวลาในการเข้าถึงข้อมูลค่อนข้างมาก ขาดประสิทธิภาพและไม่ตรงกับความต้องการของผู้ใช้ ต่อมาจึงมีผู้คิดค้นวิธีการนำเสนอผลลัพธ์แบบใหม่ โดยการนำเอาผลลัพธ์มาจัดกลุ่ม (Clustering) เพื่ออำนวยความสะดวกให้กับผู้ใช้ในการเลือกพิจารณาดูผลลัพธ์ (เว็บไซต์ Clusty.com ปัจจุบัน yippy.com) ซึ่งจะแบ่งผลลัพธ์การค้นหาออกเป็นหมวดหมู่ต่างๆ ซึ่งจะส่งผลให้ผู้ใช้สามารถเลือกดูผลลัพธ์ตามหมวดหมู่ที่ตนเองต้องการได้ทันที (วรสิทธิ์ ชูชัยวัฒนา : 2555, น.81-82) ระบบสืบค้นในยุคปัจจุบัน คือ กูเกิ้ล (Google.com) มีการนำเอาเทคนิคต่างๆ มาผสมผสานกันเพื่อให้ได้ผลลัพธ์การค้นคืนที่มีประสิทธิภาพมากยิ่งขึ้น โดยการนำเอาเทคนิค Query Dependent Ranking หรือ Similarity Ranking คือ เทคนิคการนำคำค้นไปเปรียบเทียบกับคำในเอกสาร และเทคนิค Query Independent Ranking หรือ Static Ranking คือ เทคนิคการนำเอาปัจจัยที่เกี่ยวข้องกับเอกสารเข้ามามาพิจารณาร่วมด้วย ตัวอย่างเช่น คุณภาพของเอกสาร (Document Quality) การเชื่อมโยงระหว่างเอกสารที่อยู่ในเครือข่าย (Location) ประวัติการค้นคืน (Query Log) ความเกี่ยวพันธ์กับผู้ใช้ (User Relevance) (ขวัญเรือน โสอุบล : 2557, น.1) และมีการนำปัจจัยเรื่องความใหม่ของเอกสาร (Freshness Documents) ความรวดเร็วในการแสดงผลลัพธ์และหน้าเว็บไซต์ที่รองรับการใช้งานบน Smart Devices หรือ Mobile Friendly เข้ามาร่วมในการพิจารณาการเรียงลำดับผลลัพธ์การค้นคืนอีกด้วย ซึ่งพบว่าให้ผลการเรียงลำดับผลลัพธ์การค้นคืนเป็นที่น่าพึงพอใจกับผู้ใช้มากยิ่งขึ้น

ในการศึกษางานวิจัยนี้ทำการทดลองสร้างตัวแบบ DRU Internet Search เพื่อพิสูจน์ตัวแบบในการเรียงลำดับผลลัพธ์การค้นคืนด้วยเทคนิคการเรียงลำดับแบบ Query Independent Ranking ผสมผสานกับความเชื่อมของเอกสารในเครือข่าย (Location) ให้ผลลัพธ์ดีกว่าการเรียงลำดับแบบ Query Dependent Ranking เพียงอย่างเดียว ในการสร้างดัชนีต้นแบบของทั้ง 2 วิธี จะใช้ตัวอย่างเอกสารจากอินทราเน็ต มหาวิทยาลัยราชภัฏธนบุรี และมุ่งเน้นไปที่ความเชื่อมโยงของเอกสารในเครือข่ายเป็นหลัก

1.2 วัตถุประสงค์ของงานวิจัย

1. เพื่อสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต
2. เพื่อวัดประสิทธิผลของตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต

1.3 สมมติฐานของการวิจัย

ถ้ามีการนำข้อมูลของเอกสารบนระบบอินทราเน็ตซึ่งได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) และหมวดหมู่ของเอกสาร (Category) มาเข้ากระบวนการสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนโดยใช้เทคนิค Hybrid Ranking ซึ่งได้แก่การผสมผสานระหว่าง Query Dependent Ranking มาผสมผสานกับ Ranking ที่สร้างจากความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) คือ Query Independent Ranking จะให้ผลลัพธ์การค้นคืนที่ดีกว่าการเรียงลำดับผลลัพธ์การค้นคืนแบบ Query Dependent Ranking เพียงอย่างเดียว

1.4 ประโยชน์ที่คาดว่าจะได้รับ

สามารถนำตัวแบบที่ได้จากการวิจัยมาปรับปรุงการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรี ซึ่งจะส่งผลให้การเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนระบบอินทราเน็ตมีประสิทธิผลและตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น

1.5 ขอบเขตของการศึกษา

1. ด้านเนื้อหา

เก็บรวบรวมข้อมูลสารสนเทศจากระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏ-ธนบุรี ซึ่งประกอบด้วยเว็บไซต์คณะฯและหน่วยงานรวมจำนวน 19 เว็บไซต์

2. ด้านประชากร

ประชากรตัวอย่างเป็นนักศึกษาระดับปริญญาตรี บุคลากรและอาจารย์ภายในมหาวิทยาลัยราชภัฏธนบุรี จำนวน 35 ตัวอย่าง เพื่อประเมินผลลัพธ์การค้นคืนเอกสาร

3. ขอบเขตด้านเวลา

ช่วงเวลาที่เก็บข้อมูลคือ เดือนมกราคม พ.ศ. 2559 – เดือนมีนาคม พ.ศ. 2560

1.6 เหตุผลเชิงวิชาการที่สนใจศึกษางานวิจัย

1. ต้องการศึกษาวิธีการวัดประสิทธิผลของการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนระบบอินทราเน็ต

2. ต้องการปรับปรุงคุณภาพในระบบค้นคืนเอกสารบนระบบอินทราเน็ต

3. เพื่อนำเสนอโมเดลตัวแบบผสมผสานสำหรับการเรียงลำดับผลลัพธ์การค้นคืนเอกสารบนระบบอินทราเน็ต

1.7 ประเด็นปัญหาและคำถามวิจัย

1. รูปแบบของการสร้างดัชนีของเอกสารมีผลต่อการเรียงลำดับผลลัพธ์การค้นคืนอย่างไร

2. การให้ค่าน้ำหนักของฟิลด์เอกสารที่แตกต่างกัน (Field Boost) จะส่งผลให้ผลลัพธ์ของการเรียงลำดับผลลัพธ์การค้นคืนดีขึ้นได้อย่างไร

3. การจัดอันดับที่สร้างจากความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) เข้ามามีส่วนในการเรียงลำดับผลลัพธ์การค้นคืน จะส่งผลให้ผลลัพธ์การค้นคืนมีประสิทธิผลและตรงความกับต้องการของผู้ใช้หรือไม่

1.8 นิยามศัพท์

เสิร์ชเอนจิน (search engine) หรือโปรแกรมค้นหา คือ โปรแกรมที่ช่วยในการสืบค้นข้อมูล โดยเฉพาะข้อมูลบนอินเทอร์เน็ตโดยครอบคลุมทั้งข้อความ รูปภาพ ภาพเคลื่อนไหว เพลง ซอฟต์แวร์ แผนที่ ข้อมูลบุคคล กลุ่มข่าว และอื่น ๆ ซึ่งแตกต่างกันไปแล้วแต่โปรแกรมหรือผู้ให้บริการแต่ละรายเสิร์ช เอนจินส่วนใหญ่จะค้นหาข้อมูลจากคำสำคัญ (คีย์เวิร์ด) ที่ผู้ใช้ป้อนเข้าไป จากนั้นก็จะแสดงรายการผลลัพธ์ที่คิดว่าผู้ใช้น่าจะต้องการขึ้นมา ในปัจจุบันเสิร์ชเอนจินบางตัว เช่น กูเกิล จะบันทึกประวัติการค้นหาและการเลือกผลลัพธ์ของผู้ใช้ไว้ด้วย และจะนำประวัติที่บันทึกไว้นั้นมาช่วยกรองผลลัพธ์ในการค้นหาครั้งต่อๆ ไป

Information Retrieval (IR) คือ การค้นหาข้อมูลหรือสารสนเทศซึ่งมีเป้าหมายในการค้นเป็นจำนวนมากให้ได้มาอย่างรวดเร็ว และการค้นหาข้อมูลต้องสอดคล้องกับความต้องการในการค้นหา

Crawler Based Search Engines คือ เครื่องมือการค้นหาบนอินเตอร์เน็ตแบบอาศัยการบันทึกข้อมูล และจัดเก็บข้อมูลเป็นหลัก ซึ่งจะเป็นจำพวก Search Engine ที่ได้รับความนิยมสูงสุด เนื่องจากให้ผลการค้นหาแม่นยำที่สุด และการประมวลผลการค้นหาสามารถทำได้อย่างรวดเร็ว จึงทำให้มีบทบาทในการค้นหาข้อมูลมากที่สุดในปัจจุบัน

เว็บครอว์เลอร์ (Web Crawler) เป็นบอตอินเทอร์เน็ตที่ทำงานท่องไปบนเวิลด์ไวด์เว็บ มีจุดประสงค์เพื่อทำการจัดทำดัชนีสำหรับระบบค้นคืน

คำค้นหา (Query) หมายถึง ประโยคคำค้นที่ใช้ค้นคืน ซึ่งผู้ใช้ป้อนเข้าระบบ

คำหยุด (Stop Word) หมายถึง คำที่มีอยู่ในทุกเอกสาร เช่น คำบุพบท คำสันธาน ตัวอย่างเช่น ที่ และ หรือ เป็นต้น

Term หมายถึง คำตามพจนานุกรมที่ตัดได้จาก Query หรือเอกสาร

Stemming หมายถึง การทำรากศัพท์เช่น swim, swimming, swam จะเก็บเป็นคำเดียวกันคือ swim เป็นการลดจำนวนคำศัพท์ในการจัดเก็บข้อมูล

ค่าน้ำหนักของเอกสาร (Term Weight) หมายถึง ค่าน้ำหนักที่บ่งบอกถึงความสำคัญของคำแต่ละคำที่อยู่ในคลังเอกสารจะถูกปรับค่าตามอัตราส่วนระหว่างจำนวนเอกสารทั้งหมดกับจำนวนเอกสารที่มีคำนี้ปรากฏอยู่

คลังเอกสาร (Document Corpus) หมายถึง ฐานข้อมูลหรือที่เก็บรวมรวบเอกสารของระบบค้นคืนในงานวิจัยนี้จะหมายถึงเอกสารที่อยู่บนระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรี

การวิเคราะห์คำ (Parsing) หมายถึง การวิเคราะห์เอกสาร HTML ที่ได้จากการ Crawl ตามโครงสร้าง เพื่อสกัดข้อมูลที่ต้องการให้อยู่ในรูปแบบของฟิลด์ (Field)

โทเคนไนซิ่ง (Tokenizing) หมายถึง กระบวนการประมวลผลข้อความในเอกสารเพื่อให้อยู่ในรูปแบบของคำ ซึ่งในการวิจัยนี้จะหมายถึงคำและข้อความที่เป็นภาษาไทยและภาษาอังกฤษจากเอกสารบนระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรีเท่านั้น

ลูซีน (Lucene) เป็นซอฟต์แวร์โอเพนซอร์สสำหรับใช้เป็นส่วนต่อประสานโปรแกรมประยุกต์ในการค้นคืนสารสนเทศ ลูซีนเหมาะกับการใช้งานใดที่ต้องการการสร้างดัชนีข้อความอย่างเต็มรูปแบบ (Full-text indexing) มีความสามารถในการค้นคืนข้อความแบบเต็มรูปแบบ (Full-text searching)

ThaiAnalyzerLucene เป็นซอฟต์แวร์โอเพนซอร์สที่พัฒนาโดยทีมสรรสารสำหรับใช้เป็นส่วนต่อประสานโปรแกรมประยุกต์ในการค้นคืนสารสนเทศที่เป็นภาษาไทยและภาษาอังกฤษมีหน่วยโปรแกรมที่ทำหน้าที่วิเคราะห์การเรียงตัวของอักขระภาษาไทย(Parser) และทำการสกัดคำเพื่อนำไปสร้างดัชนี (Index)

อินทราเน็ต (Intranet) หมายถึง เครือข่ายคอมพิวเตอร์เชื่อมโยงการสื่อสารด้วยระบบโปรโตคอล (TCP/IP) ซึ่งใช้ในการสื่อสารแลกเปลี่ยนข้อมูลภายในองค์กร

Judgment Score หมายถึง คะแนนหรือเกณฑ์ที่ได้จากความเกี่ยวข้องระหว่างเอกสารกับคำค้น (Query) โดยแบ่งเกณฑ์หรือคะแนนออกเป็น 5 ระดับ คือ 0 – 4 โดยคะแนนเท่ากับ 0 คือเอกสารไม่มีความเกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 4 คือ เอกสารมีความเกี่ยวข้องกับคำค้นมากที่สุด

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่สำคัญ แนวคิด องค์ความรู้และงานวิจัยที่เกี่ยวข้องสำหรับการดำเนินการวิจัยและประเมินผล ซึ่งแบ่งออกเป็น 2 ส่วนคือทฤษฎีและงานวิจัยที่เกี่ยวข้องกับระบบค้นคืนในด้านต่างๆ ดังต่อไปนี้

2.1 ทฤษฎี

2.1.1 ตัวแบบการค้นคืนแบบปริภูมิเว็กเตอร์ (Vector Space Model)

แนวความคิดของเว็กเตอร์คือ การใช้เว็กเตอร์แต่ละมิติ (Dimension) เป็นตัวแทนของเอกสารและคำค้น จากสมการที่ 2.1 แทนเอกสาร

$$D_i = (d_{i1}, d_{i2}, \dots, d_{it},) \tag{2.1}$$

เมื่อ

D_i แทนเว็กเตอร์ของ Index Term

t แทนจำนวน Index Term เช่น คำ (Words) สเตม (Stems) วลี (Phrases) และอื่นๆ

d_{ij} แทนค่าน้ำหนักของ Term ที่ตำแหน่ง j

เมื่อคลังเอกสารมีจำนวน n เอกสาร เขียนปริภูมิเว็กเตอร์ด้วยเมตริกค่าน้ำหนักของคำตามภาพที่ 2.1 ได้ดังนี้

$$\begin{array}{cccccc} & T_1 & T_2 & \dots & T_t \\ D_1 & d_{11} & d_{21} & \dots & d_{t1} \\ D_2 & d_{12} & d_{22} & \dots & d_{t2} \\ \vdots & \vdots & & & \\ D_n & d_{1n} & d_{2n} & \dots & d_{tn} \end{array}$$

ภาพที่ 2.1 ปริภูมิเว็กเตอร์เอกสาร

ในลักษณะเดียวกัน แทนคำค้น Q ด้วยเว็กเตอร์ของ Term Weight เขียนเซตของคำในเอกสาร หรือคำในคำค้นได้ตามสมการที่ 2.2

$$Q = (q_1, q_2, \dots, q_t,) \tag{2.2}$$

เมื่อ Q แทนคำค้น

t แทนค่าน้ำหนักของคำในเอกสารหรือในคำค้น

จากภาพที่ 2.2 เป็นตัวอย่างของปริภูมิเว็กเตอร์ของเอกสารกับจำนวน Term ที่ปรากฏอยู่ในแต่ละเอกสาร เมื่อแต่ละแถวคือค่าน้ำหนักของคำ (Term) และแต่ละคอลัมน์คือเอกสาร เมื่อนำเว็กเตอร์มาใช้แทนเอกสารและคำ สามารถหมุนแกนทั้งสองได้ตามความเหมาะสม ตัวอย่างเอกสาร D_3 แทนด้วยเว็กเตอร์(1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1)และคำค้น "ทุนการศึกษาต่อเนื่อง 2559" แทนด้วยเว็กเตอร์ Q (0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1) เพื่อให้เข้าใจง่ายขึ้นแทนเว็กเตอร์ด้วยภาพ 3 มิติได้ตามภาพที่ 2.3

D1 รับสมัครนักศึกษาทุนเรียนดี เกรดเฉลี่ย 3.00 สำหรับนักศึกษารุ่น 58 ภาคปกติ

D2 ผลการคัดเลือกนักศึกษาเข้ารับทุนต่อเนื่องภาคเรียนที่ 2/2558

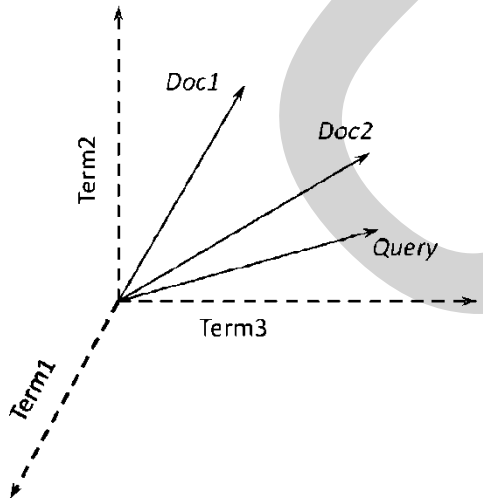
D3 รับสมัครนักศึกษาทุนต่อเนื่อง ปีการศึกษา 2559

Term	**Document**		
	D1	**D2**	**D3**
รับสมัคร	1	0	1
นักศึกษา	2	1	1
ทุน	1	0	1
เรียน	1	0	0
ดี	1	0	0
เกรดเฉลี่ย	1	0	0
3.00	1	0	0
สำหรับ	1	0	0

ภาพที่ 2.2 Term –document matrix ของเอกสาร

Term	Document		
	D1	D2	D3
58	1	0	0
ภาค	1	0	0
ปกติ	1	0	0
ผล	0	0	0
การคัดเลือก	0	1	0
เข้า	0	1	0
รับ	0	1	0
ทุนการศึกษา	0	1	0
ต่อเนื่อง	0	1	1
ปีการศึกษา	0	1	1
2558	0	1	0
2559	0	0	1

ภาพที่ 2.2 (ต่อ)



ภาพที่ 2.3 เว็กเตอร์ของเอกสารและคำค้น

ซึ่งในความเป็นจริงแล้วมิติของทั้งเอกสารและ Term เองนั้น มีปริมาณมหาศาล ทวีคูณมากกว่าจำนวนเอกสารเกินกว่าที่จะแสดงออกมาเป็นภาพสามมิติได้ เพื่อให้ง่ายต่อการคำนวณ จึงต้องแทนแต่ละมิติด้วยจุด (Point) แล้ววัดระยะห่าง (Distance) ระหว่างมุมของเว็กเตอร์ คิดได้จากสมการที่ 2.3 เรียกค่านี้ว่า Similarity Measure หรือ Cosine Similarity Ranking เป็นผลรวมของ Dot Product ระหว่าง Term Weight ของเอกสารกับคำค้น โดยทำ Normalized ค่าคะแนนนี้ด้วย Product Length ของเว็กเตอร์ทั้งสอง นั่นหมายความว่าถ้าระยะห่างมีค่าเข้าใกล้ศูนย์หรือเป็นศูนย์ แสดงว่าคำที่อยู่ในเอกสารสองเอกสารหรือคำค้นไม่มีความเกี่ยวข้องระหว่างกัน

$$Cosine(D_i, Q) = \frac{\sum_{j=1}^{t} d_{ij} . q_j}{\sqrt{\sum_{j=1}^{t} {d_{ij}}^2 . \sum_{j=1}^{t} {q_j}^2}} \tag{2.3}$$

ปัจจัยที่มีผลกับ Vector Space Model คือ Term ที่ปรากฏอยู่ในเอกสารและจำนวน Term ที่ตรงกับคำค้น ซึ่งจะเห็นว่าการที่เอกสารมีความยาวที่มากกว่าย่อมมีจำนวน Term ที่มากกว่า เพื่อลดผลกระทบที่เกิดขึ้น จำเป็นต้องนำความยาวเอกสารมาพิจารณาเพิ่มคือ Term Frequency และจำนวนเอกสารที่ Term นั้นปรากฏ เรียกว่าค่า Term Weights ค่าน้ำหนักของคำนี้คิดจาก $tf_{ik} . idf_i$ พิจารณาเป็น 2 ค่าด้วยกันคือ

Term Frequency (*tf*) ค่าความถี่ของคำในเอกสาร บ่งบอกถึงความสำคัญของคำที่อยู่ในเอกสารนั้น คำนวณได้จากสมการที่ 2.4

$$tf_{ik} = \frac{f_{ik}}{\sum_{j=1}^{t} f_{ij}} \tag{2.4}$$

เมื่อ

tf_{ik} แทน Term Frequency Weight ของคำ kในเอกสาร D_i

f_{ik} แทน จำนวนครั้งที่คำ k ปรากฏในเอกสาร D_i

Inverse document frequency (*idf*) เป็นการพิจารณาถึงความสำคัญของคำที่อยู่ในคลังเอกสาร โดยดูจาก Term นั้นปรากฏอยู่ในเอกสารใดบ้าง ตามสมการที่ 2.5 จะเห็นว่าค่าความสำคัญของ Term จะถูกลดทอนลงและมีค่าเข้าใกล้ศูนย์ เมื่อคำนั้นปรากฏอยู่ในทุกเอกสาร ซึ่งหมายความว่าคำนั้นจะไม่มีประโยชน์ต่อการสืบค้น

$$idf_i = log\frac{N}{n_k} \tag{2.5}$$

2.1.1 Normalized Discounted Cumulative Gain (NDCG)

เป็นการวัดประสิทธิภาพของผลลัพธ์การค้นคืนเอกสารของระบบแนะนำ ระบบค้นคืน เว็บสืบค้น และแอปพลิเคชันที่เกี่ยวข้อง โดยใช้เกรดเป็นเกณฑ์ให้คะแนนกับเอกสารที่เกี่ยวข้องและให้ความสำคัญกับเอกสารที่อยู่ในลำดับต้นๆ ตามสมการที่ 2.6

$$DCG_P = \sum_{i=1}^{P} \frac{(2^{rel_i} - 1)}{log_2(1+i)} \tag{2.6}$$

เมื่อ

P แทนจำนวนผลลัพธ์การค้นคืน

rel_i แทนคะแนนที่ได้จาก Judgment Score ความเกี่ยวข้องระหว่างเอกสารกับคำค้น ในงานวิจัยนี้แบ่งระดับของคะแนนหรือเกณฑ์ออกเป็น 5 ระดับ (5 Point Scale) คือ 0 – 4 โดย 0 คือ เอกสารไม่มีความเกี่ยวข้องกับคำค้น และ 4 คือเอกสารมีความเกี่ยวข้องกับคำค้นมากที่สุด ตามลำดับ $log_2 i$ แทนปัจจัยที่ทำให้คะแนนของเอกสารในตำแหน่งต่างๆ ถูกลดทอนลงตามอัตราส่วนการเปรียบเทียบค่า NDCG Perfect แทนด้วย IDCG (Ideal DCG) คือค่าที่มากที่สุดที่สามารถเป็นไปได้เป็นลำดับการค้นคืนที่ผู้ใช้แต่ต้องการ และเป็นการเรียงลำดับเอกสารที่มีความเกี่ยวข้องกับคำค้นมากที่สุด ถึงน้อยที่สุดคำนวณได้ตามสมการที่ 2.7

$$NDCG_P = \frac{DCG_P}{IDCG_P} \tag{2.7}$$

2.1.2 Mean Average Precision (MAP)

การวัดประสิทธิภาพโดยการหาค่าเฉลี่ยความถูกต้องเป็นการประเมินเอกสารที่ได้จากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยเพียงใด จะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2 หมายถึงเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึงเอกสารมีความเกี่ยวข้องกับคำค้น ดังสมการที่ 2.8

$$MAP = \frac{\sum_{q=1}^{Q} AveP\,(q)}{Q} \tag{2.8}$$

2.2 งานวิจัยที่เกี่ยวข้อง

Chen, Hearst and Lin (1999) ได้ศึกษาและพัฒนาระบบ Cha-Cha : ระบบสำหรับการจัดอันดับผลการค้นหาบนอินทราเน็ต ระบบถูกออกแบบมาเพื่อให้ผู้ใช้ที่มีความหลากหลายของทักษะการใช้คอมพิวเตอร์สามารถใช้งานได้สะดวกจากอินเตอร์เฟซ (Interface) ที่คุ้นเคย และยังสามารถใช้งานได้ดีกับคอมพิวเตอร์ที่มีประสิทธิภาพต่ำ การเชื่อมต่อแบนด์วิดธ์ต่ำและการใช้เว็บเบราเซอร์รุ่นเก่าเนื่องจากการออกแบบอินเตอร์เฟซของระบบลดการใช้กราฟิคของคอมพิวเตอร์ ในส่วนของผลการค้นหาเว็บไซต์เว็บจะทำการบันทึกเส้นทางที่สั้นที่สุดในแง่ของการเชื่อมโยงหลายมิติจากหน้าเว็บเซิร์ฟเวอร์ไปยังหน้าเว็บทุกหน้าภายในอินทราเน็ตเพื่อให้ได้ผลลัพธ์ที่รวดเร็ว ในส่วนของการสร้างดัชนีระบบจะสร้างจากเนื้อหาทั้งหมด (Full text) ของหน้าเว็บและแอตทริบิวต์ข้อมูลที่อยู่ในเมตาดาต้าไฟล์ (Metadata Files) ทำให้การเข้าถึงผลลัพธ์มีประสิทธิภาพมากยิ่งขึ้น

พิมพ์รำไพ เปรมสมิทธ์ (2545) ได้ศึกษาเกี่ยวกับโปรแกรมค้นหา (Search Engine) : การสืบค้นและการประเมิน โดยเฉพาะในเรื่องของประสิทธิภาพการค้นหา Gordon and Pathak (1999: น. 146-147) เสนอลักษณะ 7 ประการที่จะนำมาทำในการประเมินเป็นไปอย่างถูกต้องแม่นยำและให้เนื้อหาสาระ และมีเป้าหมายที่จะประเมินโปรแกรมค้นหาโดยการค้นหาจริง ลักษณะดังกล่าวได้แก่

1. การสืบค้นควรจะเกิดมาจากความต้องการสารนิเทศจริงของผู้ใช้

2. ถ้าทำการทดลองสืบค้นในหัวข้อที่ผู้อื่นระบุไว้ก็ต้องพยามสืบค้นให้ครอบคลุมให้มากที่สุดตามบริบทของความต้องการของผู้นั้น

3. ควรทำการสืบค้นให้มากเพียงพอที่จะสามารถนำมาประเมินประสิทธิภาพของโปรแกรมค้นหาได้

4. ควรทดลองกับโปรแกรมค้นหาที่สำคัญๆ ให้มากที่สุด

5. การวิเคราะห์ประสิทธิภาพของโปรแกรมค้นหาแต่ละโปรแกรมควรจะใช้ลักษณะพิเศษของแต่ละโปรแกรม ซึ่งหมายความว่าไม่จำเป็นต้องใช้กลยุทธ์การสืบค้นเดียวกันในแต่ละโปรแกรมเพราะอาจจะไม่ได้ใช้งานคุณสมบัติพิเศษที่มีในระบบ

6. ผู้ที่ต้องการสารนิเทศจะต้องตัดสินความตรงต่อความต้องการของผลการค้นหา

7. การทดลองที่ดีซึ่งหมายถึงการทดลองที่ทำตามการออกแบบอย่างดีมีการใช้เกณฑ์เป็นที่ยอมรับในการค้นคืนสารนิเทศและการใช้เทคนิคทางสถิติในการประเมินโปรแกรมค้นหาอาจไม่มีการดำเนินการตามลักษณะที่ควรจะเป็นครบทั้ง 7 ประการ แต่สิ่งที่การประเมินจะใช้ก็คือ การ

ตัดสินความตรงต่อความต้องการผลการค้นหาและให้คะแนนผลการสืบค้นดังเช่น คะแนนสูงสุด 5 คะแนน ให้กับการค้นได้เอกสารเนื้อหาทั้งหมด หรือข้อมูลที่ครอบคลุมเกี่ยวกับเรื่องที่ค้นให้ 4 คะแนน สำหรับบทความหรือเพจที่เกี่ยวข้องกับเรื่องให้ 3 คะแนน สำหรับเอกสารที่ชี้แนะไปยังเอกสารที่อยู่ในระดับ 5 หรือ 4 ให้ 2 คะแนน สำหรับโฮมเพจทั่วๆไปให้ 1 คะแนน สำหรับเอกสารที่มีการกล่าวถึงเรื่องที่ค้นบ้าง และ 0 คะแนน สำหรับเอกสารที่ไม่มีเนื้อหาที่ต้องการ จากนั้นก็นำคะแนนไปคำนวณอัตราความถูกต้องตามความต้องการ

Sato, Sakai and Uehara (2004) ได้ศึกษาเกี่ยวกับความสดใหม่ของเอกสารตามเกณฑ์การให้คะแนนสำหรับการดึงข้อมูลในระบบค้นคืน โดยในบทความนี้ได้นำเสนอเกณฑ์การให้คะแนนความสดใหม่ของเอกสารโดยการประเมินค่า FTF-IDF ซึ่งเป็นวิธีการที่ใช้ FTF (Fresh Term Frequency multiplied by Inverse Document Frequency) TF คือความสดใหม่ของเนื้อหาบนเอกสาร แทน TF-FTF โดยกำหนดให้เอกสารที่มีความสดใหม่กว่ามีความสำคัญมากกว่าเอกสารเก่า จากการประเมินค่า FTF-IDF จากเว็บไดอารี่ภาษาญี่ปุ่นและใช้คำภาษาญี่ปุ่นทั้งหมดจะถูกรวมอยู่ในดัชนีและทดลองวิธีที่ง่ายในการตรวจสอบคำหยุด (stop word) ที่จะพิจารณาคำที่มี IDF ค่าน้อยกว่าหรือเท่ากับ 1.0 และพบ Stop Word 87 คำ และจากการรวบรวมข้อมูลจากเว็บไดอารี่ภาษาญี่ปุ่นในเดือนมกราคม ค.ศ. 2004 พบยอดรวมของ TF เป็น 2,437,341 คำและลดลงเป็น 1,034,149 คำ หลังจากมีการวิเคราะห์ Stop Word ซึ่งมีผลทำให้ค่าเฉลี่ยของ TF ลดลงเป็น 15.63 % จาก 36.84 % เนื่องจากคำที่ไม่มีความหมายที่จะค้นหาในเอกสารส่วนใหญ่จะถูกลบออก และในสมุดบันทึกโดย Dr.Tatsumi Hosokawa ได้มีการยืนยันว่าเอกสารที่เนื้อหาสดใหม่กว่าจะจัดอันดับให้อยู่ด้านบนในการแสดงผลลัพธ์ เนื่องจากคุณค่าของเอกสารไม่ได้อยู่ที่เนื้อหาของเอกสารเพียงอย่างเดียวเท่านั้น แต่เวลาและความสดใหม่ของเอกสารก็มีความสำคัญเช่นเดียวกัน

Hang, Yunbo, Jun, Yunhua, Shenjie and Dmitriy (2005) ได้ศึกษาวิธีการใหม่ในการสืบค้นบนระบบอินทราเน็ต (Intranet) บนพื้นฐานของการสกัดข้อมูลโดยการการวิเคราะห์ความต้องการของการสืบค้นข้อมูลจึงแบ่งหมวดหมู่ของความต้องการของระบบค้นคืนดังนี้ เวลา When (time) สถานที่ Where (place) เหตุผล Why (reason) คำนิยาม What is (definition) ผู้เชี่ยวชาญ Who knows about (expert) ใคร Who is (person) วิธีใช้ How to (manual) ความเกี่ยวพันธ์ Tell me about (relevance) กลุ่ม (Group) บุคคล (Person) สินค้า (Product) เทคโนโลยี (Technology) บริการ (Services) บันทึกการค้นหา (Query Log) หมวดหมู่ของเอกสาร (Categories) คำค้นจากผู้ใช้ (Query) ผลลัพธ์ (Result) เอกสารที่ผู้ใช้คลิก (Clicked) เครือข่ายของผู้ค้นหา (Network) และการสำรวจความต้องการของผู้ใช้ (Survey) จากการตรวจสอบปัญหาที่เกิดขึ้นในการสืบค้นบนอินทราเน็ตจึงได้ข้อสรุปคือ

1. การวิเคราะห์ผลการสำรวจและการวิเคราะห์การบันทึกการค้นหา (Search Log Data) พบว่าการสืบค้นบนอินทราเน็ตมีความต้องการโดยสามารถแบ่งออกเป็นลำดับชั้น

2.ได้นำเสนอแนวทางใหม่ในการค้นหาบนระบบอินทราเน็ตโดยใช้เทคนิคระบุชนิดของข้อมูล

3.ได้มีการพัฒนาระบบที่เรียกว่า “Information Desk”

ในการบริการสืบค้นข้อมูลสี่ประเภทของข้อมูล – การหาคำจำกัดความยาวหน้าแรกของกลุ่มหรือข้อมูลส่วนบุคคลหรือพนักงาน และข้อมูลผู้เชี่ยวชาญในหัวข้อนั้นๆ ระบบถูกนำไปใช้กับระบบอินทราเน็ตของไมโครซอฟท์และได้รับเข้าใช้บริการประมาณ 500 คนต่อเดือน แสดงให้เห็นว่าวิธีการที่นำเสนอมีประสิทธิภาพและระบบช่วยให้พนักงานสามารถหาข้อมูลที่ต้องการได้ 4 ประเภทของการค้นหาจากการใช้เทคโนโลยีการสกัดข้อมูลได้รับการใช้ในการสกัดฟีวส์และสรุปข้อมูลล่วงหน้า และการพัฒนาเทคโนโลยีที่มีประสิทธิภาพสูงสำหรับการทำเหมืองข้อมูล

ชูชาติ หฤไชยะศักดิ์ (2548) ได้ศึกษาและพัฒนาระบบค้นคืนข้อมูลโดยไลบรารี่ลูซีนในขั้นตอนการประมวลผลข้อความไม่ว่าจะในระหว่างการสร้างดัชนีหรือการค้นคืนจะต้องมีการใช้ Analyzer ให้เหมาะกับลักษณะของข้อความซึ่งรวมถึงภาษาที่ใช้ สำหรับภาษาไทยนั้นเนื่องจากภาษาไทยมีการเขียนแบบต่อเนื่องโดยไม่มีการเว้นวรรคตอนที่แน่นอน ทีมสรรสารจึงได้พัฒนา ThaiAnalyzer ซึ่งสามารถวิเคราะห์ได้ทั้งภาษาไทยและภาษาอังกฤษ ในกรณีทั่วไปการสร้างดัชนีหรือการค้นคืนก็ควรจะใช้ Analyzer แบบเดียวกัน ตัวอย่างต่อไปนี้แสดงให้เห็นถึงการใช้ ThaiAnalyzer ในการวิเคราะห์คิวรี่สำหรับ QueryParser รวมทั้งการใช้ Query ที่เหมาะสมกับภาษาไทย

ตัวอย่างข้อความ: ข้อความที่ใช้ในการทดสอบมี 5 ชุดเป็นภาษาไทยทั้งหมด ซึ่งเมื่อใช้ ThaiAnalyzer จะมีการแบ่งคำ (Tokenizing) และส่งไปทำการสร้างดัชนี ดังนี้

ข้อความชุดที่ 1 : รัฐบาลไทยร่วมสนับสนุนการท่องเที่ยวในประเทศ

[รัฐบาล] [ไทย] [ร่วม] [สนับสนุน] [การ] [ท่องเที่ยว] [ใน] [ประเทศ]

ข้อความชุดที่ 2 : เที่ยวทั่วไทยในแบบอเมซิ่งไทยแลนด์

[เที่ยว] [ทั่ว] [ไทย] [ใน] [แบบ] [อ] [เม] [ซิ่งไทย] [แลนด์]

ข้อความชุดที่ 3 : นักท่องเที่ยวจากต่างประเทศเดินทางมาประเทศไทย

[นักท่องเที่ยว] [จาก] [ต่าง] [ประเทศ] [เดินทาง] [มา] [ประเทศ] [ไทย]

การตัดคำแม้จะเป็นคำเดียวกันแต่ถ้าปรากฏอยู่ในข้อความแวดล้อมที่ต่างกันก็อาจจะให้ผลที่ต่างกันได้ เช่น คำว่า “ท่องเที่ยว” ในข้อความชุดที่ 1 และข้อความชุดที่ 2

การตัดคำที่ไม่ปรากฏอยู่ในพจนานุกรม (Unknown word) อาจจะมีความผิดพลาดได้ เช่น คำว่า “อเมซิ่ง” ซึ่งอาจทำให้การตัดคำผิดพลาดไปถึงคำที่อยู่ถัดไป

ข้อความชุดที่ 4 : การทดสอบข้อความภาษาอังกฤษเมื่อใช้ ThaiAnalyzer

“The XY&Z Corporation: -xyz@example.com”

[xy] [z] [corporation] [xyz] [example.com]

ข้อความชุดที่ 5 : การทดสอบข้อความภาษาอังกฤษและภาษาไทยเมื่อใช้ ThaiAnalyzer

“กทม.ร่วมสนับสนุนการท่องเที่ยว Amazing Thailand 2005”

[กทม] [ร่วม] [สนับสนุน] [การ] [ท่องเที่ยว] [Amazing] [Thailand] [2005]

จากผลที่ได้จะเห็นว่า ThaiAnalyzer สามารถวิเคราะห์ข้อความที่เป็นภาษาไทยได้ดีทำให้ ThaiAnalyzer เหมาะสำหรับการใช้งานการค้นคืนเอกสารที่เป็นภาษาไทยและภาษาอังกฤษ นอกจากทำให้ผลลัพธ์การค้นคืนที่ถูกต้องและรวดเร็วแล้วยังช่วยประหยัดเนื้อที่ในการจัดเก็บฐานข้อมูลดัชนี

ศิรัตน์ ศิรนานนท์ (2549) ได้ศึกษาการสืบค้นสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้ โดยนำเสนอการทดสอบประสิทธิภาพของระบบการค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์และผลสะท้อนกลับจากผู้ใช้ โดยจะเปรียบเทียบกับระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์และระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์ของคำ ซึ่งในเทคนิคปริภูมิเวกเตอร์จะมีการแปลงเอกสารและข้อสอบถามให้อยู่ในรูปของเวกเตอร์ ส่วนเทคนิคกฎความสัมพันธ์เป็นเทคนิคของการทำเหมืองข้อมูล โดยหาความสัมพันธ์ของคำที่เกิดขึ้นพร้อมกันบ่อยครั้งในเอกสาร เพื่อเพิ่มคำที่มีความสัมพันธ์กับคำในข้อสอบถามก่อนนำไปใช้ดึงเอกสาร ส่วนเทคนิคผลสะท้อนกลับจากผู้ใช้คือเทคนิคที่ใช้ผลสะท้อนกลับจากผู้ใช้ในการปรับข้อสอบถามให้เข้าใกล้กลุ่มเอกสารที่เกี่ยวเนื่องกับข้อสอบถามมากยิ่งขึ้น งานวิจัยนี้เป็นงานวิจัยเชิงทดลอง โดยใช้เอกสารนิตยสาร TIME จำนวน 425 เอกสารและข้อสอบถามจำนวน 83 ข้อสอบถามทดลองเปรียบเทียบประสิทธิภาพของระบบค้นคืนเอกสาร โดยการคำนวณค่าเฉลี่ยฮาร์โมนิคของระบบค้นค้นเอกสารทั้ง 3 รูปแบบ ดังกล่าวข้างต้น จากการวิเคราะห์ผลการทดลองสรุปได้ว่าระบบค้นคืนเอกสารที่ใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์ของคำสามารถทำให้ประสิทธิภาพดีขึ้นกว่าการใช้เทคนิคปริภูมิเวกเตอร์ แต่เมื่อใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับกฎความสัมพันธ์และผลสะท้อนกลับจากผู้ใช้ทำให้ประสิทธิภาพของระบบการค้นคืนเอกสารมากกว่าการใช้เทคนิคปริภูมิเวกเตอร์เพียงอย่างเดียว แต่ต่ำกว่าการใช้เทคนิคปริภูมิเวกเตอร์ร่วมกับเทคนิคการใช้กฎความสัมพันธ์ของคำ

พิมลพรรณ ไชยนันท์ (2548) ได้ศึกษาถึงบทบาทของเสิร์ชเอนจินที่มีบริการสืบค้นด้วยภาษาไทยในการคัดเลือกเนื้อหาที่นำเสนอให้แก่ผู้ใช้บริการ โดยใช้ระเบียบวิธีวิจัยเชิงคุณภาพ คือ การศึกษาเอกสารการสัมภาษณ์เชิงลึกและระเบียบวิธีวิจัยเชิงปริมาณ คือ การทดลองสืบค้นข้อมูล ผลการวิจัยสรุปได้คือ

1. เว็บไซต์เสิร์ชเอนจินที่ทำการศึกษาส่วนใหญ่เป็นการให้บริการเว็บไซต์ในลักษณะของเว็บท่าซึ่งเสิร์ชเอนจินเป็นบริการหลักอย่างหนึ่งในเว็บไซต์

2. ข้อตกลงการให้บริการของเสิร์ชเอนจินระหว่างประเทศที่ทำการศึกษาส่วนใหญ่จะมีใจความหลักที่คล้ายคลึงกัน คือการปฏิเสธความรับผิดชอบและการรับประกันต่อเนื้อหาที่ผู้ใช้ได้รับผ่านบริการของเว็บไซต์ ส่วนเสิร์ชเอนจินของไทยยังไม่มีการกำหนดข้อตกลงการให้บริการที่ชัดเจน

3. เสิร์ชเอนจินที่มีกระบวนการทำงานในลักษณะที่เป็นเสิร์ชเอนจินแบบสมบูรณ์จะใช้วิธีการคัดเลือกและรวบรวมข้อมูลผ่านกระบวนการหลัก 4 ประเภท คือ 1) การเก็บรวบรวมข้อมูลอัตโนมัติโดยเว็บครอเลอร์ 2) การลงทะเบียนเพิ่มชื่อไว้ในฐานข้อมูลของเสิร์ชเอนจินโดยเว็บไซต์ต่างๆ 3) การจ่ายเงินเพื่อให้เว็บไซต์ถูกจัดเก็บไว้ในฐานข้อมูล และ 4) การใช้ฐานข้อมูลของเว็บไดเร็กทอรี่

4. การจัดอันดับของผลลัพธ์การสืบค้นของเสิร์ชเอนจินจะใช้หลักเกณฑ์ในการจัดอันดับที่คล้ายคลึงกัน คือ จำนวนลิงค์ Anchor Text โครงสร้างของเว็บเพจความทันสมัยของเว็บเพจ คุณภาพและปริมาณของเนื้อหาในเว็บเพจ การจ่ายเงินและความนิยมของเว็บเพจ

5. เสิร์ชเอนจินที่ให้บริการระหว่างประเทศจะไม่มีการกลั่นกรองเนื้อหาก่อนทำการจัดเก็บไว้ในฐานข้อมูล แต่มีการจัดกลไกไว้สำหรับให้ผู้ใช้ตั้งค่าการกลั่นกรองเนื้อหาได้เองซึ่งสามารถกลั่นกรองเนื้อหาได้เฉพาะเนื้อหาที่มีการแสดงออกทางเพศอย่างโจ่งแจ้งเท่านั้นส่วนเสิร์ชเอนจินของไทยจะมีการกลั่นกรองเนื้อหาในส่วนของการจัดเก็บและคัดเลือกเนื้อหาไว้ในฐานข้อมูลโดยมีเจ้าหน้าที่เป็นผู้จัดการดูแลเนื้อหา

Stenmark (2006) ได้ศึกษาและวิเคราะห์เนื้อหาจากบันทึกของเครื่องมือค้นหาบนอินทราเน็ต โดยการทำความเข้าใจพฤติกรรมการค้นหาบนอินทราเน็ตจากการเปรียบเทียบแฟ้มบันทึกการสืบค้นของบริษัทขนาดใหญ่ 3 ช่วงเวลาที่แตกต่างกันได้แก่ปี 2000 ปี 2002 และปี 2004 ทำให้สามารถวิเคราะห์ความต้องการข้อมูลจากการสืบค้นข้อมูลของพนักงานในบริษัท ซึ่งทำให้ทราบว่าความต้องการข้อมูลของพนักงานจะเปลี่ยนแปลงไปตามเวลาและจากการวิเคราะห์ความถี่ของคำค้นแสดงให้เห็นถึงความต้องการและทำให้พบว่าผู้ใช้สืบค้นข้อมูลบนอินทราเน็ต (Intranet) มากกว่าการสืบค้นจากเว็บสาธารณะ และส่วนใหญ่คำค้นและเงื่อนไขจะถูกแทนที่เป็นคำค้น

ใหม่ๆ ในแต่ละปีเนื่องจากความต้องการข้อมูลมีความผันแปรอยู่ตลอดเวลา ดังนั้นจึงต้องมีการติดตามความต้องการสารสนเทศทั้งในปัจจุบันและอนาคตและพร้อมที่จะปรับปรุงให้ข้อมูลมีความสอดคล้องกันกับความต้องการของผู้ใช้อยู่ตลอดเวลา

Kharazmi, Nejad, and Abolhassani (2009) ได้ศึกษาการค้นหาจากความสดใหม่โดยการปรับปรุงประสิทธิภาพของเครื่องมือค้นหาโดยใช้เทคนิคการทำเหมืองข้อมูล ความสดใหม่เป็นหนึ่งในตัวชี้วัดประสิทธิภาพการทำงานของเครื่องมือค้นหาจึงได้พัฒนาโปรแกรมรวบรวมข้อมูลที่ชื่อ IFCrawlerโดยสร้างฟังก์ชันของการทำเหมืองข้อมูลและกฎที่ใช้ในการรวบรวมด้านเวลาเข้าไปในระบบ ซึ่งมีประสิทธิภาพที่ดีในการวัดข้อมูลใหม่ๆ และใช้รวบรวมข้อมูลใหม่ๆ เป็นระยะไม่ที่ไม่ห่างกันเช่น รวบรวมข้อมูลสัปดาห์หรือเดือนละครั้งเพื่อเป็นการปรับปรุงเนื้อหาและดัชนีที่มีอยู่และสามารถรวบรวมเอกสารใหม่ๆ เพื่อใช้ในการค้นหา ซึ่งทำให้เห็นว่าการทำเหมืองข้อมูลสามารถปรับปรุงวิธีการรวบรวมข้อมูลใหม่และสามารถปรับปรุงประสิทธิภาพของโปรแกรมค้นหาได้อย่างมีประสิทธิภาพ

Vaughan and Zhang (2007) ได้ศึกษาการวัดอัตราความใหม่ข้อมูลของดัชนีเครื่องมือค้นหาเว็บ จากวิเคราะห์เว็บค้นหาที่นิยม เช่น Google, Yahoo, MSN และพบว่า Google จะดำเนินการโดยแสดงผลลัพธ์ที่ดีที่สุดในหน้าเว็บและมีการปรับปรุงความสดใหม่ของข้อมูลอยู่เป็นประจำทุกวัน แต่ MSN สามารถที่จะปรับปรุงหน้าทั้งหมดภายในเวลาช่วงน้อยกว่า 20 วัน ในแง่ของรูปแบบการจัดทำดัชนีพวกเขาพบว่าวิธีการที่แตกต่างกันในเครื่องมือที่แตกต่างกัน ในขณะที่ MSN แสดงให้เห็นรูปแบบการปรับปรุงที่ชัดเจน ส่วน Google จะแสดงค่าผิดปกติบางอย่างและขั้นตอนการปรับปรุงของดัชนี ส่วน Yahoo ยังไม่มีความชัดเจนในการสร้างดัชนี สรุปได้ว่าคุณภาพของดัชนีที่แตกต่างกันมีผลกระทบต่อผลลัพธ์ที่แตกต่างเช่นเดียวกัน

วิภาพร กุศลชูกุล (2552) ได้ศึกษาการจัดเก็บและค้นคืนกรณีทดสอบและผลของการทดสอบโดยใช้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ โดยนำเสนอการจัดเก็บและค้นคืนกรณีทดสอบและผลของการทดสอบโดยอาศัยโครงสร้างของเอกสารและเพิ่มวิธีการค้นคืนโดยใช้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ของกรณีทดสอบ และผลของการทดสอบด้วยการเปลี่ยนแปลงเทอมในคิวรีและการเปลี่ยนแปลงค่าน้ำหนักของเทอมในคิวรี ด้วยวิธีการค้นคืน โดยใช้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้แบบเวคเตอร์สเปซ (Vector Space) ซึ่งในการค้นคืนโดยใช้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้ จะอาศัยส่วนต่อประสานกับผู้ใช้เพื่อให้ผู้ใช้สามารถเลือกเอกสารที่ค้นคืนได้และสามารถเลือกคำบนฮิสโทแกรมของคำที่ค้นคืนได้พร้อมทั้งสามารถกำหนดน้ำหนักให้กับคำในแต่ละส่วนประกอบได้ ในงานวิทยานิพนธ์ใช้ค่าเรียกคืนและค่าความแม่นยำในการวัดประสิทธิผลของระบบค้นคืนกรณีทดสอบและผลของการทดสอบของ 3 กรณี ดังนี้

1. การค้นคืนโดยไม่ให้ผลป้อนกลับ

2. การค้นคืนที่ให้ผลป้อนกลับโดยการสร้างข้อคำถามใหม่ด้วยการพิจารณาข้อคำถามเดิม

3. การค้นคืนที่ให้ผลป้อนกลับโดยการสร้างข้อคำถามใหม่ด้วยการไม่พิจารณาข้อคำถามเดิม

จากผลการทดลองที่ได้ให้ค่าความแม่นยำแตกต่างกันดังนี้ ในการค้นคืนกรณีทดสอบด้วยกรณีที่ (2) เทียบกับ (1) และ (3) เทียบกับ (1) และ (3) เทียบกับ (2) ให้ผลค่าความแม่นยำเพิ่มขึ้นร้อยละ 2.51, 2.57 และ 2.97 ตามลำดับ และในการค้นคืนผลของการทดสอบด้วยกรณีที่ (2) เทียบกับ (1) ให้ค่าความแม่นยำเพิ่มขึ้นร้อยละ 7.91 และให้ค่าความแม่นยำลดลงด้วยกรณีที่ (3) กับ (1) และ (3) เทียบกับ (2) ร้อยละ 33.30 และ 37.47 ตามลำดับ

Jomsri, Sanguansintukul and Choochaiwattana (2010) ได้การศึกษาการเรียงลำดับผลลัพธ์การค้นคืนบนความวิจัยโดยใช้ Similarity Ranking ร่วมกับเวลาเผยแพร่บทความ (Post time) โดยใช้ข้อมูลจาก CiteULike และสร้างดัชนีจาก Tag Title และ Abstract (TTA) ในการทดลองได้กำหนดค่าน้ำหนักระหว่าง Similarity กับ Static Rank เป็น 50:50 80:20 และ 90:10 ผลลัพธ์ที่ได้จากการประเมินโดยใช้ NDCG ของเอกสาร 15 ลำดับแรกพบว่า CSTRank (90:10) มีค่า NDCG สูงสุด

Lincheng (2011) ได้ศึกษาการสืบค้นข้อมูลขนาดใหญ่ด้วยเทคนิคแบบ Full Text Searching โดยใช้ไลบรารี่ลูซีน (Lucene) เพื่อทดสอบประสิทธิภาพของเครื่องมือค้นหาโดยจะใช้ Analyzer ที่สามารถรองรับการใช้งานหลายภาษาที่เป็นมาตรฐานที่ใช้ในการวิเคราะห์คำคือ 1.SimpleAnalyzer 2.StandardAnalyzer และหลักของการสร้างดัชนีในลูซีนจะใช้พื้นฐานของการดึงข้อความทั้งหมดคือการสร้างดัชนีจากข้อความโดยวิเคราะห์ความหมายของคำเพื่อรองรับการสืบค้นที่มีประสิทธิภาพมากขึ้นจากการทดสอบประสิทธิภาพของระบบค้นหาที่ได้ดำเนินการทดลองจำลองบางส่วน ผลการทดสอบจากการดำเนินการกับคอมพิวเตอร์ Pentium duo 1.86GHz, RAM 2048 MB พื้นที่ค้นหาที่มีประมาณ 530,000 ระเบียน ที่อยู่ในเครือข่ายคอมพิวเตอร์องค์กร (Intranet) จากการใช้แบบสอบถาม (Query) ไปยังระบบผลการค้นหาจะใช้เวลาในการตอบสนองประมาณ 3000 มิลลิวินาทีและเพื่อทดสอบประสิทธิภาพการทำงานจึงเปรียบเทียบการสอบถามโดยตรงกับฐานข้อมูลโดยการดำเนินการประโยค SQL เช่น "Computer" หรือ "Network"จากข้อมูลประมาณ 550,000 ระเบียน จะใช้เวลาการตอบสนองคือ 200 มิลลิวินาทีแต่ถ้ามีการสืบค้นจากฐานข้อมูลโดยตรงจะใช้เวลาตอบสนองถึง 750 มิลลิวินาที เห็นได้ชัดว่าประสิทธิภาพการทำงานของลูซีนมีประสิทธิภาพมากว่าการสืบค้นจากฐานข้อมูลโดยตรง

วรสิทธิ์ ชูชัยวัฒนา (2555) ได้นำเสนอแนวคิดและเทคนิคการปรับปรุงประสิทธิผลของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหา โดยเริ่มจากการกล่าวถึงสถาปัตยกรรมของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหา และอธิบายฟังก์ชันการทำงานของส่วนประกอบต่างๆ ของระบบ มีการอภิปรายถึงแนวทางและเทคนิคสำหรับการปรับปรุงคุณภาพของการจัดทำดัชนีการเรียงลำดับผลลัพธ์การค้นคืน รวมทั้งการออกแบบส่วนแสดงผลลัพธ์การค้นคืน การปรับปรุงประสิทธิผลการทำงานของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหาสามารถทำได้ด้วยการปรับปรุงความสามารถในการทำงานของส่วนประกอบต่างๆของระบบ ไม่ว่าจะเป็นการปรับปรุงคุณภาพของดัชนีด้วยการวิเคราะห์โครงสร้างของเว็บเพจ การนำเอาบันทึกการใช้งานของผู้ใช้มาเป็นข้อมูลตอบกลับ หรือการนำเอาข้อมูลอื่นๆ นอกเหนือจากข้อมูลที่ปรากฏในเอกสารหรือเว็บเพจมาใช้ในการสร้างดัชนี นอกจากนั้นแล้วการคิดค้นและพัฒนาวิธีการเรียงลำดับผลลัพธ์การค้นหาและการปรับปรุงแนวทางในการออกแบบส่วนแสดงผลลัพธ์การค้นหาก็มีบทบาทที่สำคัญในการปรับปรุงประสิทธิผลของระบบค้นคืนสารสนเทศ

ขวัญเรือน โสอุบล และ วรสิทธิ์ ชูชัยวัฒนา (2557) ได้ศึกษาการสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม โดยพิจารณาความเหมือนระหว่างคำสืบค้นและดัชนีของเอกสารและยังพิจารณาปัจจัยที่เกี่ยวข้องกับคุณภาพของบทความวิจัยและแหล่งตีพิมพ์ คุณภาพของบทความวิจัยพิจารณาจากจำนวนของบรรณานุกรม จำนวนของการถูกอ้างอิงโดยบทความอื่น และคุณภาพของแหล่งตีพิมพ์การทดลองได้ถูกจัดขึ้นโดยมีการใช้ NDCG และ MAP เป็นตัววัดสำหรับการประเมินประสิทธิผลของตัวแบบการเรียงลำดับผลลัพธ์การค้นคืน จากผลการประเมินด้วยค่าเฉลี่ย NDCG พบว่า Hybrid2 มีการเรียงลำดับผลลัพธ์การค้นคืนในลำดับที่ 2 ถึง 5 สูงที่สุด ซึ่งเป็นวิธีการนำข้อมูลทางบรรณานุกรมเข้ามาเป็นปัจจัยในการเรียงลำดับใหม่ ได้แก่ จำนวนการอ้างอิงเอกสาร (References) และการถูกอ้างอิงจากเอกสารอื่น (Citation) เป็นตัวแปรที่ใช้วัดคุณภาพของบทความวิจัย จะเห็นว่าเมื่อจำนวนของการอ้างอิงสูงจะส่งผลให้สามารถดันลำดับของบทความขึ้นไปในลำดับต้นๆ ได้ แต่ค่านี้ตัวเลขที่ได้เป็นค่าคงที่ที่จะไม่มีการเปลี่ยนแปลง ส่วนค่าการถูกอ้างอิงจากบทความวิจัยอื่นๆ โดยค่านี้เป็นตัวแปรที่ส่งผลให้สามารถดันลำดับของเอกสารขึ้นไปอยู่ในลำดับต้นๆ ได้เช่นกัน และเมื่อเวลาผ่านไปจำนวนการถูกอ้างอิงถึงนี้จะต้องเพิ่มขึ้นอย่างต่อเนื่องด้วย เมื่อมีบทความใหม่ๆ เผยแพร่ตีพิมพ์ออกมาใหม่แล้วมี Reference ถึงอย่างต่อเนื่องส่วนที่สองคือคุณภาพของแหล่งตีพิมพ์กำหนดให้บทความวิจัยที่ตีพิมพ์ในวารสารวิชาการมีค่าน้ำหนักมากกว่างานประชุมวิชาการเนื่องจากกระบวนการคัดกรอง ตรวจสอบจากผู้เชี่ยวชาญและขั้นตอนในการประเมินที่เน้นคุณภาพมากกว่า และสุดท้ายจำนวนการอ้างอิงถึงมายังบทความวิจัยที่แหล่งตีพิมพ์นี้เป็นผู้จัดพิมพ์ เมื่อระยะเวลาเพิ่มขึ้นแล้วจำนวนการอ้างอิงถึงบทความวิจัยของแหล่งตีพิมพ์ดังกล่าว

เพิ่มขึ้นอย่างต่อเนื่อง ซึ่งสอดคล้องกับการวัดคุณภาพของบทความวิจัยในส่วนแรกส่วนของการให้ค่าน้ำหนักข้อมูลแต่ละฟิลด์ (Field Boost) เข้ามาเป็นปัจจัยเสริมร่วมกับการทำ Similarity Ranking นั้นพบว่าถ้า Term ที่อยู่ในFieldที่เพิ่มค่าน้ำหนักมากๆ จะมีจำนวน Term น้อยซึ่งจะส่งผลให้ค่าคะแนนของ Similarity Score นั้นสูงตามไปด้วยเนื่องจากมีการนำความยาวของฟิลด์มาคิดร่วมด้วย ตัวอย่างเช่น ชื่อบทความวิจัยเป็นหัวข้อหลักที่มีความสำคัญมากสุดให้ค่าน้ำหนักเท่ากับ 3 เป็นฟิลด์ที่มี Term น้อย และค่าน้ำหนักสูงสุดจึงส่งผลให้ค่า Similarity Score สูงกว่าแบบไม่ใช่ Field Boost ซึ่งเมื่อนำปัจจัยทั้งหมดร่วมพิจารณาแล้วทำให้ Hybrid2 มีค่า NDCG ของเอกสารในลำดับต้นๆ ของลิสต์ดีกว่าดัชนีอื่นๆค่าเฉลี่ยความถูกต้อง (Mean Average Precision: MAP) เป็นการวัดค่าแบบไบนารีคือ จริง จะหมายถึงบทความเกี่ยวข้องกับคำค้น และเท็จซึ่งจะหมายถึงบทความที่ค้นคืนออกมาไม่เกี่ยวข้องกับคำค้น เมื่อพิจารณาดูแล้วการประเมินเป็นแบบคะแนน 5 ช่วง คือ 4 – 0 จึงตัดค่าของความถูกต้องที่ 3 โดยช่วงคะแนนที่ 3-4 และ 0-2 พบว่า Hybrid2 ให้ค่าเฉลี่ยความถูกต้องสูงสุดซึ่งสอดคล้องกับการหาค่าเฉลี่ย NDCG สูงแสดงว่าบทความที่แสดงอยู่ในลำดับต้นๆ นั้นตรงกับความต้องการของผู้ใช้และค่าเฉลี่ยความถูกต้องสูงตามไปด้วยเช่นกันการประเมินผลพิจารณาเฉพาะคุณภาพของบทความวิจัยและแหล่งตีพิมพ์ เมื่อบทความมีการเผยแพร่ไปได้ช่วงเวลาหนึ่งจะทำให้ปริมาณการอ้างอิงถึงมีเพิ่มมากขึ้นส่งผลลำดับการค้นคืนเปลี่ยนไปด้วย ตรงส่วนนี้ทำให้การค้นคืนไม่สามารถค้นคืนบทความวิจัยที่เพิ่งถูกตีพิมพ์ออกมาใหม่ขึ้นมาอยู่ในลำดับต้นๆ ได้โดยเฉพาะอย่างยิ่งบทความที่เกี่ยวข้องกับเทคโนโลยีใหม่ การคิดค้นประดิษฐ์วิธีการใหม่ๆ ที่ยังไม่เคยมีใครนำมาใช้

Kaushik, Gaur and Singh (2014) ได้ศึกษาการเพิ่มประสิทธิภาพของเครื่องมือค้นหาโดยการใช้แคช(Cache) เพื่อบันทึกการค้นหาโดยนำเสนอวิธีการที่จะทำให้กระบวนการค้นหาในเครื่องมือค้นหาได้อย่างรวดเร็ว โดยการใช้แคช (Cache)ในการเพิ่มประสิทธิภาพเครื่องมือค้นหา (Search Engine) เนื่องจากแคชเป็นเทคนิคที่มีประสิทธิภาพในเครื่องมือค้นหาสำหรับการปรับปรุงเวลาตอบสนอง และลดระยะเวลาในการประมวลผลแบบสอบถามและลดการใช้แบนด์วิธเครือข่าย แคชจะกระทำโดยใช้บันทึกประวัติของแบบสอบถามเพื่อให้การใช้งานเครื่องมือค้นหามีการตอบสนองอย่างรวดเร็วโดยมุ่งเน้นไปที่การค้นหาจากคำสั่งเชื่อมโยงจากการบันทึกแบบสอบถาม (Query Logs) ของเครื่องมือค้นหานอกจากนี้เครื่องมือค้นหาดำเนินการเชื่อมโยงเอกสารที่เกี่ยวข้องไปยังแบบสอบถามล่วงหน้าจากการบันทึกประวัติการค้นหาของแคชซึ่งจะสามารถช่วยลดระยะเวลาในการค้นหาได้อย่างมีประสิทธิภาพการพัฒนาแคชเพื่อใช้บันทึกแบบสอบถามจะบันทึกแบบสอบถามที่กำหนดโดยผู้ใช้ที่ใช้ในการค้นหาบ่อยครั้งและในการทำงานของแคชจะใช้กฎความสัมพันธ์ของการทำเหมืองข้อมูล (Data Mining) ที่มีให้สำหรับเครื่องมือค้นหาเมื่อผู้ใช้ทำการสอบถามแคชจะสามารถเรียกข้อมูลทุกเพจที่เกี่ยวข้องจากฐานข้อมูลได้อย่างรวดเร็ว ดังนั้นการใช้

แคชเพื่อช่วยเพิ่มประสิทธิภาพของเครื่องมือค้นหายังเป็นเทคนิคที่นิยมใช้กับระบบสืบค้นในปัจจุบันอีกด้วย

สุธาสินี ยกระดับชั้น และ มหศักดิ์ เกตุฉ่ำ (2557) ได้ศึกษาและพัฒนาระบบช่วยเหลือและแก้ไขปัญหาการใช้งานเว็บไซต์สำหรับบุคลากรสายวิชาการ มหาวิทยาลัยราชภัฏสวนสุนันทา ในการรวบรวมปัญหาและแนวทางการแก้ไขปัญหาเว็บไซต์ (Moodle) โดยการประยุกต์ใช้ Case-Based Reasoning ในการสืบค้นแนวทางแก้ไขปัญหาในอดีตมาเป็นแนวทางการแก้ไขปัญหาใหม่ที่เกิดขึ้น โดยการค้นคืนเอกสารใช้วิธีการเปรียบเทียบความเหมือนของคำในเอกสารในรูปแบบของเวกเตอร์(Vector) และการให้น้ำหนักคำ (Term Weighting) ด้วยวิธี TF-IDF (Term Frequency - Inverse Document Frequency)จากผลการทดสอบประสิทธิภาพการค้นคืนข้อมูลของระบบได้ค่าความแม่นยำ (Precision) เท่ากับ 0.80 ค่าความครบถ้วน (Recall) เท่ากับ 0.63 และ ค่า F-Measure เท่ากับ 0.71 จึงสรุปได้ว่า ระบบสามารถใช้งานได้ในระดับดีจากการประเมินคุณภาพของระบบโดยผู้เชี่ยวชาญได้ค่าเฉลี่ยเท่ากับ 4.43 และค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.41 จากผู้ดูแลระบบได้ค่าเฉลี่ยเท่ากับ 4.25 และค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.31 จากผู้ใช้งานได้ค่าเฉลี่ยเท่ากับ 4.30 และค่าส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.45 ซึ่งสรุปได้ว่าระบบที่พัฒนาขึ้นมีคุณภาพอยู่ในระดับดี

จากงานวิจัยที่เกี่ยวข้องมีการนำข้อมูลและส่วนประกอบอื่นๆ ที่อยู่นอกเหนือจากเอกสารมาเป็นส่วนหนึ่งในการสร้างดัชนี (Index)และการเรียงลำดับผลลัพธ์การค้นคืน เช่น การพิจารณาคุณภาพของเอกสาร (Document Quality)การใช้บันทึกการค้นหา (Query Log) ความใหม่ของเอกสาร (Freshness Documents) เครือข่ายและที่อยู่ของเอกสาร (Location) การให้ค่าน้ำหนักส่วนต่างๆ ของเอกสารที่แตกต่างกัน (Weighting Schema) การใช้ Analyzer ให้เหมาะกับเอกสาร และการประเมินโปรแกรมค้นหาโดยใช้ค่าเฉลี่ย NDCG และ(Mean Average Precision: MAP) เป็นต้น ซึ่งสามารถช่วยเพิ่มประสิทธิผลให้การเรียงลำดับผลลัพธ์การค้นคืนตรงกับความต้องการของผู้ใช้มากยิ่งขึ้นในการวิจัยนี้จึงนำเอาแนวความคิดที่ได้มาประยุกต์ใช้ในการสร้างเป็นตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนเอกสารบนระบบอินทราเน็ตซึ่งกล่าวในบทถัดไป

บทที่ 3
ระเบียบวิจัย

การศึกษาวิธีการสร้างตัวแบบผสมผสานสำหรับการเรียงลำดับผลลัพธ์การค้นคืนเอกสารบนระบบอินทราเน็ตโดยใช้เทคนิค Hybrid Ranking โดยนำ Query Dependent Ranking ที่ดึงข้อมูลจากฟิลด์ทั้งหมด 4 ฟิลด์ ได้แก่ ชื่อเอกสาร (Title) รายละเอียดของเอกสาร (Detail) ชื่อหน่วยงาน (Department) และหมวดหมู่ของเอกสาร (Category) มาผสมผสานกับ Ranking ที่สร้างจากการเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) คือ Query Independent Ranking จะให้ผลลัพธ์การค้นคืนที่ดีกว่าการเรียงลำดับผลลัพธ์การค้นคืนแบบ Query Dependent Ranking นั้นมีรายละเอียดในการศึกษาวิจัยสามารถแบ่งเป็นขั้นตอนได้ดังต่อไปนี้

3.1 การเก็บรวบรวมข้อมูลและการวิเคราะห์ข้อมูล
3.2 การสร้างดัชนี
3.3 การสร้างตัวแบบ
3.4 การทดลอง
3.5 การประเมินผล
3.6 เครื่องมือที่ใช้

3.1 การเก็บรวบรวมและการวิเคราะห์ข้อมูล (Data Preparation)

กระบวนการเตรียมข้อมูลสำหรับนำเข้าในขั้นตอนถัดไปคือการสร้างดัชนี (Indexing) ของระบบ ค้นคืน เนื่องจากดัชนีเป็นสิ่งจำเป็นสำหรับการค้นคืนสารสนเทศ เพราะดัชนีเป็นตัวชี้ไปยังข้อมูลและแหล่งข้อมูล ผู้ใช้ดัชนีสามารถสืบค้นสารสนเทศที่ต้องการได้ตรงประเด็นและชี้แหล่งเก็บข้อมูลสารสนเทศได้อย่างถูกต้อง รวดเร็ว ประหยัดเวลาและเพิ่มประสิทธิภาพในการค้นหาซึ่งมีขั้นตอนดังนี้

3.1.1 การครอลว์ข้อมูล (Crawl)

Crawler ทำหน้าที่อ่านข้อมูลเอกสารที่มีรูปแบบโครงสร้างอยู่ในลักษณะของ HTML จากระบบอินทราเน็ต (Intranet) ภายในมหาวิทยาลัยราชภัฏธนบุรี ลักษณะของข้อมูลในเว็บเพจที่

ได้จะมีลักษณะเป็นสตรีมไบต์และนำเข้ากระบวนการวิเคราะห์คำต่อไปเพื่อตัดเอาเฉพาะเนื้อหาสำคัญ

3.1.2 การวิเคราะห์และการคัดกรองข้อมูล (Parsing)

ข้อมูลที่ได้จาก Crawler มีรูปแบบโครงสร้างอยู่ในลักษณะของ HTML Element หรือ HTML Tags โดย Parser จะทำการวิเคราะห์ข้อมูลที่อยู่ภายใต้ Element ที่ต้องการ เพื่อสกัด (Extraction) ข้อมูลที่เป็นข้อความภายในเอกสาร HTML ออกมาเท่านั้นโดยไม่สกัดข้อความที่อยู่ในรูปแบบโครงสร้างอื่น เช่น เอกสารในรูปแบบ .doc และ .pdf เป็นต้น การสกัดข้อมูลแต่ละครั้งจะถูกจัดเก็บลงในฐานข้อมูล MySQL

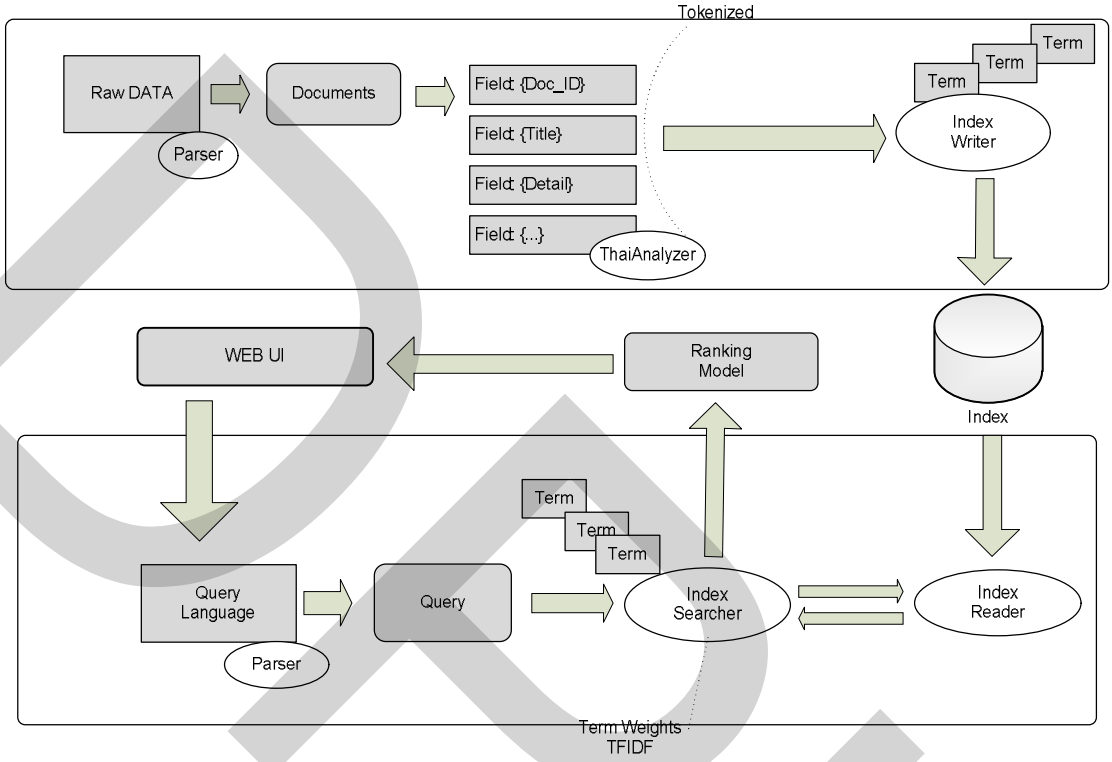
3.1.3 การวิเคราะห์จัดเก็บข้อมูล

การสร้างดัชนีดำเนินการรวบรวมข้อมูลจากระบบอินทราเน็ตภายในมหาวิทยาลัยราช-ภัฏธนบุรี ประกอบด้วยเว็บไซต์ของหน่วยงานต่างๆ ในระหว่างเดือนมกราคม พ.ศ. 2559 – มีนาคม พ.ศ.2560 ประกอบด้วยเอกสารจำนวน 2,892 เอกสาร ซึ่งในแต่ละหน่วยงานจะมีเอกสารประชาสัมพันธ์ผ่านเว็บไซต์ซึ่งสามารถแบ่งประเภทของเอกสารต่างๆ เช่น ข่าวประชาสัมพันธ์ ข่าวกิจกรรม ข่าวอบรม/สัมมนา ประกาศจากมหาวิทยาลัย ประกาศจากหน่วยงาน และเอกสารอื่นๆ ที่เกี่ยวข้องในการปฏิบัติงานภายในมหาวิทยาลัยราชภัฏธนบุรี เป็นต้น ในการทดลองวิจัยครั้งนี้ได้กำหนดฟิลด์ (Field) ของข้อมูลที่ใช้ในการสร้างดัชนีได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category) ซึ่งข้อมูลแต่ละส่วนจะถูกแยกเก็บในฟิลด์ (Field) บนฐานข้อมูล MySQL

3.2 การสร้างดัชนี (Indexing)

ดัชนีเป็นกระบวนการวิเคราะห์ที่เกี่ยวข้องกับเนื้อหาของเอกสาร โดยการอธิบายเนื้อหาของเอกสารออกมาเป็นคำหรือวลีสั้นๆ เพื่อชี้ไปยังตำแหน่งที่อยู่ของเอกสาร โดยการแปลงข้อมูลจากเอกสารบนระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรี โดยการวิเคราะห์โครงสร้างของเว็บเพจและส่วนประกอบอื่นๆที่เกี่ยวข้องเพื่อให้ระบบค้นคืนสามารถเข้าถึงข้อมูลและสืบค้นได้อย่างรวดเร็ว มีประสิทธิภาพและนำมาใช้เป็นฐานข้อมูลเอกสารหรือคลังเอกสาร (Document Corpus) ซึ่งในขั้นตอนของการสร้างดัชนีการวิจัยในครั้งนี้จะใช้ไลบรารีลูซีน (Lucene) และ ThaiAnalyzer เป็นเครื่องมือที่ช่วยทำหน้าที่ในการวิเคราะห์เอกสาร ข้อความ และคำที่อยู่ในเอกสาร ซึ่งข้อมูลส่วนใหญ่เป็นคำและข้อความภาษาไทย จากนั้นจะนำมาจำแนกข้อมูลที่ได้จากกระบวนการวิเคราะห์คำ (Parsing) ออกเป็นฟิลด์ (Field) ต่างๆ โดยที่เอกสารที่จะนำมาสร้างดัชนี

ต้องนำมาผ่านในส่วนสำหรับประมวลผลข้อความ (Text Processing Module) ก่อนเพื่อสกัดเอาคำที่สำคัญไปสร้างดัชนี (ชูชาติ หฤไชยะศักดิ์ : 2548)



ภาพที่ 3.1 ขั้นตอนการวิเคราะห์คำเพื่อสร้างและค้นคืนผ่านดัชนี

จากภาพที่ 3.1 ข้อมูลแต่ละฟิลด์จะถูกนำเข้ามาเพื่อผ่านกระบวนการตัดคำ (Tokenized) จัดเก็บลงในคลังเอกสารรูปแบบของ Inverted Index ทำให้การค้นคืนมีประสิทธิภาพและยืดหยุ่นมากยิ่งขึ้นเพื่อเข้าสู่กระบวนการสร้างส่วนติดต่อกับผู้ใช้และสร้างตัวแบบสำหรับการทดสอบ และประเมินผลในขั้นตอนถัดไป

ในส่วนแรกเป็นการสร้างดัชนีข้อมูลนำเข้าจะได้จากฐานข้อมูลทีละเอกสาร และแยกออกเป็นฟิลด์ (Field) เพื่อแบ่งแยกข้อมูลออกเป็นหมวดหมู่ที่ชัดเจน จากนั้น ThaiAnalyzer จะนำเอกสารมาตัดคำ (Tokenized) เพื่อคำนวณหาค่า Term Weight ให้ Index Writer เขียนลงในคลังเอกสารข้อมูลแต่ละฟิลด์ แสดงตามตารางที่ 3.1 ส่วนที่สองเป็นการอ่านหรือการค้นคืนคำที่ได้จากผู้ใช้งาน จะต้องผ่านกระบวนการเช่นเดียวกับการนำเข้า แต่สิ่งที่ได้จาก Index Reader คือรายการของเอกสารที่ Hit กับคำค้นเข้าสู่ตัวแบบเพื่อเรียงลำดับผลลัพธ์การค้นคืนใหม่ ซึ่งจะอธิบายการคำนวณในหัวข้อถัดไป

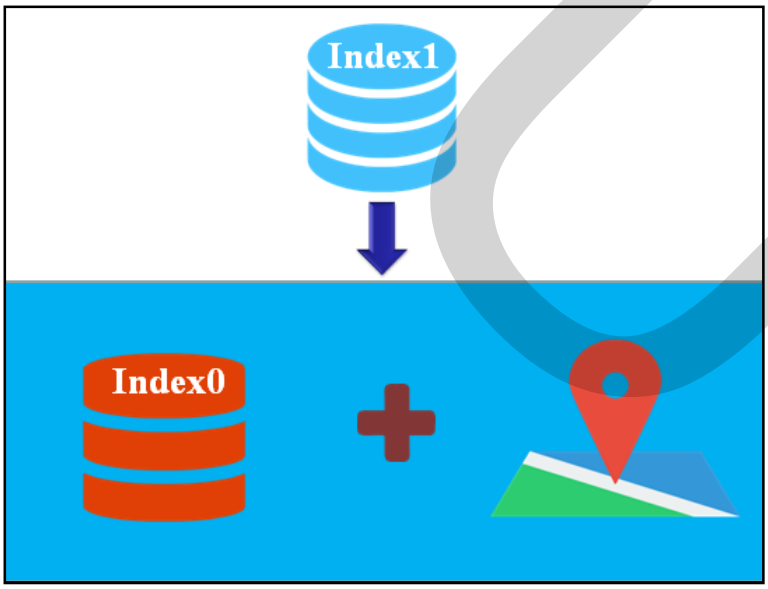
3.3 การสร้างตัวแบบ

ในการศึกษาวิจัยในครั้งนี้ ได้ทำการทดลองสร้างต้นแบบดัชนีทั้งหมด 2 ตัวแบบ ดังนี้

3.3.1 Index0 แทน Full-Text Index ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category) ดังภาพที่ 3.2

ภาพที่ 3.2 ตัวแบบ Index0

3.3.2 Index1 แทน Full-Text Index + Location ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category) และความเชื่อมโยงของเอกสารในเครือข่าย (Location) ดังภาพที่ 3.3



ภาพที่ 3.3 ตัวแบบ Index1

ซึ่งจากตัวแบบทั้ง 2 วิธีสามารถคำนวณได้จากสมการที่ 3.1 โดยค่าคะแนนที่ได้หมายถึง Similarity Measure ระหว่างแต่ละ Team ใน Query เปรียบเทียบกันแต่ละเอกสารและ Hybrid Model โดยการนำ Similarity Feature จาก Index0 ผสมผสานกับความเชื่อมโยงของเอกสารในเครือข่าย (Location) คือ Index1

$$Sim(q,d) = \sum_{t\ in\ q} (tf(t\ in\ d) \times idf(t)^2 \times b(t.field\ in\ d) \times ln(q)) \times c(q,d) \times qN(q) \quad (3.1)$$

เมื่อ

$tf(t\ in\ d)$ แทน Term Frequency

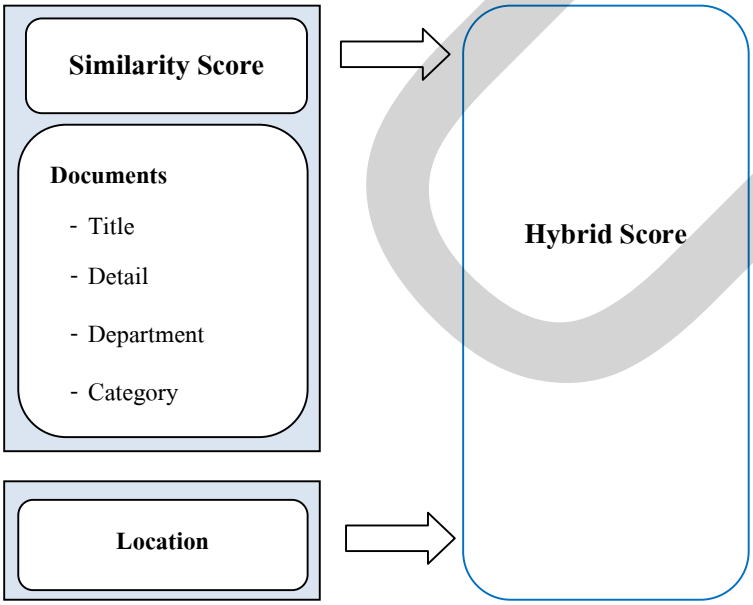
$idf(t)$ แทน Inverse Document Frequency

$ln(q)$ แทน $lenNorm$ คือNormalized ของ Field คิดจากจำนวน Term ใน Field

$c(q,d)$ แทน $coord(q,d)$ จำนวน Term ใน Query ที่ปรากฏในเอกสาร

qN แทน $queryNorm(q)$ค่า Normalized ของคะแนนแต่ละ Query Term

ข้อมูลเอกสารในงานวิจัยนี้เรียกว่า Document Quality เมื่อพิจารณาถึงฟิลด์ข้อมูลที่เป็นตัวแปรสำคัญ ตามภาพที่ 3.4



ภาพที่ 3.4 กรอบแนวคิดการสร้างตัวแบบ

จากภาพที่ 3.2 กำหนดให้ความสัมพันธ์ระหว่าง Similarity Feature กับ Document Feature ของเอกสารเพื่อคำนวณค่า Hybrid Score ตามสมการที่ 3.2

$$Hybrid\ Score\ = Sim(\propto) + Doc(1-\propto)_{+LT}(1-\propto) \quad (3.2)$$

เมื่อ

$\propto$ แทน ค่าน้ำหนัก

Sim แทน Similarity Score ของคำค้น

Doc แทน Document Score คิดจากค่าเฉลี่ยจากการวัดค่าคุณภาพของเอกสาร ร่วมกับชื่อเอกสาร ประเภทเอกสาร รายละเอียดเอกสารและหน่วยงาน

LT แทนความเชื่อมโยงของเอกสารในเครือข่าย

ตารางที่ 3.1 ฟิลด์ข้อมูลที่ใช้ทำดัชนี

ลำดับ	ฟิลด์	รายละเอียด	ดัชนี	ประเภท
1.	Doc_ID	หมายเลขเอกสาร	Not Analyzed	Numeric
2.	Title	ชื่อเอกสาร	Tokenized	String
3.	Detail	รายละเอียดเอกสาร	Tokenized	String
4.	Category	หมวดหมู่ของเอกสาร	Tokenized	String
5.	Date_Public	วันเดือนปีที่เผยแพร่	Not Analyzed	Numeric
6.	Doc_Url	URL ของเอกสาร	Not Analyzed	String
7.	Groupname	ระดับของหน่วยงาน	Not Analyzed	String
8.	Department	ชื่อหน่วยงาน	Tokenized	String
9.	IPDEPT	ไอพีของหน่วยงาน	Tokenized	String

3.4 การทดลอง

การทดสอบเพื่อพิสูจน์ตัวแบบที่สร้างขึ้นจะสามารถเรียงลำดับผลลัพธ์การค้นคืนที่ดีขึ้นตามสมมุติฐาน จึงจัดทำระบบ DRU Intranet Search เป็นหน้าเว็บ GUI ให้ผู้ใช้ติดต่อกับระบบค้นคืนในการทดสอบตัวแบบ โดยนักศึกษา บุคลากร และอาจารย์ภายในมหาวิทยาลัยราชภัฏธนบุรี จำนวน 35 คนจำนวนคำค้นทั้งหมด 105 คำสืบค้น โดยทำการมอบหมายงานให้ผู้ทดสอบแต่ละคนใส่คำค้นที่ต้องการเป็นคำ หรือประโยคใดๆ ในระบบ DRU Intranet Search ระบบจะสืบค้นข้อมูลจากดัชนีจาก Index0 และ Index1 เพื่อคำนวณหาค่า Similarity Score และนำคะแนนที่ดีที่สุดนำเข้าสู่ตัวแบบเพื่อผสมผสานกับความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) เพื่อเข้าสู่กระบวนการประมวลผลภายใต้ตัวแบบเพื่อหาค่า Hybrid Score อีกครั้ง โดยก่อนที่จะแสดงผลให้ผู้ทดสอบประเมิน ระบบจะตรวจสอบเอกสารที่ได้ในแต่ละดัชนีที่เป็นเอกสารเดียวกัน โดยระบบจะรวมผลลัพธ์ให้เหลือเพียงเอกสารเดียว ทั้งนี้เพื่อไม่ให้มีการแสดงผลลัพธ์การคืนค้นซ้ำ หน้าเว็บที่แสดงผลลัพธ์จะแสดงรายละเอียดของเอกสาร ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) หมวดหมู่ของเอกสาร (Category) ชื่อหน่วยงาน (Department) วันที่ประชาสัมพันธ์ (Date_Public) โดยผู้ทดสอบจะต้องอ่านรายละเอียดทั้งหมดและให้คะแนนเอกสาร (Judgment Score) ที่กำลังพิจารณาว่ามีความเกี่ยวข้อง (Relevance) กับคำค้นมากน้อยเพียงใด ซึ่งคะแนนที่ได้ดังกล่าวจะนำไปประเมินผลตัวแบบโดยการทดสอบมีขั้นตอนดังนี้

1. ผู้ทดสอบระบุคำค้นที่ต้องการในหน้าเว็บ โดยระบุเป็นคำหรือประโยคทั้งภาษาไทยหรือภาษาอังกฤษที่ต้องการค้นหา

2. การค้นคืนเอกสาร 20 ลำดับแรกของแต่ละดัชนี ดังรายละเอียดต่อไปนี้

3. การค้นคืนเอกสาร 20 ลำดับของดัชนี Index0 จะเรียงลำดับผลลัพธ์จากเทคนิค Query Dependent Ranking หรือ Similarity Ranking

4. การค้นคืนเอกสาร 20 ลำดับของดัชนี Index1 จะเรียงลำดับผลลัพธ์จากเทคนิค Query Independent Ranking หรือ Static Ranking และผสมผสานกับความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location)

5. จากนั้นนำผลลัพธ์ของทั้ง 2 Index มาผสมผสานกัน (Combination) และระบบจะหาค่าคะแนนที่ดีที่สุดของการเรียงลำดับผลลัพธ์การค้นคืน

6. การแสดงผลลัพธ์จะแสดงแบบสุ่ม (Random) ดังนั้นผู้ทดสอบจะไม่ทราบว่าผลลัพธ์การค้นคืนมาจากดัชนีตัวแบบใด และผู้ทดสอบจะไม่ทราบว่าผลลัพธ์นั้นอยู่ในลำดับใด ทั้งนี้เพื่อป้องกันการเกิดความลำเอียงในการประเมินผล

7. ผู้ทดสอบให้คะแนน (Judgment Score) แต่ละเอกสารโดยพิจารณาความเกี่ยวข้องระหว่างคำค้นกับผลลัพธ์การค้นคืนโดยคะแนนอยู่ระหว่าง 0 ถึง 4 ตามตารางที่ 3.2

ตารางที่ 3.2 Judgments Score

คะแนน	คำอธิบาย
4	มีความเกี่ยวข้องกันอย่างมาก (Very Relevant)
3	มีความเกี่ยวข้อง (Relevant)
2	มีความเกี่ยวข้องกันบางส่วน (Somewhat Relevant)
1	มีความเกี่ยวข้องกันเป็นส่วนน้อย (Only Slightly Relevant)
0	ไม่มีความเกี่ยวข้องกัน (Non-Relevant)

3.5 การประเมินผล

เมื่อได้ผล Judgment Score ของเอกสารจากการประเมินของผู้ทดสอบจากนั้นจะถูกนำมาประเมินผล 2 แบบ คือแบบแรกคิดค่า Normalized Discounted Cumulative Gain (NDCG) เป็นการประเมินลำดับผลลัพธ์การค้นคืนที่ได้มีประสิทธิผล (Effectiveness) โดยนำเอกสารทั้งหมดเรียงลำดับตาม Judgment Score นำมาคำนวณเป็น DCG Perfect หรือ Ideal DCG คะแนนที่ได้มีความหมาย คือ คำค้น (Query) มีความเกี่ยวพันธ์ (Relevance) กับผลลัพธ์ของเอกสารนั้นๆ ที่ตำแหน่ง k เมื่อกำหนดให้คำค้น q และเซตของเอกสารจากการสืบค้น ซึ่งคะแนนของเอกสารในแต่ละตำแหน่งสามารถคำนวณได้จากผลลัพธ์การค้นคืนของเอกสารลำดับแรกจนถึงเอกสารลำดับสุดท้าย ส่วนที่สองคือการวัดประสิทธิภาพคือการหาค่าเฉลี่ยความถูกต้องของการค้นหาแต่ละครั้งเรียกว่า Mean Average Precision (MAP) เป็นการประเมินผลลัพธ์ของเอกสารที่ได้จากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยเพียงใด โดยจะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2 หมายถึงผลลัพธ์การค้นคืนของเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึง ผลลัพธ์ของเอกสารมีความเกี่ยวข้อง (Relevance) กับคำค้น (Query)

3.6 เครื่องมือที่ใช้ในการวิจัย

วิธีการดำเนินการวิจัยเป็นวิธีการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรีมีขั้นตอนต่างๆ ตั้งแต่การเก็บรวบรวมข้อมูล การวิเคราะห์ข้อมูล การสร้างดัชนีและระบบค้นคืนสำหรับทดลองดัชนีตัวแบบ มีเครื่องมือที่ใช้ในการวิจัยดังนี้

1. Crawler โปรแกรมที่พัฒนาด้วย Java Application เพื่ออ่านข้อมูลบนเว็บ และจัดเก็บข้อมูลที่ต้องการลงในฐานข้อมูล
2. MySQL ฐานข้อมูลสำหรับเก็บข้อมูลที่ได้จากการวิเคราะห์คำ
3. Lucene จาวาไลบรารีสำหรับการสร้างดัชนีเอกสารที่เป็นภาษาไทยและภาษาอังกฤษเพื่อใช้เป็นคลังเอกสารในระบบค้นคืน
4. ThaiAnalyzer ใช้ในการวิเคราะห์คิวรี่สำหรับ Query Parser รวมทั้งการใช้ Query ที่เหมาะสมกับภาษาไทย
5. Java server pages (JSP) ภาษา java application เพื่อใช้สร้างหน้า GUI ระบบค้นคืนเพื่อใช้ในการทดสอบตัวแบบ

บทที่ 4
ผลการศึกษา

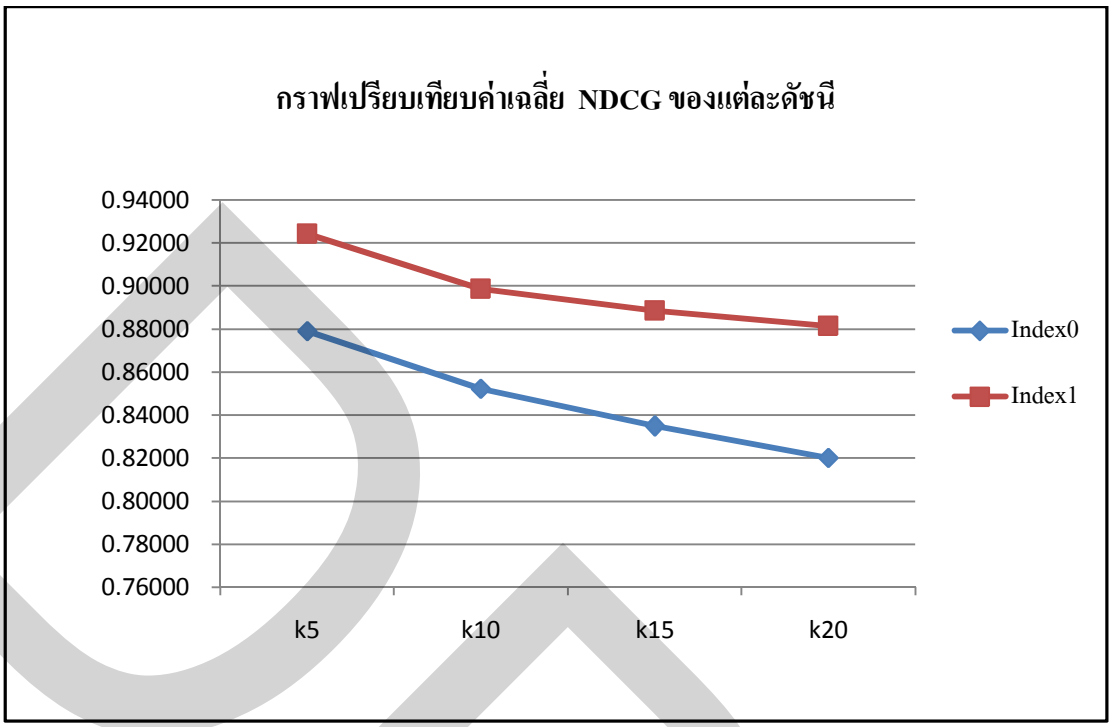
จากผลการทดลองระบบ DRU Intranet Search เพื่อประเมิน Judgment Score ที่ได้จากผู้ทดสอบ และนำไปคำนวณค่า DCG Perfect Score จากผลการทดสอบเบื้องต้นมีผู้ทดสอบทั้งสิ้น 35 ผู้ทดสอบ โดยมีจำนวนคำค้นทั้งหมด 105 คำสืบค้น จากการสร้างตัวแบบของทั้ง 2 วิธี ได้ผลดังนี้

4.1 ค่าเฉลี่ย NDCG

เป็นการประเมินลำดับผลลัพธ์การค้นคืนที่ได้มีประสิทธิผล (Effectiveness) โดยนำเอกสารทั้งหมดเรียงลำดับตาม Judgment Score นำมาคำนวณเป็น DCG Perfect หรือ Ideal DCG คะแนนที่ได้มีความหมาย คือ คำค้น (Query) มีความเกี่ยวพันธ์ (Relevance) กับผลลัพธ์ของเอกสารนั้นๆ ที่ตำแหน่ง k เมื่อกำหนดให้คำค้น q และเซตของเอกสารจากการสืบค้น ซึ่งคะแนนของเอกสารในแต่ละตำแหน่งสามารถคำนวณได้จากผลลัพธ์การค้นคืนของเอกสารลำดับแรกจนถึงเอกสารลำดับสุดท้าย ตามสมการที่ 2.7 ซึ่งกล่าวไปแล้วในบทที่ 2

ตารางที่ 4.1 ค่าเฉลี่ย NDCG

k	Index0	Index1
5	0.87918	0.92443
10	0.85239	0.89877
15	0.83485	0.88866
20	0.82020	0.88155



ภาพที่ 4.1 เปรียบเทียบค่าเฉลี่ย NDCG ของแต่ละดัชนี

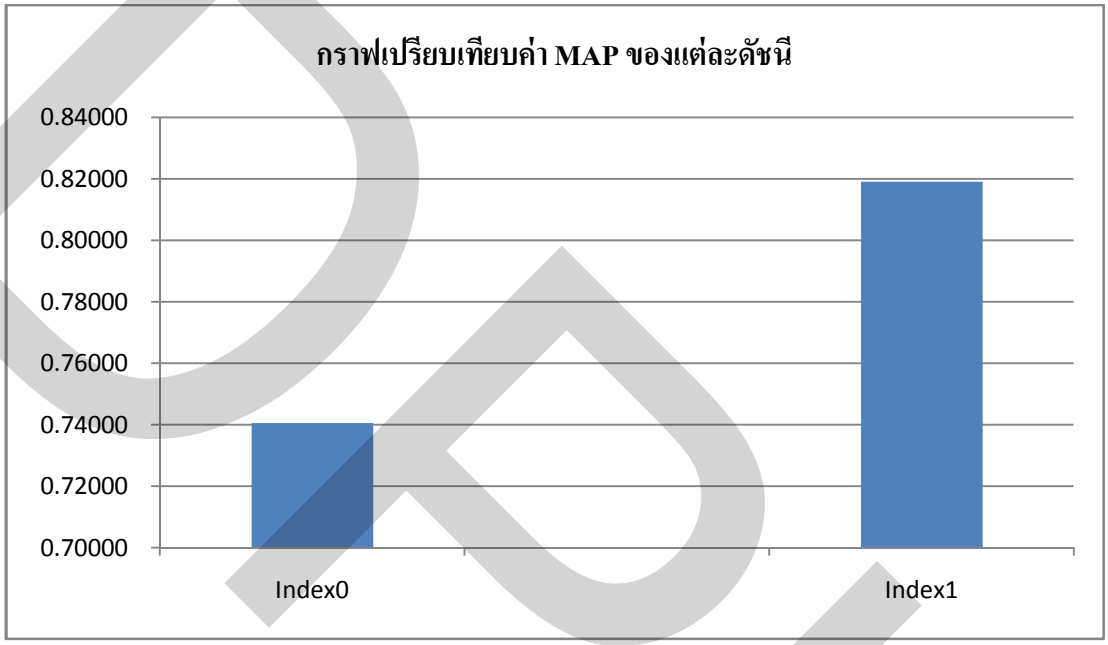
เมื่อพิจารณาจากกราฟแสดงดังภาพที่ 4.1 พบว่า NDCG ของ Index0 ที่ตำแหน่งเอกสาร K5 เท่ากับ 0.87918 และ Index1 ที่ตำแหน่งเอกสาร K=5 เท่ากับ 0.92443 ดังตารางที่ 4.1 เมื่อพิจารณาค่าเฉลี่ย NDCG ของ Index0 และ Index1 จะสังเกตเห็นว่าทุกช่วงของตำแหน่ง K5 ถึง K20 ของ Index1 ให้ผลลัพธ์การค้นคืนเอกสารมีประสิทธิผลดีกว่า Index0 โดยพิจารณาผลลัพธ์จากค่าเฉลี่ย NDCG ของทั้ง 2 ตัวแบบ

4.2 ค่า MAP

เป็นการหาค่าเฉลี่ยความถูกต้องของการค้นหาแต่ละครั้ง เรียกว่า Mean Average Precision (MAP) ตามสมการที่ 2.8 ซึ่งได้กล่าวไปแล้วในบทที่ 2 เป็นการประเมินผลลัพธ์ของเอกสารที่ได้จากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยเพียงใด โดยจะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2 หมายถึงผลลัพธ์การค้นคืนของเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึง ผลลัพธ์ของเอกสารมีความเกี่ยวข้อง (Relevance) กับคำค้น (Query)

ตารางที่ 4.2 ค่า MAP

Index0	Index1
0.74043	0.81892



ภาพที่ 4.2 เปรียบเทียบค่า MAP ของแต่ละดัชนี

เมื่อพิจารณาจากกราฟแสดงดังภาพที่ 4.2 พบว่าค่า MAP ของ Index0 = 0.74043 และ Index1 = 0.81892 ดังตารางที่ 4.2 สรุปได้ว่า Index1 สามารถค้นคืนเอกสารได้ถูกต้องมากกว่า Index0 ที่คะแนน 0.07849 โดยพิจารณาจากผลลัพธ์ค่า MAP ของทั้ง 2 ตัวแบบ

จากการประเมินประสิทธิผลของตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนเอกสารบนอินทราเน็ตทั้ง 2 ตัวแบบ จึงสรุปได้ว่า ค่าเฉลี่ย NDCG และค่า MAP เป็นไปในทิศทางเดียวกัน คือ ตัวแบบ Index1 ให้ผลลัพธ์การค้นคืนเอกสารมีประสิทธิผลดีกว่า Index0 ซึ่งสามารถดูตัวอย่างการประเมินทั้ง 2 วิธีได้ในภาคผนวก ฉ

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต และการศึกษาวิจัยพบว่า ถ้ามีการนำข้อมูลของเอกสารบนระบบอินทราเน็ตซึ่งได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) และหมวดหมู่ของเอกสาร (Category) มาเข้ากระบวนการสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนโดยใช้เทคนิค Hybrid Ranking ซึ่งได้แก่การผสมผสานระหว่าง Query Dependent Ranking มาผสมผสานกับ Ranking ที่สร้างจากความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) คือ Query Independent Ranking จะให้ผลลัพธ์การค้นคืนที่ดีกว่าการเรียงลำดับผลลัพธ์การค้นคืนแบบ Query Dependent Ranking เพียงอย่างเดียว ดังนั้นในการเพิ่มประสิทธิภาพและประสิทธิผลในการสร้างตัวแบบจึงต้องมีการนำปัจจัยอื่นๆ เข้ามามีส่วนในการสร้างดัชนีและเรียงลำดับผลลัพธ์การค้นคืน เช่น การนำปัจจัยความใหม่ของเอกสาร การนำปัจจัยของการบันทึกการค้นหา การลดระยะเวลาและขนาดของการสร้างดัชนี และการพัฒนา Crawler เป็นต้น จากผลการทดลองสามารถสรุป อภิปรายผลการดำเนินการวิจัย ปัญหาและอุปสรรค ข้อจำกัดของงานวิจัย และข้อเสนอแนะ โดยมีรายละเอียดดังต่อไปนี้

5.1 อภิปรายผล

จากผลการทดลองระบบ DRU INTRANET SEARCH และประเมินด้วยค่าเฉลี่ย NDCG โดยการนำเทคนิค Query dependent Ranking ผสมผสานกับความเชื่อมโยงของเอกสารในเครือข่ายและเรียงลำดับผลลัพธ์การค้นคืนด้วยเทคนิค Query Independent Ranking โดยมีผู้ร่วมทดสอบจำนวนทั้งสิ้น 35 คน และจำนวนคำสืบค้นทั้งสิ้น 105 คำค้น พบว่าตัวแบบ Index1 มีการเรียงลำดับผลลัพธ์การค้นคืนเอกสาร 20 อันดับแรกดีที่สุด โดยที่เอกสารที่มีความเชื่อมโยงภายในเครือข่ายและเอกสารที่เกี่ยวข้องกับหน่วยงานของผู้ใช้จะถูกดันผลลัพธ์การค้นคืนให้อยู่ในอันดับต้นๆ ได้

ค่าความถูกต้อง (Mean Average Precision : MAP) เป็นการหาค่าเฉลี่ยความถูกต้องของการสืบค้นแต่ละครั้งจากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยเพียงใด ใน

การศึกษาวิจัยนี้จะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2 หมายถึงผลลัพธ์การค้นคืนของเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึง ผลลัพธ์ของเอกสารมีความเกี่ยวข้องกับคำค้น พบว่าค่าเฉลี่ย MAP ของ Index1 ให้ค่าเฉลี่ยความถูกต้องของการค้นคืนเอกสารได้ถูกต้องมากกว่า Index0 ที่คะแนน 0.07849 โดยพิจารณาจากผลลัพธ์ค่าเฉลี่ย MAP ของทั้ง 2 ตัวแบบ

จากผลการทดลองสรุปได้ว่า ตัวแบบ Index1 ให้ประสิทธิผลของการค้นคืนเอกสารบนอินทราเน็ตดีกว่าตัวแบบ Index0 ทั้งในด้านมุมมองและความสอดคล้องกับพฤติกรรมของผู้ใช้ที่ให้ความสำคัญกับเอกสารที่เกี่ยวข้องกับหน่วยงานที่ตนเองสังกัดมากกว่าผลลัพธ์ของเอกสารอื่นๆ ดังนั้นตัวแบบที่ศึกษาวิจัยจึงเหมาะสำหรับการประเมินคุณภาพเอกสารจากความเชื่อมโยงของเอกสารในเครือข่าย (Location)

5.2 ปัญหาและอุปสรรค

การวิจัยในครั้งนี้ได้ใช้ตัวอย่างข้อมูลบนอินทราเน็ตมหาวิทยาลัยราชภัฏธนบุรีในการพัฒนาตัวแบบการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต พบปัญหาดังนี้

1. โครงสร้างเว็บไซต์ของแต่ละหน่วยงานมีโครงสร้างของหน้าเว็บที่แตกต่างกันโดยสิ้นเชิง
2. เอกสารมีความซ้ำซ้อนกันค่อนข้างสูง

ซึ่งปัญหาดังกล่าวมีผลทำให้เสียเวลาในการปรับแต่ง Crawler และคัดกรองเอกสารเพื่อเก็บลงในฐานข้อมูลและสร้างดัชนี

5.3 ข้อจำกัดของงานวิจัย

จากการทดลองการค้นคืนเอกสารในงานวิจัยนี้ มีข้อจำกัดบางประการดังนี้

1. ผลการทดลองของงานวิจัยนี้ ได้จากการประเมิน Judgment Score จากผู้ทดสอบทั้งสิ้น 35 คน และจำนวนคำสืบค้นทั้งสิ้น 105 คำค้น เท่านั้น เนื่องจากมีเวลาจำกัด
2. Crawler จะไม่อ่านเอกสารที่อยู่ในรูปแบบ .PDF แต่จะอ่านเฉพาะเอกสารที่อยู่ในรูปแบบ HTML เท่านั้น

5.4 ข้อเสนอแนะ

ในการพัฒนาระบบค้นคืนของเอกสารบนอินทราเน็ตในครั้งต่อไปควรมีการนำปัจจัยอื่นๆ ที่ไม่เกี่ยวข้องกับเอกสารมาเป็นส่วนประกอบในการสร้างดัชนีและเพิ่มประสิทธิภาพและประสิทธิผลในการเรียงลำดับผลลัพธ์การค้นคืน ดังนี้

1. ควรนำปัจจัยของความใหม่ของเอกสาร (Freshness Documents) เข้ามาเป็นส่วนประกอบในการสร้างตัวแบบเพื่อให้เอกสารที่ความใหม่กว่าแสดงผลลัพธ์การค้นคืนอยู่ในอันดับต้นๆ

2. ควรนำบันทึกการค้นหา (Query Log) เข้ามาเป็นส่วนประกอบในการสร้างตัวแบบเพื่อใช้เป็นปัจจัยในการช่วยเพิ่มประสิทธิภาพและความรวดเร็วในการแสดงผลลัพธ์การค้นคืน

3. การลดระยะเวลาและขนาดของการสร้างดัชนีจะช่วยทำให้การแสดงผลลัพธ์การค้นคืนมีความรวดเร็วมากยิ่งขึ้น

4. พัฒนา Crawler ที่สามารถเก็บรวบรวมเอกสารที่มีความแตกต่างกันของโครงสร้างเอกสารบนเว็บไซต์

5. เพิ่มการทดสอบทางสถิติเพื่อหานัยสำคัญของผลลัพธ์การค้นคืนของแต่ละตัวแบบ

บรรณานุกรม

บรรณานุกรม

ภาษาไทย

ขวัญเรือน โสอุบล และวรสิทธิ์ ชูชัยวัฒนา. (2557). ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม. *การประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ 10 (NCCIT 2014)*. (หน้า 150-155). ภูเก็ต.

ชูชาติ หฤไชยะศักดิ์. (2548). การพัฒนาโปรแกรมสำหรับค้นคืนสารสนเทศภาษาไทย. *National and Computer Technology Center (NECTEC),* 2537.

พิมพ์รำไพ เปรมสมิทธ์. (2545). โปรแกรมค้นหา (Search Engine): การสืบค้นและการประเมิน. *วารสารบรรณารักษศาสตร์, 22(2),* 1-14.

พิมลพรรณ ไชยนันท์. (2548). *บทบาทของเสิร์ชเอนจินที่มีบริการสืบค้นด้วยภาษาไทยในการคัดเลือกเนื้อหา* (วิทยานิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

วรสิทธิ์ ชูชัยวัฒนา. (2555). การปรับปรุงประสิทธิผลของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหา: แนวคิดและเทคนิค.*วารสารวิชาการสมาคมสถาบันอุดมศึกษานอกชนแห่งประเทศไทย, 3(1)*, 73-83.

วิภาพร กุศลชูกุล. (2552). *การจัดเก็บและค้นคืนกรณีทดสอบและผลของการทดสอบโดยใช้ผลป้อนกลับที่ตรงประเด็นจากผู้ใช้* (วิทยานิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

สุธาสินี ยกระดับชั้น และ มหศักดิ์ เกตุฉ่ำ. (2557). ระบบช่วยเหลือและแก้ไขปัญหาการใช้งานเว็บไซต์สำหรับบุคลากรสายวิชาการ มหาวิทยาลัยราชภัฏสวนสุนันทาโดยใช้เทคนิคการให้เหตุผลตามกรณีเป็นหลัก. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.

ศิรัตน์ ศิรนานนท์. (2549). *การค้นคืนสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้* (วิทยานิพนธ์ปริญญามหาบัณฑิต). กรุงเทพฯ: จุฬาลงกรณ์มหาวิทยาลัย.

ภาษาอังกฤษ

Chen, M., Hearst, M.A., Hong, J.I., and Lin, J. (1999). Cha-Cha: A System for Organizing Intranet Search Results. *USENIX Symposium on Internet Technologies and Systems*.

Gordon, M. and Pathak, P. (1999) Finding Information on the World Wide Web: The Retrieval Effectiveness of Search Engines. *Information Processing and Management, 35,* 141-180.

Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W. (2010). A framework for tag-based research paper recommender system: an IR approach. *In: Proceedings of the 24th international conference on Advanced information networking and applications (WAINA*). (pp. 103–108). Perth.

Hang L., Yunbo C., Jun X., Yunhua H., Shenjie L., and Dmitriy M. (2005). A new approach to intranet search based on information extraction. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (CIKM '05). ACM, New York, USA, 460-468.

Kaushik, P., Gaur, S., and Singh, M. (2014, 5-7 March 2014). Use of query logs for providing cache support to the search engine.*Paper presented at the 2014 International Conference on Computing for Sustainable Global Development (INDIACom)*. (pp. 819-824). New Delhi, India.

Kharazmi, S., Nejad, A. F., and Abolhassani, H. (2009, 9-12 Nov. 2009). Freshness of Web search engines: Improving performance of Web search engines using data mining techniques. *Paper presented at the 2009 International Conference for Internet Technology and Secured Transactions, (ICITST)*.(pp.1-7). London.

Lincheng, S. (2011, 27-29 May 2011). A large-scale full-text search engine using DotLuence.*Paper presented at the 2011 IEEE 3rd International Conference on Communication Software and Networks*. (pp.793-795). China.

Sato, N., Sakai, Y., and Uehara, M. (2005).The evaluations of FTF-IDF scoring for fresh information retrieval. *19th International Conference on Advanced Information Networking and Applications (AINA'05)* (pp.635-640). Sendai: Tohoku University

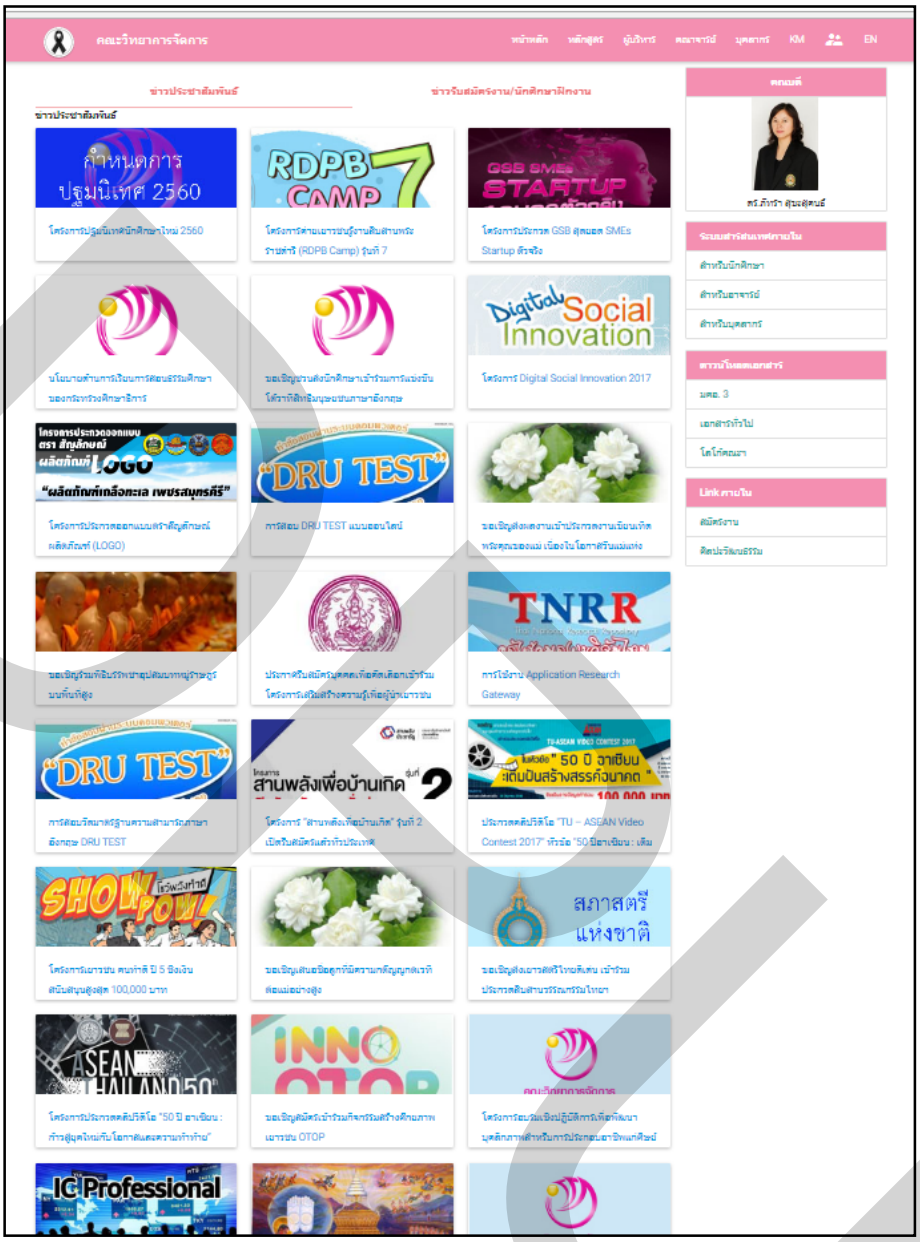
Shin, Y., Lim, J., and Park, J. (2012).Joint Optimization of Index Freshness and Coverage in Real-Time Search Engines.*IEEE Transactions on Knowledge and Data Engineering, 24(12),* 2203-2217.

Stenmark, D. (2006). What are you searching for? A content analysis of intranet search engine logs. *Proceedings of 29th Information Systems Research Seminar in Scandinavia*. Elsinore.

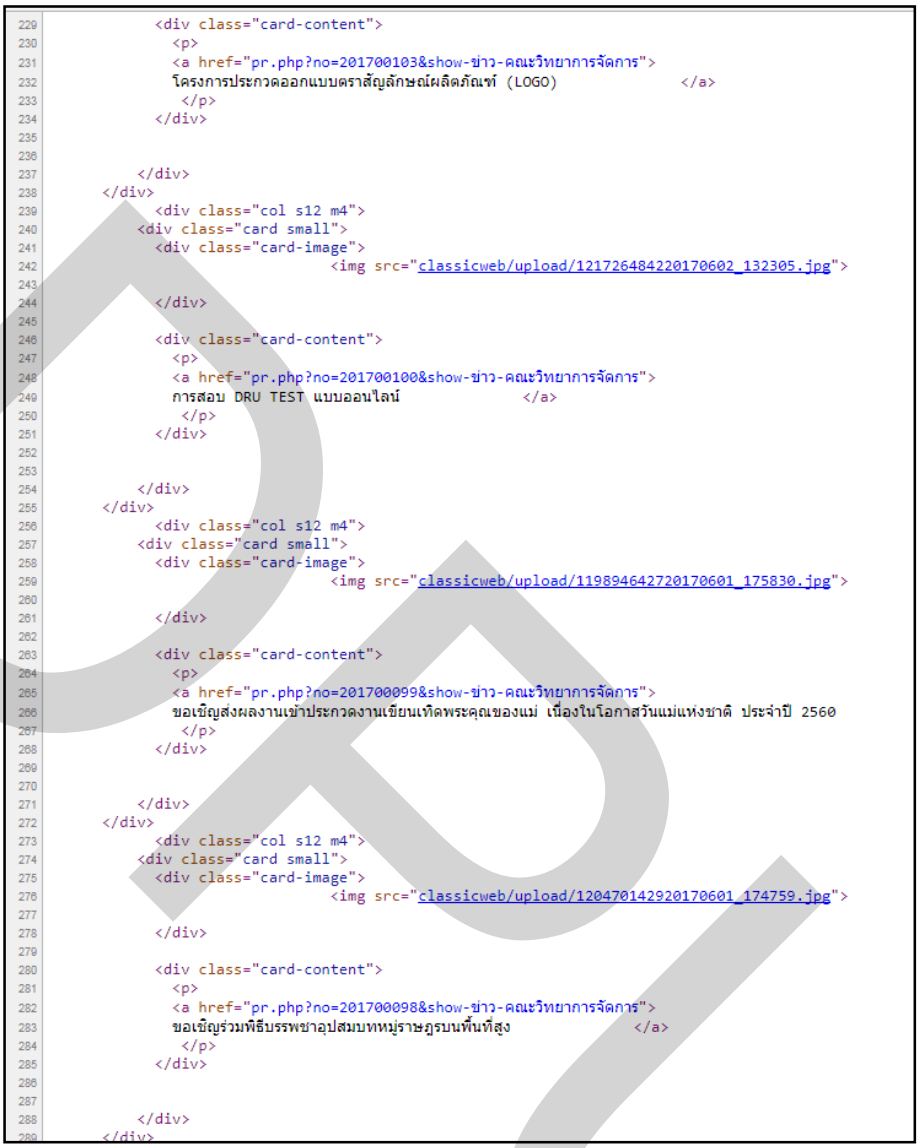
Vaughan, L. and Zhang, Y. (2007), Equal Representation by Search Engines? A Comparison of Websites across Countries and Domains. *Journal of Computer-Mediated Communication, 12*, 888–909.

ภาคผนวก

ภาคผนวก ก
ตัวอย่างการเตรียมคลังเอกสาร

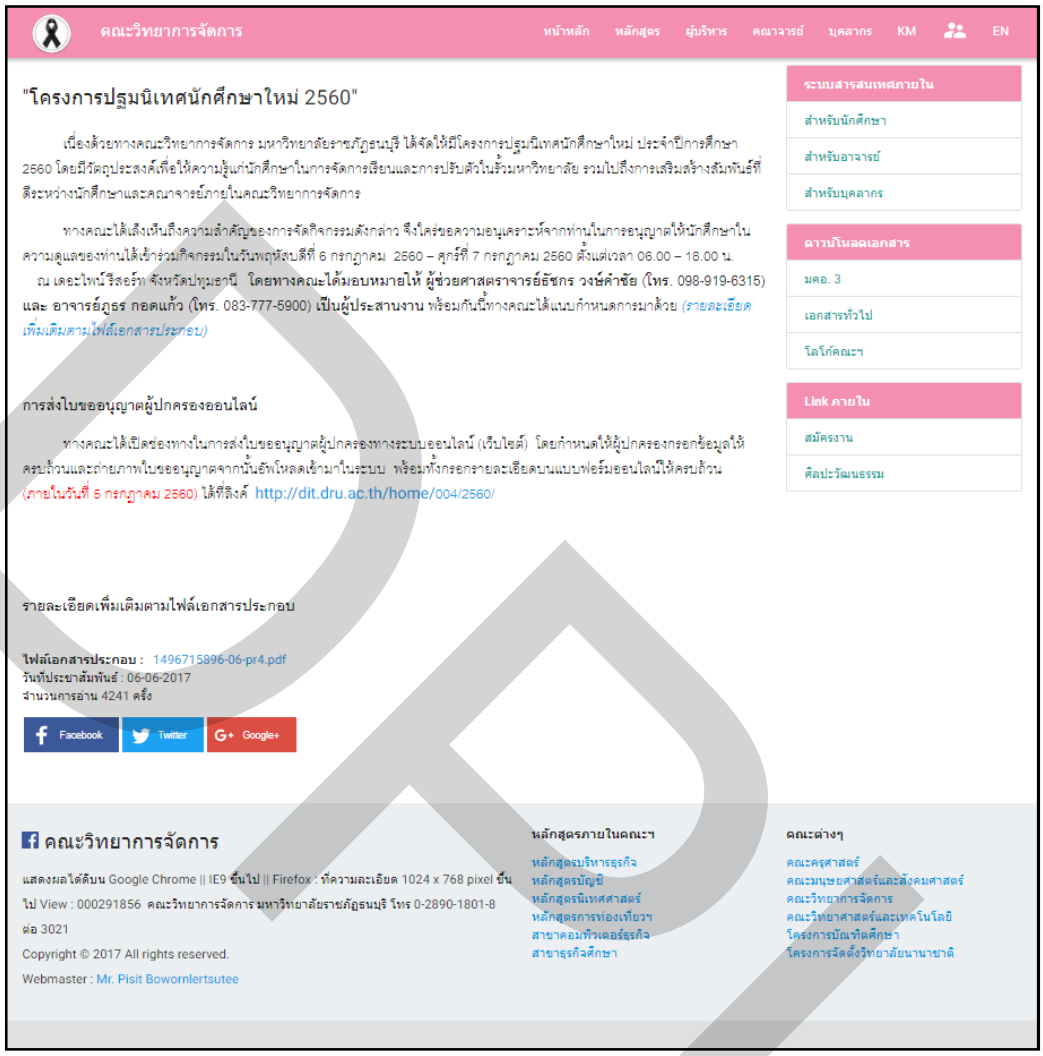


ภาพที่ ก.1 ตัวอย่างหน้ารายการเอกสารบนอินทราเน็ต มหาวิทยาลัยราชภัฏธนบุรี



```
            <div class="card-content">
              <p>
              <a href="pr.php?no=201700103&show-ข่าว-คณะวิทยาการจัดการ">
              โครงการประกวดออกแบบตราสัญลักษณ์ผลิตภัณฑ์ (LOGO)                </a>
               </p>
            </div>


        </div>
      </div>
            <div class="col s12 m4">
        <div class="card small">
          <div class="card-image">
                              <img src="classicweb/upload/12172648422017060 2_132305.jpg">

          </div>

          <div class="card-content">
            <p>
            <a href="pr.php?no=201700100&show-ข่าว-คณะวิทยาการจัดการ">
            การสอบ DRU TEST แบบออนไลน์                </a>
             </p>
          </div>


        </div>
      </div>
            <div class="col s12 m4">
        <div class="card small">
          <div class="card-image">
                              <img src="classicweb/upload/119894642720170601_175830.jpg">

          </div>

          <div class="card-content">
            <p>
            <a href="pr.php?no=201700099&show-ข่าว-คณะวิทยาการจัดการ">
            ขอเชิญส่งผลงานเข้าประกวดงานเขียนเทิดพระคุณของแม่ เนื่องในโอกาสวันแม่แห่งชาติ ประจำปี 2560
             </p>
          </div>


        </div>
      </div>
            <div class="col s12 m4">
        <div class="card small">
          <div class="card-image">
                              <img src="classicweb/upload/120470142920170601_174759.jpg">

          </div>

          <div class="card-content">
            <p>
            <a href="pr.php?no=201700098&show-ข่าว-คณะวิทยาการจัดการ">
            ขอเชิญร่วมพิธีบรรพชาอุปสมบทหมู่ราษฎรบนพื้นที่สูง                </a>
             </p>
          </div>


        </div>
      </div>
```

ภาพที่ ก.2 ตัวอย่างหน้าข้อมูลเอกสารบนอินทราเน็ต มหาวิทยาลัยราชภัฏธนบุรี



ภาพที่ ก.3 ตัวอย่างหน้ารายละเอียดเอกสาร

ภาพที่ ก.4 ตัวอย่างหน้าข้อมูลเอกสาร

ภาคผนวก ข

การออกแบบตารางฐานข้อมูล

ตารางที่ ข.1 ตาราง Article เก็บรายละเอียดข้อมูลเอกสาร

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1.	Doc_ID	ลำดับเอกสาร	Integer	PK
2.	Title	ชื่อเอกสาร	Text	
3.	Detail	รายละเอียดเอกสาร	Text	
4.	Category	หมวดหมู่ของเอกสาร	Text	
5.	Date_Public	วันเดือนปีที่เผยแพร่	Date	
6.	Doc_Url	URL ของเอกสาร	Text	
7.	Groupname	ระดับของหน่วยงาน	Text	
8.	Department	ชื่อหน่วยงาน	Text	
9.	IPDEPT	ไอพีของหน่วยงาน	Text	

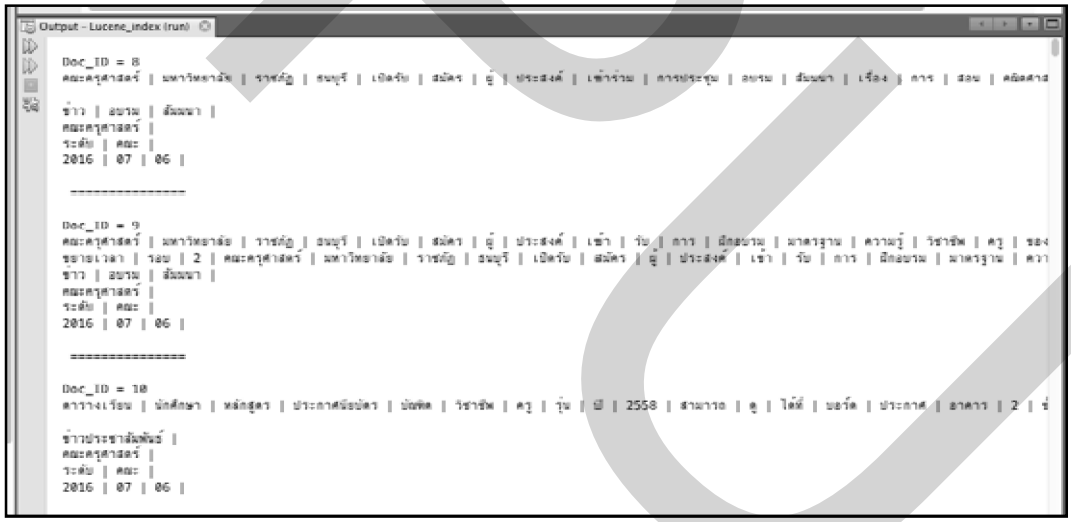
ตารางที่ ข.2 ตาราง JD_SCORE เก็บข้อมูล Judgment Score จากการประเมินระบบ

ลำดับ	แอทริบิวต์	ความหมาย	ชนิด	คีย์
1.	Jd_id	ลำดับ	Integer	PK
2.	Uid	รหัสผู้ประเมิน	Integer	
3.	Doc_id	ลำดับเอกสาร	Integer	FK
4.	Sim_score	คะแนน Similarity Score	Double	
5.	IR_model	รหัสตัวแบบดัชนี	Integer	
6.	Doc_no	ลำดับการประเมินเอกสารแต่ละครั้ง	Integer	
7.	Jd_score	คะแนน Judgment Score	Integer	
8.	Datesave	วันที่ประเมิน	Timestamp	

ภาคผนวก ค

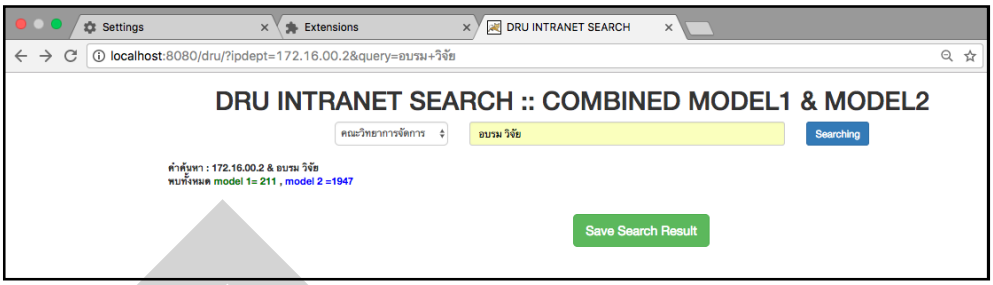
ตัวอย่างการแบ่งคำ (Tokenizing) และส่งไปทำการสร้างดัชนี

ภาพที่ ค.1 ตัวอย่างการแบ่งคำ (Tokenizing) และส่งไปทำการสร้างดัชนี

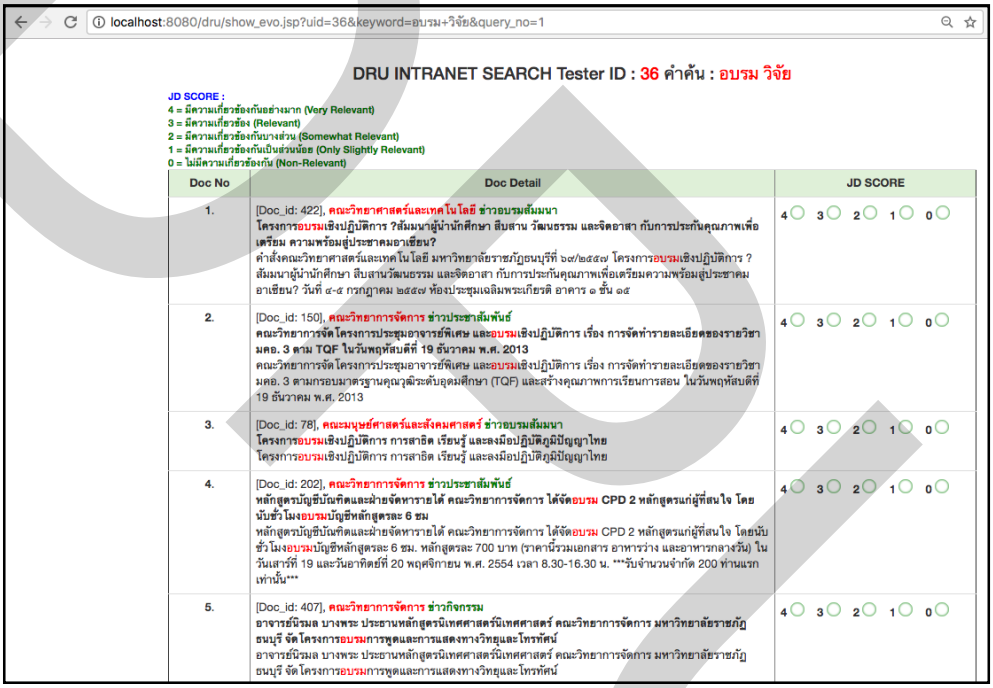


ภาพที่ ค.1 (ต่อ)

ภาคผนวก ง
ตัวอย่างหน้าจอระบบค้นคืนเอกสารบนอินทราเน็ต



ภาพที่ ง.1 หน้าจอสำหรับสืบค้น



ภาพที่ ง.2 ตัวอย่างหน้าจอแสดงผลลัพธ์การค้นคืน

ภาคผนวก จ
ตัวอย่างผลการประเมินจากผู้ทดสอบ

ตารางที่ จ.1 ตัวอย่างการประเมิน Judgment Score

Location	Title	Doc_Detail	SimScore	JudgeScore
สถาบันวิจัยและพัฒนา	ประชุมชี้แจงกรอบวิจัย ประจำปีงบประมาณ 2559	สำนักงานคณะกรรมการวิจัยแห่งชาติ ขอเชิญเข้าร่วมประชุมชี้แจงกรอบวิจัย ประจำปีงบประมาณ 2559 โดยผู้สนใจสามารถดูรายละเอียดเพิ่มเติมได้ที่ www.nrct.go.th และ www.nrms.go.th มีจำนวน 4 ครั้ง ดังนี้ 1) 6 กรกฎาคม 2559 ณ กรุงเทพมหานคร 2) 9 กรกฎาคม 2558 ณ จังหวัดเชียงราย 3) 21 กรกฎาคม 2558 ณ จังหวัดอุบลราชธานี 4) 24 กรกฎาคม 2558 ณ จังหวัดสุราษฎร์ธานี หรือสอบถามเพิ่มเติมได้ที่ 0-2579-2284	0.19082718	1
คณะวิทยาการจัดการ	หลักสูตรบัญชีบัณฑิตและฝ่ายจัดหารายได้ คณะวิทยาการจัดการ ได้จัดอบรม CPD 2 หลักสูตรแก่ผู้ที่สนใจ โดยนับชั่วโมงอบรมบัญชีหลักสูตรละ 6 ชม	หลักสูตรบัญชีบัณฑิตและฝ่ายจัดหารายได้ คณะวิทยาการจัดการ ได้จัดอบรม CPD 2 หลักสูตรแก่ผู้ที่สนใจ โดยนับชั่วโมงอบรมบัญชีหลักสูตรละ 6 ชม. หลักสูตรละ 700 บาท (ราคานี้รวมเอกสาร อาหารว่าง และอาหารกลางวัน) ในวันเสาร์ที่ 19 และวันอาทิตย์ที่ 20 พฤศจิกายน พ.ศ. 2554 เวลา 8.30-16.30 น. ***รับจำนวนจำกัด 200 ท่านแรกเท่านั้น***	0.32434222	2
คณะวิทยาศาสตร์และเทคโนโลยี	โครงการอบรมเชิงปฏิบัติการ ?การจัดการความรู้และการปฏิบัติที่เป็นเลิศ (Knowledge mannewagement and Best Practice)?	กำหนดการ ?การจัดการความรู้และแนวทางปฏิบัติที่ดี (Knowledge mannewagement and Best Practice)? คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏธนบุรี วันพุธที่ 2 กรกฎาคม พ.ศ. 2557 ณ ห้อง 1712 อาคาร 1 คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏธนบุรี	0.08101243	2
คณะวิทยาศาสตร์และเทคโนโลยี	โครงการอบรมเทคโนโลยีมือถือ	โครงการอบรมเทคโนโลยีมือถือ วันที่ 20 ? 21 กุมภาพันธ์ พ.ศ. 2557 ณ. ห้องปฏิบัติการอินเทอร์เน็ต อาคาร 2 ชั้น 11 มหาวิทยาลัยราชภัฏธนบุรี มหาวิทยาลัยราชภัฏธนบุรี เป็น Advance Course android on Google Maps API version 2 ความรู้พื้นฐานที่ควรมีคือ เบสิกแอนดรอนด์ บน android Studio	0.14439006	2
คณะวิทยาการจัดการ	โครงการอบรมเชิงปฏิบัติการเรื่องการพัฒนาผลงานทางวิชาการของอาจารย์คณะวิทยาการจัดการ	ขอเชิญชวนคณาจารย์คณะวิทยาการจัดการทุกท่าน เข้าร่วมโครงการอบรมเชิงปฏิบัติการเรื่อง การพัฒนาผลงานทางวิชาการ บรรยายโดย รศ.วรารัตน์ เขียวไพรี ในวันศุกร์ที่ 7 มีนาคม 2557 เวลา 9:00 - 12:00 น. ณ ห้องประชุมศรีเจริญ มหาวิทยาลัยราชภัฏธนบุรี	0.29669234	3

ตารางที่ จ.1 (ต่อ)

Location	Title	Doc_Detail	SimScore	JudgeScore
คณะวิทยาการจัดการ	โครงการอบรม เรื่อง ?นักธุรกิจรุ่นใหม่ เริ่มต้นธุรกิจอย่างไรให้สำเร็จ	เมื่อวันจันทร์ที่ 20 กุมภาพันธ์ 2560 เวลา 13.00 ? 16.00 น. สาขาการจัดการ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ ได้จัดโครงการอบรม เรื่อง ? นักธุรกิจรุ่นใหม่ เริ่มต้นธุรกิจอย่างไรให้สำเร็จ? วิทยากร : คุณกันตินันท์ ศรีสุริยประภา โดยมีอาจารย์ธัชกร ภัทรพันปี รองผู้อำนวยการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ เป็นประธานในพิธี ซึ่งภายในงานมีอาจารย์และนักศึกษาให้ความสนใจเข้าร่วมโครงการเป็นจำนวนมาก ผู้จัดงาน : อาจารย์พัทธนันท์ มั่งมี อาจารย์สาขาการจัดการ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ	0.2838458	2
คณะวิทยาศาสตร์และเทคโนโลยี	โครงการอบรมการใช้ชีวิตอย่างมีคุณภาพ สำหรับนักศึกษาใหม่ ประจำปีการศึกษา 2557	คำสั่งคณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏธนบุรีที่ ๖๙/๒๕๕๗ โครงการอบรมเชิงปฏิบัติการ ? สัมมนาผู้นำนักศึกษา สืบสานวัฒนธรรม และจิตอาสากับการประกันคุณภาพเพื่อเตรียมความพร้อมสู่ประชาคมอาเซียน? วันที่ ๔-๕ กรกฎาคม ๒๕๕๗ ห้องประชุมเฉลิมพระเกียรติ อาคาร ๑ ชั้น ๑๕	0.13666023	2
คณะวิทยาการจัดการ	โครงการอบรมและการจัดการแข่งขันผสมเครื่องดื่ม ครั้งที่ 12	คณะวิทยาการจัดการ จัดโครงการอบรมและการจัดการแข่งขันผสมเครื่องดื่ม ครั้งที่ 12 เพื่อสู่ความเป็นเลิศทางอุตสาหกรรมบริการ และการท่องเที่ยว ในวันศุกร์ที่ 13 มกราคม 2560 ณ โรงอาหารอาคาร 2 มีวัตถุประสงค์ 1. เพื่อให้เยาวชนและประชาชนทั่วไปที่สนใจได้ร่วม กิจกรรรมในรูปแบบของการจัดงานการบรรยายวิชาการ การแข่งขัน และการประกวดต่างๆ 2.เพื่อให้คณะได้มีส่วนร่วมในกิจกรรมต่างๆ กับ ชุมชนและประชาชนทั่วไป 3. เพื่อบริการวิชาการแก่ประชาชนทั่วไปในเรื่องที่น่าสนใจต่างๆ 4. เพื่อบูรณาการกับการเรียนการสอนรายวิชา บาร์และเครื่องดื่ม 5. เป็นการพัฒนาทักษะทางวิชาชีพเพื่อรองรับการเข้าสู่ประชาคมอาเซียน 6.เพื่อเป็นการประชาสัมพันธ์หลักสูตรอุตสากรรม บริการและการท่องเที่ยว โดยมีผู้ช่วยศาสตราจารย์ ดร.ยุวลักษณ์ เวชวิทยาขลัง อธิการบดีกล่าวเปิดงาน และมีอาจารย์รวีวรรณ วิโรจน์วรรณ และอาจารย์นุชรา แสวงสุข ผู้รับผิดชอบโครงการ	0.27742258	2

ตารางที่ จ.1 (ต่อ)

Location	Title	Doc_Detail	SimScore	JudgeScore
	โครงการอบรมนักศึกษา เรื่อง ?การจัดทำธุรกิจจำลอง?	"เมื่อวันอังคาร ที่ 20 ธันวาคม 2559 ที่ผ่านมา คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี จัดโครงการอบรมนักศึกษาเรื่อง ?การจัดทำธุรกิจจำลอง? โดยมี อาจารย์ ดร. จิระพงค์ เรืองกุน ประธานหลักสูตรบริหารธุรกิจ เป็นประธานกล่าวเปิดงาน ซึ่งมีอาจารย์ปฐมพงษ์ บำเริบ เป็นวิทยากรบรรยายให้ความรู้เกี่ยวกับการทำธุรกิจจำลองในครั้งนี้ ณ ห้อง 9214 ณ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ"	0.29026908	2
คณะวิทยาการจัดการ	เมื่อวันพุธ ที่ 20 สิงหาคม 2557 ที่ผ่านมา อาจารย์ ดร. ปรียาวดี ผลเอนก คณะวิทยาการจัดการ จัดอบรม ?เทคนิคการเขียนบทความวิจัยและบทความวิชาการเพื่อตีพิมพ์ในวารสารวิชาการ โดยมี ผศ.ดร.กาญจนา บุญภักดิ์ รองค	เมื่อวันพุธ ที่ 20 สิงหาคม 2557 ที่ผ่านมา อาจารย์ ดร. ปรียาวดี ผลเอนก คณะวิทยาการจัดการ จัดอบรม ?เทคนิคการเขียนบทความวิจัยและบทความวิชาการเพื่อตีพิมพ์ในวารสารวิชาการ โดยมี ผศ.ดร.กาญจนา บุญภักดิ์ รองคณบดีกำกับดูแลงานด้านวิจัยและบริหารองค์ความรู้ คณะครุศาสตร์อุตสาหกรรม สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง เป็นวิทยากรบรรยาย เพื่อให้คณาจารย์และบุคลากรของคณะมีการจัดการความรู้อย่างเป็นระบบ มีการถ่ายทอดความรู้และเรียนรู้ร่วมกัน ช่วยเพิ่มศักยภาพในการวิจัยและการสอน ณ ห้องประชุมคณะวิทยาการจัดการ อาคาร 3 ชั้น 4	0.10060328	3
คณะวิทยาการจัดการ	คณะวิทยาการจัดการ ได้จัดโครงการอบรมเชิงปฏิบัติการ	"เมื่อวันจันทร์ที่ 30 มกราคม 2560 เวลา 09.00 -16.00 น. คณะวิทยาการจัดการ ได้จัดโครงการอบรมเชิงปฏิบัติการ ?การจัดการความรู้ธุรกิจระหว่างประเทศต่อการจับกลุ่มเศรษฐกิจอาเซียน? ณ ห้องประชุมอาคาร 2 ชั้น 8 โดยมี อาจารย์ ดร.ภัทรา สุขะสุคนธ์ คณบดี เป็นประธานในพิธีเปิด ซึ่งมีอาจารย์และนักศึกษาให้ความสนใจเข้าร่วมโครงการดังกล่าวเป็นจำนวนมาก"	0.322925	2
คณะวิทยาการจัดการ	เมื่อวันพุธ ที่ 20 สิงหาคม 2557 ที่ผ่านมา อาจารย์ ดร. ปรียาวดี ผลเอนก คณะวิทยาการจัดการ จัดอบรม ?เทคนิคการเขียนบทความวิจัยและบทความวิชาการเพื่อตีพิมพ์ในวารสารวิชาการ โดยมี ผศ.ดร.กาญจนา บุญภักดิ์ รองค	เมื่อวันพุธ ที่ 20 สิงหาคม 2557 ที่ผ่านมา อาจารย์ ดร. ปรียาวดี ผลเอนก คณะวิทยาการจัดการ จัดอบรม ?เทคนิคการเขียนบทความวิจัยและบทความวิชาการเพื่อตีพิมพ์ในวารสารวิชาการ โดยมี ผศ.ดร.กาญจนา บุญภักดิ์ รองคณบดีกำกับดูแลงานด้านวิจัยและบริหารองค์ความรู้ คณะครุศาสตร์อุตสาหกรรม สถาบันเทคโนโลยี พระจอมเกล้าเจ้าคุณทหารลาดกระบัง เป็นวิทยากรบรรยาย เพื่อให้คณาจารย์และบุคลากรของคณะมีการจัดการความรู้อย่างเป็นระบบ มีการถ่ายทอดความรู้และเรียนรู้ร่วมกัน ช่วยเพิ่มศักยภาพในการวิจัยและการสอน ณ ห้องประชุมคณะวิทยาการจัดการ อาคาร 3 ชั้น 4	0.6525013	4

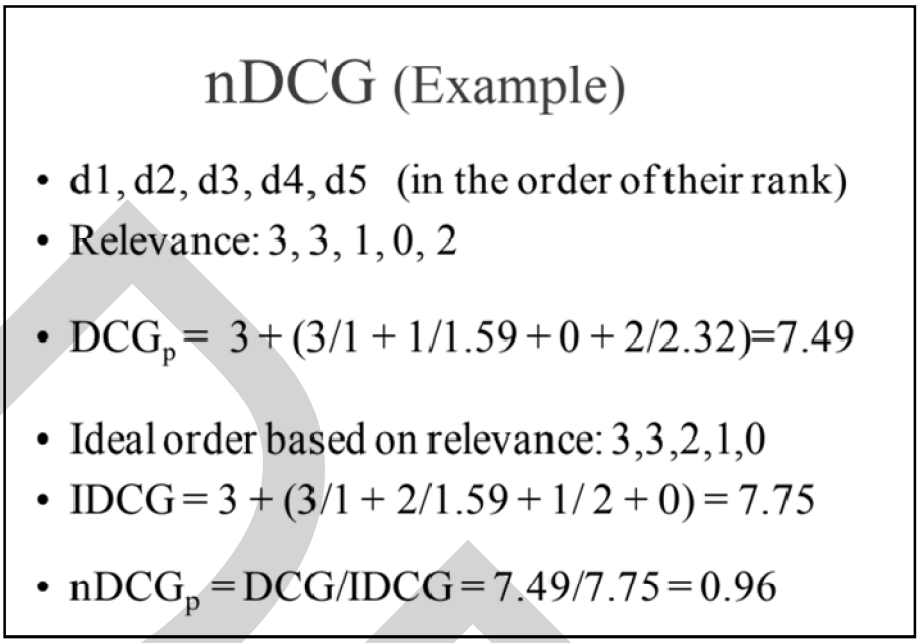
ตารางที่ จ.1 (ต่อ)

Location	Title	Doc_Detail	SimScore	JudgeScore
คณะวิทยาการจัดการ	โครงการอบรม เรื่อง ? วัยรุ่นไทยรู้และเข้าใจความปลอดภัยในเพศสัมพันธ์ (Save Sex)	"เมื่อวันจันทร์ที่ 20 กุมภาพันธ์ 2560 เวลา 8.30 ? 12.00 น. สาขาการจัดการ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ ได้จัดโครงการอบรม เรื่อง ? วัยรุ่นไทยรู้และเข้าใจความปลอดภัยในเพศสัมพันธ์ (Save Sex)? วิทยากร : คุณจิตติมา ภาณุเตชะ ผอ.มูลนิธิสร้างความเข้าใจเรื่องสุขภาพหญิง(สคส.) โดยมีอาจารย์ธัชกร ภัทรพันปี รองผู้อำนวยการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ เป็นประธานในพิธี ซึ่งภายในงานมีอาจารย์และนักศึกษาให้ความสนใจเข้าร่วมโครงการเป็นจำนวนมาก ผู้จัดงาน : อาจารย์พัทธนันท์ มั่งมี อาจารย์ สาขาการจัดการ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี สมุทรปราการ "	0.2838458	2
คณะวิทยาการจัดการ	เชิญนักศึกษาที่สนใจและต้องฝึกงาน บริษัท เอ็มไอแ◌ิอดไวส์ซอรี่ เป็นบริษัทวิจัยทางด้านการตลาด เปิดรับนศ. ฝึกงาน	เชิญนักศึกษาที่สนใจและต้องฝึกงาน บริษัท เอ็มไอแ◌ิอดไวส์ซอรี่ เป็นบริษัทวิจัยทางด้านการตลาด เปิดรับนศ. ฝึกงาน ทางบริษัทได้มีโครงการรับนักศึกษาฝึกงาน ในระหว่างภาคเรียน เพื่อเสริมทักษะ ความชำนาญ และประสบการณ์ในการทำงาน ให้สามารถนำความรู้ทางวิชาการมาประยุกต์ใช้ในงานได้อย่างมีประสิทธิภาพ โครงการนี้เริ่มตั้งแต่เดือน ตุลาคม 2554 ลักษณะงาน : การประมวลผลและตรวจสอบข้อมูล เช่น การลงรหัสข้อมูล การคีย์ข้อมูล การทำแผนภูมิ เป็นต้น คุณสมบัติผู้เข้าร่วมโครงการโดยย่อ คือ เป็น นศ.อุดมศึกษา ปี 3-4 มีทัศนคติ ความกระตือรือร้น และความรับผิดชอบที่ดีในการทำงาน และมีค่าเบี้ยเลี้ยงให้ด้วย	0.32091364	3
คณะวิทยาการจัดการ	โครงการอบรม เรื่อง ? การบริหารทรัพยากรมนุษย์ ยุค 4.0	คณะวิทยาการจัดการ จัดโครงการอบรม เรื่อง ?การบริหารทรัพยากรมนุษย์ ยุค 4.0? ในวันพุธที่ 11 มกราคม 2560 เวลา 12.00-17.00 น. ณ ห้องประชุมศรีเจริญ อาคาร 3 ชั้น 5 มีวัตถุประสงค์ 1.เพื่อแลกเปลี่ยนประสบการณ์ด้านการบริหาร ทรัพยากรมนุษย์ ให้นักศึกษา เยาวชน และบุคคลทั่วไป ได้รับความรู้จากวิทยากรที่มีความเชี่ยวชาญ 2.2 เปิดโอกาสในการสร้าง เครือข่ายความร่วมมือระหว่างหลักสูตรบริหารธุรกิจกับองค์กรภายนอก โดยมีอาจารย์จิระพงค์ เรืองกุน อาจารย์ประพัฒน์ เขียวประภัสสร อาจารย์เฉลิมชัย ศุขไพบูลย์ รองศาสตราจารย์วรารัตน์ เขียวไพรี และอาจารย์วสุธิดา นักเกษม นำนักศึกษาเข้า ร่วมกิจกรรมในครั้งนี้	0.3100785	2

ตารางที่ จ.1 (ต่อ)

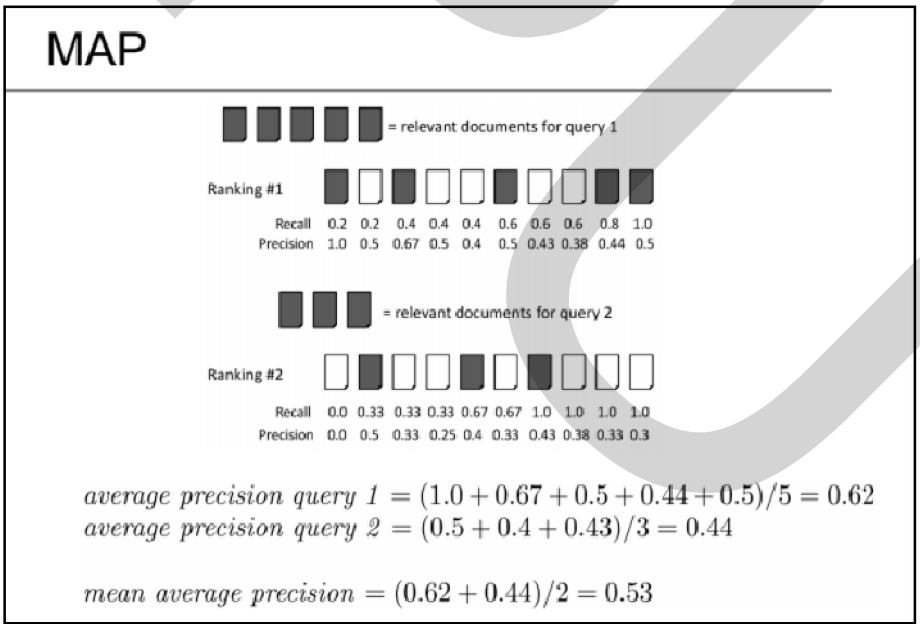
Location	Title	Doc_Detail	SimScore	JudgeScore
คณะวิทยาการจัดการ	โครงการอบรมเชิงปฏิบัติการ เรื่องการเขียนบทความวิชาการ เพื่อตีพิมพ์ในวารสารระดับชาติและนานาชาติ	คณะวิทยาการจัดการ จัดโครงการอบรมเชิงปฏิบัติการเรื่องการเขียนบทความวิชาการ เพื่อตีพิมพ์ในวารสารระดับชาติและนานาชาติ ระหว่างวันที่ 14-16 พฤศจิกายน 2559 ซึ่งมีอาจารย์ สุทธิชัย ฉายเพชรากร รองอธิการบดีมหาวิทยาลัยราชภัฏธนบุรี เป็นประธานกล่าวเปิดการอบรม โดยรองศาสตราจารย์ ดร. พยอม วงศ์สารศรี เป็นวิทยากรบรรยายในครั้งนี้ เพื่อให้คณาจารย์และบุคลากรได้รับความรู้เพื่อเป็นแนวทางในการเขียนบทความวิชาการที่สามารถนำไปสู่การตีพิมพ์ในวารสารระดับชาติและนานาชาติและได้เผยแพร่ต่สาธารณชน จัดขึ้น ณ ห้องประชุมเฉลิมพระเกียรติ อาคาร 2 ชั้น 8 มหาวิทยาลัยราชภัฏธนบุรี	0.2838458	2
คณะวิทยาการจัดการ	ขอเชิญส่งบทความวิจัยเข้าประกวดในโครงการประกวดบทความวิจัยด้านพัฒนบริหารศาสตร์ ประจำปี 2013	ขอเชิญส่งบทความวิจัยเข้าประกวดในโครงการประกวดบทความวิจัยด้านพัฒนบริหารศาสตร์ ประจำปี 2013 ระดับชาติ (อาจารย์ นักวิชาการ และบุคคลทั่วไป) ระดับบัณฑิตศึกษา (นักศึกษา ปริญญาเอก และโท) 10 สาขาวิชา ระดับชาติ แต่ละสาขาวิชา ดีเยี่ยม เกียรติบัตร พร้อมเงินรางวัล สาขาวิชาละ 30,000 บาท ดี เกียรติบัตร พร้อมเงินรางวัล สาขาวิชาละ 15,000 บาท ชมเชย เกียรติบัตร รางวัลระดับบัณฑิตศึกษา แต่ละสาขาวิชา ระดับปริญญาเอก ดีเยี่ยม เกียรติบัตรพร้อมเงินรางวัล สาขาวิชาละ 15,000 บาท ดี เกียรติบัตรพร้อมเงินรางวัล สาขาวิชาละ 7,500 บาท ชมเชย เกียรติบัตร ระดับปริญญาโท ดีเยี่ยม เกียรติบัตรพร้อมเงินรางวัล สาขาวิชาละ 10,000 บาท ดี เกียรติบัตรพร้อมเงินรางวัล สาขาวิชาละ 5,000 บาท ชมเชย เกียรติบัตร สามารถ ดาวน์โหลดจากเอกสารแนบ และได้ที่ www.nida.ac.th/researchaward2013 โดยส่งบทความวิจัยฉบับสมบูรณ์ทางไปรษณีย์ หรือ email : nidaresearch@hotmail.com ภายในวันที่ 7 มกราคม 2012	0.42589822	3
คณะวิทยาศาสตร์และเทคโนโลยี	โครงการอบรมเชิงปฏิบัติการเรื่องการประกอบอาชีพเสริมด้วยคอมพิวเตอร์ จากหน้าจอลงสู่ผลิตภัณฑ์ด้วย Photoshop	มื่อวันเสาร์ที่ 13 ธันวาคม 2557 สาขาวิชาเทค โนโลยีสารสนเทศ จัดครงการอบรมเชิงปฏิบัติการ เรื่องการประกอบอาชีพเสริมด้วยคอมพิวเตอร์ด้วย โปรแกรม Photoshop เบื้องต้น ให้กับบุคคลทั่ว ไปที่สนใจ	0.08101243	1

ภาคผนวก ฉ
ตัวอย่างการคำนวณค่าเฉลี่ย NDCG
ค่า Mean Average Precision : MAP



ภาพที่ ฉ.1 ตัวอย่างการคำนวณค่าเฉลี่ย NDCG

แหล่งที่มา : http://people.cs.georgetown.edu/~nazli/classes/ir-Slides/Evaluation-13.pdf



ภาพที่ ฉ.2 ตัวอย่างการคำนวณค่า MAP

แหล่งที่มา : http://www.cs.cornell.edu/courses/cs4300/2013fa/lectures/metrics-2-4pp.pdf

ภาคผนวก ช

บทความการประชุมวิชาการ

โครงการประชุมวิชาการบัณฑิตศึกษาระดับชาติและนานาชาติ ครั้งที่ 6

เรื่อง“สหวิทยาการสร้างสรรค์เพื่อการพัฒนาที่ยั่งยืน”

และโครงการประชุมวิชาการระดับชาติและนานาชาติ“สหวิทยาการสร้างสรรค์เพื่อการพัฒนาที่ยั่งยืน”

มหาวิทยาลัยศิลปากร

ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืน ของเอกสารบนอินทราเน็ต : กรณีศึกษา มหาวิทยาลัยราชภัฏธนบุรี

A MODEL FOR RANKING SEARCH RESULT OF DOCUMENTS ON THE INTRANET. : CASE STUDY OF DHONBURI RAJAPHAT UNIVERSITY.

พิศิษฐ์ บวรเลิศสุธี (Pisit Bowornlertsutee) และ วรสิทธิ์ ชูชัยวัฒนา (Worasit Choochaiwattana)
หลักสูตรวิศวกรรมเว็บและการพัฒนาแอปพลิเคชันบนอุปกรณ์พกพา
วิทยาลัยครีเอทีฟดีไซน์ แอนด์ เอ็นเตอร์เทนเมนต์เทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิตย์
pisit.b@dru.ac.th, worasit.cha@dpu.ac.th

บทคัดย่อ

Query Dependent Ranking หรือ Similarity Ranking เป็นเทคนิคสำหรับเรียงลำดับผลลัพธ์การค้นคืนโดยการเปรียบเทียบคำค้นและดัชนีของเอกสาร โดยไม่ได้พิจารณาถึงปัจจัยอื่นๆ เช่น คุณภาพของเอกสาร (Document Quality) ความเชื่อมโยงของเอกสารภายในเครือข่าย (Location) เป็นต้น ซึ่งแตกต่างจากเทคนิค Query Independent Ranking หรือ Static Ranking จะนำคุณภาพและส่วนประกอบอื่นๆ ของเอกสารเข้ามามีส่วนในการพิจารณาเรียงลำดับผลลัพธ์การค้นคืน ในงานวิจัยนี้ได้นำเสนอตัวแบบการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต ระหว่างเทคนิค Query Dependent Ranking และเทคนิค Query Independent Ranking โดยการนำความเชื่อมโยงของเอกสารภายในเครือข่ายเป็นส่วนประกอบในการเรียงลำดับผลลัพธ์การค้นคืน ซึ่งเป็นการใช้ Similarity Feature ประกอบด้วย ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) และความเชื่อมโยงของเอกสารในเครือข่าย (Location) มาใช้ในการสร้างดัชนี (Index) จากผลการทดลองเบื้องต้นด้วยผู้ทดสอบจำนวน 35 ผู้ทดสอบ และคำค้นทั้งหมดจำนวน 105 คำสืบค้น พบว่า การเรียงลำดับโดยใช้เทคนิค Query Independent Ranking ผสมผสานกับความเชื่อมโยงของเอกสารภายในเครือข่ายให้ผลลัพธ์การค้นคืนเอกสาร 20 อันดับแรกดีกว่าการเรียงลำดับผลลัพธ์การค้นคืนโดยใช้เทคนิค Query Dependent Ranking เพียงอย่างเดียว ซึ่งสรุปได้ว่า ผู้ใช้ให้ความสำคัญกับผลลัพธ์การค้นคืนของเอกสารที่อยู่ใกล้ตัวและมีความเกี่ยวข้องกับหน่วยงานที่สังกัดมากกว่าเอกสารทั่วๆ ไป

คำสำคัญ : การเรียงลำดับผลลัพธ์การค้นคืน, ระบบค้นคืนเอกสาร, อินทราเน็ต, ตัวแบบ

Abstract

Query dependent ranking or Similarity Ranking is a technique for Re-Ranking Model search results by comparing keywords and indexes of documents without considering other factors such as document quality, location etc. It is different from query independent ranking or static ranking, it brings quality and other components. The document takes part in the re-ranking of retrieval results. In this paper, we present a model for re-ranking of document retrieval on an intranet. There is a combination of query dependent ranking techniques and the query independent ranking technique, which links the documents within the network. This will use a similarity feature, including title, detail, department name, and Location to index. There are 35 testers and 105 keywords searched. re-ranking using the query independent ranking technique combined with the linking of documents within the location. The top 20 search results are better than queries using only the query dependent ranking technique. It can be concludes that users location more importance on the search results of documents that are closer to oneself and more relevant to the agencies than the general documents.

Key Word (s): Ranking, Search Result, Intranet Search, Model

1. บทนำ

ในปัจจุบันเทคโนโลยีสารสนเทศและอินเทอร์เน็ตถูกพัฒนาไปอย่างรวดเร็ว ทำให้ปริมาณสารสนเทศต่างๆ ถูกเผยแพร่บนระบบอินเทอร์เน็ต (WWW) และบนระบบอินทราเน็ต (Intranet) ภายในองค์กรต่างๆ อย่างมากมายมหาศาล ซึ่งข้อมูลส่วนใหญ่เป็นข้อมูลสารสนเทศที่มีความสำคัญในด้านต่าง ๆ เช่น ข่าวสาร การศึกษา การวิจัย เป็นต้น การพัฒนาระบบสืบค้นภายในองค์กรเป็นที่ต้องการมากยิ่งขึ้น ทั้งนี้เพื่ออำนวยความสะดวกในการสืบค้นข้อมูลที่เป็นประโยชน์สำหรับใช้ในการปฏิบัติงานต่างๆ และผลจากการที่ข้อมูลมีปริมาณเพิ่มมากขึ้นอยู่ตลอดเวลาจึงส่งผลให้การสืบค้นข้อมูลเกิดปัญหาและใช้เวลาในการคัดกรองข้อมูลที่ตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น และผู้ใช้ส่วนใหญ่ขาดความรู้ความเข้าใจเกี่ยวกับการใช้คำค้น (Query) ที่เหมาะสมสำหรับการสืบค้น ซึ่งมีผลทำให้ระบบไม่สามารถค้นคืนข้อมูลที่ตรงกับความต้องการของผู้ใช้ได้อย่างแท้จริง เนื่องจากระบบจะแสดงผลลัพธ์การค้นคืน (Result) เฉพาะเอกสารที่ตรงกับคำค้นเท่านั้น(ศิรัตน์ ศิรนานนท์ : 2549, น.1) และอีกปัญหาที่พบบ่อยในระบบสืบค้นข้อมูลภายในมหาวิทยาลัย คือ เอกสารที่มีประกาศภายในมหาวิทยาลัย เอกสารจะมีการเปลี่ยนแปลงหลายครั้งและประชาสัมพันธ์บนเว็บไซต์ทั้งหมด เมื่อทำการสืบค้นผลลัพธ์ของเอกสารที่ได้จะไม่เรียงลำดับผลลัพธ์การค้นคืนตามความต้องการของผู้ใช้ ซึ่งในบางครั้งทำให้ผู้ใช้เกิดความสับสนในการเลือกดูผลลัพธ์เพื่อนำไปใช้ประโยชน์ ปัญหาต่อมาที่พบได้บ่อยในการพัฒนาระบบสืบค้นเกิดจากการสร้างดัชนี (Index) ที่ไม่มีคุณภาพ ขาดเทคนิคการวิเคราะห์องค์ประกอบที่เกี่ยวข้องกับสถาปัตยกรรมและโครงการสร้างของเว็บเพจ (Web Page) และส่วนประกอบอื่นๆ ที่ปรากฏอยู่นอกเหนือเอกสารมาเป็นส่วนประกอบในการสร้างดัชนี (วรสิทธิ์ ชูชัยวัฒนา : 2555) ซึ่งปัญหาดังกล่าวก็ส่งผลให้การเรียงลำดับผลลัพธ์การค้นคืนไม่ตรงกับความต้องการของผู้ใช้เช่นเดียวกัน

การพัฒนาระบบสืบค้น (Search Engine) ภายในองค์กรส่วนใหญ่นิยมพัฒนาโดยใช้เทคนิคการทำ Full Text Search หรือ (Full Text Indexing) คือการสืบค้นจากคำที่มีทั้งหมดในเอกสาร โดยจะนำคำค้น (Query) ไปเปรียบเทียบกับเอกสารทั้งหมดที่มีอยู่ในฐานข้อมูล (Database) ซึ่งเป็นที่นิยมและมีการใช้งานในฐานข้อมูลบรรณานุกรมออนไลน์มาตั้งแต่ปี ค.ศ.1990 เช่นเว็บสืบค้นข้อมูล AltaVista ใช้เทคนิคการสืบค้นข้อมูลแบบ Full Text Search โดยการสร้างดัชนีจากส่วนหนึ่งของเอกสารบนหน้าเว็บไซต์ที่มีอยู่ทั้งหมดในฐานข้อมูล เมื่อผู้ใช้ทำการสืบค้นระบบก็จะทำการนำคำค้นไปเปรียบเทียบกับคำทั้งหมดที่มีอยู่ในคลังเอกสาร (Document Corpus) และแสดงผลลัพธ์การค้นคืนออกมา ซึ่งเกิดปัญหา คือ ผู้ใช้จะเสียเวลาในการเข้าถึงข้อมูลค่อนข้างมาก ขาดประสิทธิภาพและไม่ตรงกับความต้องการของผู้ใช้ ต่อมาจึงมีผู้คิดค้นวิธีการนำเสนอผลลัพธ์แบบใหม่ โดยการนำเอาผลลัพธ์มาจัดกลุ่ม (Clustering) เพื่ออำนวยความสะดวกให้กับผู้ใช้ในการเลือกพิจารณาดูผลลัพธ์ (เว็บไซต์ Clusty.com ปัจจุบัน yippy.com) ซึ่งจะแบ่งผลลัพธ์การค้นหาออกเป็นหมวดหมู่ต่างๆ ซึ่งจะส่งผลให้ผู้ใช้สามารถเลือกดูผลลัพธ์ตามหมวดหมู่ที่ตนเองต้องการได้ทันที (วรสิทธิ์ ชูชัยวัฒนา : 2555, น.81 -82) ระบบสืบค้นในยุคปัจจุบัน คือ กูเกิ้ล (Google.com) มีการนำเอาเทคนิคต่างๆ มาผสมผสานกันเพื่อให้ได้ผลลัพธ์การค้นคืนที่มีประสิทธิภาพมากยิ่งขึ้น โดยการนำเอาเทคนิค Query Dependent Ranking หรือ Similarity Ranking คือ เทคนิคการนำคำค้นไปเปรียบเทียบกับคำในเอกสาร และเทคนิค Query Independent Ranking หรือ Static Ranking คือ เทคนิคการนำเอาปัจจัยที่เกี่ยวข้องกับเอกสารเข้ามามาพิจารณาร่วมด้วย ตัวอย่างเช่น คุณภาพของเอกสาร (Document Quality) การเชื่อมโยงระหว่างเอกสารที่อยู่ในเครือข่าย (Location) ประวัติการค้นคืน (Query Log) ความเกี่ยวพันธ์กับผู้ใช้ (User Relevance) (ขวัญเรือน โสอุบล : 2557, น.1) และมีการนำปัจจัยเรื่องความใหม่ของเอกสาร (Freshness Documents) ความรวดเร็วในการแสดงผลลัพธ์ และหน้าเว็บไซต์ที่รองรับการใช้งานบน Smart Devices หรือ Mobile Friendly เข้ามาร่วมในการพิจารณาการเรียงลำดับผลลัพธ์การค้นคืนอีกด้วย ซึ่งพบว่าให้ผลการเรียงลำดับผลลัพธ์การค้นคืนเป็นที่น่าพึงพอใจกับผู้ใช้มากยิ่งขึ้น

ในการศึกษางานวิจัยนี้ทำการทดลองสร้างตัวแบบ DRU Internet Search เพื่อพิสูจน์ตัวแบบในการเรียงลำดับผลลัพธ์การค้นคืนด้วยเทคนิคการเรียงลำดับแบบ Query Independent Ranking ผสมผสานกับความเชื่อมของเอกสารในเครือข่าย (Location) ให้ผลลัพธ์ดีกว่าการเรียงลำดับแบบ Query Dependent Ranking เพียงอย่างเดียว ในการสร้างดัชนีต้นแบบของทั้ง 2 วิธีจะใช้ตัวอย่างเอกสารจากอินทราเน็ต มหาวิทยาลัยราชภัฏธนบุรี และมุ่งเน้นไปที่ความเชื่อมโยงของเอกสารในเครือข่ายเป็นหลัก

ในส่วนที่ 2 กล่าวถึงวัตถุประสงค์ของการวิจัย ส่วนที่ 3 กล่าวถึงการทบทวนวรรณกรรมที่เกี่ยวข้องกับระบบสืบค้น (Search Engine) ที่ผ่านมา ส่วนที่ 4 กล่าวถึงการนำเสนอตัวแบบประกอบด้วยภาพรวมของระบบตั้งแต่ขั้นตอนการเก็บข้อมูลเพื่อสร้างดัชนีต้นแบบไปจนถึงการสร้างตัวแบบ DRU Intranet Search ส่วนที่ 5 กล่าวถึงการออกแบบทดลองและประเมินผลระบบค้นคืนของแต่ละดัชนี และการอภิปรายผลในหัวข้อสุดท้าย

2. วัตถุประสงค์ของการวิจัย

2.1 เพื่อสร้างตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต

2.2 เพื่อวัดประสิทธิผลของตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต

3. ทบทวนวรรณกรรม

ในปัจจุบันกระบวนการสืบค้นข้อมูลได้ถูกพัฒนาโดยนักวิจัยหลายกลุ่ม เพื่อให้ผลลัพธ์การค้นคืนมีประสิทธิภาพและประสิทธิผลเพิ่มมากยิ่งขึ้น จากงานวิจัยที่ผ่านมามีการรายงานการคิดค้นวิธีการเรียงลำดับผลการค้นคืน โดยการพิจารณาความใหม่ของเอกสาร (Freshness Documents) ทำให้ได้ผลการค้นคืนที่เป็นปัจจุบันมากขึ้น (Nohuyoshi Sato, 2004), (SadegeKharazmi, et al, 2009), (Lewandowski, et al, 2009) และบันทึกการสืบค้น (Query Log) เป็นอีกหนึ่งข้อมูลสำคัญที่นิยมนำมาใช้เป็นตัวช่วยเพิ่มประสิทธิภาพในการเรียงลำดับผลลัพธ์การค้นคืนรวมทั้งการนำส่วนประกอบและปัจจัยอื่นๆที่อยู่นอกเหนือจากเอกสารมาเป็นส่วนหนึ่งในการสร้างดัชนี (วรสิทธิ์ ชูชัยวัฒนา, 2555) และสามารถแสดงให้เห็นถึงความต้องการข้อมูลที่เปลี่ยนแปลงไปตามเวลาอีกด้วย (Dick Stenmark, 2006) อีกทั้งการพัฒนาเครื่องมือบันทึกการสืบค้นยังสามารถยกระดับคุณภาพการค้นคืนได้อีกด้วย เช่น การใช้แคช (Cache) ช่วยบันทึกการค้นหาเป็นต้น (PragyaKaushik, 2014) นอกจากนี้ภาษาที่ใช้ในการสืบค้นก็เป็นสิ่งสำคัญที่จะเพิ่มความถูกต้องแม่นยำของผลลัพธ์การค้นคืนที่ได้ ดังนั้น จึงมีการนำ Analyzer มาประยุกต์ใช้ในขั้นตอนการประมวลผล ซึ่งสามารถรองรับการใช้งานได้หลายภาษา (Sun Lincheng, 2011) และในส่วนของการค้นคืนเอกสารที่เป็นภาษาไทยได้มีการพัฒนา ThaiAnalyzer ในการวิเคราะห์คำ (ชูชาติ หฤไชยะ-ศักดิ์, 2548) และการประเมินโปรแกรมค้นหาโดยใช้ค่าเฉลี่ย NDCG และ MAP เป็นต้น ซึ่งสามารถช่วยเพิ่มประสิทธิผลให้การเรียงลำดับผลลัพธ์การค้นคืนตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น (Information Retrieval , Guy. 2009) (ขวัญเรือน โสอุบล : 2557) ในการศึกษาวิจัยนี้จึงนำเอาแนวความคิดที่ได้มาประยุกต์ใช้ในการสร้างเป็นตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนเอกสารบนอินทราเน็ตด้วยเทคนิค Query Dependent Ranking และ Query Independent Ranking โดยมุ่งเน้นไปที่ความเชื่อมโยงของเอกสารในเครือข่าย (Location) เพื่อให้ผลลัพธ์การค้นคืนมีประสิทธิผลและตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น

4. การนำเสนอตัวแบบ

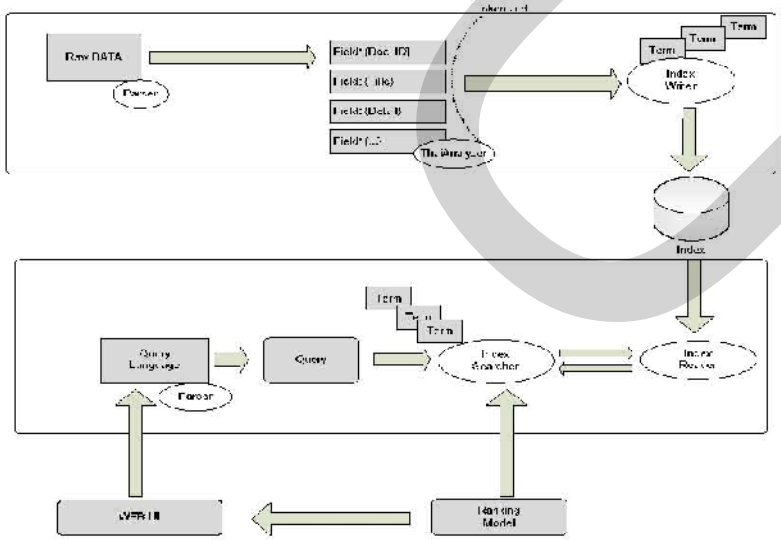
การศึกษาวิจัยมีการสร้างระบบ DRU Intranet Search ขึ้นเพื่อพิสูจน์ตัวแบบ มีขั้นตอนการทำงาน แสดงดังภาพที่ 1

4.1 Crawler

เป็นโปรแกรมที่ทำหน้าที่เก็บข้อมูลจากอินทราเน็ต เพื่อวิเคราะห์ กรองรายละเอียดของเอกสารและเก็บลงในฐานข้อมูล ในการศึกษาวิจัยในครั้งนี้ได้เก็บข้อมูลจากระบบอินทราเน็ตภายในมหาวิทยาลัยราชภัฏธนบุรี ประกอบด้วยเว็บไซต์ของหน่วยงานต่างๆ ในระหว่างเดือนมกราคม พ.ศ. 2559 มีนาคม พ.ศ. 2560 ซึ่งในแต่ละหน่วยงานจะมีเอกสารประชาสัมพันธ์ผ่านเว็บไซต์ซึ่งสามารถแบ่งประเภทของเอกสารต่างๆ เช่น ข่าวประชาสัมพันธ์ ข่าวกิจกรรม ข่าวอบรม/สัมมนา ประกาศจากมหาวิทยาลัย ประกาศจากหน่วยงาน และเอกสารอื่นๆ ที่เกี่ยวข้องในการปฏิบัติงานภายในมหาวิทยาลัยราชภัฏธนบุรี เป็นต้น

4.2 การเตรียมเอกสารสำหรับสร้างดัชนี

การสร้างดัชนีดำเนินการรวบรวมข้อมูล ในการทดลองวิจัยครั้งนี้ได้กำหนดฟิลด์ (Field) ของข้อมูลที่ใช้ในการสร้างดัชนีได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category) ซึ่งข้อมูลแต่ละส่วนจะถูกแยกเก็บในฟิลด์ (Field) บนฐานข้อมูล MySQL และนำมาใช้เป็นฐานข้อมูลเอกสารหรือคลังเอกสาร (Document Corpus) ซึ่งในขั้นตอนของการสร้างดัชนีการวิจัยในครั้งนี้จะใช้ไลบรารีลูซีน (Lucene) สำหรับคำนวณค่าความถี่ของคำในเอกสาร (Team) ที่ปรากฏอยู่ในเอกสารทั้งหมด และจัดเก็บอยู่ในรูปแบบ (Vector) เพื่อให้การค้นคืนมีประสิทธิภาพและรวดเร็ว และ Standard ที่ใช้คือ ThaiAnalyzer เป็นเครื่องมือที่ช่วยทำหน้าที่ในการวิเคราะห์เอกสาร ข้อความ และคำที่อยู่ในเอกสารซึ่งข้อมูลส่วนใหญ่เป็นคำและข้อความภาษาไทย จากนั้นจะนำมาจำแนกข้อมูลที่ได้จากกระบวนการวิเคราะห์คำ (Parsing) ออกเป็นฟิลด์ (Field) ต่างๆ โดยที่เอกสารที่นำมาสร้างดัชนีต้องผ่านในส่วนสำหรับประมวลผลข้อความ (Text Processing Module) ก่อนเพื่อสกัดเอาคำที่สำคัญไปใช้ในการสร้างดัชนี



ภาพที่ 1 ขั้นตอนการทำงานของระบบ DRU Intranet Search

4.3 การสร้างตัวแบบ

ในการศึกษาวิจัยในครั้งนี้ ได้ทำการทดลองสร้างต้นแบบดัชนีทั้งหมด 2 ตัวแบบ ดังนี้

1. Index0 แทน Full-Text Index ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category)

2. Index1 แทน Full-Text Index + Location ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) ชื่อหน่วยงาน (Department) หมวดหมู่ของเอกสาร (Category) และความเชื่อมโยงของเอกสารในเครือข่าย (Location)

5. การออกแบบทดลองและประเมินผล

การทดสอบเพื่อพิสูจน์ตัวแบบที่สร้างขึ้นจะสามารถเรียงลำดับผลลัพธ์การค้นคืนที่ดีขึ้นตามสมมุติฐาน จึงจัดทำระบบ DRU Intranet Search เป็นหน้าเว็บ GUI ให้ผู้ใช้ติดต่อกับระบบค้นคืนในการทดสอบตัวแบบ โดยนักศึกษา บุคลากร และอาจารย์ภายในมหาวิทยาลัยราชภัฏธนบุรีจำนวน 35 คน จำนวนคำค้นทั้งหมด 105 คำสืบค้น โดยทำการมอบหมายงานให้ผู้ทดสอบแต่ละคนใส่คำค้นที่ต้องการเป็นคำ หรือประโยคใดๆ ในระบบ DRU Intranet Search ระบบจะสืบค้นข้อมูลจากดัชนีจาก Index0 และ Index1 เพื่อคำนวณหาค่า Similarity Score และนำคะแนนที่ดีที่สุดนำเข้าสู่ตัวแบบเพื่อผสมผสานกับความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location) เพื่อเข้าสู่กระบวนการประมวลผลภายใต้ตัวแบบเพื่อหาค่า Hybrid Score อีกครั้ง โดยก่อนที่จะแสดงผลให้ผู้ทดสอบประเมิน ระบบจะตรวจสอบเอกสารที่ได้ในแต่ละดัชนีที่เป็นเอกสารเดียวกัน โดยระบบจะรวมผลลัพธ์ให้เหลือเพียงเอกสารเดียว ทั้งนี้เพื่อไม่ให้มีการแสดงผลลัพธ์การคืนค้นซ้ำ หน้าเว็บที่แสดงผลลัพธ์จะแสดงรายละเอียดของเอกสาร ได้แก่ ชื่อเอกสาร (Title) รายละเอียดเอกสาร (Detail) หมวดหมู่ของเอกสาร (Category) ชื่อหน่วยงาน (Dcpartmcnt) วันที่ประชาสัมพันธ์ (Datc_Public) โดยผู้ทดสอบจะต้องอ่านรายละเอียดทั้งหมดและให้คะแนนเอกสาร (Judgment Score) ที่กำลังพิจารณาว่ามีความเกี่ยวข้อง (Relevance) กับคำค้นมากน้อยเพียงใด ซึ่งคะแนนที่ได้ดังกล่าวจะนำไปประเมินผลตัวแบบโดยการทดสอบมีขั้นตอนดังนี้

1. ผู้ทดสอบระบุคำค้นที่ต้องการในหน้าเว็บ โดยระบุเป็นคำหรือประโยคทั้งภาษาไทยหรือภาษาอังกฤษที่ต้องการค้นหา

2. การค้นคืนเอกสาร 20 ลำดับแรกของแต่ละดัชนี ดังรายละเอียดต่อไปนี้

2.1 การค้นคืนเอกสาร 20 ลำดับของดัชนี Index0 จะเรียงลำดับผลลัพธ์จากเทคนิค Query Dependent Ranking หรือ Similarity Ranking

2.2 การค้นคืนเอกสาร 20 ลำดับของดัชนี Index1 จะเรียงลำดับผลลัพธ์จากเทคนิค Query Independent Ranking หรือ Static Ranking และผสมผสานกับความเชื่อมโยงของเอกสารที่อยู่ในเครือข่าย (Location)

2.3 จากนั้นนำผลลัพธ์ของทั้ง 2 Index มาผสมผสานกัน (Combination) และระบบจะหาค่าคะแนนที่ดีที่สุดของการเรียงลำดับผลลัพธ์การค้นคืน

2.4 การแสดงผลลัพธ์จะแสดงแบบสุ่ม (Random) ดังนั้นผู้ทดสอบจะไม่ทราบว่าผลลัพธ์การค้นคืนมาจากดัชนีตัวแบบใด และผู้ทดสอบจะไม่ทราบว่าผลลัพธ์นั้นอยู่ในลำดับใด ทั้งนี้เพื่อป้องกันการเกิดความลำเอียงในการประเมินผล

3. ผู้ทดสอบให้คะแนน (Judgment Score) แต่ละเอกสารโดยพิจารณาความเกี่ยวข้องระหว่างคำค้นกับผลลัพธ์การค้นคืน โดยคะแนนอยู่ระหว่าง 0 ถึง 4 ตามตารางที่ 1

ตารางที่ 1 Judgments Score

คะแนน	คำอธิบาย
4	มีความเกี่ยวข้องกันอย่างมาก (Very Relevant)
3	มีความเกี่ยวข้อง (Relevant)
2	มีความเกี่ยวข้องกันบางส่วน (Somewhat Relevant)
1	มีความเกี่ยวข้องกันเป็นส่วนน้อย (Only Slightly Relevant)
0	ไม่มีความเกี่ยวข้องกัน (Non-Relevant)

เมื่อได้ผล Judgment Score ของเอกสารจากการประเมินของผู้ทดสอบจากนั้นจะถูกนำมาประเมินผล 2 แบบ คือแบบแรกคิดค่า Normalized Discounted Cumulative Gain (NDCG) เป็นการประเมินลำดับผลลัพธ์การค้นคืนที่ได้มีประสิทธิผล (Effectiveness) โดยนำเอกสารทั้งหมดเรียงลำดับตาม Judgment Score นำมาคำนวณเป็น DCG Perfect หรือ Ideal DCG คะแนนที่ได้มีความหมาย คือ คำค้น (Query) มีความเกี่ยวพันธ์ (Relevance) กับผลลัพธ์ของเอกสารนั้นๆ ที่ตำแหน่ง k เมื่อกำหนดให้คำค้น q และเซตของเอกสารจากการสืบค้น ซึ่งคะแนนของเอกสารในแต่ละตำแหน่งสามารถคำนวณได้จากผลลัพธ์การค้นคืนของเอกสารลำดับแรกจนถึงเอกสารลำดับสุดท้าย ตามสมการที่ (1) ส่วนที่สองคือการวัดประสิทธิภาพคือการหาค่าเฉลี่ยความถูกต้องของการค้นหาแต่ละครั้ง เรียกว่า Mean Average Precision (MAP) ตามสมการที่ (2) เป็นการประเมินผลลัพธ์ของเอกสารที่ได้จากการค้นคืนถูกต้องตรงกับความต้องการของผู้ใช้มากน้อยเพียงใด โดยจะตัดคะแนนความถูกต้องจาก 0 ถึง 4 ด้วยค่าคะแนนเท่ากับ 3 ถ้าเอกสารที่ได้คะแนนเท่ากับ 0 ถึง 2 หมายถึงผลลัพธ์การค้นคืนของเอกสารนั้นไม่เกี่ยวข้องกับคำค้น และคะแนนเท่ากับ 3 ถึง 4 หมายถึง ผลลัพธ์ของเอกสารมีความเกี่ยวข้อง (Relevance) กับคำค้น (Query)

$$NDCG_q = \sum_{i=1}^{k} \frac{\left(2^{r(j)} - 1\right)}{log_2(1+i)} \quad (1)$$

เมื่อ j แทนตำแหน่งของเอกสาร และ $r(j)$ แทนจำนวนตัวเลข ซึ่งเป็นค่าคะแนน (Judgment Score) ที่ได้จากผู้ทดสอบ

NDCG แทนค่าคะแนนความเกี่ยวพันธ์ (Relevance) ของผลลัพธ์การค้นคืนเอกสารจากลำดับ (Ranking) แรกไปจนถึงลำดับสุดท้าย

$$MAP = \frac{\sum_{q=1}^{Q} AP\,(q)}{Q} \quad (2)$$

เมื่อ Q แทนจำนวนเอกสารที่ค้นคืนออกมาได้ทั้งหมด

MAP แทนค่าเฉลี่ยความถูกต้องของการค้นคืนเอกสาร

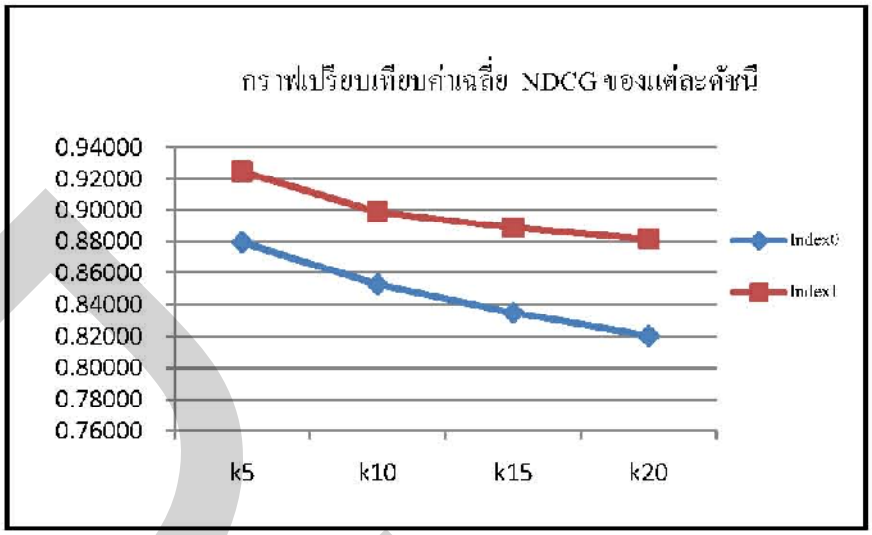
5.1 ผลการวิจัย

จากผลการทดลองระบบ DRU Intranet Search เพื่อประเมิน Judgment Score ที่ได้จากผู้ทดสอบ และนำไปคำนวณค่า DCG Perfect Score จากผลการทดสอบเบื้องต้นมีผู้ทดสอบ 35 ผู้ทดสอบ โดยมีจำนวนคำค้นทั้งหมด 105 คำสืบค้น

เมื่อพิจารณาจากกราฟแสดงดังภาพที่ 2 พบว่า NDCG ของ Index0 ที่ตำแหน่งเอกสาร K5 เท่ากับ 0.87918 และ Index1 ที่ตำแหน่งเอกสาร K=5 เท่ากับ 0.92443 ดังตารางที่ 2 เมื่อพิจารณาค่าเฉลี่ย NDCG ของ Index0 และ Index1 จะสังเกตเห็นว่าทุกช่วงของตำแหน่ง K5 ถึง K20 ของ Index1 ให้ผลลัพธ์การค้นคืนเอกสารได้ดีกว่า Index0 โดยพิจารณาผลลัพธ์จากค่าเฉลี่ย NDCG

ตารางที่ 2 ค่าเฉลี่ย NDCG

k	Index0	Index1
5	0.87918	0.92443
10	0.85239	0.89877
15	0.83485	0.88866
20	0.82020	0.88155

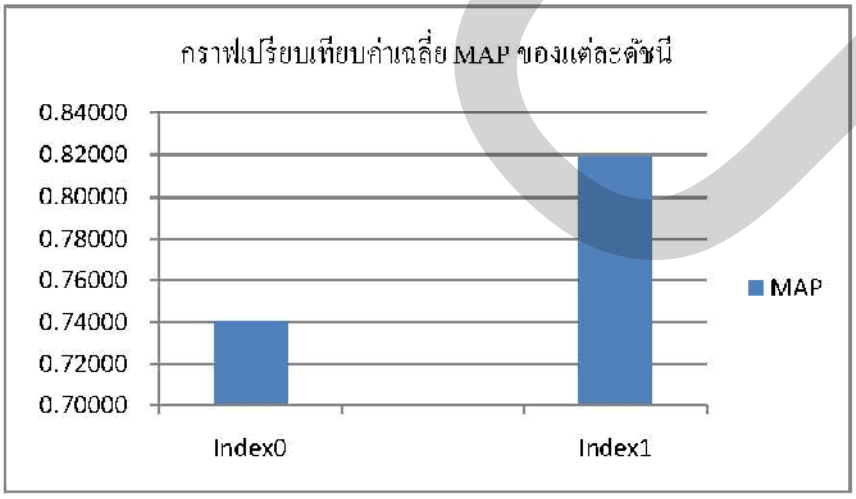


กราฟที่ 1 เปรียบเทียบค่าเฉลี่ย NDCG ของแต่ละดัชนี

พิจารณาจากกราฟที่ 2 พบว่าค่าเฉลี่ย MAP ของ Index0 = 0.74043 และ Index1 = 0.81892 ดังตารางที่ 3 พิจารณาได้ว่า Index1 ให้ค่าเฉลี่ยความถูกต้องของการค้นคืนเอกสารได้ถูกต้องมากกว่า Index0 ที่ค่าเฉลี่ย 0.07849

ตารางที่ 3 ค่าเฉลี่ย MAP

Index0	Index1
0.74043	0.81892



กราฟที่ 2 เปรียบเทียบค่าเฉลี่ย MAP ของแต่ละดัชนี

6. อภิปรายผล

จากผลการประเมินด้วยค่าเฉลี่ย NDCG พบว่า Index1 มีการเรียงลำดับผลลัพธ์การค้นคืนเอกสาร 20 อันดับแรกดีที่สุด ซึ่งเป็นการนำ Similarity Feature มาผสมผสานกับความเชื่อมโยงของเอกสารในเครือข่าย และเรียงลำดับผลลัพธ์การค้นคืนด้วยเทคนิค Query Independent Ranking จะเห็นได้ว่าการพัฒนาระบบค้นคืนเอกสารบนอินทราเน็ตผู้ใช้จะให้ความสำคัญกับผลลัพธ์ของเอกสารที่อยู่ใกล้ตัวหรือเอกสารที่อยู่ในเครือข่าย และสนใจเอกสารที่เกี่ยวข้องกับหน่วยงานมากกว่าเอกสารทั่วๆไป

เนื่องจากการวิจัยในครั้งนี้ได้ใช้ตัวอย่างข้อมูลบนอินทราเน็ต มหาวิทยาลัยราชภัฏธนบุรีในการพัฒนาตัวแบบการเรียงลำดับผลลัพธ์การค้นคืนบนอินทราเน็ต แต่เนื่องจากโครงสร้างเว็บไซต์ของแต่ละหน่วยงานมีโครงสร้างของหน้าเว็บที่แตกต่างกันและมีการประชาสัมพันธ์ข่าวซ้ำๆ กันหลายครั้ง ทำให้เสียเวลาในการปรับแต่ง Crawler และคัดกรองเอกสารเพื่อเก็บข้อมูลเอกสารลงในฐานข้อมูล

ในการพัฒนาระบบค้นคืนของเอกสารบนอินทราเน็ตในครั้งต่อไปควรมีการนำปัจจัยอื่นๆ ที่ไม่เกี่ยวข้องกับเอกสาร (Query Independent Ranking) เช่น ความใหม่ของเอกสาร บันทึกการค้นหา เป็นต้น มาเป็นส่วนหนึ่งของการเรียงลำดับผลลัพธ์การค้นคืนของเอกสารบนอินทราเน็ต และการลดระยะเวลาและขนาดของการสร้างดัชนีจะทำให้การแสดงผลลัพธ์การค้นคืนมีความรวดเร็วมีประสิทธิภาพและประสิทธิผลดีขึ้น และตรงกับความต้องการของผู้ใช้มากยิ่งขึ้น

รายการอ้างอิง

ศิรัตน์ ศิรนานนท์.(2549). “การค้นคืนสารสนเทศโดยใช้กฎความสัมพันธ์ร่วมกับผลสะท้อนกลับจากผู้ใช้.”วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิตสาขาวิชาการพัฒนาซอฟแวร์ด้านธุรกิจ ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย.

วรสิทธิ์ ชูชัยวัฒนา . (2555). “การปรับปรุงประสิทธิผลของระบบค้นคืนสารสนเทศและโปรแกรมการค้นหา :แนวคิดและเทคนิค.” **วารสารวิชาการสมาคมสถาบันอุดมศึกษาเอกชนแห่งประเทศไทย** ฉบับวิทยาศาสตร์และเทคโนโลยี ปีที่ 3 ฉบับที่ 1 เดือน มกราคม-มิถุนายน 2557.

ขวัญเรือน โสอุบล. (2557) “ตัวแบบสำหรับการเรียงลำดับผลลัพธ์การค้นคืนในระบบค้นคืนบทความวิจัยโดยการใช้ข้อมูลทางบรรณานุกรม.”วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิตสาขาวิชาวิศวกรรมเว็บ คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยธุรกิจบัณฑิตย์.

ชูชาติ หฤไชยะศักดิ์. (2548)“การพัฒนาโปรแกรมสำรหับค้นคืนสารสนเทศภาษาไทย”**National and Computer Technology Center (NECTEC)**, 2537.

Information Retrieval , Guy. (2009).**Information Retrieval**, Cambridge University Press Cambridge, England

Nohuyoshi Sato, (2004). The Evaluations of FTF-IDF Scoring for Fresh Information Retrieval, **Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA’05),** Tohoku University, Sendai,Miyagi 981-8555 Japan.

Hang Li, (2005). A New Approach to Intranet Search Based on Information Extraction, DmitriyMeyerzon**Microsoft Corporation One Microsoft Way, Redmond, WA,** USA, 98052

Dick Stenmark, (2006). What are you searching for? A content analysis of intranet search engine logs, **University of Gothenburg Göteborg, VaestraGoetaland,** Sweden

SadeghKharazmi, (2009).Freshness of Web search engines: Improving performance of Web search engines using data mining techniques, **Faculty of Computer Engineering, Payame Noor University; WI Lab, Department of Computer Engineering, Sharif University of Technology** Tehran, Iran

Lewandowski, et al (2009).Joint Optimization of Index Freshness and Coverage in Real-Time Search Engines, **Published in: IEEE Transactions on Knowledge and Data Engineering (Volume: 24, Issue: 12, Dec. 2012), Department of Information Science, Heinrich-Heine-University Düsseldorf,** Germany

Sun Lincheng, (2011). A large-scale full-text search engine using DotLuence, **2011 IEEE 3rd International Conference on Communication Software and Networks,** Xi'an, China

Pragya Kaushik, (2014). Use of query logs for providing cache support to the search engine, **2014 International Conference on Computing for Sustainable Global Development (INDIACom)** New Delhi, India

4. ผลการพิจารณา

☐ สมควรให้นำเสนอได้โดยไม่ต้องแก้ไข

☑ สมควรให้นำเสนอได้ โดยให้ปรับแก้ไข

☑ ปรับแก้แล้วไม่ต้องส่งมาให้พิจารณา

☐ ปรับแก้แล้วส่งมาให้พิจารณาอีกครั้ง (ภายใน 3 วัน นับจากวันที่ได้รับการติดต่อให้แก้ไขบทความ)

ข้อเสนอแนะเพื่อการปรับแก้

☐ ไม่สมควรให้นำเสนอ เนื่องจาก

ในกรณีที่ท่านพิจารณาบทความวิจัยแล้วให้ 32 คะแนนขึ้นไป ซึ่งผ่านเกณฑ์การประเมินฯ บทความนั้นจะได้รับการพิจารณาให้ได้รับรางวัลบทความวิจัยดีเด่นและใบประกาศเกียรติคุณ ท่านเห็นว่าบทความวิจัยดังกล่าวสมควรได้รับรางวัลบทความวิจัยดีเด่นหรือไม่

☐ สมควร ☑ ไม่สมควร

ประวัติผู้เขียน

ชื่อ-นามสกุล	พิศิษฐ์ บวรเลิศสุธี
ประวัติการศึกษา	ปีการศึกษา 2555 สำเร็จการศึกษาระดับปริญาตรี สาขาคอมพิวเตอร์ธุรกิจ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี
ตำแหน่งและสถานที่ทำงานปัจจุบัน	นักวิชาการคอมพิวเตอร์ คณะวิทยาการจัดการ มหาวิทยาลัยราชภัฏธนบุรี ตั้งอยู่เลขที่ 172 ถนนอิสรภาพ แขวง วัดกัลยาณ์ เขตธนบุรี กรุงเทพฯ 10600