การลำเอียงด้วยความใกล้ชิดด้านเวลาในการคำนวณเพจแร็งค์ส่วนบุคคล

นายกานต์กมล ทองทิพย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

# TIME-PROXIMITY BIASING IN PERSONALIZED PAGERANK COMPUTATION

Mr. Kankamol Tongtip

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2012

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การลำเอียงด้วยความใกล้ชิดด้านเวลาในการคำนวณเพจแร็งค์ส่วนบุคคล |
| โดย | นายกานต์กมล ทองทิพย์ |
| สาขาวิชา | วิทยาศาสตร์คอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม | ดร. บัณฑิต มนัสเกษมศักดิ์ |

---

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.บุญสม เลิศหิรัญวงศ์)

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.พิษณุ คนองชัยยศ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(อาจารย์ ดร.บัณฑิต มนัสเกษมศักดิ์)

.....กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง)

กานต์กมล ทองทิพย์ : การลำเอียงด้วยความใกล้ชิดด้านเวลาในการคำนวณเพจแร็งค์ส่วนบุคคล. (Time-Proximity Biasing in Personalized PageRank Computation) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ.ดร.อรรถสิทธิ์ สุรฤกษ์, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : ดร.บัณฑิต มนัสเกษมศักดิ์, 45 หน้า.

ปัจจุบันการวิเคราะห์ฐานข้อมูลเว็บที่จัดเก็บมาได้เพียงชุดเดียวเริ่มไม่มีประสิทธิภาพที่เพียงพอสำหรับการจัดการเครื่องมือสืบค้นเพื่อให้ได้ผลลัพธ์ค้นคืนที่เหมาะสม โดยเฉพาะอย่างยิ่งในกระบวนการจัดเรียงลำดับเว็บ ซึ่งโดยลักษณะการเปลี่ยนแปลงของเว็บนั้น ทำให้อัลกอริทึมจัดเรียงลำดับที่อิงตามเส้นเชื่อมโยงแบบดั้งเดิมจำนวนมากมักให้ความสำคัญกับเว็บเพจเก่ามากจนเกินไป อีกทั้งยังไม่อาจรับรู้ถึงความสำคัญของเว็บเพจใหม่ เนื่องจากเว็บเพจเก่าย่อมมีเวลาสั่งสมจำนวนเส้นเชื่อมโยงเข้าหาหรือถูกอ้างอิงมากกว่าเว็บเพจใหม่นั่นเอง วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการจัดเรียงลำดับเว็บส่วนบุคคล ที่อิงตามเส้นเชื่อมโยงร่วมกับข้อมูลเชิงเวลา ที่สกัดจากประวัติความเคลื่อนไหวของเว็บเพจ แบบจำลองความใกล้ชิดด้านเวลา ด้วยฟังก์ชันเคอเนลที่แตกต่างกัน ถูกนำเสนอเพื่อประเมินความเกี่ยวข้องกันระหว่างเว็บเพจ ซึ่งจะถูกนำไปใช้ในกระบวนถ่ายทอดย้อนกลับ สำหรับในการคำนวณค่าคะแนนความลำเอียงด้านเวลาของเว็บเพจในท้ายที่สุด ค่าคะแนนดังกล่าวจะถูกกำหนดเป็นเวกเตอร์ความลำเอียง ในการคำนวณเพจแร็งค์ส่วนบุคคล จากการทดลองบนฐานข้อมูลเว็บจริงที่ได้จากอินเตอร์เน็ตอาร์ไคว์ฟ แสดงให้เห็นว่าแนวคิดของวิทยานิพนธ์ฉบับนี้ได้เพิ่มประสิทธิภาพการจัดเรียงลำดับผลลัพธ์ค้นคืนของเพจแร็งค์ได้ดียิ่งขึ้น เมื่อพิจารณาตามความพึงพอใจของผู้ใช้งาน

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2555 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม .....

# # 5271406421 : MAJOR COMPUTER SCIENCE
KEYWORDS : TEMPORAL-BIASED PERSONALIZED PAGERANK TIME-PROXIMITY MODEL TEMPORAL ANALYSIS PAGERANK COMPUTATION WEB RANKING ALGORITHM.

KANKAMOL TONGTIP : TIME-PROXIMITY BIASING IN PERSONALIZED PAGERANK COMPUTATION. ADVISOR : ASST. PROF. ATHASIT SURARERKS Ph.D., CO-ADVISOR : BUNDIT MANASKASEMSAK, D.Eng., 45 pp.

Today, an analysis on only a single crawled snapshot of World Wide Web becomes not efficient enough for a search engine administration, especially a web ranking procedure, to provide appropriate search results. By the dynamic nature of the Web, many traditional link-based ranking algorithms, like PageRank, suffer from over granting stale pages an authority and also fail to recognize important new ones since the former have had much time to accumulate in-links (i.e., referrers) than the latter. In this Thesis, we propose a web personalized link-based ranking scheme that incorporates temporal information extracted from historical page activities. A time-proximity model based on several kernel functions is introduced to estimate page relatedness that is subsequently employed in inverse propagation for calculating temporal biased scores of web pages. These scores finally act as a bias vector used in personalized PageRank computation. Experiments conducted on a real-world web data collected from the Internet Archive show that our approach improves upon PageRank in ranking of search results with respect to human users' preference.

Department : Computer Engineering Student's Signature .....
Field of Study : Computer Science Advisor's Signature .....
Academic Year : 2012 Co-advisor's Signature .....

# กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดีด้วยความช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.อรรถสิทธิ์ สุรฤกษ์ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ และอาจารย์ ดร.บัณฑิต มนัสเกษมศักดิ์ ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนร่วม ที่ซึ่งได้มอบความรู้ คำแนะนำ ตรวจทานเพื่อแก้ไขส่วนบกพร่องของงานวิจัย ตลอดจนการตรวจทานแก้ไขวิทยานิพนธ์ให้มีความสมบูรณ์ นอกจากนี้ผู้เขียนยังได้รับความกรุณาจากห้องปฏิบัติการวิจัยทางวิศวกรรมระบบนับได้เชิงทฤษฎี ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย และห้องปฏิบัติการวิจัยข้อมูลและฐานความรู้ขนาดใหญ่ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ ที่กรุณาให้การสนับสนุนในด้านสถานที่ ทรัพยากร และเครื่องมือ อีกทั้งผู้ช่วยศาสตราจารย์ ดร.พิษณุ คนองชัยยศ ผู้ซึ่งเป็นประธานกรรมการสอบวิทยานิพนธ์ รวมถึงผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง อาจารย์กรรมการสอบวิทยานิพนธ์ผู้ทรงคุณวุฒิจากภายนอก ที่ได้ให้คำแนะนำ รวมทั้งข้อเสนอแนะที่เป็นประโยชน์เพื่อนำมาใช้ปรับปรุงวิทยานิพนธ์ให้เกิดความสมบูรณ์มากยิ่งขึ้น

ผู้เขียนขอกราบขอบพระคุณบิดา มารดา ที่ได้สนับสนุนด้านทุนทรัพย์ในการศึกษาและคอยเป็นกำลังใจและเป็นแรงบันดาลใจให้ข้าพเจ้าเสมอมา

# สารบัญ

หน้า

หน้า

# สารบัญตาราง

หน้า

# สารบัญภาพ

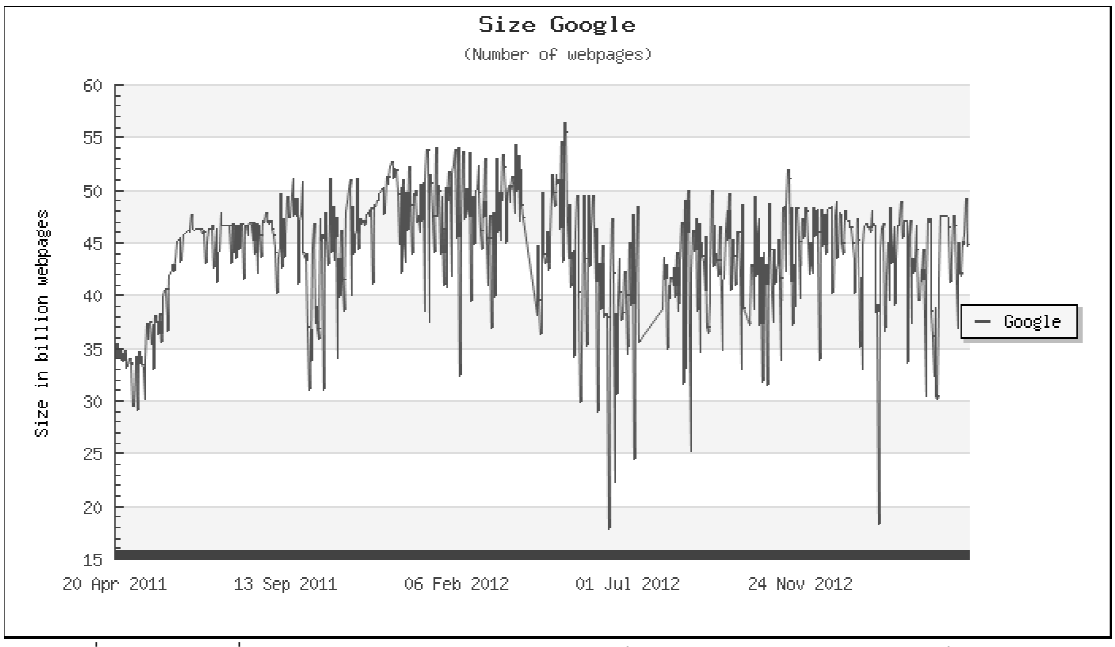
หน้า

# บทที่ 1

# บทนำ

## 1.1. ความเป็นมาและความสำคัญของปัญหา

อัลกอริทึมจัดเรียงลำดับเว็บเพจโดยการวิเคราะห์เส้นเชื่อมโยงได้รับการพิสูจน์แล้วว่า มีประสิทธิภาพในแง่ของการจัดเรียงลำดับผลลัพธ์ค้นคืน อาทิเช่น อัลกอริทึมเพจแร็งค์ (PageRank algorithm) [*18*] ซึ่งพิจารณาโครงสร้างการเชื่อมโยง (link structure) เพื่อคำนวณการส่งผ่านค่าคะแนนความสำคัญระหว่างเว็บเพจ โดยเว็บเพจที่ถูกอ้างอิงจากเว็บเพจอื่นจำนวนมากจะมีค่าคะแนนมากและควรถูกจัดอยู่ลำดับต้นของผลลัพธ์ค้นคืน อย่างไรก็ดี จุดอ่อนของอัลกอริทึมดังกล่าวคือการคำนวณค่าคะแนนโดยอาศัยเว็บเพจที่จัดเก็บมาได้ (crawled snapshot) เพียงชุดเดียวแล้ววิเคราะห์บนโครงสร้างการเชื่อมโยงนั้น ซึ่งจุดนี้เองทำให้เพจแร็งค์มักจะให้คะแนนสูงกับเว็บเพจที่ถูกสร้างขึ้นนานแล้ว ในขณะที่จะให้คะแนนน้อยกับเว็บเพจที่ถูกสร้างขึ้นใหม่ [*1, 6*] ถึงแม้ว่าเว็บเพจใหม่นั้นจะเป็นเว็บเพจที่มีคุณภาพก็ตาม ทั้งนี้เนื่องจากเว็บเพจที่ถูกสร้างขึ้นใหม่มักยังไม่เป็นที่รู้จักและถูกอ้างอิงถึงน้อยนั่นเอง ดังนั้นจึงมักเกิดกรณีเว็บเพจที่มีคุณภาพและทันสมัยแต่ถูกลำเอียงให้ได้รับคะแนนความสำคัญน้อยกว่าเว็บเพจที่เก่าจนข้อมูลล้าสมัยไปแล้วได้




รูปที่ 1.1 การเปลี่ยนแปลงจำนวนโดยประมาณของเว็บเพจจากฐานข้อมูลดรรชนีเว็บเพจของ google[1]

จากรูปที่ 1.1 จะเห็นได้ว่าโดยธรรมชาติของเว็บมักมีลักษณะที่เปลี่ยนแปลงอยู่ตลอดเวลา เนื้อหา (content) และเส้นเชื่อมโยง (hyperlink) ของเว็บเพจจึงอาจถูกสร้างขึ้นใหม่ ถูก

[1] www.worldwidewebsize.com

ลบทิ้ง หรือถูกแก้ไขปรับปรุง ซึ่งโดยปกติแล้ว ผู้ดูแลเว็บเพจก็มักจะแก้ไขปรับปรุงเพื่อให้เนื้อหาและเส้นเชื่อมโยงมีความเป็นปัจจุบัน (up-to-dateness) ยิ่งไปกว่านั้น เว็บเพจที่มีข้อมูลทันสมัยก็มักได้รับความสนใจและต้องการเข้าถึงจากผู้เยี่ยมชมมากกว่าเว็บเพจที่มีข้อมูลเก่า ตัวอย่างเช่น เว็บเพจที่นำเสนอข่าวสารใหม่ย่อมน่าสนใจกว่าเว็บเพจที่นำเสนอข่าวสารเก่า หรือเว็บเพจของงานประชุมวิชาการที่กำลังจะจัดขึ้นย่อมน่าสนใจกว่าเว็บเพจของงานประชุมวิชาการที่ผ่านไปแล้ว เป็นต้น ด้วยเหตุผลดังกล่าว เครื่องมือสืบค้นควรคำนึงถึงปัจจัยในเชิงเวลา (temporal factor) เพื่อให้สามารถค้นคืนและจัดเรียงลำดับผลลัพธ์ได้ตรงตามความต้องการของผู้ใช้งานมากยิ่งขึ้น

**1.2. วัตถุประสงค์ของการวิจัย**

งานวิจัยนี้นำเสนออัลกอริทึมจัดเรียงลำดับเว็บเพจที่คำนึงถึงมุมมองด้านเวลา (temporal aspect) โดยมุ่งปรับปรุงประสิทธิภาพของอัลกอริทึมเพจแร็งค์เดิม ซึ่งงานที่นำเสนอเริ่มจากการศึกษาและวิเคราะห์ประวัติการเปลี่ยนแปลงของเว็บเพจ จากนั้นได้ออกแบบแบบจำลองความใกล้ชิดด้านเวลา (time-proximity model) ที่ใช้ฟังก์ชันเคอเนล (kernel function) ห้าแบบ [*7, 8, 17, 19*] เพื่อประเมินค่าความใกล้ชิดอันเนื่องจากการเปลี่ยนแปลงระหว่างสองเว็บเพจที่มีเส้นเชื่อมโยงถึงกัน โดยค่าความใกล้ชิดที่ได้นั้นจะถูกใช้ในกระบวนการคำนวณการส่งผ่านค่าคะแนนย้อนกลับ (inverse score propagation) เพื่อสร้างเป็นเวกเตอร์ถ่วงน้ำหนักเวลา (temporal bias vector) ที่ถูกนำไปเป็นค่าความลำเอียงของการคำนวณเพจแร็งค์ส่วนบุคคล (personalized PageRank computation) อีกทีหนึ่ง

**1.3. ขอบเขตของการวิจัย**

1.3.1. พัฒนาระบบด้วยอาร์ปาเช่ลูซีน (apache lucence)

1.3.2. ฐานข้อมูลเว็บเพจทดลองนำมาจากยูอาร์แอลในหัวข้อการท่องเที่ยวไทย จากโอดีพี (Open Directory Project: ODP) และข้อมูลของเว็บเพจจากอินเตอร์เน็ตอาร์ไคว์ฟ (Internet archive)

1.3.3. ใช้ค่าถ่วงน้ำหนักทีเอฟไอดีเอฟ (tf-idf) สำหรับการจัดเรียงลำดับพื้นฐาน

1.3.4. เลือกเฉพาะผลลัพธ์ค้นคืนจากค่าถ่วงน้ำหนักทีเอฟไอดีเอฟในยี่สิบอันดับแรก

1.3.5. ใช้มาตรวัดผลการทดลองเอ็นดีซีจี (Normalized Discounted Cumulative Gain: NDCG)

1.3.6. เปรียบเทียบโดยใช้ค่าเฉลี่ยเอ็นดีซีจีในผลลัพธ์ค้นคืนห้าอันดับแรกและสิบอันดับแรกกับอัลกอริทึมเพจแร็งค์แบบดั้งเดิม (PageRank: PR)

**1.4. ประโยชน์ที่คาดว่าจะได้รับ**

งานวิจัยชิ้นนี้ได้นำเสนออัลกอริทึมเพจแร็งค์ส่วนบุคคลตามความลำเอียงด้านเวลา (Temporal biased in Personalized PageRank: TPPR) ในการจัดเรียงลำดับเว็บเพจซึ่งใช้มุมมอง

ด้านเวลามาช่วยในการคำนวณ โดยมุ่งปรับปรุงประสิทธิภาพของอัลกอริทึมเพจแร็งค์แบบดั้งเดิม นอกจากนี้ยังได้นำเสนอวิธีการหาค่าความใกล้ชิดกันระหว่างสองเพจที่มีเส้นเชื่อมโยงกันผ่านประวัติการปรับปรุงเว็บเพจเพื่อที่จะสกัดเวกเตอร์ความลำเอียงด้านเวลาของแต่ละเว็บเพจ

อัลกอริทึมที่พัฒนาขึ้นมาสามารถจัดเรียงลำดับกับเว็บเพจจริงบนอินเตอร์เน็ต เพราะสามารถเพิ่มประสิทธิภาพการจัดเรียงลำดับเว็บเพจผลลัพธ์ค้นคืน ซึ่งสอดคล้องตามความพึงพอใจของผู้ใช้งานได้มากกว่าการจัดเรียงโดยเพจแร็งค์ดั้งเดิม

**1.5. วิธีดำเนินการวิจัย**

1.5.1. ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง

1.5.2. เตรียมชุดข้อมูลเว็บเพจทดลอง

1.5.3. ออกแบบแบบจำลองเว็บพิจารณาบนแกนเวลา เพื่อทำการเก็บประวัติการเปลี่ยนแปลงของเว็บเพจ

1.5.4. หาค่าถ่วงน้ำหนักความใกล้ชิดด้านเวลาระหว่างเว็บเพจ จากการคำนวณค่าข้อมูลเชิงเวลาโดยใช้แบบจำลองความใกล้ชิดด้านเวลา

1.5.5. คำนวณหาค่าเวกเตอร์ความลำเอียงเชิงเวลาโดยใช้อินเวิร์สเพจแร็งค์

1.5.6. คำนวณค่าคะแนนเพจแร็งค์โดยนำเวกเตอร์ความลำเอียงที่ได้จากขั้นตอนก่อนหน้ามาคำนวณในอัลกอริทึมเพจแร็งค์ดั้งเดิม

1.5.7. เปรียบเทียบผลการจัดเรียงลำดับเว็บเพจโดยอัลกอริทึมที่นำเสนอ (ทีพีพีอาร์) และอัลกอริทึมเพจแร็งค์แบบดั้งเดิม

1.5.8. วิเคราะห์ผลการวิจัย

1.5.9. สรุผผลงานวิจัย

**1.6. ลำดับขั้นตอนในการเสนอผลการวิจัย**

วิทยานิพนธ์นี้แบ่งเนื้อหาออกเป็นห้าบท ดังรายละเอียดต่อไปนี้

บทที่ 1 นำเสนอ ความเป็นมาและความสำคัญของปัญหา วัตถุประสงค์ของการวิจัย ขอบเขตของการวิจัย คำจำกัดความที่ใช้ในการวิจัย ประโยชน์ที่คาดว่าจะได้รับ และวิธีดำเนินการวิจัย

บทที่ 2 นำเสนอแยกออกเป็นสองส่วนย่อย ดังนี้

1. ทฤษฎีที่เกี่ยวข้อง ได้แก่ แบบจำลองเว็บกราฟ ทรานซิชันเมทริกซ์ อินเวิร์สทรานซิชันเมทริกซ์ อัลกอริทึมเพจแร็งค์ อัลกอริทึมอินเวิร์สเพจแร็งค์ และเพจแร็งค์ส่วนบุคคล

2. งานวิจัยที่เกี่ยวข้อง

บทที่ 3 วิธีดำเนินการวิจัยจะแบ่งออกเป็นสองส่วนคือ การออกแบบอัลกอริทึมเพจแร็งค์ส่วนบุคคลตามความลำเอียงด้านเวลา และวิธีดำเนินการทดลอง

บทที่ 4 นำเสนอผลการทดลอง

บทที่ 5 สรุปและวิเคราะห์ผลการวิจัย

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

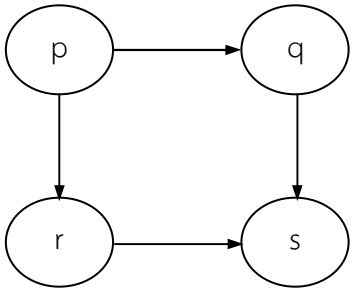
ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้อง โดยในส่วนของทฤษฎีที่เกี่ยวข้องจะเริ่มอธิบายจากแบบจำลองเว็บกราฟเพื่อให้เข้าใจในการแทนฐานข้อมูลเว็บด้วยกราฟแบบมีทิศทาง จากนั้นจะอธิบายถึงการใช้เมทริกซ์เพื่อแสดงถึงเว็บกราฟแบบมีทิศทาง ต่อมาจะอธิบายถึงอัลกอริทึมเพจแร็งค์ซึ่งในงานวิจัยนี้จะนำมาเป็นอัลกอริทึมเปรียบเทียบกับอัลกอริทึมที่นำเสนอ สุดท้ายจะกล่าวถึงอัลกอริทึมอินเวิร์สเพจแร็งค์ที่ในงานวิจัยชิ้นนี้จะนำมาใช้ในการคำนวณเวกเตอร์ความลำเอียง

ในส่วนของงานวิจัยที่เกี่ยวข้องจะนำเสนองานวิจัยที่นำมุมมองเว็บเพจในแต่ละด้านมาปรับปรุงอัลกอริทึมเพจแร็งค์

## 2.1. ทฤษฎีที่เกี่ยวข้อง

### 2.1.1. แบบจำลองเว็บกราฟ (web graph model)

สำหรับข้อมูลเว็บชุดหนึ่ง สามารถถูกแทนด้วยแบบจำลองกราฟแบบมีทิศทาง (directed graph) $G = (V, E)$ เรียกว่า "เว็บกราฟ" (web graph) โดยที่เซตของจุดยอด (vertex) $V$ แทนเว็บเพจจำนวน $N$ เว็บเพจ และเซตของขอบ (edge) $E \subset V^2$ แทนเส้นเชื่อมโยงระหว่างเว็บเพจ โดยที่หากปรากฏเส้นเชื่อมโยงจำนวนมากกว่าหนึ่งเส้นระหว่างคู่ของเว็บเพจเดียวกันแล้ว เราจะแทนด้วยขอบเพียงเส้นเดียวในกราฟ และจะไม่พิจารณาเส้นเชื่อมโยงในกรณีที่อ้างอิงเข้าหาตัวเอง รูปที่ 2.1 แสดงตัวอย่างเว็บกราฟอย่างง่าย




รูปที่ 2.1 แบบจำลองเว็บกราฟอย่างง่าย

จากรูปที่ 2.1 เราสามารถแสดงแบบจำลองเว็บกราฟด้วยเมทริกซ์ขนาด $N \times N$ ได้ตามจำนวนจุดยอดของเว็บกราฟ ซึ่งจากรูปตัวอย่างจะสามารถเขียนแทนได้ด้วยเมทริกซ์ขนาด $4 \times 4$ ดังที่จะอธิบายในหัวข้อถัดไป

### 2.1.2. ทรานซิชันเมทริกซ์ (transition matrix) และอินเวิร์สทรานซิชันเมทริกซ์ (inverse transition matrix)

ในทางปฏิบัติ เว็บกราฟแบบมีทิศทางสามารถถูกแสดงได้ด้วยเมทริกซ์ กล่าวคือ เมื่อนิยามให้ $O(p)$ คือจำนวนเว็บเพจที่ถูกอ้างอิงโดยเว็บเพจ $p$ เรียกว่า "จำนวนการเชื่อมโยงออก"

(out-degree) และนิยามให้ $I(p)$ คือจำนวนเว็บเพจที่อ้างอิงมายังเว็บเพจ $p$ เรียกว่า "จำนวนการเชื่อมโยงเข้า" (in-degree) แล้ว เมทริกซ์แสดงเว็บกราฟสามารถถูกพิจารณาได้สองลักษณะ ได้แก่

ลักษณะที่ 1. ทรานซิชันเมทริกซ์

ทรานซิชันเมทริกซ์คือเมทริกซ์ที่มีค่าในแต่ละจุดบนเมทริกซ์แสดงถึงค่าความน่าจะเป็นในการเปลี่ยนแปลงจากจุดยอดหนึ่งไปยังอีกจุดยอดหนึ่ง โดยค่าความน่าจะเป็นดูได้จากจำนวนเส้นขอบที่ชี้ออกจากจุดยอดนั้น

กำหนดให้ $A$ แทนทรานซิชันเมทริกซ์ขนาด $N \times N$ โดยที่แถวและคอลัมน์คือจุดยอดซึ่งนิยามโดย ให้ $a_{qp}$ เป็นจุดบนเมทริกซ์ $A$ ที่แสดงถึงความสัมพันธ์ระหว่างเพจ $q$ และ เพจ $p$ ถ้ามีขอบจากจุดยอด $p$ ไปยัง $q$ แล้วให้ $a_{qp}$ มีค่าเป็น $\frac{1}{O(p)}$ ในทางกลับกันถ้าไม่มีขอบจากจุดยอด $p$ ไปยัง $q$ ให้มีค่าเป็น 0 ดังสมการที่ 1

$$A(q,p) = \begin{cases} \frac{1}{O(p)} & \text{ถ้า } (p,q) \in E \\ 0 & \text{กรณีอื่นๆ} \end{cases} \quad (1)$$

ดังนั้นจากตัวอย่างเว็บกราฟรูปที่ 2.1 เราสามารถเขียนทรานซิชันเมทริกซ์ $A$ ขนาด $4 \times 4$ โดยที่แถวและคอลัมน์แสดงถึงแต่ละจุดบนเว็บกราฟ ได้ดังรูปที่ 2.2

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

รูปที่ 2.2 ตัวอย่างทรานซิชันเมทริกซ์

ลักษณะที่ 2. อินเวิร์สทรานซิชันเมทริกซ์

อินเวิร์สทรานซิชันเมทริกซ์มีลักษณะคล้ายกันกับทรานซิชันเมทริกซ์เพียงแต่ ค่าความน่าจะเป็นในการเปลี่ยนผ่านในแต่ละจุดจะมีมุมมองที่แตกต่างกัน โดยในทรานซิชันเมทริกซ์จะคำนึงถึงจำนวนเส้นเชื่อมโยงออก แต่อินเวิร์สทรานซิชันเมทริกซ์จะคำนึงถึงเส้นเชื่อมโยงเข้า

กำหนดให้ $B$ แทนอินเวิร์สทรานซิชันเมทริกซ์ขนาด $N \times N$ ให้ $b_{pq}$ เป็นจุดบนเมทริกซ์ $B$ ถ้ามีขอบจากจุดยอด $p$ ไปหา $q$ แล้วให้ $b_{pq}$ มีค่าเป็น $\frac{1}{I(q)}$ แต่ถ้าไม่มีขอบจาก $p$ ชี้ไปยัง $q$ ให้มีค่าเป็น 0 ดังสมการที่ 2

$$B(p,q) = \begin{cases} \frac{1}{I(q)} & \text{ถ้า } (p,q) \in E \\ 0 & \text{กรณีอื่นๆ} \end{cases} \quad (2)$$

ดังนั้นจากตัวอย่างเว็บกราฟรูปที่ 2.1 เราสามารถเขียนอินเวอร์สทรานซิชันเมทริกซ์ $B$ ขนาด $4\times4$ โดยที่แถวและคอลัมน์แสดงถึงแต่ละจุดบนเว็บกราฟ ได้ดังรูปที่ 2.3

$$B=\begin{pmatrix}0 & 1 & 1 & 0\\ 0 & 0 & 0 & \frac{1}{2}\\ 0 & 0 & 0 & \frac{1}{2}\\ 0 & 0 & 0 & 0\end{pmatrix}$$

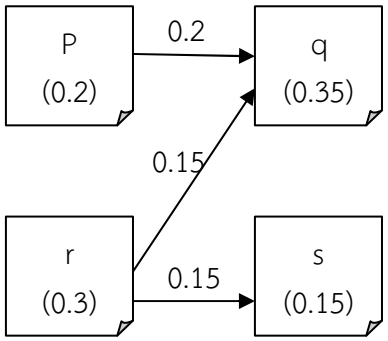
รูปที่ 2.3 ตัวอย่างอินเวิร์สทรานซิชันเมทริกซ์

### 2.1.3. อัลกอริทึมเพจแร็งค์ (PageRank Algorithm: PR)

เพจแร็งค์เป็นอัลกอริทึมจัดเรียงลำดับเว็บเพจที่มีประสิทธิภาพสูงซึ่งถูกคิดค้นขึ้นโดยผู้ก่อตั้งบริษัทกูเกิล (Google) [*5, 18*] เพจแร็งค์จะกำหนดค่าคะแนนความสำคัญ (authoritative score) ให้กับเว็บเพจโดยวิเคราะห์โครงสร้างการเชื่อมโยงตามเว็บกราฟ สำหรับแนวคิดพื้นฐานของเพจแร็งค์คือ เพจที่มีลำดับคะแนนความสำคัญสูงคือเพจที่มีผลรวมของคะแนนเพจแร็งค์ของเพจที่สร้างเส้นเชื่อมโยงเข้ามาสูง ไม่ว่าจะเป็นเพจที่มีเส้นเชื่อมโยงเข้ามามากหรือเพจที่มีเส้นเชื่อมโยงเข้าน้อยแต่มาจากเพจที่มีคะแนนความสำคัญสูง ตัวอย่างเช่น ถ้ามีเส้นเชื่อมโยงจากเว็บเพจ $p$ ไปยังเว็บเพจ $q$ แล้ว เราอาจพิจารณาได้ว่าผู้ดูแลเว็บเพจ $p$ สนใจในเนื้อหาและเห็นความสำคัญของเว็บเพจ $q$ ซึ่งเสมือนเป็นการส่งค่าคะแนนความสำคัญหรือ "ค่าเพจแร็งค์" (PageRank score) ไปยังเพจนั้น โดยคะแนนที่ถูกส่งออกไปจะขึ้นอยู่กับจำนวนเส้นเชื่อมโยงออกไปนั่นเองดังสมการที่ 3 เพจแร็งค์อย่างง่ายและสามารถแสดงการส่งผ่านค่าเพจแร็งค์ของเพจดังแสดงในรูปที่ 2.4 การส่งผ่านค่าคะแนนเพจแร็งค์อย่างง่าย

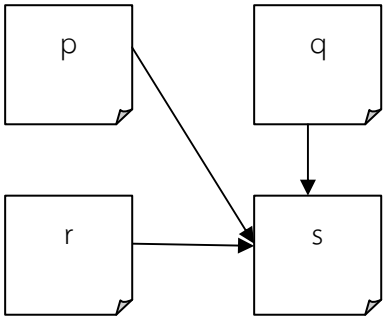
$$r(q)=\sum_{p(p,q)\in E}\frac{r(p)}{O(p)} \tag{3}$$

สมการที่ 3 ให้ $r(q)$ เป็นค่าคะแนนเพจแร็งค์ของเพจ $q$ และ $r(p)$ เป็นค่าคะแนนเพจแร็งค์ของเพจ $p$ ที่มีเส้นเชื่อมโยงเข้าหาเพจ $q$ ส่วน $O(p)$ คือจำนวนเส้นเชื่อมโยงที่ชี้ออกจากเพจ $q$ ดังนั้นหากเว็บเพจใดถูกอ้างอิงถึงมากก็ย่อมมีค่าเพจแร็งค์มากด้วยนั่นเอง



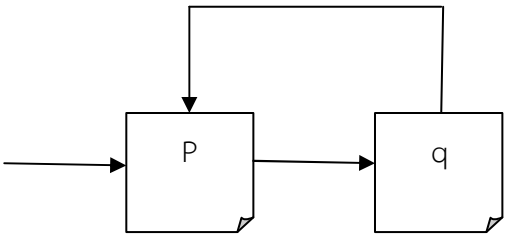

รูปที่ 2.4 ตัวอย่างการการส่งผ่านค่าเพจแร็งค์อย่างง่าย

จากสมการที่ 3 เพจแร็งค์แบบง่ายพบว่ามีปัญหาเกี่ยวกับการคำนวณค่าเพจแร็งค์อยู่สองปัญหาคือ แร็งค์ลีก (rank leak) และ แร็งค์ซิงก์ (rank sink)




รูปที่ 2.5 ตัวอย่างการเกิดปัญหาแร็งค์ลีก

พิจารณาจากรูปที่ 2.5 เว็บเพจใดที่มีเส้นเชื่อมโยงชี้เข้ามาแต่ไม่มีการสร้างเส้นเชื่อมโยงชี้ออกไปยังเพจอื่นเลยทำให้เพจนี้จะไม่มีการส่งผ่านค่าเพจแร็งค์ต่อไปยังเพจอื่น ซึ่งเหตุการณ์นี้เรียกว่าแร็งค์ลีก และเพจที่ไม่มีการส่งผ่านค่าเพจแร็งค์หรือมีเส้นเชื่อมโยงชี้ออกไปจะถูกเรียกว่าแด็งกลิ้งเพจ (dangling page) ซึ่งจะทำให้ค่าเพจแร็งค์ที่ถูกส่งผ่านมายังแด็งกลิ้งเพจหายไปจากระบบ เพราะจะไม่มีการส่งผ่านค่าเพจแร็งค์ออกไปจากเพจนี้ ทำให้ในการคำนวณค่าเพจแร็งค์ในแต่ละรอบจะทำให้ค่าเพจแร็งค์ของเพจโดยรวมต่ำลง




รูปที่ 2.6 ตัวอย่างการเกิดปัญหาแร็งค์ซิงก์

พิจารณาจากรูปที่ 2.6 เว็บเพจสองเพจที่สร้างเส้นเชื่อมโยงเข้าหากันแต่ไม่มีเส้นเชื่อมโยงออกไปยังเพจอื่นเลย และสมมติว่ามีการเส้นเชื่อมโยงเข้ามาหาเพจใดเพจหนึ่งในกลุ่มนี้ ดังนั้นในขณะที่ทำการคำนวณซ้ำค่าเพจแร็งค์ของเพจก็จะทำให้เกิดการวนซ้ำซึ่งจะไม่มีการส่งผ่านค่าเพจแร็งค์ออกมาจากเว็บเพจกลุ่มนี้เลยเนื่องจากไม่มีเส้นเชื่อมโยงชี้ออกไปยังเพจอื่น ซึ่งเหตุการณ์นี้เรียกว่าแร็งค์ซิงก์ จะทำให้เว็บที่เป็นกลุ่มแร็งค์ซิงก์มีค่าเพจแร็งค์สูง เนื่องจากไม่มีการส่งค่าคะแนนออกจากกลุ่มเลย

จึงได้มีการนำเสนอแร็งค์ซอร์ส (rank source) ขึ้นมาเพื่อแก้ปัญหาเหล่านี้ โดยในการแก้ปัญหาแร็งค์ลีกได้ใช้แนวคิดของการเดินแบบสุ่ม (random walk) บนกราฟแบบมีทิศทาง โดยจะทำการสร้างเส้นเชื่อมโยงเทียมจากแด็งกลิ้งเพจอ้างอิงไปยังทุกเพจ ดังนั้นค่าความน่าจะเป็นในการเปลี่ยนเพจจากแด็งกลิ้งเพจไปยังเพจอื่นจะเท่ากับ $\frac{1}{N}$ ซึ่งวิธีนี้จะทำให้เกิดการส่งผ่านค่าเพจแร็งค์ออกจากแด็งกลิ้งเพจได้ ส่วนการแก้ปัญหาแร็งค์ซิงก์ได้ใช้แนวคิดของการกระโดดแบบสุ่ม (random

jump) บนกราฟแบบมีทิศทาง กล่าวคือค่าความน่าจะเป็นในการกระโดดไปยังเพจใดบนกราฟจะมีค่าเท่ากัน ซึ่งโดยทั่วไปค่าความน่าจะเป็นที่จะกระโดดไปยังเพจอื่นหรืออยู่ที่เพจเดิมจะมีค่าเท่ากับ $\frac{1}{N}$ โดยจะทำการเพิ่มค่าความน่าจะเป็นในการกระโดดนี้เข้าไปในทุกเพจ ซึ่งจะทำให้สามารถแก้ปัญหาแร็งค์ซิงก์ได้

จากการแก้ปัญหาที่กล่าวมาทั้งสองวิธี ทำให้สามารถเขียนสมการที่ 3 ได้ใหม่ดังนี้

$$r(q) = \alpha \sum_{p:(p,q)\in E} \frac{r(p)}{O(p)} + (1-\alpha)\frac{1}{N} \tag{4}$$

เมื่อสัมประสิทธิ์ $\alpha$ คือตัวประกอบการเสื่อมสลาย (decay factor) หรืออาจจะกล่าวได้ว่าเป็นค่าถ่วงน้ำหนักระหว่างคะแนนที่มาจากเส้นเชื่อมโยงเข้าของเว็บเพจและคะแนนที่มาจากการกระโดดแบบสุ่ม ซึ่งมักถูกกำหนดให้มีค่าเท่ากับ 0.85 [*11*] และ $N$ คือจำนวนของเพจทั้งหมดที่มีบนเว็บกราฟ นอกจากนี้การคำนวณเพจแร็งค์ดังกล่าวยังสมมูลกันกับการหาไอเกนเวกเตอร์หลัก (principal eigenvector) $\vec{r}$ ของปัญหา $\vec{r} = \lambda A\vec{r}$ ในระบบไอเกน (eigensystem) บนทรานซิชันเมทริกซ์ $A$ ด้วยค่าไอเกน (eigenvalue) $\lambda$ = 1 [*9*] เมื่อ $\vec{r}$ คือเวกเตอร์แทนค่าเพจแร็งค์ของทุกเว็บเพจ ดังนั้นจากสมการที่ 4 สามารถถูกแสดงได้ใหม่เป็นดังนี้

$$\vec{r} = \alpha A\vec{r} + (1-\alpha)\vec{d} \tag{5}$$

โดยที่ $\vec{d}$ คือค่าคะแนนคงที่ของแต่ละเว็บเพจที่ถูกคำนวณส่งต่อมาจากเว็บเพจทั้งหมด ซึ่งในกรณีของอัลกอริทึมเพจแร็งค์แบบดั้งเดิม ค่านี้จะถูกกำหนดให้เป็นค่าคะแนนกระจายแบบเท่ากัน (uniform distribution) นั่นคือ $[1/N]_{N\times 1}$ นอกจากนี้ตามสมการที่ 5 ยังสามารถใช้วิธีการเพาเวอร์ (power method) วิธีการจาโคบี (Jacobi method) ในการคำนวณ [*9*] ได้

ดังที่ได้กล่าวมาแล้วว่าค่าไอเกนของการคำนวณเพจแร็งค์มีค่าเท่ากับ 1 และทรานซิชันเมทริกซ์ $A$ เป็นเมทริกซ์ด้านเท่าและผลรวมแต่ละคอลัมน์เท่ากับ 1 (column-stochastic matrix) ดังนั้นจึงสามารถรับประกันได้ว่าวิธีการเพาเวอร์คำนวณค่าเพจแร็งค์ออกมาเป็นบวกเสมอ วิธีการเพาเวอร์เป็นวิธีการคำนวณหาค่าประมาณของค่าไอเกนเวกเตอร์ของเมทริกซ์ด้านเท่าซึ่งในที่นี้หมายถึงเวกเตอร์คะแนนเพจแร็งค์ของทรานซิชันเมทริกซ์ขนาด $n\times n$ โดยมีการคำนวณเป็นรอบ ดังนั้นความซับซ้อนเชิงเวลาของการคำนวณเพจแร็งค์ในแต่ละรอบจึงเท่ากับ $O(n^2)$ ซึ่งหมายความว่าความซับซ้อนเชิงเวลาในการคำนวณคะแนนเพจแร็งค์จะขึ้นอยู่กับจำนวนของเว็บเพจบนเว็บกราฟ แต่อย่างไรก็ตามในงานวิจัย [*11*] ได้เสนอวิธีการคำนวณเพจแร็งค์ด้วยเทคนิคนาอีฟ (Naive technique) ตัวอย่างรหัสเทียมในการคำนวณค่าเพจแร็งค์แสดงดังรูปที่ 2.7

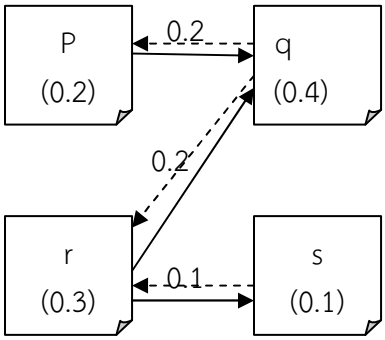
$$\forall_s Source[s] = \frac{1}{N}$$
$$while(residual > \tau)\{$$
$$\quad \forall_d Dest[d] = 0$$
$$\quad while(not\ \ Links.eof())\{$$
$$\quad\quad Links.read(source, o(source), dest_1, dest_2, ..., dest_n)$$
$$\quad\quad for\ \ j = 1\ \ to\ \ o(source)$$
$$\quad\quad\quad Dest[dest_j] = Dest[dest_j] + \frac{Source[source]}{o(source)}$$
$$\quad \}$$
$$\quad \forall_d Dest[d] = \alpha \times Dest[d] + (1-\alpha)\frac{1}{N}\ \ /*uniform\ \ vector\ \ or\ \ personalized\ \ vector*/$$
$$\quad residual = \|Source - Dest\|\ \ /*recompute\ \ theshold*/$$
$$\quad Source = Dest$$
$$\}$$

รูปที่ 2.7 ตัวอย่างรหัสเทียม (pseudo code) ของอัลกอริทึมเพจแร็งค์

จากรูปที่ 2.7 ตัวอย่างรหัสเทียมในการคำนวณค่าเพจแร็งค์ [*11*] โดย $Source[\,]$ และ $Dest[\,]$ จะเป็นเวกเตอร์สำหรับการเก็บค่าคะแนนเพจแร็งค์ของทุกเว็บเพจต้นทางและทุกเว็บเพจปลายทาง ตามลำดับ ในการคำนวณแต่ละรอบ $Links$ จะหมายถึงเว็บกราฟ เมื่อพิจารณาแต่ละเพจต้นทาง $source$ ซึ่งมีจำนวนเส้นเชื่อมโยงออกเป็น $o(source)$ โดยที่เชื่อมโยงไปยังแต่ละเพจปลายทาง $dest_j$ ดังนั้นสำหรับเพจต้นทางหนึ่งๆ จะมีความซับซ้อนเชิงเวลาของการคำนวณเพจแร็งค์เป็น $O(o(source))$ และหากพิจารณาทุกเพจต้นทาง ความซับซ้อนเชิงเวลาของการคำนวณเพจแร็งค์จะเท่ากับ $O(\sum_{i=1}^{N} o(source)) = \Theta(|\ Links\ |)$ กล่าวคือจำนวนเส้นเชื่อมโยงทั้งหมดนั่นเองคิดต่อหนึ่งรอบการคำนวณ และสำหรับ $residual$ คือค่าตัวแปรค่าขีดแบ่งสำหรับกำหนดจำนวนรอบในการคำนวณคะแนนเพจแร็งค์

### 2.1.4. อัลกอริทึมอินเวิร์สเพจแร็งค์ (Inverse PageRank Algorithm)

อินเวิร์สเพจแร็งค์เป็นรูปแบบการคำนวณเพจแร็งค์อีกประเภทหนึ่งที่ถูกนำเสนอในส่วนหนึ่งของงานวิจัย [*10*] ซึ่งจากการทดลองได้สังเกตพบว่าเพจที่ดีย่อมมีโอกาสน้อยที่จะชี้ไปยังเพจที่ไม่ดี ซึ่งก็มีเหตุผลเนื่องจากเพจที่ไม่ดีจะทำให้เกิดการจัดลำดับที่ไม่ถูกต้องด้วยและไม่ให้ข้อมูลที่เป็นประโยชน์กับผู้ใช้ ดังนั้นผู้ดูแลเว็บเพจที่ดีย่อมไม่มีเหตุผลที่จะสร้างเพจที่ชี้ไปยังเพจที่ไม่ดี ดังนั้นเว็บเพจใดก็ตามที่ชี้ไปยังเว็บเพจที่ดีหลายเพจก็เป็นไปได้ว่าจะเป็นเพจที่ดีตามไปด้วย โดยจะเห็นว่าแนวคิดนี้จะเป็นการมองย้อนกลับของการส่งผ่านคะแนนของเพจแร็งค์ดังรูปที่ 2.8 โดยเส้นประหมายถึงการส่งผ่านคะแนนจากเพจที่ถูกอ้างอิงไปยังเพจที่อ้างอิง




รูปที่ 2.8 ตัวอย่างการส่งผ่านค่าอินเวิร์สเพจแร็งค์อย่างง่าย

ซึ่งแนวคิดของอินเวิร์สเพจแร็งค์นั้นจะตรงกันข้ามกับเพจแร็งค์ กล่าวคือ อินเวิร์สเพจแร็งค์จะคำนวณค่าคะแนนแบบย้อนกลับซึ่งกลับกับเพจแร็งค์ปกติ นั่นคือถ้าสมมติให้เว็บเพจ $p$ มีเส้นเชื่อมโยงไปยังเว็บเพจ $q$ และสมมติให้เว็บเพจ $q$ มีค่าอินเวิร์สเพจแร็งค์ (inverse PageRank score) ที่สูงแล้ว เว็บเพจ $p$ นั้นก็ควรจะมีค่าอินเวิร์สเพจแร็งค์ที่สูงตามด้วย เนื่องจากเว็บเพจ $p$ ได้ชี้หรืออ้างอิงไปยังเว็บเพจที่ดีนั่นเอง หรืออาจกล่าวได้อีกนัยหนึ่งว่า เว็บเพจใดจะมีค่าอินเวิร์สเพจแร็งค์สูงก็ต่อเมื่อมีเส้นเชื่อมโยงออกไปยังเว็บเพจอื่นจำนวนมาก และหรือเชื่อมโยงไปยังเว็บเพจที่มีค่าอินเวิร์สเพจแร็งค์สูง

อัลกอริทึมอินเวิร์สเพจแร็งค์มีความคล้ายคลึงกับอัลกอริทึมเพจแร็งค์ดั้งเดิม เพียงแต่แตกต่างกันตรงที่จะใช้อินเวิร์สทรานซิชันเมทริกซ์ $A$ ดังสมการที่ 2 แทนการใช้ทรานซิชันเมทริกซ์ $B$ ตามปกติ ถ้ากำหนดให้ $\vec{s}$ เป็นเวกเตอร์แทนค่าอินเวิร์สเพจแร็งค์ของทุกเว็บเพจแล้ว อัลกอริทึมอินเวิร์สเพจแร็งค์สามารถคำนวณได้ดังสมการต่อไปนี้

$$\vec{s} = \alpha B\vec{s} + (1-\alpha)\vec{d} \qquad (6)$$

เช่นเดียวกับสมการที่ 5 ตัวแปร $\alpha$ และ $\vec{d}$ คือค่าสัมประสิทธิ์ของความเสื่อมสลายและเวกเตอร์การกระจายตัวของคะแนน ตามลำดับ ซึ่งสามารถกำหนดให้เป็นค่าเดียวกันกับที่ใช้ในอัลกอริทึมเพจแร็งค์แบบดั้งเดิม

### 2.1.5. เพจแร็งค์ส่วนบุคคล (Personalized PageRank)

จากสมการที่ 5 เวกเตอร์ $\vec{d}$ ที่เพิ่มเข้ามาเพื่อแก้ปัญหาแร็งค์ซิงก์ จากการทดลองในงานวิจัยที่ [*18*] พบว่าเวกเตอร์นี้เป็นตัวแปรที่มีผลต่อค่าคะแนนเพจแร็งค์เช่นกัน โดยค่าความน่าจะเป็นของเว็กเตอร์นี้จะหมายถึงพฤติกรรมที่ผู้ใช้จะเลือกเปลี่ยนเพจไปยังเพจไหนก็ได้ ดังนั้นเราจึงสามารถให้ค่าคะแนนกระจายแบบเท่ากันทุกเพจหรือให้ค่าคะแนนกระจายลำเอียง (bias) ตามความสนใจส่วนบุคคลของผู้ใช้ (personalized) ได้

ซึ่งงานวิจัยก่อนหน้า [*4, 15, 12*] รวมถึงงานวิจัยนี้ ได้นำเสนอวิธีการหาค่าคะแนนกระจายแบบไม่เท่ากันกัน (non-uniform distribution) เพื่อกำหนดความลำเอียงให้กับบางเว็บเพจ ซึ่งวิธีการนี้ถูกเรียกว่าการทำเพจแร็งค์ส่วนบุคคล (personalized PageRank) เนื่องจากเป็นการให้ค่า

ความลำเอียงในแต่ละเพจแบบไม่เท่ากัน ซึ่งวิธีการหาค่าความลำเอียงแต่ละแบบจะนำเสนอต่อไปในส่วนของงานวิจัยที่เกี่ยวข้อง

## 2.2. งานวิจัยที่เกี่ยวข้อง

ถึงแม้ว่าอัลกอริทึมเพจแร็งค์ดั้งเดิม [*18*] จะมีประสิทธิภาพและประสบความสำเร็จในการประยุกต์ใช้งาน ในหลากหลายสาขา แต่ก็มีงานวิจัยจำนวนมากได้ศึกษาและพยายามพัฒนาปรับปรุงเพจแร็งค์ร่วมกับปัจจัยในหลายแง่มุม

### 2.2.1. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search

Topic-sensitive PageRank [*12*] นำเสนอวิธีการลำเอียงเพจแร็งค์ด้วยมุมมองในเชิงหัวเรื่องของเนื้อหา (topic of content) ซึ่งการแบ่งหัวเรื่องในงานวิจัยนี้ได้ใช้การแบ่งกลุ่มจากโอดีพี ซึ่งแบ่งออกเป็นสิบหกกลุ่มหัวเรื่อง โดยลำดับแรกให้คำนวณกลุ่มของเวกเตอร์เพจแร็งค์ส่วนบุคคลด้วยการใช้หัวเรื่อง (precomputed) ก่อนซึ่งจะได้เวกเตอร์ความลำเอียงเพจแร็งค์ส่วนบุคคลจำนวนสิบหกเวกเตอร์ และหลังจากนั้นจึงนำเวกเตอร์ความลำเอียงไปคำนวณค่าคะแนนความสำคัญของเพจแร็งค์ตามคำค้นอีกที

เทคนิคการลำเอียงด้วยการใช้หัวเรื่องทำได้โดย ให้ $i$ เป็นยูอาร์แอลของเว็บเพจ และ $T_j$ เป็นจำนวนของเว็บเพจยูอาร์แอลที่อยู่ในหมวดมู่ $j$ ซึ่งนำมาจากโอดีพี ให้เวกเตอร์ $\vec{d}$ แทนด้วยเวกเตอร์เพจแร็งค์ส่วนบุคคล และให้เวกเตอร์ $\vec{v}_j$ แทนการลำเอียงเพจแร็งค์ส่วนบุคคลด้วยหัวเรื่องในหมวดหมู่ $j$ ซึ่งสามารถแทนเวกเตอร์ $\vec{d}$ ด้วยเวกเตอร์ $\vec{v}_j$ ได้และเวกเตอร์ $\vec{v}_j$ คำนวณได้ดังสมการที่ 7

$$\vec{v}_{ji} = \begin{cases} \dfrac{1}{|T_j|}, & i \in T_j \\ 0, & i \notin T_j \end{cases} \tag{7}$$

โดยจะนำมาคำนวณหาค่าเพจแร็งค์จากเวกเตอร์ทั้งหมด 16 หัวเรื่องจากที่ได้กล่าวมาแล้ว และคำนวณค่าเพจแร็งค์ปกติด้วยเพื่อที่จะใช้ในการเปรียบเทียบ หลังจากนั้นในขณะที่ผู้ใช้ทำการค้นหา (query-time) จะมีการทำการจัดเรียงลำดับด้วยผลรวมของค่าเพจแร็งค์ในแต่ละเวกเตอร์หัวเรื่องตามความน่าจะเป็นของคำค้น ว่ามีความน่าจะเป็นของคำค้นที่ผู้ใช้ใช้ค้นหาว่าจะอยู่หัวเรื่องไหนในสิบหกหัวเรื่องที่ได้ทำการแบ่งไว้ ค่าคะแนนสุดท้ายสำหรับการจัดเรียงลำดับ $s_{qd}$ (query-sensitive importance score) สามารถคำนวณได้ดังสมการที่ 8

$$s_{qd} = \sum_j P(c_j \mid q') \cdot rank_{jd} \tag{8}$$

โดย $P(c_j|q')$ คือค่าความน่าจะเป็นของคำค้น $q'$ ที่จะอยู่ในเอกสาร $d$ ที่อยู่ในหมวด $c_j$ และ $rank_{jd}$ คือค่าคะแนนเพจแร็งค์ของเอกสาร $d$ ที่ลำเอียงด้วยเวกเตอร์หัวเรื่อง $j$ ซึ่งจากผลการทดลองงานวิจัยชิ้นนี้ได้แสดงให้เห็นว่าสามารถคำนวณการจัดลำดับของเพจได้ดีขึ้นเมื่อเทียบกับการคำนวณจากเพจแร็งค์แบบดั้งเดิม ในแง่ของความเหมือน (similarity) คือมาตรวัดเคซิม (Ksim) และโอซิม (Osim)

### 2.2.2. A framework to compute page importance based on user behaviors

งานวิจัยนี้ [16] เกี่ยวกับการเรียนรู้จากพฤติกรรมของผู้ใช้งาน (users' behavior) ผ่านแฟ้มบันทึกเหตุการณ์ (log file) ซึ่งแตกต่างจากเพจแร็งค์แบบดั้งเดิมที่ใช้การคำนวณค่าคะแนนความสำคัญจากกราฟโครงสร้างเส้นเชื่อมโยง โดยผู้วิจัยนำเสนอแนวคิดที่ว่าข้อมูลของเส้นเชื่อมโยงไม่เพียงพอที่จะนำมาคำนวณค่าคะแนนความสำคัญของเพจได้อีกต่อไป ดังนั้นจึงได้นำพฤติกรรมของผู้ใช้งานจริงมาเป็นข้อมูลแทน

โดยได้นำเสนอวิธีการดังนี้ อันดับแรกวิเคราะห์การใช้งานของผู้ใช้จากบันทึกเหตุการณ์เพื่อออกแบบจำลอง และเมื่อได้แบบจำลองแล้วจึงนำมาสร้างอัลกอริทึมเบราส์แร็งค์ (BrowseRank) ที่ใช้สำหรับการคำนวณค่าคะแนนความสำคัญ

| | | |
|---|---|---|
| http://aaa.bbb.com/ | 2009-01-01, 21:33:05 | INPUT |
| http://aaa.bbb.com/1.htm | 2009-01-01, 21:34:11 | CLICK |
| http://ccc.ddd.org/index.htm | 2009-01-01, 21:34:52 | INPUT |
| http://eee.fff.edu/ | 2009-01-01, 21:39:03 | CLICK |
| – | – | – |
| http://eee.fff.edu/ | 2009-01-01, 22:00:45 | CLICK |

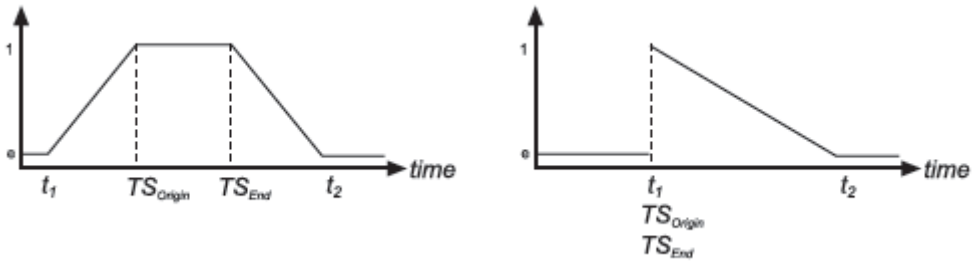
รูปที่ 2.9 ตัวอย่างแฟ้มบันทึกพฤติกรรมผู้ใช้

จากรูปที่ 2.9 แฟ้มบันทึกเหตุการณ์ที่นำมาวิเคราะห์จะประกอบด้วย ยูอาร์แอล เวลา และพฤติกรรม โดยพฤติกรรมจะแยกออกเป็น สองแบบคือ "INPUT" หมายถึงการเปลี่ยนเพจโดยใช้ยูอาร์แอลโดยตรง และ "CLICK" หมายถึงการเปลี่ยนเพจด้วยการตามเส้นเชื่อมโยง โดยจากแฟ้มบันทึกเหตุการณ์จะนำมาสกัดข้อมูลการเปลี่ยนจากเพจหนึ่งไปยังอีกเพจหนึ่งและเวลาที่ใช้ในแต่ละเพจ ข้อมูลทั้งสองแบบนี้จะถูกนำมาสร้างเป็นเว็บกราฟพฤติกรรมของผู้ใช้ (user browsing graph) และนำไปคำนวณเป็นคะแนนความสำคัญของเว็บเพจ

งานวิจัยนี้ได้ทดลองโดยนำข้อมูลพฤติกรรมของผู้ใช้จากเครื่องมือค้นหาเชิงพาณิชย์สกัดยูอาร์แอลได้ประมาณหนึ่งพันล้านยูอาร์แอล และอัลกอริทึมเปรียบเทียบคืออัลกอริทึมเพจแร็งค์แบบดั้งเดิม และอัลกอริทึมเพจแร็งค์ร่วมกับพฤติกรรมของผู้ใช้ โดยใช้มาตรวัดความแม่นยำ (Precision) ผลลัพธ์ค้นคืนในสามและห้าอันดับแรก กับมาตรวัดเอ็นดีซีจี (NDCG) ผลลัพธ์ค้นคืนในสามและห้าอันดับ ปรากฏว่าอัลกอริทึมเบราส์แร็งค์ให้คะแนนดีกว่าอัลกอริทึมเพจแร็งค์ทั้งสองแบบ

### 2.2.3. Time-aware authority ranking

Time-aware authority ranking [3] มีแนวคิดว่าผู้ใช้งานย่อมสนใจในเพจที่มีค่าคะแนนความสำคัญมากแต่ก็คำนึงถึงความทันสมัยของข้อมูลด้วยดังนั้น เวลาซึ่งหมายถึงความทันสมัยของข้อมูลเพจรวมถึงโครงสร้างเส้นเชื่อมโยงจึงเป็นปัจจัยสำคัญในการคำนวณค่าคะแนนความสำคัญของเพจ ในงานวิจัยชิ้นนี้ได้นำเสนออัลกอริทึมทีแร็งค์ (T-Rank) และอัลกอริทึมทีแร็งค์แบบเบา (T-Rank light) ซึ่งได้นำข้อมูลทางด้านเวลาเข้ามาใช้สองตัวแปรคือ ตัวแปรที่ 1 ความทันสมัยของข้อมูล (freshness) เช่น เวลาในการสร้าง เวลาในการแก้ไขข้อมูล และตัวแปรที่ 2 กิจกรรม (Activity) เช่น ความถี่ในการปรับปรุงเพจ โดยมีแนวคิดในเรื่องช่วงเวลาความสนใจ (temporal interest) ดังรูปที่ 2.9




รูปที่ 2.10 แสดงตัวอย่างแนวคิดช่วงเวลาความสนใจของผู้ใช้

จากรูปที่ 2.10 แสดงถึงค่าความทันสมัยเมื่อผู้ดูแลเว็บเพจมีการปรับปรุงเว็บเพจในแต่ละช่วงเวลาของความสนใจ โดยที่ $TS_{origin}$ คือจุดเวลาเริ่มต้นเหตุการณ์ $TS_{end}$ เป็นจุดเวลาสิ้นสุดเหตุการณ์ โดยที่ $TS_{origin} \le TS_{end}$ และให้ $t_1$ และ $t_2$ เป็นจุดเวลาสนใจ โดยที่ $t_1 \le TS_{origin} \le TS_{end} \le t_2$ จากรูปจะเห็นได้ว่าความสนใจจะแบ่งได้เป็นสองกรณีคือ กราฟทางซ้ายหมายถึงเหตุการณ์ที่เราสนใจเราสามารถคาดการณ์ได้ว่าจะเกิดขึ้นและสิ้นสุดเมื่อไหร่เช่น การจัดการแข่งขันกีฬาโอลิมปิก ค่าความทันสมัยจะเพิ่มขึ้นทีละน้อยจากจุดเวลาเริ่มต้นสนใจ $t_1$ จนไปสูงสุดที่เวลาเวลาที่เกิดเหตุการณ์ $TS_{origin}$ จนถึงเวลาสิ้นสุดเหตุการณ์ $TS_{end}$ และค่าความทันสมัยจะเริ่มลดลงจนถึงจุดเวลาสิ้นสุดความสนใจ $t_2$ ส่วนกราฟทางขวา หมายถึงเหตุการณ์ที่สนใจเกิดขึ้นแบบไม่สามารถคาดการณ์ได้ เช่น เหตุการณ์แผ่นดินไหว ดังนั้นค่าความทันสมัยจะมีค่าสูงสุดที่เวลา $t_1, TS_{origin}, TS_{end}$ เนื่องจากเป็นจุดเวลาเริ่มต้นสนใจ เริ่มต้นเหตุการณ์ และสิ้นสุดเหตุการณ์โดยเป็นจุดเวลาเดียวกันทั้งหมด และค่าความทันสมัยจะเริ่มลดลงทันทีจนถึงจุดเวลาสิ้นสุดความสนใจ $t_2$

การสกัดความทันสมัยของเว็บเพจ กำหนดให้ $f$ เป็นฟังกชันความทันสมัย $ts$ เป็นเวลาการปรับปรุงข้อมูลของเว็บเพจ คำนวณค่า f ได้ดังสมการที่ 9 โดยให้ค่า $e > 0$ เพื่อป้องกันค่า $f$ เป็น 0

$$f(ts) = \begin{cases} \frac{1-e}{TS_{origin} - t_1} \cdot (ts - t_1) + e, & t_1 \le ts \le TS_{origin} \\ 1, & TS_{origin} \le ts \le TS_{end} \\ \frac{e-1}{t_2 - TS_{end}} \cdot (ts - TS_{end}) + 1, & TS_{end} \le ts \le t_2 \\ e, & otherwise \end{cases} \quad (9)$$

นำฟังก์ชัน $f$ มาใช้กับเว็บเพจ $p$ จะได้ว่า $ts$ แทนด้วยจุดเวลาการปรับปรุงเว็บเพจหรือเส้นเชื่อมโยง $TS_{\mathrm{mod}}$ หรือจุดเวลาเว็บเพจหรือเส้นเชื่อมโยงถูกสร้าง $TS_{creation}$ ส่วนในการหาค่ากิจกรรมของเว็บเพจซึ่งหมายถึงความถี่ในการปรับปรุงเว็บเพจหรือเส้นเชื่อมโยงจะแทนด้วยฟังก์ชัน $a(p)$ ดังสมการที่ 10 โดยที่การเปลี่ยนแปลงนั้นต้องอยู่ในช่วง $t_1$ และ $t_2$

$$a(p) = \sum f(ts) \quad with \quad ts \in \quad (TS_{\mathrm{mod}(p)} \cap [t_1, t_2]) \cup \{TS_{creation(p)}\} \quad (10)$$

หลังจากนั้นจะนำค่าความทันสมัยและกิจกรรมมาช่วยในการคำนวณเพจแร็งค์ โดยได้นำเสนอแบ่งออกเป็น สองอัลกอริทึมคือ ทีแร็งค์ไลท์ ซึ่งเป็นการนำค่าความทันสมัยและค่ากิจกรรมไปแทนในเวกเตอร์ $\vec{d}$ ในการคำนวณเพจแร็งค์เพียงอย่างเดียว ส่วนอีกหนึ่งอัลกอริทึมคือ ทีแร็งค์ ซึ่งเป็นการนำค่าความทันสมัยและค่ากิจกรรมมาใช้ในการคำนวณทั้งในส่วนเวกเตอร์ $\vec{d}$ และ ทรานซิชันเมทริกซ์ $A$ ตามสมการของเพจแร็งค์ที่ได้กล่าวไว้แล้ว

งานวิจัยนี้ทดลองบนฐานข้อมูลเว็บเพจที่เก็บจากเว็บเพจที่เกี่ยวข้องกับโอลิมปิกปี ค.ศ. 2004 ประมาณสองแสนเว็บเพจ และใช้อาสามัครในการให้คะแนนการจัดเรียงลำดับผลลัพธ์ค้นคืนในสิบอันดับแรก ในแต่ละคำค้นและหลังจากนั้นจึงนำมาเปรียบเทียบกับการจัดลำดับในสิบอันดับแรกด้วย เพจแร็งค์ ทีแร็งค์ และทีแร็งค์ไลท์ ในแง่ของความเหมือนทั้งเคซิมและโอซิม พบว่าค่าคะแนนความเหมือนของการจัดลำดับผลลัพธ์ค้นคืนในสิบอันดับแรกด้วยอัลกอริทึมทีแร็งค์และทีแร็งค์ไลท์ดีกว่าเพจแร็งค์ในเกือบทุกคำค้น ทำให้สรุปได้ว่าการนำค่าความทันสมัยและค่ากิจกรรมมาช่วยในการจัดเรียงลำดับผลลัพธ์ค้นหา ช่วยให้การจัดเรียงเป็นที่ถูกใจผู้ใช้งานมากกว่าการจัดเรียงด้วยเพจแร็งค์ธรรมดา

### 2.2.4. Adding the temporal dimension to search - A case study in publication search

TimedPageRank [*20*] ได้นำมุมมองด้านเวลามาผนวกใช้กับอัลกอริทึมเพจแร็งค์ได้นำเสนออัลกอริทึมไทม์เพจแร็งค์ (TimedPageRank) ซึ่งมีแนวคิดในการแบ่งตัวแปรที่มีผลต่อค่าคะแนนความสำคัญออกเป็นสองกลุ่มคือ ตัวแปรเนื้อหา (content factor) และตัวแปรความมีชื่อเสียง (reputation factor) ตัวแปรเนื้อหาคือ ตัวแปรที่เป็นส่วนของเนื้อหาในเอกสารที่มีดีกรีความเกี่ยวข้องกับคำค้นของผู้ใช้ ส่วนตัวแปรความมีชื่อเสียงคือ ตัวแปรที่ช่วยในการจัดลำดับผลลัพธ์ค้นคืนของเอกสารที่เกี่ยวข้องกับคำค้น ซึ่งในงานวิจัยนี้ได้ยกตัวอย่างการค้นคืนบทความวิจัยทางวิทยาศาสตร์ ซึ่งหมายถึงจำนวนการอ้างอิง ความมีชื่อเสียงของผู้แต่ง และความมีชื่อเสียงของวารสารที่ตีพิมพ์ ในงานวิจัยนี้จะใช้เฉพาะตัวแปรความมีชื่อเสียง และได้ศึกษาถึงมุมมองด้านเวลา (วันที่

ตีพิมพ์งานวิจัย, วันที่มีการอ้างอิงถึงงานวิจัย) ที่มีผลต่อตัวแปรความมีชื่อเสียงซึ่ง โดยได้ทดลองกับระบบการค้นหาของฐานข้อมูลบทความวิจัยทางวิทยาศาสตร์ ซึ่งงานวิจัยนี้ได้นำเสนอค่าถ่วงน้ำหนักที่เรียกว่าอัตราการเสื่อมสลาย (decay rate) คือค่าถ่วงน้ำหนักของงานวิจัยโดยวัดจากความแตกต่างณ เวลาปัจจุบันกับวันที่ตีพิมพ์ และได้นำมาใช้กับอัลกอริทึมเพจแร็งค์ หลังจากนั้นได้นำเสนอตัวแปรทันสมัย (trend factor) โดยสกัดจากการเปลี่ยนแปลงของจำนวนการอ้างอิงงานวิจัยและได้นำมาใช้ร่วมกับค่าคะแนนเพจแร็งค์ที่ใช้ค่าถ่วงน้ำหนักอัตราการเสื่อมสลายในแต่ละเอกสารงานวิจัย จึงได้นำเสนอเป็นอัลกอริทึมไทม์เพจแร็งค์ และจากการทดลองพบว่าทำให้มีประสิทธิภาพเพิ่มขึ้น

สำหรับงานวิจัยชิ้นนี้มีความแตกต่างจากงานวิจัยก่อนหน้า [*3*] อยู่สองประการคือประการแรก ปัจจัยในเชิงเวลา (temporal factor) ถูกกำหนดจากแบบจำลองความใกล้ชิดด้านเวลาแทนที่จะใช้ค่าความสดใหม่ (freshness) ของเว็บเพจและเส้นเชื่อมโยงที่ได้จากการพิจารณาจุดเวลาของการเปลี่ยนแปลงครั้งล่าสุด (last modification timestamp) โดยตรง ซึ่งแนวคิดของงานวิจัยอัลกอริทึมทีแร็งค์จะสนใจเฉพาะข้อมูลเชิงเวลาของเว็บเพจตนเองเท่านั้น ไม่สนใจข้อมูลเชิงเวลาของเว็บเพจที่ได้ทำการเส้นเชื่อมโยงอ้างอิงไว้ ทำให้อาจจะขาดข้อมูลสำคัญที่เกี่ยวข้องไป ประการที่สองค่าคะแนนความสำคัญ (authoritative score) สุดท้ายของเว็บเพจถูกคำนวณโดยใช้เวกเตอร์ความลำเอียงเชิงเวลาที่ได้จากอินเวิร์สเพจแร็งค์ แทนที่จะแก้ไขโดยตรงในทรานซิชันเมทริกซ์ด้วยมุมมองเชิงเวลา นอกจากนี้งานวิจัยนี้ยังแตกต่างจากงานวิจัยอัลกอริทึมไทม์แร็งค์ [*20*] เนื่องจากในงานวิจัยนี้ได้ทำการคำนวณบนฐานข้อมูลเว็บเพจจริงบนอินเตอร์เน็ต แต่งานวิจัยก่อนหน้านั้น ได้นำเสนอการทดลองบนฐานข้อมูลบทความวิจัยทางวิทยาศาสตร์ ซึ่งฐานข้อมูลทั้งสองมีความแตกต่างกันอยู่หลายประการ

# บทที่ 3

# วิธีดำเนินการวิจัย

ในบทนี้จะนำเสนออัลกอริทึมเพจแร็งค์ส่วนบุคคลตามความลำเอียงด้านเวลา โดยเราจะหาค่าความลำเอียงเชิงเวลามาใช้ลำเอียงในการคำนวณค่าเพจแร็งค์ของเว็บเพจ หรือจะกล่าวได้ว่านำค่าความลำเอียงเชิงเวลาของเว็บเพจมาใช้แทนค่าการกระจายตัวแบบเท่ากันของเวกเตอร์ $\vec{d}$ ในสมการเพจแร็งค์แบบดั้งเดิมนั่นเอง ซึ่งแบ่งได้เป็นสี่ขั้นตอน ได้แก่

ขั้นที่ 1. ออกแบบแบบจำลองเว็บพิจารณาบนแกนเวลา

ขั้นที่ 2. คำนวณค่าข้อมูลเชิงเวลาโดยใช้แบบจำลองความใกล้ชิดด้านเวลา

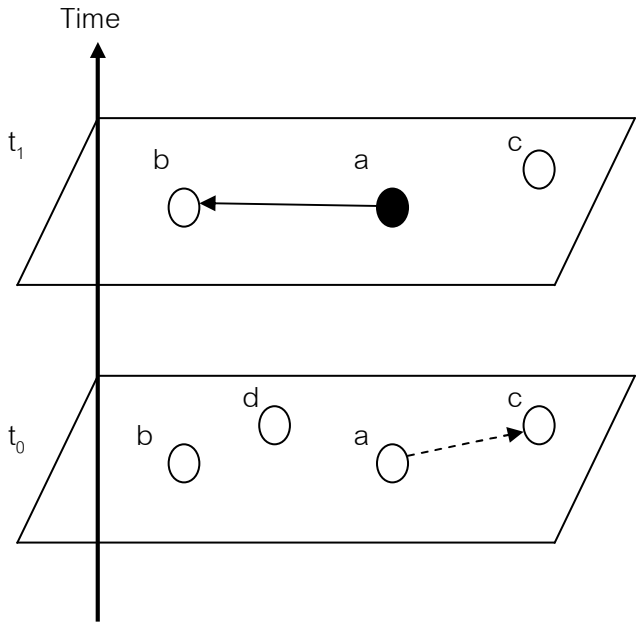
ขั้นที่ 3. ระบุเวกเตอร์ความลำเอียงเชิงเวลาโดยใช้อินเวิร์สเพจแร็งค์ และ

ขั้นที่ 4. คำนวณค่าคะแนนเพจแร็งค์โดยนำเวกเตอร์ความลำเอียงที่ได้จากขั้นตอนก่อนหน้ามาคำนวณในอัลกอริทึมเพจแร็งค์ดั้งเดิม

## 3.1. อัลกอริทึมเพจแร็งค์ส่วนบุคคลตามความลำเอียงด้านเวลา

### 3.1.1. แบบจำลองเว็บพิจารณาบนแกนเวลา

แบบจำลองเว็บที่นำเสนอนี้เป็นการจัดเก็บเว็บกราฟบนแกนเวลา ณ จุดเวลาที่แตกต่างกัน $\{t_0, t_1, ..., t_n\}$ ดังนั้นการเปลี่ยนแปลงของเว็บเพจพิจารณาได้จากความแตกต่างของเนื้อหาและเส้นเชื่อมโยงระหว่างสองจุดเวลาที่ติดกัน แสดงดังรูปที่ 3.1

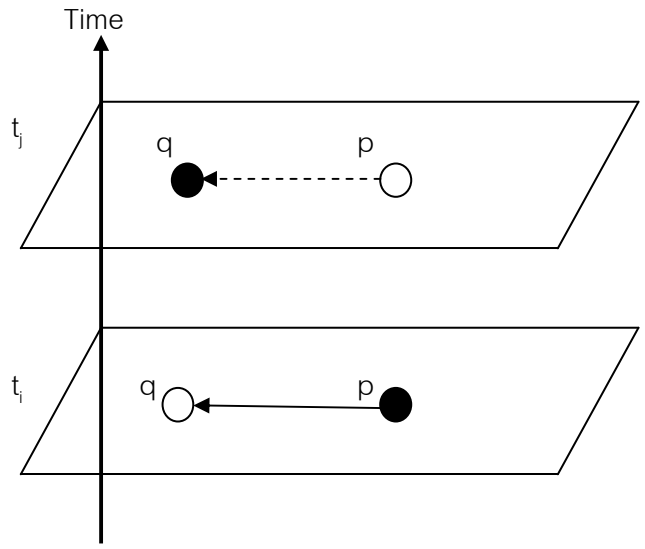



รูปที่ 3.1 ตัวอย่างแบบจำลองเว็บพิจารณาบนสองจุดเวลา

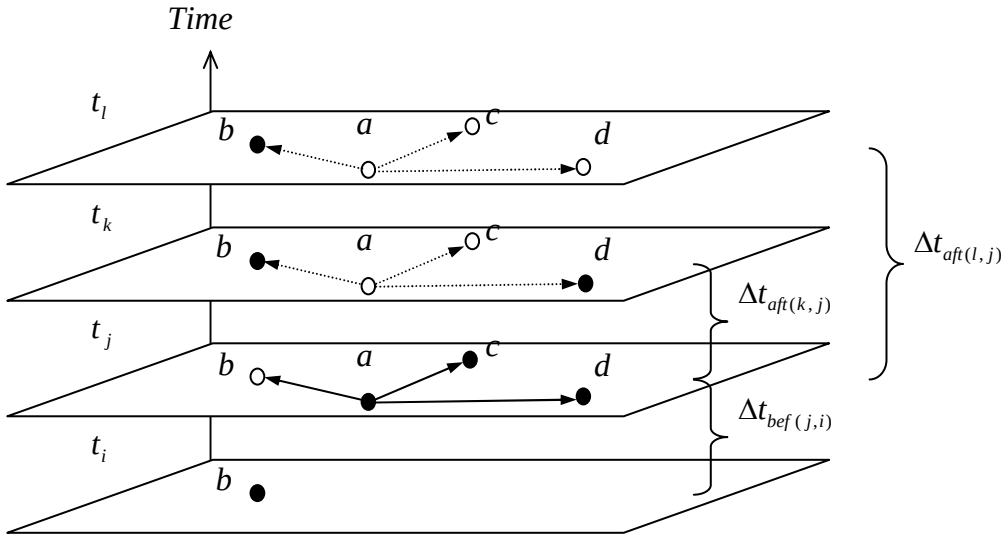
จากรูปที่ 3.1 ให้จุดยอด (vertex) และขอบแบบมีทิศทาง (directed edge) ระหว่างจุดยอดแทนหนึ่งเว็บเพจ และหนึ่งเส้นเชื่อมโยงระหว่างเว็บเพจ ตามลำดับ โดยที่จุดยอดทึบหมายถึงเว็บเพจมีการเปลี่ยนแปลง เช่น ถูกสร้างขึ้นใหม่ หรือถูกแก้ไขปรับปรุง ณ จุดเวลานั้น สำหรับจุดยอดโปร่งจะหมายถึงเว็บเพจไม่มีการเปลี่ยนแปลง ณ จุดเวลานั้น ในทำนองเดียวกัน เส้นขอบทึบหมายถึงเส้นเชื่อมโยงระหว่างเว็บเพจมีการเปลี่ยนแปลง และเส้นขอบประคือไม่มีการเปลี่ยนแปลง และถ้าหากเว็บเพจและเส้นเชื่อมโยงถูกลบทิ้งไปก็จะไม่ปรากฏในเว็บกราฟ ณ จุดเวลานั้น โดยการเปลี่ยนแปลงที่เราสนใจ ณ จุดเวลาใด คือการเปลี่ยนแปลงของเว็บเพจทั้งการปรับปรุงเนื้อหาหรือการเปลี่ยนแปลงเส้นเชื่อโยงอ้างอิงไปยังเว็บเพจอื่น จากรูปตัวอย่างที่ 3.1 เราสามารถบอกได้ว่าเว็บเพจ $a$ มีการเปลี่ยนแปลงที่จุดเวลา $t_1$ เมื่อเทียบกับจุดเวลา $t_0$ เนื่องจากมีการสร้างเส้นเชื่อมโยงชี้ไปยังเว็บเพจ $b$ และลบเส้นเชื่อมโยงที่อ้างอิงไปยังเว็บเพจ $c$ ส่วนเว็บเพจ $d$ ณ จุดเวลา $t_1$ เว็บเพจได้โดนลบทิ้งไปแล้ว ซึ่งในจุดนี้เราจะไม่พิจารณาอีกต่อไป

### 3.1.2. แบบจำลองความใกล้ชิดด้านเวลา (Time-Proximity model)

พิจารณาเว็บเพจ $p$ และเว็บเพจ $q$ ถ้าสมมติให้ ณ จุดเวลา $t_i$ ผู้ดูแลเว็บเพจ $p$ ได้สร้างเส้นเชื่อมโยงไปยังเว็บเพจ $q$ จากเหตุการณ์นี้อาจกล่าวได้ว่าที่เวลา $t_i$ เว็บเพจ $p$ มีความน่าสนใจมากกว่าเว็บเพจ $q$ เนื่องจากเว็บเพจ $p$ มีความเป็นปัจจุบัน (up-to-dateness) กว่า อย่างไรก็ดี เนื่องจากธรรมชาติของเว็บมักมีการเปลี่ยนแปลงอยู่ตลอดเวลา ดังนั้นจึงเป็นไปได้ว่าที่เวลา $t_j > t_i$ เว็บเพจ $q$ อาจถูกแก้ไขปรับปรุง โดยที่ผู้ดูแลเว็บเพจ $p$ ไม่ได้รับรู้ด้วย ซึ่งในกรณีนี้ส่งผลให้เว็บเพจ $p$ มีเนื้อหาที่ล้าสมัยแล้ว ตามตัวอย่างรูปที่ 3.2




รูปที่ 3.2 ตัวอย่างความสัมพันธ์ด้านเวลาของเพจ $p$ และเพจ $q$ พิจารณาบนสองจุดเวลา




รูปที่ 3.3 ตัวอย่างแบบจำลองความใกล้ชิดด้านเวลาระหว่างเว็บเพจ

แบบจำลองความใกล้ชิดด้านเวลาที่นำเสนอนี้ถูกใช้เพื่อคำนวณค่าความใกล้ชิดกันของจุดเวลาที่มีการเปลี่ยนแปลงของสองเว็บเพจใด ที่มีเส้นเชื่อมโยงถึงกัน พิจารณาจากรูปที่ 3.3 สมมติว่าต้องการทราบค่าความใกล้ชิดด้านเวลาระหว่างเว็บเพจ $a$ และ $b$ ณ จุดเวลา $t_l$ เนื่องจากเว็บเพจ $a$ นั้นได้ถูกสร้างขึ้นและมีเส้นเชื่อมโยงไปยังเว็บเพจ $b$ ณ จุดเวลา $t_j$ โดยที่เว็บเพจ $b$ ได้ถูกสร้างขึ้นก่อนแล้ว ณ จุดเวลา $t_i$ และหลังจากนั้นเว็บเพจ $b$ ถูกแก้ไขปรับปรุงสองครั้ง (ณ จุดเวลา $t_k$ และ $t_l$) โดยที่ผู้ดูแลเว็บเพจ $a$ ไม่สามารถรับรู้ได้ ดังนั้นการคำนวณค่าความใกล้ชิดด้านเวลาจึงสามารถแบ่งพิจารณาได้เป็นสองช่วงเวลา ได้แก่ ช่วงเวลาก่อนที่เว็บเพจ $a$ จะถูกสร้างขึ้นมานั่นคือ $\Delta t_{bef(j,i)}$ และช่วงเวลาหลังจากที่เว็บเพจ $a$ ถูกสร้างขึ้นมาแล้วจนถึงจุดเวลาที่เว็บเพจ $b$ ถูกเปลี่ยนแปลงครั้งล่าสุดนั่นคือ $\Delta t_{aft(l,j)}$ ซึ่งช่วงเวลาแรกบ่งบอกถึงความทันสมัยของเว็บเพจ $a$ เมื่อเทียบกับเนื้อหาที่ถูกแก้ไขปรับปรุงครั้งสุดท้ายของเว็บเพจ $b$ ที่ผู้ดูแลเว็บเพจ $a$ รับทราบ ส่วนช่วงเวลาที่สองบ่งบอกถึงความล้าสมัยของเว็บเพจ $a$ เมื่อเทียบกับเนื้อหาที่ถูกแก้ไขปรับปรุงครั้งสุดท้ายของเว็บเพจ $b$ ที่ผู้ดูแลเว็บเพจ $a$ ไม่สามารถทราบได้

กำหนดให้ $w_{pq}(\Delta t_{bef}, \Delta t_{aft})$ เป็นค่าถ่วงน้ำหนักที่คำนวณให้ระหว่างสองเว็บเพจ $p$ และ $q$ ภายใต้เส้นเชื่อมโยง $(p,q) \in E$ เมื่อพิจารณาจากสองช่วงเวลา $\Delta t_{bef}$ และ $\Delta t_{aft}$ และจากงานวิจัยก่อนหน้า [*7, 8, 17, 19*] ซึ่งได้มีการใช้วิธีการทางความใกล้ชิด (proximity-based method) นั้น ในที่นี้เราได้อาศัยแนวคิดดังกล่าว โดยได้ดัดแปลงฟังก์ชันเคอเนล (function kernel) จำนวนห้าฟังก์ชัน[1] ซึ่งเป็นฟังก์ชันการแจกแจงแบบต่อเนื่อง (continuous distribution) ทั้งหมดเพื่อให้ค่าถ่วงน้ำหนักระหว่างสองเว็บเพจที่คำนวณออกมาไม่ว่าจะเป็น $\Delta t_{bef}$ หรือ $\Delta t_{aft}$ ถ้ามีช่วงเวลาที่ใกล้เคียงกันค่าถ่วงน้ำหนักที่คำนวณได้จะมีค่าสูง และถ้ามีช่วงเวลาที่ห่างกันมากค่าถ่วงน้ำหนักที่คำนวณได้จะมีค่าต่ำ

[1] จุดมุ่งหมายเพื่อดูว่าฟังก์ชันเคอเนลไหนเหมาะสมกับการนำมาใช้งานที่สุด

ฟังก์ชันเคอเนลที่นำมาใช้ได้แก่ Circle ดังสมการที่ 11, Cosine ดังสมการที่ 12, Gaussian ดังสมการที่ 13, Laplace ดังสมการที่ 14 และ Triangle ดังสมการที่ 15 ตามลำดับ

$$w_{cicle}\left(\Delta t_{bef(j,i)},\Delta t_{aft(k,j)}\right)=\sqrt{1-\left(\frac{\beta\Delta t_{bef(j,i)}+(1-\beta)\Delta t_{aft(k,j)}}{|T|}\right)^2} \quad (11)$$

$$w_{cosine}\left(\Delta t_{bef(j,i)},\Delta t_{aft(k,j)}\right)=\frac{1}{2}\left(1+\cos\left(\frac{\left(\beta\Delta t_{bef(j,i)}+(1-\beta)\Delta t_{aft(k,j)}\right)\pi}{|T|}\right)\right) \quad (12)$$

$$w_{gaussian}\left(\Delta t_{bef(j,i)},\Delta t_{aft(k,j)}\right)=\exp\left(-\frac{\left(\beta\Delta t_{bef(j,i)}+(1-\beta)\Delta t_{aft(k,j)}\right)^2}{2|T|^2}\right) \quad (13)$$

$$w_{laplace}\left(\Delta t_{bef(j,i)},\Delta t_{aft(k,j)}\right)=\exp\left(-\frac{\sqrt{2}\left(\beta\Delta t_{bef(j,i)}+(1-\beta)\Delta t_{aft(k,j)}\right)}{|T|}\right) \quad (14)$$

$$w_{triangle}\left(\Delta t_{bef(j,i)},\Delta t_{aft(k,j)}\right)=1-\frac{\left(\beta\Delta t_{bef(j,i)}+(1-\beta)\Delta t_{aft(k,j)}\right)}{|T|} \quad (15)$$

โดยที่ $t_i < t_j < t_k$ เมื่อ $|T|$ คือจำนวนจุดเวลาทั้งหมดในแบบจำลองเว็บ และ $\beta$ คือค่าสัมประสิทธิ์ใช้ระบุนัยสำคัญของช่วงเวลาแรกและช่วงเวลาหลัง จากสมการค่าถ่วงน้ำหนักจะเห็นได้ว่าเมื่อช่วงเวลายิ่งมีขอบเขตกว้าง ค่าถ่วงน้ำหนักที่ได้ยิ่งมีค่าน้อย โดยเราจะให้ความสำคัญกับช่วงเวลาหลังมากกว่าเนื่องจากช่วงเวลาหลังจะบ่งบอกถึงความกระตือรือร้นในการปรับปรุงข้อมูลในเพจของตัวเองให้ทันสมัยและตรวจสอบเนื้อหาการเปลี่ยนแปลงของเว็บเพจปลายทางที่ได้ทำเส้นเชื่อมโยงอ้างอิงไปของผู้ดูแลเว็บเพจ ดังนั้นค่าของ $\beta$ ควรอยู่ระหว่าง $[0,0.5]$ ท้ายที่สุดค่าถ่วงน้ำหนักที่ได้จะถูกทำให้เป็นมาตรฐาน (normalization) เนื่องจากเราจะต้องนำไปใช้แทนค่าในอินเวิร์สทรานซิชันเมทริกซ์ กล่าวคือ ผลรวมของค่าถ่วงน้ำหนักของทุกเส้นเชื่อมโยงที่ออกจากเว็บเพจ $p$ มีค่าเท่ากับ 1[2] ดังนั้นการทำให้ค่าถ่วงน้ำหนักเป็นมาตรฐานจึงได้ตามสมการต่อไปนี้

$$w'_{pq}\left(\Delta t_{bef},\Delta t_{aft}\right)=\frac{w_{pq}\left(\Delta t_{bef},\Delta t_{aft}\right)}{\sum_{r:(p,r)\in E} w_{pr}\left(\Delta t_{bef},\Delta t_{aft}\right)} \quad (16)$$

[2] เนื่องจากเราจะนำค่าถ่วงน้ำหนักที่ได้จากเส้นเชื่อมโยงไปใช้ในการคำนวณอินเวิร์สเพจแร็งค์ ในส่วนของอินเวิร์สทรานซิชันเมทริกซ์ ซึ่งเป็นคอลัมน์เมทริกซ์แบบสุ่ม (column-stochastic matrix) ซึ่งผลรวมในคอลัมน์จะเท่ากับ 1

### 3.1.3. เวกเตอร์ความลำเอียงเชิงเวลา (Temporal Bias Vector)

ในขั้นตอนนี้ เราจะนำค่าถ่วงน้ำหนักที่ได้จากขั้นตอนที่สองมาใช้เพื่อประเมินค่าความลำเอียง (bias/preference) ของแต่ละเว็บเพจโดยพิจารณาตามประวัติการเปลี่ยนแปลงของเว็บเพจนั้น โดยได้ตั้งสมมติฐานว่า เมื่อผู้ดูแลเว็บเพจได้แก้ไขปรับปรุงเว็บเพจใดก็ตาม จะถือว่าผู้ดูแลได้ตรวจสอบความถูกต้องและความเป็นปัจจุบันของเนื้อหา รวมถึงตรวจสอบความเกี่ยวข้องของทุกเว็บเพจที่ถูกเชื่อมโยงอ้างอิงไป ดังนั้นอาจกล่าวได้ว่า เว็บเพจที่มีเส้นเชื่อมโยงออกไปยังเว็บเพจอื่นจำนวนมาก จะเป็นเว็บเพจที่มีความกระตือรือร้น (activeness) อยู่เสมอและควรได้รับค่าความลำเอียงมากกว่า เนื่องจากผู้ดูแลเว็บเพจนั้นย่อมต้องใช้ความพยายามเพื่อตรวจสอบเว็บเพจที่อ้างอิงไปทั้งหมดนั่นเอง สำหรับแนวทางคำนวณค่าความลำเอียงนี้จะอาศัยอัลกอริทึมอินเวิร์สเพจแร็งค์ อย่างไรก็ดี ในงานวิจัยนี้ได้แก้ไขเปลี่ยนแปลงอินเวิร์สทรานซิชันเมทริกซ์เดิมจากสมการที่ 2 โดยพิจารณาตามค่าถ่วงน้ำหนักเชิงเวลาเป็นดังนี้

$$B'(p,q)=\begin{cases}\dfrac{w'_{pq}(\Delta t_{bef},\Delta t_{aft})}{\sum_{r:(r,q)\in E} w'_{rq}(\Delta t_{bef},\Delta t_{aft})} & \text{ถ้า } (p,q)\in E \\ 0 & \text{กรณีอื่นๆ}\end{cases} \tag{17}$$

จากอัลกอริทึมอินเวิร์สเพจแร็งค์สมการที่ 6 เราจะคำนวณเวกเตอร์ $\vec{s}$ หรือที่เรียกว่า "เวกเตอร์ความลำเอียงเชิงเวลา" ได้โดยแทนอินเวิร์สทรานซิชันเมทริกซ์ $B$ ด้วย $B'$ โดยเราสามารถเขียนสมการใหม่ได้ตามสมการที่ 18 ดังนี้

$$\vec{s}=\alpha B'\vec{s}+(1-\alpha)\vec{d} \tag{18}$$

เวกเตอร์ความลำเอียงเชิงเวลาที่เราคำนวณได้จากขั้นตอนนี้ เราจะนำไปใช้ในการลำเอียงสมการเพจแร็งค์แบบดั้งเดิม โดยนำไปแทนที่ในส่วนของเวกเตอร์เพจแร็งค์ส่วนบุคคลที่ในการคำนวณเพจแร็งค์แบบดั้งเดิมจะมีค่าคงที่ ซึ่งจะกล่าวถึงในหัวข้อถัดไป

### 3.1.4. การจัดลำดับด้วยเวลา (Time-aware Ranking)

ขั้นตอนสุดท้ายนี้ เราจะคำนวณค่าคะแนนของเว็บเพจทั้งหมดโดยคำนึงถึงเวลาร่วมกับอัลกอริทึมเพจแร็งค์แบบดั้งเดิมตามสมการที่ 5 ซึ่งในที่นี้ อาศัยแนวคิดของอัลกอริทึมเพจแร็งค์ส่วนบุคคล (personalized PageRank algorithm) [*4*] โดยแทนที่เวกเตอร์การกระจายตัวของคะแนนแบบเท่ากันเวกเตอร์ $\vec{d}$ ด้วยเวกเตอร์ความลำเอียงเชิงเวลา $\vec{s}$ จากขั้นตอนที่แล้ว ซึ่งจะได้สมการเพจแร็งค์ส่วนบุคคลตามความลำเอียงด้านเวลาตามสมการที่ 19 ดังนี้

$$\vec{r}=\alpha A\vec{r}+(1-\alpha)\vec{s} \tag{19}$$

สำหรับความซับซ้อนเชิงเวลาของอัลกอริทึมที่นำเสนอสามารถแยกพิจารณาเป็นสามส่วนตามขั้นตอนที่กล่าวถึงข้างต้น ดังนี้

1. ขั้นตอนการคำนวณค่าความใกล้ชิดด้านเวลา ในส่วนนี้เป็นการพิจารณาแบบจำลองเชิงเวลาไปพร้อมๆ กับการคำนวณค่าความใกล้ชิดระหว่างคู่เว็บเพจใดๆ ซึ่งจะเห็นได้ว่าหนึ่งเส้นการเชื่อมโยง (หนึ่งคู่เว็บเพจ) จะถูกพิจารณาและคำนวณเพียงหนึ่งครั้งเท่านั้นตามสมการ 11-15 ถ้ากำหนดให้ $|Links|$ แทนจำนวนเส้นเชื่อมโยงทั้งหมดในเว็บกราฟแล้ว ค่าความซับซ้อนเชิงเวลาในขั้นตอนนี้จะมีค่าเท่ากับ $\Theta(|Links|)$

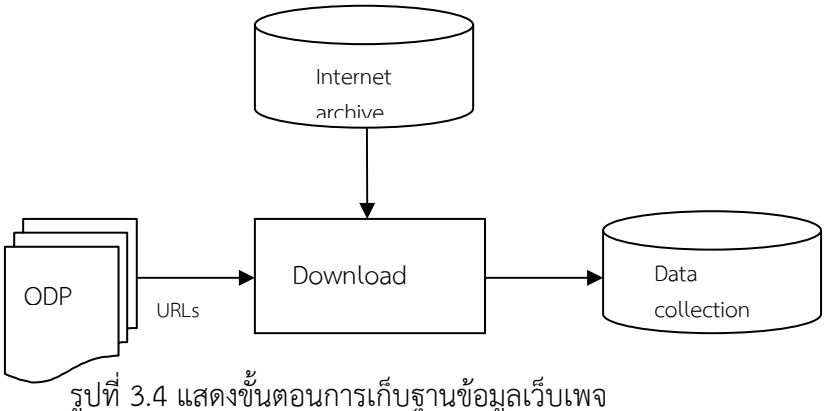
2. ขั้นตอนการคำนวณเวกเตอร์ความลำเอียงเชิงเวลา ในส่วนนี้ได้ปรับเปลี่ยนค่าในอินเวิร์สทรานซิชันเมทริกซ์ซึ่งยังคงอาศัยวิธีการคำนวณตามอัลกอริทึมเพจแร็งค์เดิม ค่าความซับซ้อนเชิงเวลาในขั้นตอนนี้จึงมีค่าเท่ากับค่าความซับซ้อนเชิงเวลาของอัลกอริทึมเพจแร็งค์ตามที่ได้กล่าวไว้ในหัวข้อ 2.1.3 ซึ่งมีค่าเท่ากับ $\Theta(|Links|)$

3. ขั้นตอนการคำนวณค่าเพจแร็งค์ที่ลำเอียงด้วยความใกล้ชิดด้านเวลา ในส่วนนี้ได้นำเวกเตอร์ความลำเอียงเชิงเวลาจากขั้นตอนก่อนหน้ามาใช้ ซึ่งยังคงอาศัยวิธีการคำนวณตามอัลกอริทึมเพจแร็งค์เดิมเช่นกัน ค่าความซับซ้อนเชิงเวลาในขั้นตอนนี้จึงมีค่าเท่ากับ $\Theta(|Links|)$

ดังนั้นจากการพิจารณาทุกขั้นตอนแล้ว ค่าความซับซ้อนเชิงเวลาของอัลกอริทึมที่นำเสนอจึงมีค่าเท่ากับ $\Theta(|Links|)$

### 3.2 วิธีดำเนินการทดลอง

ฐานข้อมูลเว็บถูกสร้างโดยเริ่มต้นจากการระบุยูอาร์แอล (URL) ที่กล่าวถึงหัวเรื่องการท่องเที่ยวในประเทศไทย (Thailand tourism topic) โดยเลือกจากโอดีพี[3] (ODP) และหลังจากนั้นจึงทำการดาวน์โหลด (download) เว็บเพจตามรายการยูอาร์แอลดังกล่าว ที่มีการเปลี่ยนแปลงในแต่ละจุดเวลาจากอินเตอร์เน็ตอาร์ไคว์ฟ[4] (Internet archive) โดยเว็บเพจที่จัดเก็บมาได้นั้นจะเริ่มตั้งแต่เดือนกุมภาพันธ์ในปี พ.ศ. 2539 ถึงเดือนธันวาคมในปี พ.ศ. 2555 ฐานข้อมูลเว็บที่ได้นี้ประกอบด้วยเว็บเพจโดยประมาณ หกหมื่นห้าพันหกร้อยเว็บเพจและมี หนึ่งล้านเก้าแสนสี่หมื่นหกพันเส้นการเชื่อมโยงสำหรับแต่ละจุดเวลา กระบวนการเตรียมฐานข้อมูลเว็บเพจตามรูปที่ 3.4




รูปที่ 3.4 แสดงขั้นตอนการเก็บฐานข้อมูลเว็บเพจ

[3] http://www.dmoz.org/

[4] http://archive.org/

ในการประเมินประสิทธิภาพของการจัดเรียงลำดับผลลัพธ์ค้นคืนของอัลกอริทึมที่นำเสนอ "ทีพีพีอาร์" (*TPPR*) เปรียบเทียบกับอัลกอริทึมเพจแร็งค์แบบดั้งเดิม "พีอาร์" (*PR*) เราได้พัฒนาระบบสืบค้นข้อมูลอิงตามลูซีน (Lucene-based searching system) และใช้ข้อมูลเว็บ ณ จุดเวลาสุดท้ายในการทำดัชนี และพิจารณาเว็บกราฟสำหรับคำนวณค่าคะแนนเพื่อการจัดลำดับ การทดลองได้ใช้คำค้น (query) จำนวน สามสิบห้าคำตามตารางที่ 3.1 โดยเลือกจากคำที่ปรากฏบ่อยที่สุดในหัวเรื่อง (title) ของเว็บเพจหลังจากที่ทำการตัดคำสต็อปเวิร์ด (stop words) ออกไปแล้ว สำหรับแต่ละคำค้นที่ถูกนำไปสืบค้นในระบบ ระบบจะค้นคืนคำตอบโดยพิจารณาเรียงลำดับตามค่าถ่วงน้ำหนักทีเอฟไอดีเอฟ (tf-idf) [*2*] มาตรฐาน ซึ่งในที่นี้จะเลือกมาเฉพาะ ยี่สิบอันดับแรก แล้วนำมาสลับลำดับและให้อาสาสมัครจำนวน แปดท่านช่วยพิจารณาแต่ละหน้าเว็บเพจของคำตอบ พร้อมทั้งระบุคะแนนตามความพึงพอใจสำหรับแต่ละคำตอบนั้น โดยแบ่งเป็น ห้าระดับตั้งแต่ ศูนย์ (ไม่พอใจ) จนถึง สี่ (พอใจมาก) ดังตารางตัวอย่างที่ 3.2

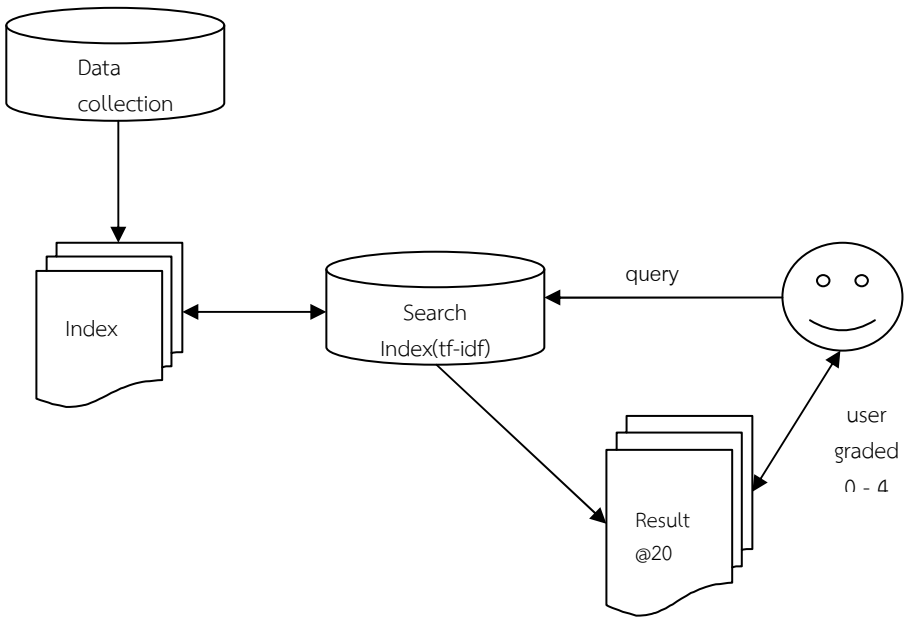
| | | | | |
|---|---|---|---|---|
| accommodation | breakfast | food | package | Samui |
| adventure | business | hostel | pattaya | scuba |
| airport | car | hotel | phuket | service |
| b2b | cruise | islands | rental | spa |
| bay | deal | krabi | resort | tour |
| beach | diving | lake | restaurant | travel |
| boutique | festival | museum | review | villa |

ตารางที่ 3.1 คำค้นที่ใช้ในการทดลองทั้งหมด สามสิบห้าคำ

| Search result | score |
|---|---|
| http://www.sawadee.com/thailand/food | 2 |
| http://www.holaspanishfood.com | 2 |
| http://gourmetfood.about.com/od/appetizersandsoups1/r | 1 |
| http://greekfood.about.com/od/festivalsholidays/a | 1 |
| http://www.afic.org | 0 |
| http://www.dtravelsround.com/site/2010/08/12/the-best-food-ever | 2 |
| http://nibbleanibble.com | 0 |
| http://homecooking.about.com/library/weekly | 0 |
| http://udisglutenfree.com/products/4/udis_gluten_free_bread | 0 |
| http://www.bakespace.com | 1 |
| http://www.bettycrocker.com | 2 |
| http://www.centralotagonz.com/Clyde_WineandFood | 1 |
| http://www.door2doorpattaya.com | 2 |
| http://www.fda.moph.go.th/project/foodsafety/v2/frontend/theme_1 | 1 |
| http://gourmetfood.about.com | 1 |
| http://www.lebienresort.com | 3 |
| http://bangkokfoodtours.com | 2 |
| http://comafoodtruck.com | 0 |
| http://spanishfood.about.com/od/discoverspanishfood/f | 0 |
| http://bangkokchefexpress.com | 2 |

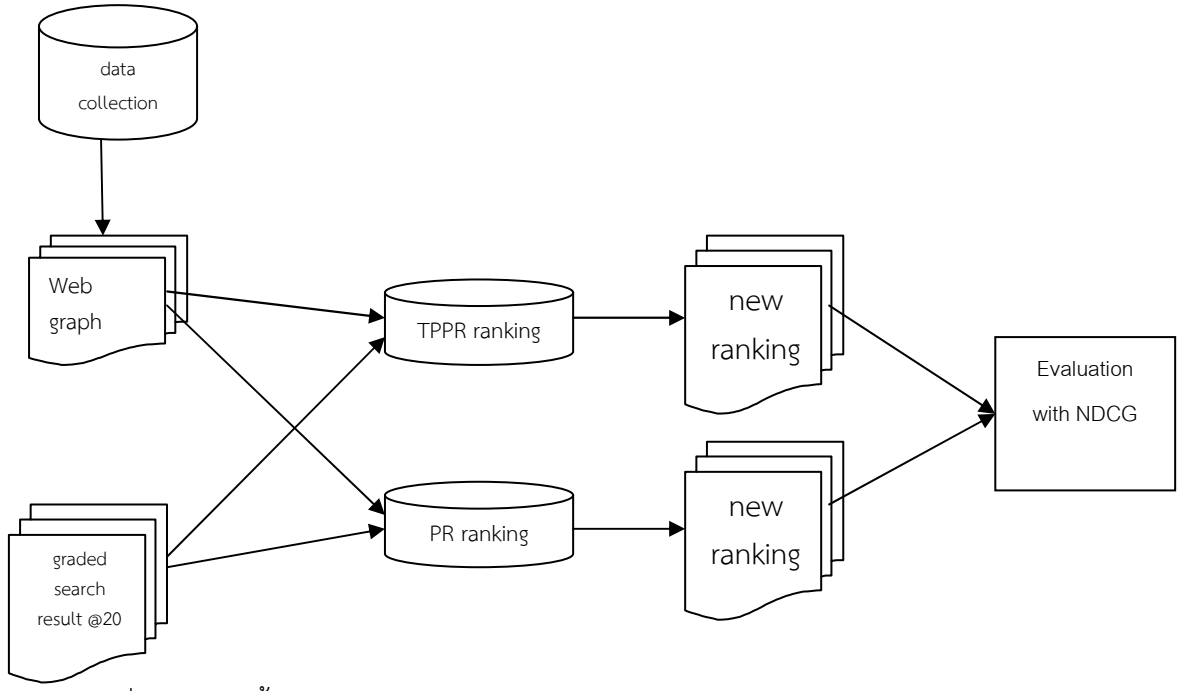
ตารางที่ 3.2 ตัวอย่างของผลการค้นคืน ยี่สิบอันดับแรกของคำค้น food จากค่าถ่วงน้ำหนักทีเอฟไอดีเอฟและผลคะแนนความพึงพอใจจากอาสาสมัคร

จากที่กล่าวมาข้างต้นกระบวนการในการจัดเตรียมฐานข้อมูลเว็บเพจและการให้คะแนนของอาสาสมัครสามารถแสดงได้ดังรูปที่ 3.5




รูปที่ 3.5 แสดงขั้นตอนการพิจารณาคะแนนความพึงพอใจด้วยอาสาสมัคร

หลังจากนั้นเราจะนำชุดคำตอบสามสิบห้าชุดคำตอบผลลัพธ์ค้นที่ผ่านการให้คะแนนความพึงพอใจจากอาสาสมัครแล้วมาจัดเรียงลำดับผลลัพธ์ค้นคืนใหม่ด้วยอัลกอริทึมทีพีพีอาร์ทั้งห้าฟังก์ชันเคอเนลและอัลอริทึมพีอาร์ ตามกระบวนการตามรูปที่ 3.6




รูปที่ 3.6 แสดงขั้นตอนการจัดเรียงลำดับผลลัพธ์ค้นคืนใหม่ด้วย ทีพีพีอาร์และ พีอาร์

จากรูปที่ 3.6 หลังจากที่เราได้จัดเรียงลำดับผลลัพธ์ค้นคืนใหม่ในแต่ละคำตอบแล้ว เราจะนำชุดคำตอบที่ได้มาคำนวณด้วยมาตรวัดนอมอลไลซ์ดิสเค้าท์คูมูเลทีฟเกน (normalized discounted cumulative gain) "เอ็นดีซีจี" (*NDCG*) [*13, 14*] เพื่อคำนวณคะแนนของการจัดเรียงลำดับรายการผลลัพธ์ค้นคืน $\tau$ ใน $k$ อันดับแรกตามสมการที่ 20 ดังนี้

$$NDCG_k(\tau) = Z_k \sum_{i=1}^{k} \frac{2^{r(i)} - 1}{\log_2(i+1)} \tag{20}$$

เมื่อ $r(i)$ คือค่าคะแนนของเอกสารในลำดับที่ $i$ ของผลลัพธ์ค้นคืน และ $Z_k$ คือค่าตัวประกอบในการทำให้เป็นมาตรฐาน (normalization factor) ของการจัดเรียงในอุดมคติ โดยปกติ $r(i)$ จะถูกกำหนดให้เป็นค่าคะแนนความเกี่ยวข้องของเว็บเพจผลลัพธ์และคำค้น อย่างไรก็ดีในการทดลองนี้จะให้เป็นค่าคะแนนความพึงพอใจของอาสาสมัครเมื่อเห็นผลลัพธ์ค้นคืนนั้น

# บทที่ 4

## ผลการทดลอง

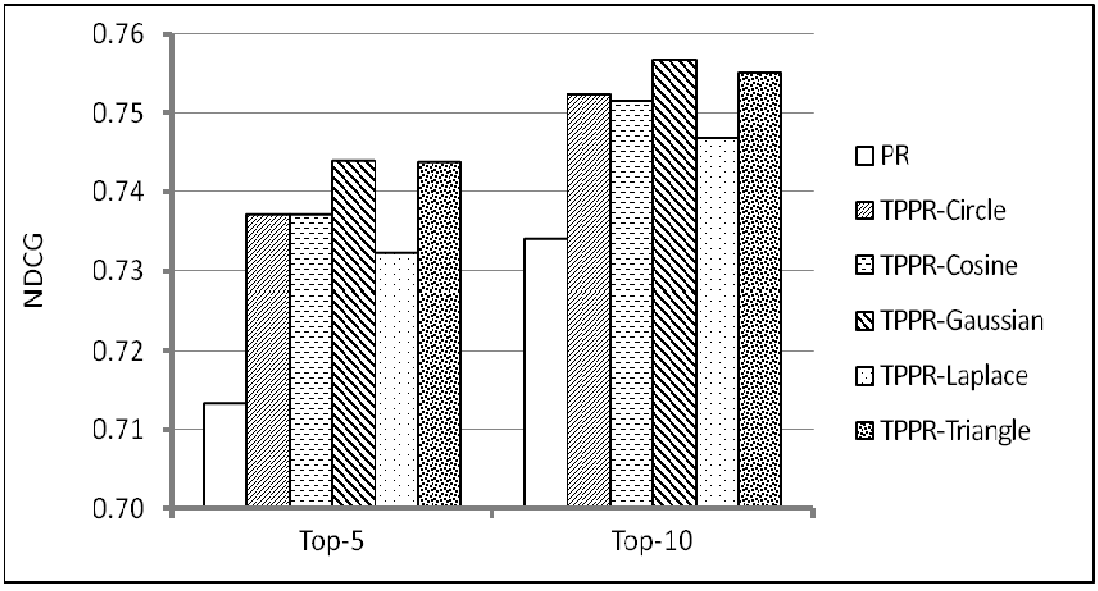
ในบทนี้ได้ทดลองเปรียบเทียบประสิทธิภาพระหว่างทีพีพีอาร์กับพีอาร์ โดยทีพีพีอาร์ได้แบ่งตามฟังก์ชันเคอเนลที่ใช้จำนวนห้าฟังก์ชันเคอเนลได้แก่ Circle, Cosine, Gaussian, Laplace และ Triangle และได้ทำการปรับค่า $\alpha$ มีค่าเท่ากับ 0.85 [*11*] และค่า $\beta$ มีค่าเท่ากับ 0.2[1] ซึ่งในส่วนนี้จะนำเสนอตัวอย่างของผลการทดลองในแต่ละอัลกอริทึม และหลังจากนั้นจะแสดงผลการเปรียบเทียบประสิทธิภาพการทดลองโดยวัดจากค่าเฉลี่ยเอ็นดีซีจีที่ผลลัพธ์ค้นคืนใน ห้าอันดับแรก และ สิบอันดับแรก

### 4.1. ผลการหาค่าเฉลี่ยเอ็นดีซีจีที่ผลลัพธ์ค้นคืนในห้าและสิบอันดับแรก

เราจะนำผลการเรียงลำดับผลลัพธ์ค้นคืนทั้ง สามสิบห้าคำค้นจากทั้งอัลกอริทึมทีพีพีอาร์และอัลกอริทึมพีอาร์มาหาค่าเฉลี่ยเอ็นดีซีจีที่ผลลัพธ์ค้นคืนใน ห้าและ สิบอันดับแรก เพื่อทดสอบความพึงพอใจของอาสาสมัครต่อผลัพธ์ค้นคืนในแต่ละอัลกอริทึม โดยได้ผลตามตารางที่ 4.1

| NDCG | PR | TPPR-Circle | TPPR-Cosine | TPPR-Gaussian | TPPR-Laplace | TPPR-Triangle |
|---|---|---|---|---|---|---|
| Top-5 | 0.713308696 | 0.737137333 | 0.73710184 | 0.743886 | 0.73218612 | 0.74381884 |
| Top-10 | 0.733954565 | 0.752408125 | 0.75151564 | 0.756801292 | 0.74692396 | 0.75506172 |

ตารางที่ 4.1 ผลการหาค่าเฉลี่ยเอ็นดีซีจีในห้าและสิบอันดับแรก




รูปที่ 4.1 กราฟแสดงผลค่าเอ็นดีซีจีเฉลี่ย ณ การจัดเรียงลำดับผลลัพธ์ค้นคืนใน ห้าและ สิบอันดับแรก ของอัลกอริทึมพีอาร์และทีพีพีอาร์แยกตามฟังก์ชันเคอเนล

[1] ค่า $\beta$ = 0.2 คือค่าที่ดีที่สุดที่ได้จากการทดลองปรับค่า $\beta$ ในหัวข้อที่ 4.2

จากรูปที่ 4.1 อัลกอริทึมทีพีพีอาร์ในทุกฟังก์ชันเคอเนลให้ค่าเอ็นดีซีจีเฉลี่ยที่มากกว่าอัลกอริทึมพีอาร์ ย่อมแสดงให้เห็นว่าตัวประกอบในเชิงเวลาส่งผลต่อการจัดเรียงลำดับที่ดีขึ้นได้ในแง่ของความพึงพอใจของผู้ใช้งาน นอกจากนี้เกาเชี่ยนเคอเนล (Gaussian kernel) ยังให้ค่าเฉลี่ยเอ็นดีซีจี ณ การจัดเรียงลำดับผลลัพธ์ค้นคืนใน สิบอันดับแรกสูงที่สุด ซึ่งแสดงว่าเคอเนลนี้สามารถให้ผลลัพธ์ค้นคืนในลำดับต้น ที่ตรงกับความสนใจของผู้ใช้งานได้มากกว่าเคอเนลอื่น รวมทั้งอัลกอริทึมพีอาร์ด้วย

เราได้นำผลการจัดเรียงลำดับผลลัพธ์ค้นคืนของคำค้น food ในอัลกอริทึมทีพีพีอาร์ของเกาเชี่ยนเคอเนลที่ค่าถ่วงน้ำหนัก $\beta$ เท่ากับ 0.2 และผลการจัดเรียงลำดับผลลัพธ์ค้นคืนของอัลกอริทึมพีอาร์มาเปรียบเทียบกันตามตารางที่ 4.2 และตารางที่ 4.3

| No. | TPPR's search result (Gaussian kernel) | Score |
|---|---|---|
| 1 | http://www.bakespace.com | 1 |
| 2 | http://gourmetfood.about.com | 1 |
| 3 | http://www.sawadee.com/thailand/food | 2 |
| 4 | http://www.holaspanishfood.com | 2 |
| 5 | http://gourmetfood.about.com/od/appetizersandsoups1/r | 1 |
| 6 | http://greekfood.about.com/od/festivalsholidays/a | 1 |
| 7 | http://www.afic.org | 0 |
| 8 | http://www.dtravelsround.com/site/2010/08/12/the-best-food-ever | 2 |
| *9* | *http://bangkokfoodtours.com* | *2* |
| 10 | http://www.door2doorpattaya.com | 2 |
| *11* | *http://nibbleanibble.com* | *0* |
| 12 | http://homecooking.about.com/library/weekly | 0 |
| 13 | http://udisglutenfree.com/products/4/udis_gluten_free_bread | 0 |
| 14 | http://www.bettycrocker.com | 2 |
| 15 | http://www.centralotagonz.com/Clyde_WineandFood | 1 |
| 16 | http://www.fda.moph.go.th/project/foodsafety/v2/frontend/theme_1 | 1 |
| 17 | http://www.lebienresort.com | 3 |
| 18 | http://comafoodtruck.com | 0 |
| 19 | http://spanishfood.about.com/od/discoverspanishfood/f | 0 |
| 20 | http://bangkokchefexpress.com | 2 |

ตารางที่ 4.2 ตัวอย่างผลการจัดเรียงลำดับผลลัพธ์ค้นคืนคำค้น food ของอัลกอริทึมทีพีพีอาร์

| No. | PR's search result | Score |
|---|---|---|
| 1 | http://www.bakespace.com | 1 |
| 2 | http://www.sawadee.com/thailand/food | 2 |
| 3 | http://gourmetfood.about.com | 1 |
| 4 | http://www.holaspanishfood.com | 2 |
| 5 | http://gourmetfood.about.com/od/appetizersandsoups1/r | 1 |
| 6 | http://greekfood.about.com/od/festivalsholidays/a | 1 |
| 7 | http://www.afic.org | 0 |
| 8 | http://www.dtravelsround.com/site/2010/08/12/the-best-food-ever | 2 |
| ***9*** | ***http://nibbleanibble.com*** | ***0*** |
| ***10*** | ***http://bangkokfoodtours.com*** | ***2*** |
| 11 | http://www.door2doorpattaya.com | 2 |
| 12 | http://homecooking.about.com/library/weekly | 0 |
| 13 | http://udisglutenfree.com/products/4/udis_gluten_free_bread | 0 |
| 14 | http://www.bettycrocker.com | 2 |
| 15 | http://www.centralotagonz.com/Clyde_WineandFood | 1 |
| 16 | http://www.fda.moph.go.th/project/foodsafety/v2/frontend/theme_1 | 1 |
| 17 | http://www.lebienresort.com | 3 |
| 18 | http://comafoodtruck.com | 0 |
| 19 | http://spanishfood.about.com/od/discoverspanishfood/f | 0 |
| 20 | http://bangkokchefexpress.com | 2 |

ตารางที่ 4.3 ตัวอย่างผลการจัดเรียงลำดับผลลัพธ์ค้นคืนคำค้น food ของอัลกอริทึมทีพีพีอาร์

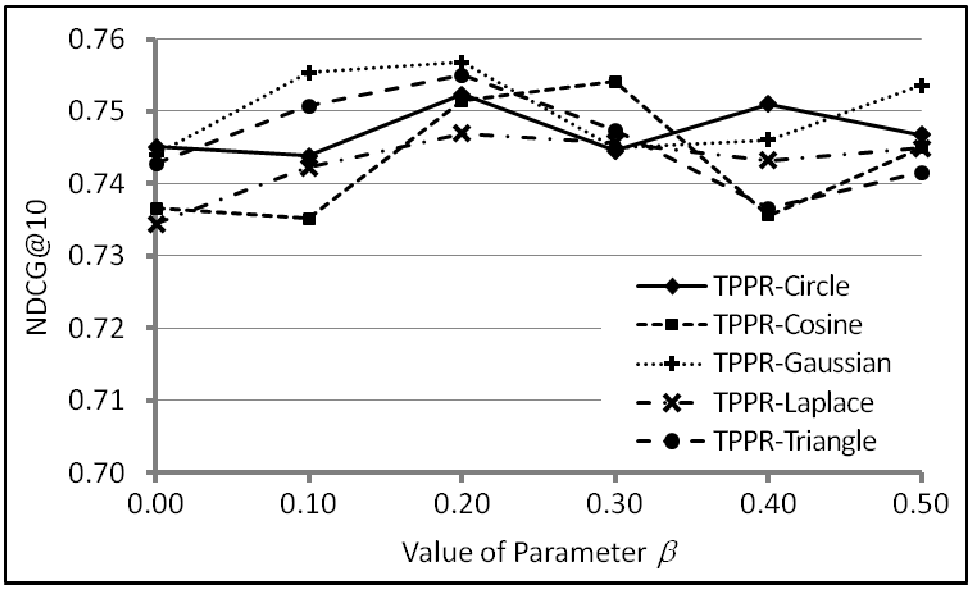
จากตารางที่ 4.2 และ 4.3 จะเห็นว่าผลการเรียงลำดับมีความแตกต่างกันโดยการจัดเรียงลำดับด้วยอัลกอริทึมทีพีพีอาร์ที่ใช้เกาเชี่ยนเคอเนล เว็บเพจ http://bangkokfoodtours.com จะถูกจัดอยู่ในลำดับที่เก้า ขณะที่การจัดเรียงลำดับด้วยอัลกอริทึมพีอาร์นั้นจะถูกจัดอยู่ในลำดับที่สิบ ซึ่งเป็นลำดับที่ต่ำกว่าเว็บเพจ http://nibbleanibble.com จากนั้นเราได้พิจารณาเปรียบเทียบเนื้อหาภายในเว็บเพจทั้งสองแสดงดังรูปที่ 4.3 โดยเว็บเพจ http://bangkokfoodtours.com อยู่ด้านซ้ายมือและเว็บเพจ http://nibbleanibble.com อยู่ด้านขวามือ


รูปที่ 4.2 ตัวอย่างเว็บเพจที่มีอันดับต่างกันในแต่ละอัลกอริทึมของผลลัพธ์ค้นคืนคำว่า food

จากรูปที่ 4.2 จะเห็นได้ว่าเว็บเพจทั้งสองที่มีการจัดเรียงลำดับที่แตกต่างกันตามที่ได้กล่าวไว้ในย่อหน้าที่แล้ว และเมื่อดูจากค่าคะแนนความพึงพอใจของเว็บเพจกับคำค้น food ที่ให้ด้วยอาสาสมัคร เว็บเพจทางซ้ายมือมีคะแนนความพึงพอใจสูงกว่า เนื่องจากเว็บเพจทางซ้ายมือมีข้อมูลที่น่าสนใจที่เกี่ยวข้องกับคำค้น food ส่วนเว็บเพจทางขวามือเป็นเว็บเพจที่ไม่มีข้อมูลที่เกี่ยวข้องแล้ว ซึ่งในตรงนี้มีความแตกต่างกันเนื่องจากเรานำความลำเอียงเชิงเวลามาใช้ในการช่วยส่งเสริมคะแนนของเว็บเพจที่มีข้อมูลทันสมัย จึงทำให้ในการจัดลำดับผลลัพธ์ค้นคืนด้วยอัลกอริทึมทีพีพีอาร์เว็บเพจทางซ้ายมือสามารถมีลำดับที่ดีกว่าเว็บเพจทางขวามือได้

### 4.2. ผลการทดสอบค่าตัวแปร $\beta$ ที่มีผลต่อการจัดลำดับ

ในส่วนนี้ เราจะศึกษาผลของการปรับค่าของตัวแปร $\beta$ ในแบบจำลองความใกล้ชิดด้านเวลาเพื่อทดสอบค่าถ่วงน้ำหนักด้านเวลา ที่อาจส่งผลต่อการจัดเรียงลำดับของทีพีพีอาร์ โดยจะปรับเปลี่ยนค่าให้อยู่ระหว่าง 0.0 ถึง 0.5 ตามที่ได้กล่าวไว้ในหัวข้อแบบจำลองความใกล้ชิดด้านเวลา หลังจากนั้นเราจะนำมาหาค่าเฉลี่ยเอ็นดีซีจี ณ การจัดเรียงลำดับผลลัพธ์ค้นคืนใน สิบอันดับแรก โดยผลการทดลองเราจะนำมาวาดกราฟดังแสดงตามรูปที่ 4.3




รูปที่ 4.3 ค่าเอ็นดีซีจีเฉลี่ย ณ การจัดเรียงลำดับผลลัพธ์ค้นคืนในสิบอันดับแรก ของอัลกอริทึมทีพีพีอาร์ โดยปรับเปลี่ยนค่า $\beta \in [0.0,0.5]$

จากรูปที่ 4.3 จะเห็นได้ว่าค่าเอ็นดีซีจีเฉลี่ยจะเพิ่มขึ้นเมื่อเปลี่ยนค่า $\beta$ จาก 0.0 ไปยัง 0.2 ทั้งนี้เนื่องจากว่าเมื่อ $\beta$ = 0 นั้น จะหมายความว่าเราไม่สนใจข้อมูลช่วงเวลาที่เว็บเพจมีการเปลี่ยนแปลงก่อนที่จะสร้างเส้นเชื่อมโยงอ้างอิงไปหาเลย ซึ่งเป็นสาเหตุให้ขาดข้อมูลในเชิงเวลาตรงส่วนนี้ไป อย่างไรก็ตาม เมื่อ $\beta$ มีค่าเพิ่มขึ้นจาก 0.2 ไปแล้วก็ไม่ได้ส่งผลให้การจัดเรียงลำดับได้ดีขึ้น ซึ่งอาจเป็นเพราะถึงจุดสมดุลในการให้น้ำหนักข้อมูลระหว่างช่วงเวลาการปรับปรุงเว็บเพจอ้างอิงก่อนและช่วงเวลาการปรับปรุงหลังการสร้างเส้นเชื่อมโยงอ้างอิงไปหา

# บทที่ 5

# สรุปผลการวิจัย และข้อเสนอแนะ

## 5.1. สรุปผลการวิจัย

งานวิจัยชิ้นนี้ได้นำเสนออัลกอริทึมสำหรับการจัดเรียงลำดับผลการค้นหา โดยได้นำมุมมองด้านเวลาเข้ามาช่วยในการลำเอียงเว็บเพจซึ่งจะให้โอกาสเว็บเพจที่มีความทันสมัยของข้อมูลมีลำดับที่ดีกว่าเว็บเพจที่ข้อมูลล้าสมัยไปแล้วในการจัดเรียงลำดับผลลัพธ์ค้นคืนและได้เปรียบเทียบประสิทธิภาพความพึงพอใจของผู้ใช้ในผลลัพธ์ค้นคืนกับอัลกอริทึมพีอาร์ซึ่งเป็นการจัดเรียงลำดับผลการค้นหาแบบใช้เส้นเชื่อมโยงเพียงอย่างเดียว

เราสามารถสรุปได้ว่าการนำมุมมองทางด้านเวลาเช่น ข้อมูลความทันสมัยของเว็บเพจ ความกระตือรือร้นในการดูแลเว็บเพจ เข้ามาช่วยในการจัดลำดับผลลัพธ์การค้นคืน ทำให้ผลการจัดอันดับเป็นที่น่าพึงพอใจแก่ผู้ใช้มากกว่า อย่างไรก็ตามการจัดลำดับด้วยความลำเอียงเชิงเวลาของเพจแร็งค์ส่วนบุคคล เป็นเพียงส่วนหนึ่งในการลำเอียงเชิงเวลาเพื่อให้โอกาสกับเพจที่มีความทันสมัยของข้อมูลได้รับการจัดลำดับได้ดีขึ้นเท่านั้น เนื่องจากในส่วนของเวกเตอร์เพจแร็งค์ส่วนบุคคลนั้นจะมีผลต่อคะแนนเพจแร็งค์สิบห้าเปอร์เซ็นต์[1] และคะแนนเพจแร็งค์ที่มาจากเส้นเชื่อมโยงเข้าจะส่งผลแปดสิบห้าเปอร์เซ็นต์ ดังนั้นจึงกล่าวได้ว่าการนำมุมมองด้านเวลาเข้ามาใช้ในการคำนวณคะแนนเพจแร็งค์จะเป็นการส่งเสริมให้เพจที่มีความทันสมัยของข้อมูลได้รับคะแนนเพจแร็งค์ที่ดีกว่าคะแนนที่มาจากอัลกอริทึมเพจแร็งค์แบบดั้งเดิมที่จะใช้เฉพาะข้อมูลของเส้นเชื่อมโยงของเว็บเพจเท่านั้น

นอกจากนี้เรายังได้ทำการทดสอบผลของค่าถ่วงน้ำหนัก $\beta$ ในการคำนวณหาค่าความใกล้ชิดด้านเวลา จากผลการทดลองเราสรุปได้ว่าข้อมูลการปรับปรุงเพจอ้างอิงหลังจากที่มีการสร้างเส้นเชื่อมโยงอ้างอิงไปหาแล้ว มีความสำคัญเพราะสิ่งนี้หมายถึงความกระตือรือร้นในการตรวจสอบข้อมูลเพจของผู้ดูแลเว็บเพจ ซึ่งจะทำให้เว็บเพจที่ดูแลมีข้อมูลที่เป็นปัจจุบันและมีความน่าสนใจแก่ผู้เยี่ยมชม

สุดท้ายเราสามารถกล่าวได้อย่างหนึ่งว่าความเป็นปัจจุบันของเนื้อหาในหน้าเว็บเพจเป็นส่วนสำคัญส่วนหนึ่งที่จะส่งเสริมความน่าสนใจของเว็บเพจต่อผู้เยี่ยมชมโดยมาก ด้วยเหตุผลนี้ เราจึงได้นำเสนออัลกอริทึมจัดเรียงลำดับเว็บอิงตามการเชื่อมโยงร่วมกับบทวิเคราะห์ในมุมมองของเวลาซึ่งประกอบด้วย 4 ขั้นตอน ได้แก่ ขั้นตอนแรก การสร้างแบบจำลองเว็บพิจารณาบนแกนเวลาโดยศึกษาพฤติกรรมการเปลี่ยนแปลงของเว็บเพจ ขั้นตอนที่สอง การสร้างแบบจำลองความใกล้ชิดด้านเวลาเพื่อหาความสัมพันธ์ของสองเว็บเพจที่อ้างอิงกัน ขั้นตอนที่สาม การคำนวณอินเวิร์สเพจแร็งค์โดยพิจารณาตามความใกล้ชิดด้านเวลา และขั้นตอนสุดท้าย การคำนวณเพจแร็งค์ส่วนบุคคลตามความลำเอียงเชิงเวลา ในการทดลองได้เก็บข้อมูลเว็บจริงจากอินเตอร์เน็ตอาร์ไคว์ฟ ซึ่งผลการทดลองก็แสดงให้เห็นว่าอัลกอริทึมที่นำเสนอนั้นมีประสิทธิภาพมากกว่าอัลกอริทึมเพจแร็งค์แบบดั้งเดิม

---

[1] เนื่องจากค่าสัมประสิทธิ์การเสื่อมสลายถูกกำหนดไว้เท่ากับ 0.85 ดังนั้นคะแนนในส่วนของเวกเตอร์เพจแร็งค์ส่วนบุคคลจึงคิดเป็นสิบห้าเปอร์เซ็นต์

**5.2. ข้อเสนอแนะ**

งานที่น่าสนใจซึ่งสามารถพัฒนาต่อยอดได้ อาทิเช่น

5.2.1. การกำหนดค่าถ่วงน้ำหนักของความใกล้ชิดด้านเวลาสำหรับเพจทรานซิชัน (page transition) แทนที่จะใช้ตัวแปรจำนวนของเส้นเชื่อมโยงของเว็บเพจ ซึ่งคาดว่าน่าจะให้ผลการจัดเรียงลำดับที่ดีขึ้น

5.2.2. นำไปทดลองบนฐานข้อมูลเว็บเพจที่มีขนาดใหญ่ขึ้น

5.2.3. นำมุมมองในด้านอื่นเช่น หัวเรื่อง พฤติกรรมผู้ใช้ผ่านแฟ้มบันทึกเหตุการณ์มาใช้ร่วมกับมุมมองด้านเวลาในการสร้างเวกเตอร์ความลำเอียงเพจแร็งค์ส่วนบุคคลซึ่งอาจจะทำให้มีประสิทธิภาพดีขึ้นเนื่องจากวิเคราะห์จากข้อมูลหลายมุมมอง เป็นต้น

5.2.4. ทดสอบกับชุดคำค้นที่เป็นวลี (phrase qurey) ซึ่งอาจจะให้ผลลัพธ์ค้นคืนที่แตกต่างออกไป

# รายการอ้างอิง


[1] Adamic, L. A., and Huberman, B. A. The web's hidden order. Communications of the ACM, pp.55–59., 2001.

[2] Baeza-Yates, R. A., and Ribeiro-Neto, B. A. Modern Information Retrieval. New York : ACM Press & Addison Wesley., 1999.

[3] Berberich, K., Vazirgiannis, M., and Weikum, G. Time-aware authority ranking. Internet Mathematics, pp.301–332., 2006.

[4] Brin, S., Motwani, R., Page, L., and Winograd, T. What can you do with a web in your pocket. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, pp.37–47., 1998.

[5] Brin, S., and Page, L. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, pp.107–117., 1998.

[6] Cho, J., and Roy, S. Impact of search engines on page popularity. In Proc. of the 13th International World Wide Web Conference 2004.

[7] Dai, N., and Davison, B. D. Freshness matters: In flowers, food, and web authority. In Proc. of the 33th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2010.

[8] Gerani, S., Carman, M., and Crestani, F. Aggregation methods for proximity-based opinion retrieval. ACM Transactions on Information Systems 2012.

[9] Golub, G. H., and Loan, C. F. V. Matrix Computations. Baltimore and London : Johns Hopkins University Press., 1996.

[10] Gyöngyi, Z., Garcia-Molina, H., and Pederson, J. Combating web spam with TrustRank. In Proc. of the 30th International Conference on Very Large Data Bases 2004.

[11] Haveliwala, T. H. Efficient computation of PageRank. Technical Report, Stanford InfoLab 1999.

[12] Haveliwala, T. H. Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search. IEEE Transactions on Knowledge and Data Engineering, pp.784–796., 2003.

[13] Järvelin, K., and Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. In Proc. of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2000.

[14] Järvelin, K., and Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, pp.422–446., 2002.

[15] Jeh, G., and Widom, J. Scaling personalized web search. In Proc. of the 12th International World Wide Web Conference 2003.

[16] Liu, Y., Liu, T. Y., Gao, B., Ma, Z., and Li, H. A framework to compute page importance based on user behaviors. Information Retrieval, pp.22–45., 2010.

[17] Lv, Y., and Zhai, C. Positional language models for information retrieval. In Proc. of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval 2009.

[18] Page, L., Brin, S., Motwani, R., and Winograd, T. The PageRank citation ranking: Bringing order to the web. Technical Report, Stanford InfoLab 1999.

[19] Petkova, D., and Croft, W. B. Proximity-based document representation for named entity retrieval. In Proc. of the 16th ACM Conference on Information and Knowledge Management 2007.

[20] Yu, P. S., Li, X., and Liu, B. Adding the temporal dimension to search—A case study in publication search. In Proc. of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence 2005.

# ประวัติผู้เขียนวิทยานิพนธ์

นายกานต์กมล ทองทิพย์ เกิดเมื่อวันที่ 23 พฤกษภาคม พ.ศ. 2527 ที่จังหวัดหนองคาย เป็นบุตรชายคนเดียว ของนายประเสริฐ ทองทิพย์ และนางฉลาด ทองทิพย์ สำเร็จการศึกษาระดับปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์ จากมหาวิทยาลัยเทคโนโลยีสุรนารี ในปี พ.ศ. 2547 และได้เข้าศึกษาต่อในระดับปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในภาคการศึกษาต้น ปีการศึกษา 2552