วิทยานิพนธ์

การคำนวณเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม

Trend Weight PageRank Computation

นายศุภกร กาญจนการุณ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

พ.ศ. 2551

ใบรับรองวิทยานิพนธ์

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

วิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

ปริญญา

วิศวกรรมคอมพิวเตอร์	วิศวกรรมคอมพิวเตอร์
สาขา	**ภาควิชา**

เรื่อง การคำนวณเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม

Trend Weight PageRank Computation

นามผู้วิจัย นายศุภกร กาญจนการุณ

ได้พิจารณาเห็นชอบโดย

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก .....

(ผู้ช่วยศาสตราจารย์อานนท์ รุ่งสว่าง, Ph.D.)

อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม .....

(ผู้ช่วยศาสตราจารย์จิตรทัศน์ ฝักเจริญผล, Ph.D.)

หัวหน้าภาควิชา .....

(ผู้ช่วยศาสตราจารย์เขมะฑัต วิภาตะวนิช, Ph.D.)

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์รับรองแล้ว

.....

(รองศาสตราจารย์กัญจนา ธีระกุล, D.Agr.)

คณบดีบัณฑิตวิทยาลัย

วันที่ 5 เดือน มิถุนายน พ.ศ. 2551

วิทยานิพนธ์

เรื่อง

การคำนวณเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม

Trend Weight PageRank Computation

โดย

นายศุภกร กาญจนการุณ

เสนอ

บัณฑิตวิทยาลัย มหาวิทยาลัยเกษตรศาสตร์

เพื่อความสมบูรณ์แห่งปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์)

พ.ศ. 2551

ศุภกร กาญจนการุณ 2551: การคำนวณเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม ปริญญาวิศวกรรมศาสตรมหาบัณฑิต (วิศวกรรมคอมพิวเตอร์) สาขาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก: ผู้ช่วยศาสตราจารย์อานนท์ รุ่งสว่าง, Ph.D. 63 หน้า

ปัจจุบันอายุของเว็บเพจเป็นองค์ประกอบหนึ่งที่สำคัญในการสืบค้นข้อมูลเนื่องจากผู้ใช้งานโดยมากมักต้องการข้อมูลที่มีถูกต้องและสดใหม่ อย่างไรก็ตาม ถึงแม้ว่าอัลกอริทึมเพจแร็งค์จะเป็นที่นิยมอย่างกว้างขวางในการใช้จัดลำดับผลลัพธ์ของการสืบค้นของระบบสืบค้นข้อมูลเว็บโดยส่วนใหญ่ แต่อัลกอริทึมดังกล่าวจะคำนวณค่าความสำคัญของเว็บเพจโดยอาศัยเพียงโครงสร้างการเชื่อมโยงเพียงอย่างเดียวเท่านั้น แต่ไม่ได้คำนึงถึงเงื่อนไขด้านเวลา อันได้แก่ ความสดใหม่ของข้อมูล ในวิทยานิพนธ์นี้ เราได้วิจัยหาแนวทางในการคำนวณค่าความสำคัญของเว็บเพจขึ้นใหม่ อัลกอริทึมที่พัฒนาขึ้นจะอิงตามอัลกอริทึมเพจแร็งค์พื้นฐาน โดยพยายามเพิ่มค่าความสำคัญให้กับเว็บเพจที่มีค่าวันเวลาแก้ไขล่าสุดใกล้เคียงปัจจุบัน และเว็บเพจที่มีแนวโน้มการเปลี่ยนแปลงบ่อยๆในช่วงเวลาที่ผ่านมา ผลการทดสอบอัลกอริทึมที่นำเสนอในเบื้องต้น เราพบผลลัพธ์ที่น่าติดตามทำวิจัยต่อ เมื่อเปรียบเทียบกับอัลกอริทึมเพจแร็งค์ดั้งเดิม

ศุภกร กาญจนการุณ — 29 / พ.ค. / 2551

ลายมือชื่อนิสิต ลายมือชื่ออาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

Supakorn Kanjanakaroon 2008: Trend Weight PageRank Computation. Master of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Assistant Professor Arnon Rungsawang, Ph.D. 63 pages.

Nowadays, web pages' freshness is an important factor in web search engine since users in general need both relevant and fresh information. However, although PageRank algorithm is widely used in ranking the search results of most search engines in the Internet, it calculates the importance of web pages based only on their hyper-links' structure, but ignores any time factor, e.g., the freshness of web pages. In this thesis, we study another way to calculate the importance of web pages. The proposed algorithm is inspired from the original PageRank one. It tries to increase the importance of web pages in accordance with both their current modification dates and their changing activities found in the past, i.e., their trend. In the preliminary experiments, we found some promising results using the proposed algorithm when comparing to the original PageRank one.

		29 / May / 2008
Student's signature	Thesis Advisor's signature	

กิตติกรรมประกาศ

ข้าพเจ้าขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง ประธานกรรมการที่ปรึกษา ที่ได้ช่วยเหลือในการวางแผนงานวิจัยในวิทยานิพนธ์ฉบับนี้ ตลอดจนการให้คำปรึกษา พร้อมทั้งให้แนวทางและความรู้เกี่ยวกับทฤษฎีต่างๆ มากมายในการทำวิจัย ข้อเสนอแนะที่เป็นประโยชน์ต่อการทำวิทยานิพนธ์ฉบับนี้ รวมถึงการตรวจแก้ไขข้อบกพร่องต่างๆ ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.จิตร์ทัศน์ ฝักเจริญผล กรรมการที่ปรึกษาวิชาเอก ที่กรุณาให้คำปรึกษาแนะนำและได้ให้ข้อเสนอแนะดีๆ ในการทำวิทยานิพนธ์ และขอขอบคุณพี่บัณฑิต มนัสเกษมศักดิ์ ที่คอยช่วยให้คำปรึกษาไม่ว่าจะเป็นเรื่องงานและเรื่องอื่นๆมากมาย และสุดท้ายขอขอบคุณสมาชิกห้องปฏิบัติการวิจัยข้อมูลและฐานความรู้ขนาดใหญ่ คุณท่านที่คอยให้คำแนะนำและกำลังใจที่ดี ให้สำเร็จลุล่วงไปด้วยดี

ขอขอบคุณพี่จุ๊ เจ้าหน้าที่ธุรการโครงการปริญญาโทที่ช่วยเหลือในการประสานงานและงานด้านเอกสารต่างๆ ให้งานเป็นไปอย่างสะดวกลุล่วงไปด้วยดี รวมถึงขอขอบคุณเจ้าหน้าที่ธุรการภาควิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเกษตรศาสตร์ทุกท่าน

คุณงามความดี หรือประโยชน์อันใดที่เกิดจากวิทยานิพนธ์ฉบับนี้ ขออุทิศให้แก่ บิดา มารดา บุพการี และผู้มีพระคุณทุกท่าน

ศุภกร กาญจนการุณ

มกราคม 2551

สารบัญ

	หน้า
สารบัญ	(1)
สารบัญตาราง	(2)
สารบัญภาพ	(3)
คำอธิบายสัญลักษณ์และคำย่อ	(5)
คำนำ	1
วัตถุประสงค์และขั้นตอนการวิจัย	3
การตรวจเอกสาร	4
อุปกรณ์และวิธีการ	33
อุปกรณ์	33
วิธีการ	34
ผลและวิจารณ์	42
ผล	42
วิจารณ์	58
สรุปและข้อเสนอแนะ	59
สรุป	59
ข้อเสนอแนะ	60
เอกสารและสิ่งอ้างอิง	61
ประวัติการศึกษา และการทำงาน	63

สารบัญตาราง

ตารางที่		หน้า
1	แสดงค่าคะแนนของเว็บเพจเมื่อคำนวณด้วยอัลกอริทึมเพจแร็งค์สมการที่ (1) โดยใช้ค่า *d* เท่ากับ 0.85	9
2	แสดงค่าคะแนนของเว็บเพจเมื่อคำนวณด้วยอัลกอริทึมเพจแรงค์สมการที่ (2) โดยใช้ค่า *d* เท่ากับ 0.85	10
3	เปรียบเทียบคุณลักษณะด้านต่างๆของอัลกอริทึม Age Based PageRank, T-Rank และ TimedPageRank	32
4	แสดงค่าแนวโน้มของเว็บเพจในแต่ละควอเตอร์	44
5	แสดงค่าแนวโน้มของเส้นการเชื่อมโยงในแต่ละควอเตอร์	44
6	แสดงเปรียบเทียบจำนวนจุดเชื่อมโยงเว็บเพจที่ได้รับสูงสุด 20 อันดับและ 50 อันดับแรกโดยใช้อัลกอริทึมเพจแร็งค์, T-Rank, TWPR ตามลำดับ	47
7	แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 20 อันดับแรก	48
8	แสดงเปรียบเทียบผลลัพธ์การจัดอันดับเว็บเพจ 50 อันดับแรก	49
9	แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับ 20 อันดับแรกระหว่าง PR และ TWPR	52
10	แสดงเปรียบเทียบผลลัพธ์การค้นคืน 20 อันดับแรกระหว่าง TWPR และ PR และ T-Rank	54

สารบัญภาพ

ภาพที่		หน้า
1	ตัวอย่างเว็บกราฟแสดงความสัมพันธ์การเชื่อมโยงของเว็บเพจ u , v , w , x , y และ z	5
2	แสดงตัวอย่าง Rank Leak	7
3	แสดงตัวอย่าง Rank Sink	7
4	แสดงความสัมพันธ์ของความสดกับอายุของเว็บเพจใดๆ ในช่วงเวลาที่มีการเปลี่ยนแปลงต่างๆ	14
5	แสดงค่าการกระจายสะสม (cumulative distribution) ของเพจ (เส้นล่าง) และไซต์ (เส้นบน) ในรูปของผลรวมสะสมเทียบกับตัวแปรอายุ (age) สำหรับชุดข้อมูล CL-2000	15
6	แสดงค่าสัดส่วนเว็บเพจที่มีการเปลี่ยนแปลงในแต่ละคาบเวลา	16
7	แสดงค่าการเปลี่ยนแปลงเพจแร็งค์ของแต่ละเว็บเพจที่เปลี่ยนไปตามอายุ	17
8	TLP แสดงสัดส่วนจำนวนเส้นการเชื่อมโยง ณ ช่วงเวลาต่างๆของเหตุการณ์ Y2K	20
9	ค่าความสดเมื่อมีการเปลี่ยนแปลง ณ จุดเวลาต่างๆพิจารณาบนกรอบ temporal window of interest ณ จุดเวลา (timestamp)	24
10	แสดง temporal window of interest กรณี $TS_{origin} = TS_{end}$	25
11	แสดง temporal window of interest กรณี $t_1 = TS_{focus}$	25
12	แสดง temporal window of interest กรณี $t_1 = TS_{origin}$	26
13	ตัวอย่างเว็บกราฟอย่างง่ายซึ่งแสดงถึงจุดเวลาแก้ไขล่าสุดสำหรับเว็บเพจและเส้นการเชื่อมโยง	27
14	แสดงภาพรวมการทำงานของอัลกอริทึมเพจแรงค์แบบถ่วงน้ำหนักตามแนวโน้ม	34
15	แสดงค่าของ AF	38
16	แสดงค่าถ่วงน้ำหนักตามแนวโน้ม	41
17	แสดงการกระจายตัวของค่าเวลาแก้ไขล่าสุดของเว็บเพจต่างๆตามปี	43
18	แสดงค่าแนวโน้มในแต่ละควอเตอร์ของเว็บเพจ x	45
19	แสดงค่าแนวโน้มในแต่ละควอเตอร์ของเส้นการเชื่อมโยง	46
20	แสดงค่าความนิยมของเว็บเพจของแต่ละอัลกอริทึม	48

สารบัญภาพ (ต่อ)

ภาพที่		หน้า
21	แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 20 อันดับแรก	49
22	แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 50 อันดับแรก	50
23	แสดงค่าคะแนนของเพจแร็งค์ของเว็บเพจตามเวลาต่างๆ	51
24	แสดงเว็บเพจ http://spiderbites.about.com/sitemap.htm ซึ่งแสดงเพียงเส้นการเชื่อมโยง	52
25	แสดงเว็บเพจ http://auto.howstuffworks.com/ ซึ่งแสดงเนื้อหาที่ค่อนข้างละเอียด	53
26	แสดงเว็บเพจ http://www.about.com/gi/pages/patent.htm ซึ่งแสดงข้อมูล patent ที่คลอบคลุมทั้งหมดของเว็บเพจ about	53
27	แสดงเว็บเพจ http://lumos.mugglenet.com/	57
28	แสดงเว็บเพจ http://www.the-leaky-cauldron.org/	57

คำอธิบายสัญลักษณ์และคำย่อ

PR	=	PageRank
APR	=	Age Based PageRank
TWI	=	Time Window of Interest
TWPR	=	Trend Weight PageRank
T-Rank	=	Time-Aware Authority Ranking

การคำนวณเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม

Trend Weight PageRank Computation

คำนำ

ในปัจจุบันเพจแร็งค์ (PageRank) (Page *et al.* 1998) เป็นอัลกอริทึมพื้นฐานสำหรับการจัดลำดับความสำคัญของเว็บเพจผลลัพธ์ที่นิยมใช้กันอย่างแพร่หลายในระบบสืบค้นข้อมูลเว็บทางธุรกิจ เพจแร็งค์จะพิจารณาให้ค่าความสำคัญแก่เว็บเพจโดยพิจารณาจากการถ่ายทอดค่าความสำคัญของเว็บเพจหนึ่งผ่านไปยังอีกเว็บเพจหนึ่งผ่านเส้นโครงสร้างการเชื่อมโยงระหว่างเว็บเพจ (Link structure) เท่านั้น ด้วยวิธีการดังกล่าวทำให้เพจแร็งค์มีข้อบกพร่องบางประการ เนื่องจากเมื่อเวลาผ่านไปเส้นการเชื่อมโยงกันของเว็บเพจบางเว็บเพจขาดการพิจารณาปรับปรุงข้อมูลจากผู้สร้าง หรือเจ้าของเว็บเพจนั้นๆ ซึ่งทำให้ข้อมูลโครงสร้างเชื่อมโยงที่มีอยู่ขาดความน่าถือ อาทิเช่น การเปลี่ยนโดเมนเนม (domain name) การคงอยู่ (Existency) หรือ การเปลี่ยนแปลงเนื้อหาของข้อมูล (Content change) ของเว็บเพจปลายทาง เป็นต้น

ในบทความวิจัยของ Amitay *et al.* (2004) ได้กล่าวถึงการศึกษาผลกระทบระหว่างเวลากับเว็บเพจ และแสดงให้เห็นว่าหากมีเหตุการณ์ใดเหตุการณ์หนึ่งที่สำคัญเกิดขึ้นจะส่งผลกระทบทำให้เกิดการเพิ่มขึ้นของกลุ่มเว็บเพจ หรือการเปลี่ยนแปลงด้านเนื้อหาของเว็บเพจใหม่ๆที่เกี่ยวข้องกับเหตุการณ์นั้นๆในช่วงเวลาดังกล่าวเป็นจำนวนมาก ดังเช่น เมื่อปีค.ศ. 2000 มีเว็บเพจที่เกิดขึ้น และมีความเกี่ยวข้องกับปัญหา Y2K เป็นจำนวนมาก และเมื่อเวลาผ่านไปเส้นการเชื่อมโยงกันของเว็บเพจในกลุ่มดังกล่าวเกิดขาดการปรับปรุงข้อมูล ทำให้เมื่อคำนวณค่าความสำคัญของเว็บเพจก็จะได้ผลลัพธ์ที่เป็นกลุ่มเว็บเพจเก่าถูกจัดอยู่ในอันดับต้นๆของการสืบค้นเมื่อเปรียบเทียบกับเว็บเพจที่เกิดขึ้นมาใหม่ซึ่งอาจมีข้อมูลใหม่และสำคัญกว่า แต่กลับถูกจัดอยู่ในอันดับท้ายๆเนื่องจากได้รับเส้นการเชื่อมโยงจากเว็บเพจอื่นค่อนข้างต่ำ เพราะยังไม่เป็นที่รู้จักโดยเว็บเพจทั่วไป

อย่างไรก็ตามได้มีงานวิจัยของ Berberich *et al.* (2004) ที่พิจารณาแก้ไขการคำนวณเพจแร็งค์พื้นฐานโดยรวมค่าองค์ประกอบทางด้านเวลาเข้าไปด้วยกันกับการคำนวณค่าความสำคัญของเว็บเพจ โดยใช้เทคนิคของการให้ค่าความสำคัญของเว็บเพจตาม “กรอบเวลาที่สนใจ” (Time Window of Interest) โดยหากค่าเวลาการแก้ไขล่าสุด (Modification date) ของเว็บเพจอยู่ในกรอบ

จากแนวความคิดการกำหนดกรอบเวลาที่สนใจนั้น นักวิจัยพบว่ามีประสิทธิภาพมากสำหรับการคัดเลือกเว็บเพจใหม่ๆให้อยู่ในลำดับต้นๆของการคำนวณได้ แต่จากการพิจารณาฐานข้อมูลเว็บเพจในปัจจุบันนั้น เราจะพบว่าที่ช่วงเวลาต่างๆอาจมีเหตุการณ์ที่สำคัญแยกย่อยได้อีก ซึ่งเหตุการณ์เหล่านี้บางครั้งเราไม่สามารถหาค่ากรอบของเวลาเริ่มต้น และสิ้นสุดของเวลาที่ผู้ใช้สนใจได้ ซึ่งจะส่งผลให้การคำนวณโดยใช้ค่ากรอบของเวลาดังกล่าวไม่สามารถใช้ได้ในทุกกรณี

ในวิทยานิพนธ์ฉบับนี้ได้เสนอวิธีการพิจารณาเพิ่มค่าความสำคัญให้แก่เว็บเพจใหม่และเว็บเพจเก่าบางเว็บเพจที่ยังมีความสำคัญอยู่โดยพิจารณาองค์ประกอบตามค่าแนวโน้ม (Trend Factor) และค่าอายุ (Aging Factor) เป็นปัจจัยทางด้านเวลาเป็นสำคัญ โดยองค์ประกอบดังกล่าวอาศัยข้อมูลพื้นฐานที่สำคัญของเว็บเพจ อันได้แก่ ค่าวันเวลาแก้ไขล่าสุด (Last-modified date) ของเว็บเพจ และค่าวันเวลาแก้ไขล่าสุดของเส้นการเชื่อมโยง ซึ่งวิธีในการหาเบื้องต้นได้ประยุกต์ใช้ตามวิธีการของ Amitay et, al. (2004) ที่ทำการหาค่าดังกล่าวจาก HTTP header field แต่อย่างไรก็ตามวิธีดังกล่าวยังมีปัญหาอยู่เนื่องจากเครื่องแม่ข่ายของบางเว็บเพจไม่คืนค่าเวลาที่ต้องการให้ ดังนั้นในวิทยานิพนธ์นี้เรายังได้พัฒนาวิธีในการหาค่าเวลาดังกล่าวเพิ่มเติม โดยวิเคราะห์จากส่วนประกอบต่างๆที่ปรากฏในแต่ละหนึ่งหน้าของเว็บเพจ ซึ่งจะประกอบไปด้วย ข้อความ(text) และเส้นการเชื่อมโยงต่างๆ (รูปภาพ) เป็นต้น โดยได้เน้นที่วิธีการหาค่าวันเวลาแก้ไขล่าสุดจากเส้นการเชื่อมโยงที่เป็นรูปภาพเป็นหลักเนื่องจากสามารถหาค่าเจอได้ค่อนข้างมากกว่าแบบข้อความ และได้นำองค์ประกอบทั้งหมดที่หาได้ไปประยุกต์ใช้ในอัลกอริทึมเพจแรงค์พื้นฐานเพื่อให้การคำนวณค่าความสำคัญของเว็บเพจได้ผลลัพธ์เป็นเว็บเพจในลำดับต้นๆที่มีข้อมูลใหม่สด ในขณะที่ยังคงให้ความสำคัญกับเว็บเพจเก่าบางเว็บเพจที่ยังคงได้รับเส้นการเชื่อมโยงสูงตามคุณลักษณะพื้นฐานของอัลกอริทึมเพจแร็งค์ดั้งเดิม

วัตถุประสงค์และขั้นตอนการวิจัย

วัตถุประสงค์ของการวิจัย

1. พัฒนาเทคนิคการหาวันเวลาแก้ไขล่าสุดของเว็บเพจ และของเส้นการเชื่อมโยงเว็บเพจ
2. พัฒนาเทคนิคสำหรับการจัดลำดับเว็บเพจโดยใช้ค่าอายุในการให้ค่าความสำคัญมากกับเว็บเพจที่อายุน้อย และลดค่าความสำคัญลงกับเว็บเพจเก่าบางเว็บเพจที่ยังคงได้รับเส้นการเชื่อมโยงสูงอยู่ อีกทั้งลดค่าความสำคัญมากกับเว็บเพจเก่าที่มีอายุมากและเส้นการเชื่อมโยงน้อย
3. พัฒนาเทคนิคการประยุกต์ใช้ค่าแนวโน้ม เพื่อการจัดลำดับความสำคัญให้แก่เว็บเพจ

ขั้นตอนการวิจัย

1. ศึกษาทฤษฎีต่างๆ ของเทคนิคการจัดลำดับเว็บเพจโดยใช้อัลกอริทึมเพจแร็งค์พื้นฐาน รวมถึงศึกษาทฤษฎีที่เกี่ยวข้อง เพื่อที่จะนำความรู้ที่ได้มาใช้ในการวิจัย
2. ศึกษาผลของการทำงานของงานวิจัยก่อนหน้า และวิเคราะห์ปัญหาและรวบรวมข้อดีและข้อด้อยต่างๆ เพื่อนำมาเป็นข้อมูลในการพัฒนาเทคนิคและอัลกอริทึมในการจัดลำดับเว็บเพจโดยมีเงื่อนไขทางด้านเวลามาเกี่ยวข้องด้วย
3. รวบรวมชุดข้อมูลทดสอบที่จะนำมาใช้ในกระบวนการทดสอบ เพื่อศึกษาผลของการจัดลำดับเว็บเพจที่ได้จากแต่ละอัลกอริทึม
4. พัฒนาเทคนิคการจัดลำดับเว็บเพจโดยคำนึงถึงองค์ประกอบทางด้านเวลาและแนวโน้ม
5. ทดสอบและวัดผลของเทคนิคการจัดลำดับเว็บเพจโดยใช้อัลกอริทึมที่ได้พัฒนาขึ้น
6. สรุปและวิเคราะห์ผลการทดลอง

การตรวจเอกสาร

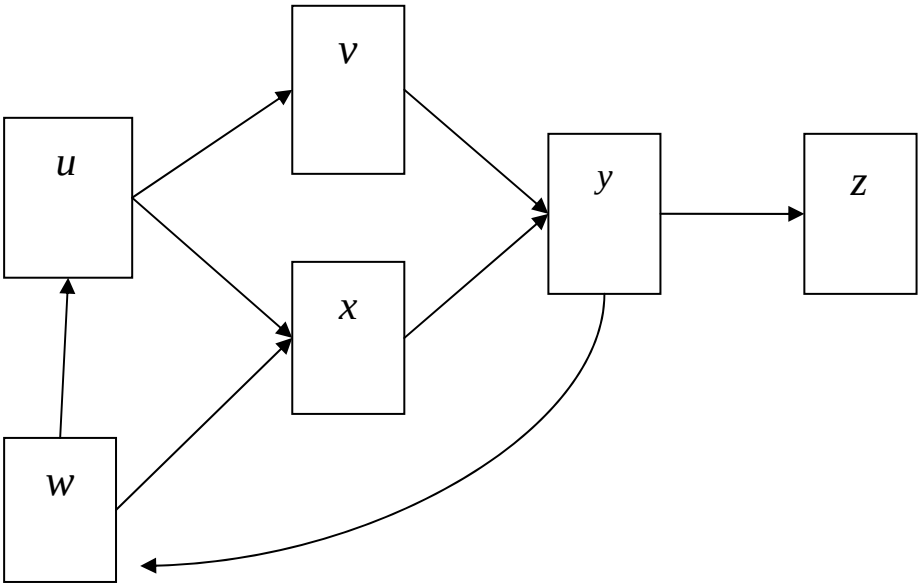
ความรู้พื้นฐานการจัดลำดับเว็บเพจ

การค้นหาข้อมูลจากเว็บเพจที่มีจำนวนมหาศาลบนอินเตอร์เน็ตเป็นสิ่งที่ท้าทายอย่างมาก ดังนั้นระบบสืบค้นข้อมูลจึงถูกพัฒนาขึ้นเพื่อใช้เป็นเครื่องมือในการค้นคืนเว็บเพจที่เกี่ยวข้องกับคำถาม (query) ของผู้ใช้ เทคนิคการจัดลำดับผลลัพธ์การค้นคืนเว็บเพจเป็นปัจจัยสำคัญประการหนึ่งสำหรับระบบสืบค้นข้อมูลบนอินเตอร์เน็ต ดังเช่น อัลกอริทึมเพจแร็งค์ (PageRank algorithm) ที่คิดค้นขึ้นโดย Page *et al.* (1998) และถูกใช้เป็นเทคนิคในการจัดลำดับตามความสำคัญของระบบสืบค้นข้อมูลกูเกิ้ล (Google) ภายในเวลาต่อมาเทคนิคดังกล่าวนี้ก็ได้รับการยอมรับ และใช้งานกันอย่างแพร่หลายมากขึ้นจนถึงปัจจุบัน แต่เนื่องจากจำนวนเว็บเพจในปัจจุบันมีอัตราการขยายตัวอย่างรวดเร็ว ซึ่งจากการศึกษาของ Amitay *et al.* (2004) พบว่าเนื้อหาที่ปรากฏอยู่บนหน้าเว็บเพจต่างๆมีรูปแบบการเปลี่ยนแปลงแก้ไขตามรูปแบบของความสนใจ ยกตัวอย่างเช่น ก่อนเกิดเหตุการณ์ก่อการร้ายตึกเวิล์ดเทรดเซ็นเตอร์ (World Trade Center Building) จะมีเว็บเพจที่แสดงข้อมูลเกี่ยวกับธุรกิจการค้าอยู่มาก แต่หลังจากเกิดเหตุการณ์ดังกล่าวมีเว็บเพจที่เกิดขึ้นใหม่โดยแสดงข้อมูลแตกต่างไปจากเดิมโดยบางเว็บเพจแสดงถึงข้อมูลผู้สูญหายและการขอความช่วยเหลือ ซึ่งจะเห็นว่าเว็บเพจที่เคยแสดงข้อมูลที่สำคัญในอดีตแต่ปัจจุบันอาจไม่ได้รับความนิยมแล้วเพราะมีเหตุการณ์อื่นที่สำคัญกว่าเกิดขึ้น

จากการศึกษาของ Baeza-Yates *et al.* (2002) ในเวลาต่อมา พบว่าเทคนิคการจัดลำดับเว็บเพจด้วยอัลกอริทึมเพจแร็งค์เป็นการเน้นให้ค่าความสำคัญกับจุดเชื่อมโยงเว็บเพจดังนั้นเว็บเพจเก่าจึงมีแนวโน้มที่จะได้รับค่าความสำคัญมากกว่าเว็บเพจที่เกิดขึ้นใหม่เนื่องจากเว็บเพจใหม่ต้องใช้เวลาในการให้เว็บเพจต่างๆสร้างจุดเชื่อมโยงเข้าหาหากมีข้อมูลที่มีความเกี่ยวข้องสัมพันธ์กัน โดยจากการวิจัยได้นำเสนอเทคนิคการแก้ปัญหาดังกล่าวโดยใช้ชื่อว่าอัลกอริทึม Age Based PageRank และในเวลาต่อมา Berberich *et al.* (2004) ได้นำเสนอกรอบวิธีการนำเวลาที่ระบุถึงจุดเวลาเริ่มต้นและสิ้นสุดช่วงที่เราสนใจเข้ามาประยุกต์ใช้กับเพจแร็งค์โดยใช้ชื่อว่า “อัลกอริทึม T-Rank” และจากการศึกษาของ Yu *et al.* (2005) นำเสนอ “อัลกอริทึม TimedPageRank” ซึ่งอัลกอริทึมดังกล่าวได้กำหนดค่าการถ่วงน้ำหนัก ที่เป็นไปตามอัตราความเสื่อมสภาพ (decay rate) ของเว็บเพจ ยกกำลังด้วยอายุ สำหรับรายละเอียดของอัลกอริทึมทั้งหมดได้กล่าวถึงในหัวข้อถัดไป

อัลกอริทึมเพจแร็งค์ (PageRank Algorithm)

อัลกอริทึมเพจแร็งค์ (Page *et al.*, 1998) เป็นอัลกอริทึมการคำนวณค่าความสำคัญของเว็บเพจบนพื้นฐานที่พิจารณาโครงสร้างหลักจากการเชื่อมโยงของเว็บเพจเหล่านั้น ในขั้นตอนแรกของการคำนวณเพจแร็งค์นั้น เราจะจำลองโครงสร้างของเว็บเพจให้อยู่ในรูปแบบของกราฟที่เรียกว่า "เว็บกราฟ" (web graph) ซึ่งจะประกอบไปด้วยโหนด (node) แทนเว็บเพจ และเอจ (edge) แทนการเชื่อมโยงระหว่างเว็บเพจ ในขั้นตอนต่อมาของการคำนวณ เราจะกำหนดค่าความสำคัญ (important/authoritative score) เริ่มต้นให้กับแต่ละโหนด และทำการคำนวณเพื่อกระจายค่าความสำคัญเหล่านั้นไปยังโหนดอื่นๆผ่านทางเอจ ซึ่งแสดงตัวอย่างได้ดังภาพที่ 1



ภาพที่ 1 ตัวอย่างเว็บกราฟแสดงความสัมพันธ์การเชื่อมโยงของเว็บเพจ u , v , w , x , y และ z

จากตัวอย่างเว็บกราฟแสดงความสัมพันธ์การเชื่อมโยงของเว็บเพจดังภาพที่ 1 พิจารณาเว็บเพจ x และ y ใดๆ เมื่อเว็บเพจ x มีการเชื่อมโยงไปยังเว็บเพจ y เขียนแทนด้วย "$x \to y$" ซึ่งเว็บเพจ y จะต้องมีความสำคัญ หรือเกี่ยวข้องอะไรบางอย่างกับเว็บเพจ x จึงทำให้ผู้สร้างเว็บเพจ x ทำการสร้างเส้นการเชื่อมโยงไปหา หรืออาจกล่าวได้ว่าผู้สร้างเว็บเพจ x ได้มอบความสำคัญบางอย่างให้กับเว็บเพจ y ดังนั้นจากแนวความคิดนี้เราสามารถคำนวณหาค่าความสำคัญของเว็บ

กำหนดให้

- $PR^{(k)}(x)$ แทนค่าความสำคัญ (Authoritative score) หรือค่าเพจแร็งค์ (PageRank score) ของเว็บเพจ x ที่คำนวณ ณ รอบที่ k
- $O(x)$ แทนจำนวนเส้นการเชื่อมโยงออกจากเว็บเพจ x ไปยังเว็บเพจใดๆ เรียกว่า "out-degree"
- $B(x)$ แทนเซตของเว็บเพจทั้งหมดที่มีการเชื่อมโยงมายังเว็บเพจ x หรือ "in-edge"

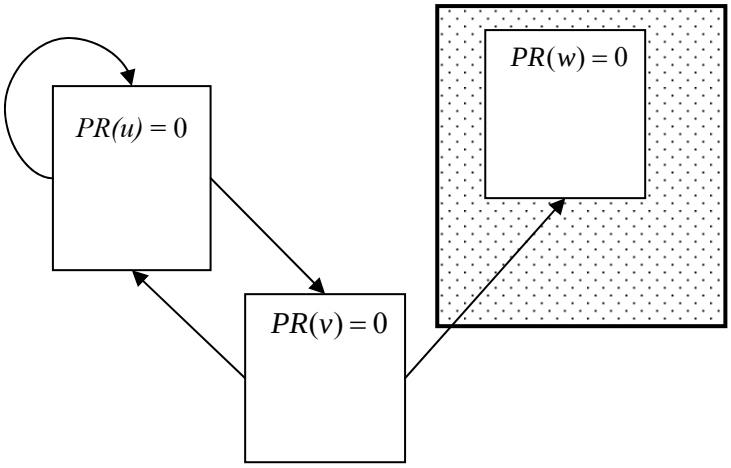
ค่าความสำคัญของเว็บเพจ y จากภาพที่ 1 มีค่าเท่ากับผลรวมของค่าความสำคัญที่ได้รับจากเว็บเพจ v และ x ซึ่งในขณะที่ค่าความสำคัญของเว็บเพจ v ได้รับจากเว็บเพจ u และเว็บเพจ x ก็ได้รับจากเว็บเพจ u และ w ตามลำดับ ซึ่งจะเห็นว่ามีการส่งมอบค่าความสำคัญกันเป็นทอดๆไป ดังนั้นอัลกอริทึมเพจแร็งค์จึงเป็นการคำนวณแบบวนซ้ำด้วยวิธีการที่เรียกว่า "พาวเวอร์เมตทอด" (power method) (Donath และ Hoffman 1972) แสดงดังสมการที่ (1)

$$PR^{(k+1)}(y) = \sum_{x \in B(y)} \frac{PR^{(k)}(x)}{O(x)} \qquad (1)$$

จากสมการที่ (1) กล่าวได้ว่า ค่าเพจแร็งค์ของ y ซึ่งคำนวณ ณ รอบที่ $k+1$ เกิดจากผลรวมของค่าเพจแร็งค์ของทุกๆ เว็บเพจ x ที่คำนวณ ณ รอบที่ k ที่มีการเชื่อมโยงมายัง y หารด้วยจำนวนการเชื่อมโยงออกของเว็บเพจ x นั้น (เนื่องจาก เว็บเพจ x ได้ส่งมอบความสำคัญไปยังทุกๆ เว็บเพจปลายทาง ในสัดส่วนที่เท่ากันตามจำนวน $O(x)$

อย่างไรก็ตาม เว็บกราฟโดยทั่วไปมักมีลักษณะอยู่ 2 ประการซึ่งเป็นปัญหาต่อการคำนวณค่าความสำคัญ ดังนี้

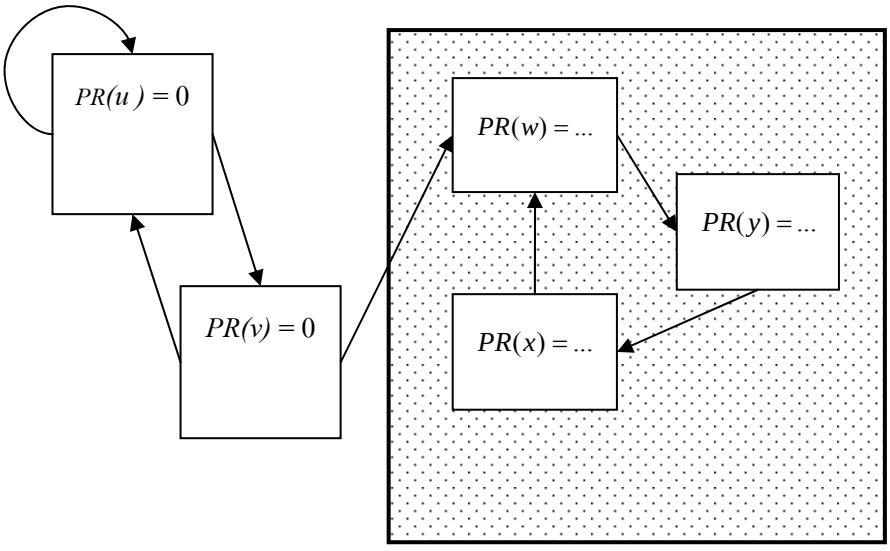
1) ค่าอันดับรั่ว (rank leak) เกิดขึ้นในกรณีที่เว็บกราฟประกอบด้วยเว็บเพจที่ไม่มีเส้นการเชื่อมโยงออก เรียกเว็บเพจดังกล่าวว่า "dangling node" แสดงได้ดังภาพที่ 2



ภาพที่ 2 แสดงตัวอย่าง Rank Leak

จากภาพที่ 2 ในการคำนวณแต่ละรอบค่าเพจแร็งค์ของทั้งระบบจะลดลงเรื่อยๆ เนื่องจากค่าคะแนนในระบบจะถ่ายทอดจากเว็บเพจ *v* ออกไปยังเว็บเพจ *w* และไม่มีการถ่ายทอดกลับเข้าไปในระบบทำให้ค่าคะแนนของเว็บเพจอื่นๆในระบบลดลงไปเรื่อยจนในที่สุดค่าเพจแร็งค์ของทุกๆเว็บเพจจะมีค่าเป็น 0 และมีค่าเป็นอนันต์สำหรับเว็บเพจ *w*

2) ค่าอันดับจม (rank sink) เกิดขึ้นในกรณีที่เว็บกราฟประกอบด้วยกลุ่มการเชื่อมโยงกันของเว็บเพจบางกลุ่ม โดยที่กลุ่มดังกล่าวไม่มีเส้นการเชื่อมโยงกลับเข้าไปในระบบหรือที่เรียกว่า “isolate cluster” แสดงได้ดังภาพที่ 3



ภาพที่ 3 แสดงตัวอย่าง Rank Sink

จากภาพที่ 3 จะเห็นว่าเส้นการเชื่อมโยงกันของเว็บเพจในกลุ่ม w, x และ y ไม่มีการสร้างเส้นการเชื่อมโยงกลับเข้าไปในระบบ (เว็บเพจ u และ v) ซึ่งจากโครงสร้างดังกล่าวหรือที่เรียกว่า "isolate cluster" จะทำให้ค่าเพจแร็งค์ของเว็บเพจอื่นๆในระบบลดลง และค่าเพจแร็งค์ในกลุ่มเว็บเพจ w, x และ y จะมีค่าคะแนนมากขึ้นไปเรื่อยๆจนผลรวมเป็นค่าอนันต์ เนื่องจากค่าเพจแร็งค์ที่ได้รับไม่มีการกระจายกลับเข้าไปยังเว็บเพจอื่นๆในระบบ

เพื่อให้รับประกันได้ว่าเมื่อคำนวณไปจนกระทั่งถึงจำนวนรอบหนึ่งๆ (ซึ่งมีค่าเป็นจำนวนมาก) แล้วจะพบว่าค่าเพจแร็งค์ของทุกเว็บเพจที่ได้นั้นลู่เข้า (convergence) เสมอ กล่าวคือ ค่าเพจแร็งค์ ณ รอบปัจจุบันและรอบที่แล้วมีค่าเกือบเท่ากัน และเพื่อแก้ไขปัญหา ค่าอันดับรั่ว ซึ่งส่งผลให้หลักการการกระจายค่าความสำคัญตกแก่เว็บเพจในกลุ่ม isolate cluster นั้นในอัลกอริทึมเพจแร็งค์จึงได้เพิ่มเส้นการเชื่อมโยงเสมือนแบบทั่วถึงกัน (virtual strongly connected hyperlinks) เข้าไปในเว็บกราฟ ดังนั้นจากสมการที่ (1) ถูกปรับเปลี่ยนเป็นดังนี้

$$PR^{(k+1)}(y) = d \sum_{x \in B(y)} \frac{PR^{(k)}(x)}{O(x)} + (1-d)\frac{1}{N} \qquad (2)$$

เมื่อ N คือ จำนวนเว็บเพจทั้งหมดในฐานข้อมูลเว็บกราฟ ซึ่งตามสมการที่ (2) เราจะถ่วงน้ำหนักของค่าทั้งสองเทอมด้วย d (เรียกว่า "damping factor" จากการทดลองฐานข้อมูลเว็บเพจโดย Page *et al.* (1998) กำหนดให้มีค่าเท่ากับ 0.85) และเพื่อให้ง่ายต่อการอ้างอิงในภายหลังจึงเขียนให้อยู่ในรูปแบบสมการทั่วไปโดยกำหนดให้ $t(x,y) = \frac{1}{O(x)}$ และ $s(y) = \frac{1}{N}$ แสดงดังสมการต่อไปนี้

$$PR^{(k+1)}(y) = d \sum_{x \in B(y)} PR^{(k)}(x) \cdot t(x,y) + (1-d) \cdot s(y) \qquad (3)$$

สำหรับเงื่อนไขในการตรวจสอบการลู่เข้าของค่าเพจแร็งค์ สามารถคำนวณได้จาก

$$\forall x : \frac{\left| PR^{(k+1)}(x) - PR^{(k)}(x) \right|}{\left| PR^{(k)}(x) \right|} \leq \delta$$

ผลต่างของค่าเพจแร็งค์ ณ รอบที่ $k+1$ และรอบที่ k หารด้วยค่าเพจแร็งค์ ณ รอบที่ k ต้องมีค่าไม่น้อยกว่าค่าเทรดโฮลด์ (threshold) δ ค่าหนึ่งที่กำหนด จึงถือว่าค่าเพจแร็งค์ของเว็บเพจทั้งหมดนั้นลู่เข้า และใช้เป็นเงื่อนไขในการหยุดคำนวณ

ในลำดับถัดไปนี้จะแสดงให้เห็นถึงตัวอย่างการคำนวณค่าเพจแร็งค์ของเว็บเพจ *u*, *v*, *w*, *x*, *y* และ *z* ตามตัวอย่างเว็บกราฟภาพที่ 1

ตารางที่ 1 แสดงค่าคะแนนของเว็บเพจเมื่อคำนวณด้วยอัลกอริทึมเพจแร็งค์สมการที่ (1) โดยใช้ค่า *d* เท่ากับ 0.85

เว็บเพจ	ค่าเพจแร็งค์					
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**	**รอบที่ 4**	**…**	**รอบที่ 18**
u	0.1666667	0.09444445	0.053518526	0.07475172		**0.0000000**
v	0.1666667	0.09444445	0.040138893	0.050999388		**0.0000000**
w	0.1666667	0.09444445	0.14383103	0.08399203		**0.0000000**
x	0.1666667	0.16527778	0.11205441	0.092995405		**0.0000000**
y	0.1666667	0.30694446	0.16557293	0.15761833		**0.0000000**
z	0.1666667	0.09444445	0.0961661	0.09243269		**0.0000000**

จากตารางที่ 1 จะเห็นว่าค่าเพจแร็งค์ของเว็บเพจ *u*, *v*, *w*, *x*, *y* และ *z* มีค่าลดลงจนเป็นศูนย์ ทั้งนี้เนื่องจากเว็บกราฟมีเว็บเพจ *z* เป็น dangling node ทำให้เกิดปรากฏการณ์ค่าอันดับรั่วดังรายละเอียดที่กล่าวในข้างต้นซึ่งเมื่อเพิ่มเส้นการเชื่อมโยงแบบทั่วถึงกันและคำนวณเพจแร็งค์ใหม่อีกครั้งตามสมการที่ (2) จะได้ผลลัพธ์การคำนวณค่าเพจแรงค์ตามตารางที่ 2

ตารางที่ 2 แสดงค่าคะแนนของเว็บเพจเมื่อคำนวณด้วยอัลกอริทึมเพจแร็งค์สมการที่ (2) โดยใช้ค่า d เท่ากับ 0.85

เว็บเพจ	ค่าเพจแร็งค์					
	รอบที่ 1	**รอบที่ 2**	**รอบที่ 3**	**รอบที่ 4**	**…**	**รอบที่ 18**
u	0.1666667	0.11944445	0.09268519	0.1286987		**0.11946108**
v	0.1666667	0.11944445	0.09268519	0.090315886		**0.09937069**
w	0.1666667	0.11944445	0.1829977	0.1806284		**0.16663508**
x	0.1666667	0.19027779	0.14344908	0.16808993		**0.17021652**
y	0.1666667	0.33194447	0.30518523	0.25163883		**0.27768186**
z	0.1666667	0.11944445	0.1829977	0.1806284		**0.16663508**

จากเทคนิคพื้นฐานในการจัดลำดับค่าความสำคัญของเว็บเพจได้มีการศึกษาต่อเพื่อปรับปรุงให้มีประสิทธิภาพมากยิ่งขึ้นในกรณีที่เว็บเพจเกิดการเปลี่ยนแปลงแก้ไขข้อมูลซึ่งสามารถศึกษารายละเอียดได้จากผลงานที่เกี่ยวข้องในลำดับถัดไป

งานวิจัยที่เกี่ยวข้อง

สำหรับในหัวข้อนี้จะกล่าวถึงงานวิจัยต่างๆที่เกี่ยวข้อง โดยงานวิจัยเหล่านี้มีจุดมุ่งหมายเพื่อปรับปรุงพัฒนาอัลกอริทึมเพจแร็งค์ (Page *et al.*, 1998) เดิมให้มีประสิทธิภาพมากยิ่งขึ้น อันได้แก่ งานวิจัยของ Baeza-Yate *et al.* (2002) ที่ได้ทำการศึกษาค่าการจัดลำดับเพจแร็งค์ของเว็บเพจที่เวลาต่างๆซึ่งพบว่าค่าคะแนนเพจแร็งค์จะมีค่าสูงสุดในช่วง 3 เดือนหลังจากตอนมกราคมปีค.ศ. 2003 จากนั้นค่าเฉลี่ยของคะแนนจะลดลงเรื่อยๆ อีกทั้งงานวิจัยของ Amitay *et al.* (2002) ที่พบว่าจำนวนการเพิ่มขึ้นของเว็บเพจมีความสอดคล้องกับเหตุการณ์สำคัญที่เกิดขึ้นในช่วงเวลาต่างๆ อัลกอริทึม T-Rank เสนอโดย Berberich *et al.* (2004) ซึ่งคำนึงถึงปัจจัยค่าความสด (freshness) และค่ากิจกรรม (activity) ของแต่ละเว็บเพจโดยกำหนดช่วงเวลาที่ให้ความสนใจแล้วคำนวณค่าปัจจัยทั้งสองเพื่อนำไปประยุกต์ใช้ในอัลกอริทึมเพจแร็งค์ และอัลกอริทึม TimedPageRank เสนอโดย Yu *et al.* (2005) เป็นอัลกอริทึมที่คำนวณค่าความสำคัญของเว็บเพจให้เป็นไปตามอัตราการเสื่อมสภาพ (decay rate) ของความสำคัญตามอายุของเว็บเพจนั้นๆ ซึ่งรายละเอียดทั้งหมดจะกล่าวถึงในลำดับถัดไป

การเปลี่ยนแปลงเว็บเพจ โครงสร้างและคุณภาพของเว็บเพจ (Web dynamic, Structure and Page Quality)

Baeza-Yate *et al.* (2002) ได้ศึกษาถึงความสัมพันธ์ระหว่างการเปลี่ยนแปลงโครงสร้างของระบบเว็บเพจว่าเกี่ยวข้องกับคุณภาพของเว็บเพจอย่างไร โดยผลลัพธ์ที่ได้จะนำไปใช้ในการจัดลำดับความสำคัญของเว็บเพจ นอกจากนี้ยังได้นำเสนอแนวทางในการปรับปรุงอัลกอริทึมเพจแร็งค์เดิม โดยได้ประยุกต์ใช้เวลาแก้ไขล่าสุด (last modified date) ของเว็บเพจที่ได้จากเครื่องแม่ข่าย (server) มาเป็นองค์ประกอบหนึ่งในการคำนวณ

ในงานวิจัยได้ยกตัวอย่างกรณีศึกษาถึงผลกระทบจากการเปลี่ยนแปลงโครงสร้างบนระบบเว็บของชิลี (Chilean web) หรือเฉพาะโดเมน (domain) ที่เป็น .cl ณ เวลาระหว่างครึ่งแรกของปีค.ศ. 2000 ซึ่งประกอบด้วย 670,000 เว็บเพจ (ประมาณ 7,500 เว็บไซต์) เขียนแทนด้วย “CL-2000” และจากการวิเคราะห์ข้อมูลที่เก็บมาสามารถสรุปเป็นโมเดลลักษณะการเปลี่ยนแปลงต่างๆได้ดังนี้

การสร้างเว็บเพจ (Creation) เมื่อเว็บเพจได้ถูกสร้างขึ้นแล้ว (เรียก “เว็บเพจใหม่”) เว็บเพจนั้นยังไม่เป็นที่รู้จักโดยเว็บเพจอื่นทั่วไป ดังนั้นจำเป็นต้องใช้เวลาเพื่อให้เว็บเพจอื่นทราบและสร้าง

การแก้ไขเว็บเพจ (Update) โดยพิจารณาจากการปรับปรุงเปลี่ยนแปลงเส้นการเชื่อมโยงที่ปรากฏบนหน้าเว็บเพจเมื่อเวลาผ่านไป ซึ่งจากการศึกษาสามารถสรุปคุณลักษณะการแก้ไขได้เป็น 2 ประการ

การแก้ไขเฉพาะเนื้อหาเพียงอย่างเดียว เว็บเพจมีการแก้ไขเพียงข้อความที่ปรากฏบนหน้าเว็บเพจเท่านั้น ซึ่งไม่กระทบต่อเส้นการเชื่อมโยง ทำให้เส้นการเชื่อมโยงต่างๆยังคงอยู่ (link valid) ทำให้การกระจายค่าความสำคัญของเว็บเพจยังคงถูกต้องตามหลักของเพจแร็งค์

การแก้ไขเส้นการเชื่อมโยง เป็นการแก้ไขเส้นการเชื่อมโยงหลักที่ปรากฏบนหน้าสำคัญของเว็บเพจเช่น หน้าต้อนรับ (Welcome page) โดยเส้นการเชื่อมโยงใหม่ที่สร้างขึ้นมีโอกาสผิดพลาดที่ชี้ไปยังเว็บเพจที่หายไปจากระบบอินเตอร์เน็ต (ยกตัวอย่างเช่น เว็บเพจปลายทางที่มีการเปลี่ยนโดเมนเนม) ส่งผลกระทบทำให้ข้อมูลการเชื่อมโยงกันของเว็บเพจต้นทางไม่ถูกต้อง (link invalid) ทำให้ค่าความสำคัญของระบบเว็บเพจกระจายไปยังเว็บเพจที่ไม่มีอยู่จริง

จากการศึกษาการแก้ไขเว็บเพจโดยงานวิจัย Baeza-Yates *et al.* (2002) พบว่าเว็บเพจส่วนใหญ่มีแนวโน้มที่จะเป็น การแก้ไขเส้นการเชื่อมโยงบนหน้าสำคัญของเว็บเพจ ดังนั้นจึงเลือกใช้โมเดลการแก้ไขเส้นการเชื่อมโยงนี้ในการพิจารณาคุณลักษณะอัตราการเปลี่ยนแปลงของเว็บเพจโดยทั่วไป

การลบเว็บเพจ (Deletion) เว็บเพจนั้นจะถูกลบ ถ้าเว็บเพจนั้นถูกนำออกไปจากระบบอินเตอร์เน็ต หรือถ้าเส้นการเชื่อมโยงที่ชี้ไปหายังเว็บเพจนั้นโดยเว็บเพจอื่นถูกลบออกไป ซึ่งหากเป็นการลบเส้นการเชื่อมโยงโดยผู้ดูแลระบบ ที่เครื่องมือเก็บข้อมูลเว็บเพจ (crawler) ไม่สามารถตรวจสอบการลบได้หรือข้ามไป ก็อาจส่งผลให้เป็นปัญหาต่อฐานข้อมูลกับระบบเครื่องมือสืบค้นข้อมูลเว็บเพจเนื่องจากเส้นการเชื่อมโยงที่ลบไปแล้วจะยังคงอยู่ในฐานข้อมูลและเมื่อนำไปคำนวณจะส่งผลให้ได้ผลลัพธ์ที่ผิดพลาดไป

จากทุกกรณีการเปลี่ยนแก้ไขข้อมูลเว็บเพจดังที่กล่าวมาแล้ว โดยทั่วไปเราสามารถจำลองฟังก์ชั่นที่แสดงถึงการมีข้อมูลเว็บเพจฉบับคัดลอกในฐานข้อมูลที่ไม่ตรงกับกับข้อมูลเว็บเพจฉบับจริงหรือ

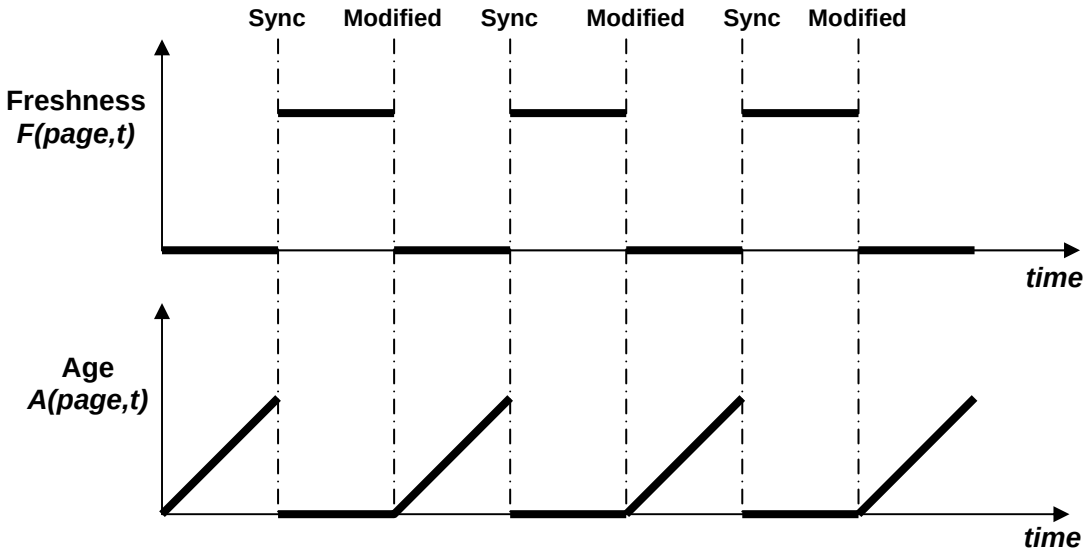
ความสด (Freshness) พิจารณากรณีเว็บเพจที่ให้ค่าเวลาการแก้ไขล่าสุดตรงกับปัจจุบัน (up-to-date) จะมีค่าความสดเป็น 1 และเว็บเพจที่ให้ค่าเวลาการแก้ไขล่าสุดไม่ตรงกับปัจจุบัน (outdated) จะมีค่าความสดเป็น 0 ซึ่งสามารถเขียนในรูปของฟังก์ชันได้ดังสมการที่ (4)

$$F(page_i ; t) = \begin{cases} 1 : & \text{ถ้า } page_i \text{ up-to-date ที่เวลา t} \\ 0 : & \text{กรณีอื่นๆ} \end{cases} \quad (4)$$

อายุ (Age) พิจารณากรณีเว็บเพจที่ให้ค่าเวลาการแก้ไขล่าสุดตรงกับปัจจุบัน (up-to-date) จะมีค่าอายุเป็น 0 และเว็บเพจที่ ให้ค่าเวลาการแก้ไขล่าสุดไม่ตรงกับปัจจุบัน (outdated) จะมีค่าอายุเพิ่มขึ้นเป็นเส้นตรงตามเวลา ซึ่งสามารถเขียนในรูปของฟังก์ชันได้ดังสมการที่ (5)

$$A(page_i ; t) = \begin{cases} 0 & : \text{ ถ้า } page_i \text{ up-to-date ที่เวลา t} \\ t - \text{modification time of } page_i & : \text{ กรณีอื่นๆ} \end{cases} \quad (5)$$

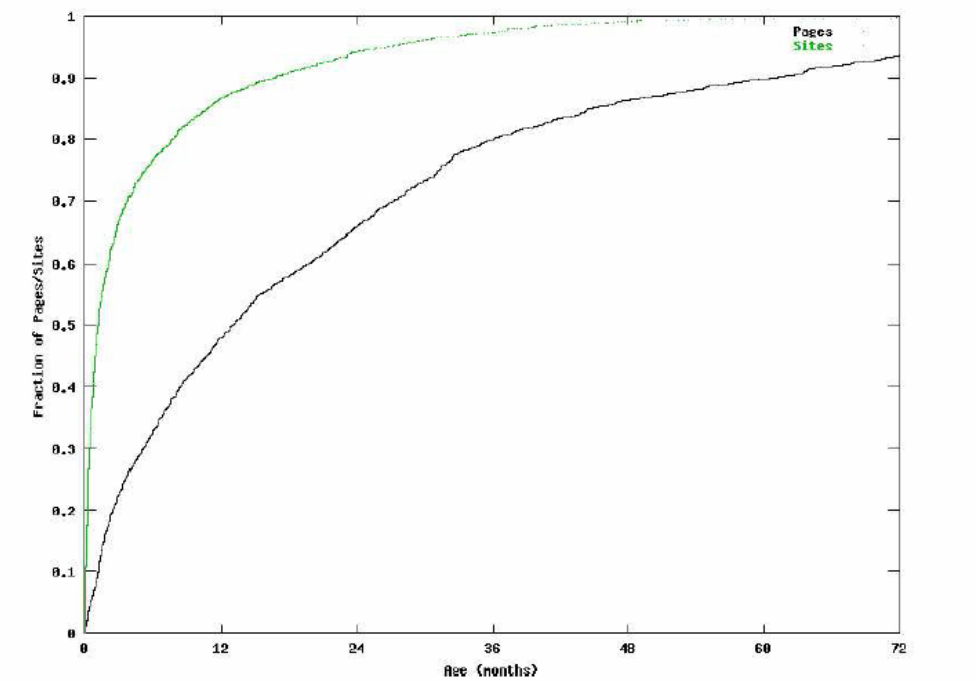
หรือจากสมการ (4) และ (5) เราสามารถเขียนแสดงความสัมพันธ์ของกราฟค่าความสดและอายุของเว็บเพจใดๆได้ดังภาพที่ 4



ภาพที่ 4 แสดงความสัมพันธ์ของความสดกับอายุของเว็บเพจใดๆ ในช่วงเวลาที่มีการเปลี่ยนแปลงต่างๆ

จากภาพที่ 4 กำหนดให้เหตุการณ์ 2 เหตุการณ์เกิดขึ้น ได้แก่ การแก้ไขของเว็บเพจบนเครื่องแม่ข่าย (Modified) โดยเว็บเพจจะได้ค่าความสดเป็น 0 และอายุจะเพิ่มขึ้นเรื่อยๆตราบเท่าที่ยังไม่มีการ Sync อีกครั้งหนึ่ง และการดาวน์โหลดข้อมูลที่แก้ไขแล้วโดย crawler (Sync) จะทำให้ค่าความสดของเว็บเพจกลับเป็นใหม่สุดมีค่าเป็น 1 ในขณะที่ ค่าอายุลดลงเป็น 0

จากข้อมูลชุด CL-2000 ซึ่งได้ตรวจสอบไปยังเครื่องแม่ข่าย เพื่อเก็บข้อมูลค่าเวลาการแก้ไขล่าสุด (last modified) ของแต่ละเว็บเพจ เพื่อนำมาคำนวณหาค่าอายุ โดยค่าอายุจะนับค่าจากเวลาการแก้ไขล่าสุดของเว็บเพจจนถึง ณ จุดพิจารณาทำการเก็บข้อมูล ซึ่งค่าเวลาการแก้ไขล่าสุดนี้ถูกส่งมาโดยเครื่องแม่ข่ายของเว็บเพจนั้นๆและเนื่องจากเว็บเพจส่วนใหญ่ยังมีอายุไม่มากดังนั้นจึงสนใจข้อมูลที่แสดงอยู่ในช่วงระยะเวลาเก่า 3 ปี (หรือ 36 เดือน) จากการเก็บข้อมูลเราสามารถแสดงอัตราส่วนความสัมพันธ์ระหว่างอายุต่อจำนวนของเว็บเพจและเว็บไซต์ ซึ่งสามารถแสดงได้ดังภาพที่ 5

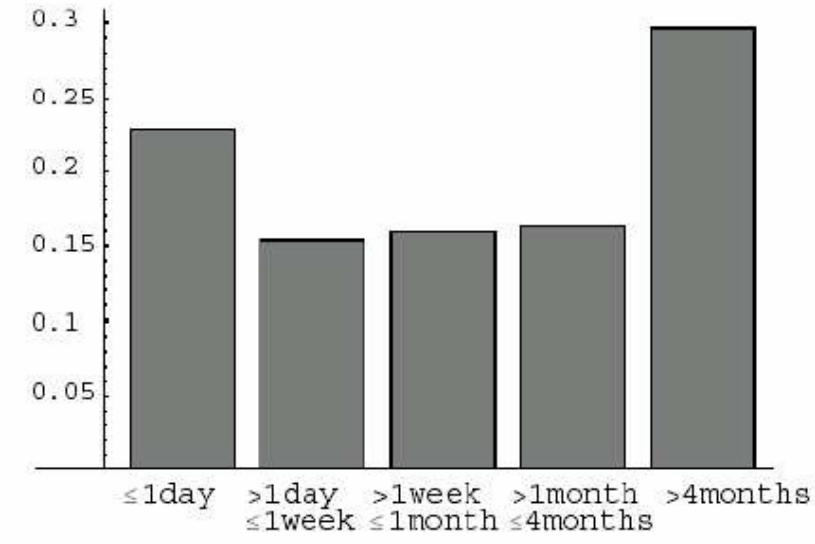


ภาพที่ 5 แสดงค่าการกระจายสะสม (cumulative distribution) ของเพจ (เส้นล่าง) และไซต์ (เส้นบน) ในรูปของผลรวมสะสมเทียบกับตัวแปรอายุ (age) สำหรับชุดข้อมูล CL- 2000 (Baeza-Yates *et al.* (2002))

จากภาพที่ 5 จะเห็นว่าจำนวนข้อมูลสะสมของทั้งเพจและไซต์มีสัดส่วนเป็นจำนวนมากในช่วงระหว่าง 12 ถึง 24 เดือนแรก ซึ่งประกอบไปด้วยเว็บเพจใหม่ (อายุน้อย) บางเว็บเพจที่อาจแสดงข้อมูลที่สำคัญแต่ยังไม่ได้รับเส้นการเชื่อมโยงมากเนื่องเกิดขึ้นใหม่ยังไม่เป็นที่รับรู้โดยเว็บเพจทั่วไป ดังนั้นจึงจำเป็นที่จะต้องทำการศึกษาองค์ประกอบของเวลากับเว็บเพจโดย Baeza-Yates *et al.* (2002) ได้เสนอแนวทางดังนี้

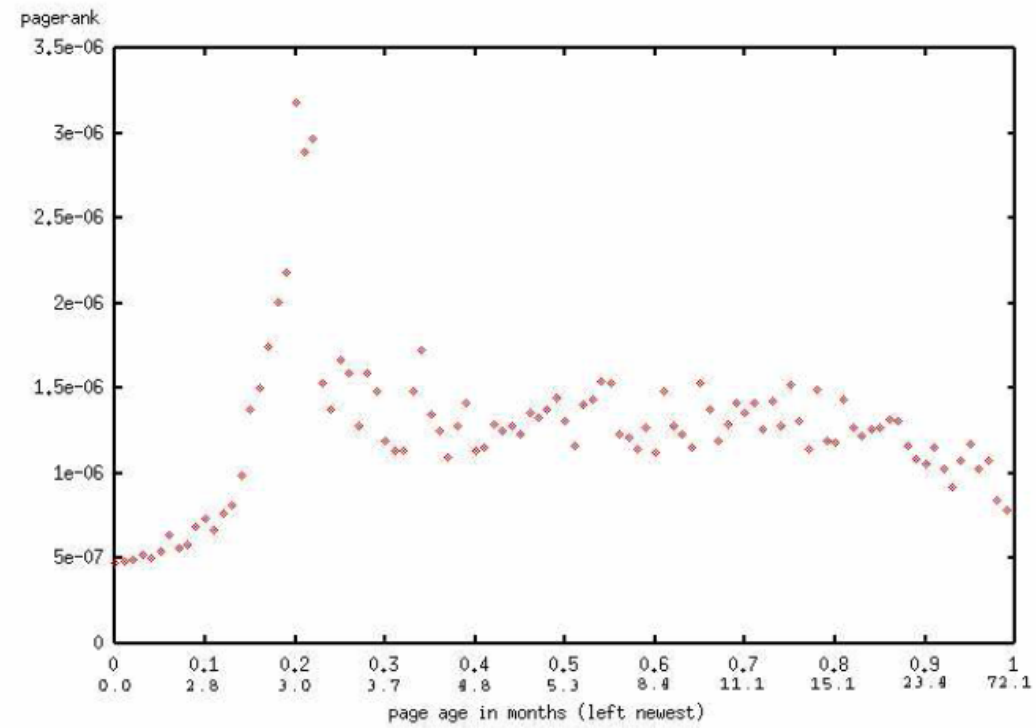
แนวทางที่ 1 ใช้การศึกษาจากข้อมูลบันทึก (log) ที่ได้จากการตรวจสอบ (monitor) เครื่องแม่ข่ายที่เป็น proxy ในองค์กรขนาดใหญ่หรือผู้ให้บริการอินเตอร์เน็ต (ISP) โดยข้อมูลบันทึกที่ได้จะแสดงวันเวลาแก้ไขล่าสุดของเว็บเพจรวมถึง บริมาณการเข้าถึง(access) ของแต่ละเว็บเพจ ซึ่งหากเลือกใช้วิธีนี้ก็จะทำให้เป็นการสะดวกที่จะสามารถตรวจสอบหาเว็บเพจที่สำคัญ (popular) สำหรับผู้ใช้ได้ แต่ก็เป็นการยากที่จะกำหนดค่าพารามิเตอร์สำหรับ ควบคุมช่วงเวลาการหมุนวนกลับไปตรวจสอบข้อมูลอีก (synchronize) เนื่องจากการเปลี่ยนแปลงของข้อมูลนั้นขึ้นอยู่กับพฤติกรรมของผู้ใช้

แนวทางที่ 2 ใช้ข้อมูลจากเครื่องมือเก็บข้อมูลเว็บเพจ (web crawler) โดยทำการเก็บข้อมูลจาก HTTP Header Field ที่ปรากฏในแต่ละเว็บเพจโดยวิธีนี้นั้นถึงแม้มีข้อเสียคือ เครื่องมือจะทำการเก็บข้อมูลในลักษณะวงกว้างไม่ระบุเจาะจงไปยังเว็บเพจใดๆซึ่งเป็นการยากที่จะเก็บเฉพาะเว็บเพจที่สำคัญ แต่มีข้อดีคือค่าพารามิเตอร์สำหรับการกำหนดช่วงเวลาของการหมุนวนกลับไปเก็บข้อมูลบนเว็บเพจต่างๆสามารถปรับจูนได้เพื่อให้ข้อมูลมีความทันสมัยล่าสุดมากที่สุด และค่าพารามิเตอร์นี้ก็ได้มีการศึกษาต่อมาโดย J. Cho *et al.* (2000) โดยได้ทำการหาอัตราการเปลี่ยนแปลงข้อมูลเว็บเพจ(Change interval) กับฐานข้อมูล 720,000 เว็บเพจในโดเมน com , edu, netorg, gov ยกตัวอย่างเช่น เว็บเพจหนึ่งอยู่ในฐานข้อมูล 50 วันมีการเปลี่ยนแปลงข้อมูลทั้งหมด 5 ครั้งดังนั้นอัตราการเปลี่ยนแปลงโดยเฉลี่ยเท่ากับ 10 วันต่อครั้ง ในภาพที่ 6 แสดงสัดส่วนเว็บเพจที่มีการเปลี่ยนแปลงในแต่ละคาบเวลาในแกนแนวตั้งหน่วยเป็น สัดส่วนของเว็บเพจ (Fraction of pages) ส่วนแกนแนวนอนแสดงความกว้างของคาบเวลา โดยจากงานวิจัย J. Cho *et al.* (2000) พบว่ากราฟช่วงเวลาของการเปลี่ยนแปลงเว็บเพจหรือช่วงเวลาการหมุนวนที่เหมาะสมสำหรับเครื่องมือเก็บข้อมูลอยู่ในช่วงระหว่างประมาณมากกว่า 4 เดือน ซึ่งมีสัดส่วนถึง 30 เปอร์เซ็นต์



ภาพที่ 6 แสดงค่าสัดส่วนเว็บเพจที่มีการเปลี่ยนแปลงในแต่ละคาบเวลา (J. Cho *et al.* (2000))

จากวิธีการเก็บข้อมูลในแนวทางที่ 2 ข้างต้น Baeza-Yates *et al.* (2002) ได้ทำการเก็บรวบรวมข้อมูลตัวอย่างประมาณ 35 ล้านเว็บเพจ ณ เดือนมกราคม ของปี 2003 เพื่อทำการศึกษาถึงการเปลี่ยนแปลงค่าของเพจแร็งค์ในช่วงเวลาต่างๆแสดงได้ดังภาพที่ 7 โดยค่าในแกนแนวตั้งแสดงค่าเพจแร็งค์และแกนแนวนอนบรรทัดล่างแสดงค่าอายุของเว็บเพจหน่วยเป็นเดือนโดยใช้การวัด



ภาพที่ 7 แสดงค่าการเปลี่ยนแปลงเพจแร็งค์ของแต่ละเว็บเพจที่เปลี่ยนไปตามอายุ (Baeza-Yates *et al.* (2002))

จากภาพที่ 7 เมื่อพิจารณาที่เว็บเพจใหม่ (new page) จะพบว่าค่าเพจแร็งค์เฉลี่ยโดยรวมของเว็บเพจในกลุ่มนี้เริ่มต้นมีค่าน้อยแต่จะค่อยๆมีค่าสูงขึ้น ซึ่งสาเหตุที่กลุ่มเว็บเพจใหม่นั้นมีค่าเพจแร็งค์เริ่มต้นน้อยอันเนื่องมาจากว่ากลุ่มเว็บเพจดังกล่าวยังไม่เป็นที่รู้จักโดยทั่วไปในระบบอินเตอร์เน็ตซึ่งจะทำให้ได้รับเส้นการเชื่อมโยงจากเว็บเพจอื่นค่อนข้างน้อยแต่เมื่อเวลาผ่านไปกลุ่มของเว็บเพจจะได้รับเส้นการเชื่อมโยงเพิ่มมากขึ้นหากเป็นเว็บเพจที่น่าสนใจ ทำให้มีค่าเพจแร็งค์ที่เพิ่มมากขึ้นเรื่อยๆถึงจุดๆหนึ่งซึ่งได้แก่ กลุ่มของเว็บเพจเก่า (old page) อายุประมาณ 3 เดือนและจะมีค่าเพจแร็งค์ที่ลดลงไปอีกเมื่อมีอายุมากขึ้นตามแกนแนวนอน (page age in months) ซึ่งสาเหตุที่ค่าเพจแร็งค์ลดลงนั้นมาจากเว็บเพจเก่าเดิมนั้นพยายามปรับปรุงแก้ไขข้อมูลตัวเองทำให้อายุของเว็บเพจกลับมา อยู่ในระหว่างช่วงอายุ 3 เดือน ทำให้เว็บเพจในกลุ่มนี้ยิ่งประกอบไปด้วยเว็บเพจเก่าที่มีค่าเพจแร็งค์ค่อนข้างสูงในขณะเดียวกันหากเว็บเพจนั้นไม่ทำการปรับปรุงเปลี่ยนแปลงแก้ไขข้อมูลให้ทันสมัยก็จะทำให้ได้รับความสนใจลดลงทำให้เส้นการเชื่อมโยงที่เคยได้รับมากลดจำนวนลง

การตรวจสอบแนวโน้มจากการวิเคราะห์ระหว่างเวลากับข้อมูล (Trend Detection Though Temporal Analysis)

ในงานวิจัยที่ศึกษาความสัมพันธ์กันระหว่างเวลากับเอกสารข้อมูลไม่ว่าจะเป็นบทความทางวิชาการหรือเว็บเพจก็ตาม Amitay *et al.*, (2004) พบว่า เวลาเป็นองค์ประกอบหนึ่งที่สำคัญกับการค้นหาข้อมูลในอินเตอร์เน็ต ตัวอย่างเหตุการณ์ที่มีความสัมพันธ์กับปัจจัยเวลา เช่น การที่ผู้ดูแลห้องสมุดจะตัดสินใจเสียค่าสมาชิกในการตอบรับวารสารงานประชุมวิชาการต่างๆ (Journal) หรือไม่นั้นมักจะดูที่ความสำเร็จและความสำคัญของ วารสารงานวิจัยนั้นๆว่ามีมากน้อยเพียงใด ซึ่งจำเป็นต้องอาศัยเวลาเป็นตัวพิสูจน์โดยถ้างานวิจัยนั้นเป็นที่สำคัญก็จะได้รับการอ้างอิงมาก หรือกรณีผู้แต่งวารสารงานวิจัย จะตัดสินใจตีพิมพ์ผลงานของตนก็มักที่จะเลือกตีพิมพ์ในงานประชุมวิชาการที่สำคัญ จากการศึกษาถึงผลกระทบของการเขียนงานวิจัยที่อ้างอิงถึงผลงานวิจัยอื่นๆ นักวิจัย Egghe (2001) และ Garfield (1998) พบว่า เมื่อเวลาผ่านไปแนวโน้มของงานวิจัยเก่าๆก็มีโอกาสที่จะได้รับความสำคัญน้อยลง อันเนื่องผลงานวิจัยใหม่ๆ ได้รับการปรับปรุงล่าสุดและมีความน่าเชื่อถือมากขึ้น นอกจากนี้ในการประเมินค่าความสำคัญของงานประชุมวิชาการ อาจคำนวณได้จากค่าเฉลี่ยความถี่ของผลงานวิจัยที่ถูกอ้างอิงต่อปี ซึ่งมักจะเปลี่ยนแปลงไปตามเวลา

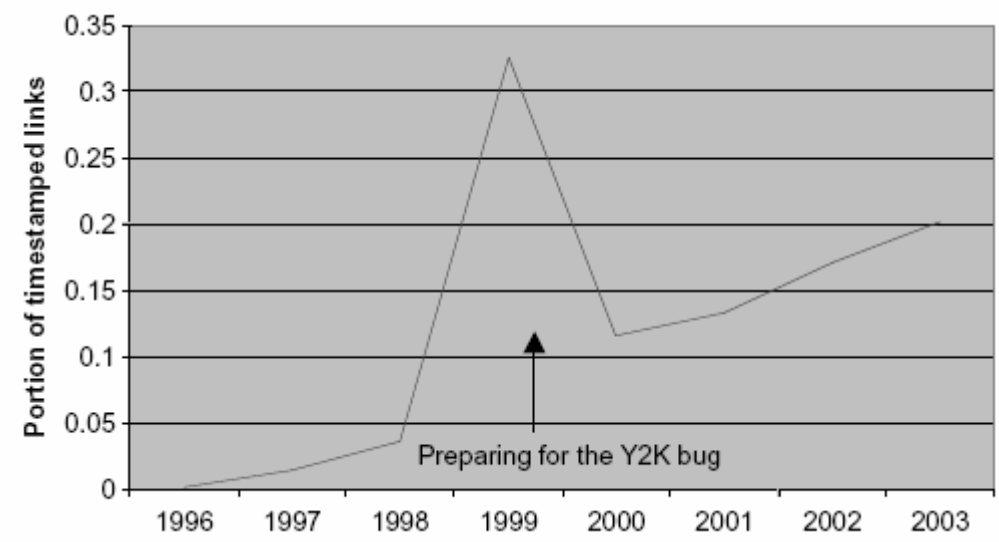
จากการศึกษาเบื้องต้นของ Amitay *et al.* (2004) กับฐานข้อมูลงานวิจัยเทียบกับฐานข้อมูลเว็บเพจพบว่าแนวโน้มหรืออัตราการอ้างอิงกันระหว่างบทความวิจัยจะไม่เหมือนกับแนวโน้มที่เกิดขึ้นกับการอ้างอิงกันของเว็บเพจ เนื่องจากเนื้อหา (content) ที่ปรากฏบนหน้าเว็บเพจเหล่านั้นจะมีการปรับเปลี่ยนไปตามเหตุการณ์จริงที่เกิดขึ้น ยกตัวอย่างเช่น เว็บไซต์ที่กล่าวถึงรัฐบาลหากเมื่อรัฐบาลเกิดการเปลี่ยนแปลงทำให้ข้อมูลรายละเอียดภายในเว็บเพจดังกล่าวนั้นเปลี่ยนตามไปด้วย แต่ยูอาร์เอล (URL) ของเว็บเพจยังคงเดิม ซึ่งความแตกต่างนี้หากเปรียบเทียบกับบทความที่ได้รับการตีพิมพ์แล้ว หากพบว่าเป็นบทความที่เก่าล้าสมัย ก็มักมีการตีพิมพ์บทความฉบับใหม่ที่มีความใหม่กว่า (fresh document) จากการศึกษาของ Adamic และ Huberman (2001) พบว่าเมื่อเวลาผ่านไปจำนวนเส้นการเชื่อมโยงมายังเว็บเพจหนึ่งๆมักมีปริมาณเพิ่มขึ้น (ได้รับการอ้างอิงเพิ่มขึ้น) ส่งผลให้การคำนวณค่าความสำคัญของเว็บเพจด้วยการใช้เทคนิคที่พิจารณาเพียงเส้นการเชื่อมโยงอย่างเดียว เช่น อัลกอริทึมเพจแร็งค์เกิดความไม่เหมาะสม จากเว็บเพจที่ปรากฏขึ้นนานแล้วมักมีจำนวนเส้น

ในการศึกษาถึงความสัมพันธ์ของคำถาม (query) ของผู้ใช้กับผลลัพธ์ที่ได้จากระบบสืบค้นข้อมูลบนอินเตอร์เน็ต ระหว่างคำถามที่สอบถามในอดีตกับปีปัจจุบัน พบว่ามีความแตกต่างกันอย่างมาก ยกตัวอย่าง เช่น คำถามว่า “The World Trade Center” (WTC) ในอดีตก่อนปีค.ศ. 2001 เกิดเหตุการณ์ก่อการร้าย ผลลัพธ์การค้นคืนจากระบบสืบค้นข้อมูลต่างๆใน 100 อันดับแรกจะแสดงถึงข้อมูลเกี่ยวกับธุรกิจ บริการ การท่องเที่ยวภายในบริเวณอาคาร WTC แต่หลังจากเหตุการณ์เมื่อวันที่ 11 กันยายน ค.ศ. 2001 กลับพบว่าผลลัพธ์จากการค้นคืนด้วยคำถามเดิมใน 100 อันดับแรกส่วนใหญ่แสดงถึงข้อมูลส่วนบุคคล ข้อมูลการขอความช่วยเหลือ และแสดงถึงการเตรียมการที่จะวางแผนซ่อมแซมอาคาร WTC ที่ถูกทำลายไป ดังนั้นจะเห็นว่าการเปลี่ยนแปลงของเหตุการณ์ต่างๆมีผลต่อการจัดลำดับและแสดงผลลัพธ์การค้นคืน ในงานวิจัยของ Amitay *et al.* (2004) ได้นำเสนอเทคนิคการจัดเก็บองค์ประกอบทางด้านเวลากับเว็บเพจเรียกว่า *“Timestamped Link Profile”* (TLP) โดยจะกำหนดจุดเวลาให้กับเส้นการเชื่อมโยง (timestamped link) ทั้งหมดเพื่อบ่งบอกถึงจุดเวลาที่เส้นการเชื่อมโยงนั้นถูกสร้างขึ้น สำหรับวิธีการคำนวณหาจุดเวลาดังกล่าวมีขั้นตอนรายละเอียดดังนี้

1) พิจารณากิจกรรมหรือเหตุการณ์ต่างๆที่ปรากฏในเนื้อหา (Content) ภายในเว็บเพจนั้นโดยสนใจเฉพาะเหตุการณ์ที่มีความสำคัญและมีผลกระทบต่อเว็บเพจอื่นๆในฐานข้อมูลโดยรวม กำหนดจุดเวลาตามเหตุการณ์นั้นๆ อย่างไรก็ตามจะเห็นว่าวิธีการนี้เป็นวิธีการที่ค่อนข้างลำบากแต่ก็มีความแม่นยำสูง

2) พิจารณาจากข้อมูลส่วนหัวของหน้าเว็บเพจ (header) ที่ตอบกลับมาจากเครื่องแม่ข่ายซึ่งเว็บเพจโดยมากมักจะระบุเวลาที่มีการเปลี่ยนแปลงล่าสุด (last-modified field) ไว้ด้วย

เนื่องจากวิธีการหาจุดเวลาของเส้นการเชื่อมโยงด้วยวิธีการที่ 1 นั้นกระทำได้ยากและใช้เวลานาน ดังนั้นในวิทยาพนธ์เล่มนี้จึงเลือกใช้วิธีที่ 2 แทน นอกจากนี้ Amitay *et al.*, (2004) ยังได้นำเสนอตัวอย่างในการกำหนดจุดเวลาให้กับเส้นการเชื่อมโยงของเว็บเพจที่กล่าวถึงปัญหา Y2K ซึ่งอาจเกิดขึ้นระหว่างปีค.ศ. 1999 ถึงปีค.ศ. 2000 และจากการเก็บรวบรวมข้อมูลทั้งหมด สามารถแสดง TLP ตามช่วงเวลาต่างได้ดังภาพที่ 8 โดยค่าในแกนตั้งแสดงถึงสัดส่วนจำนวนเส้นการเชื่อมโยงที่เกิดขึ้น และค่าในแกนนอนคือช่วงเวลาปีต่างๆ



ภาพที่ 8 TLP แสดงสัดส่วนจำนวนเส้นการเชื่อมโยง ณ ช่วงเวลาต่างๆของเหตุการณ์ Y2K (Amitay *et al.*, (2004))

จากการสังเกต TLP ของเว็บเพจพบว่าสัดส่วนของเส้นการเชื่อมโยงที่มีค่าวันเวลาแก้ไขล่าสุดอยู่ในระหว่างปี ค.ศ. 1999 มีสัดส่วนอยู่เป็นจำนวนมากเมื่อเทียบกับปีอื่นๆเช่น ค.ศ. 2003หรือ ค.ศ. 1998 อันเนื่องจากเหตุการณ์ที่ไม่ปกติกำลังจะเกิดขึ้นหรือที่เรียกว่าปัญหา Y2K โดยมีเส้นเชื่อมโยงที่มีจำนวนเพิ่มมากขึ้น จากงานวิจัยของ Amitay et al. (2004) พบว่าขณะที่เมื่อพิจารณาที่เวลาผ่านไป 4 ปี (เริ่มจากปี 2000 ถึงสิ้นสุดปี 2003) จำนวนเส้นการเชื่อมโยงจากหนึ่งในสามบนหน้าเว็บเพจจะยังคงอยู่ ทำให้ผลลัพธ์การค้นคืนมักแสดงเป็นเว็บเพจที่อยู่ในระหว่างปีค.ศ. 2000 ในลำดับแรกๆ ขณะที่เว็บเพจใหม่ที่สร้างขึ้นอาจเป็นเว็บเพจที่น่าสนใจและมีความสำคัญมากกว่า แต่กลับถูกแสดงไว้ในลำดับการค้นคืนท้ายๆ เนื่องจากมีจำนวนการอ้างอิงปริมาณน้อยกว่า ดังนั้น การสำรวจการเปลี่ยนแปลงของจำนวนเส้นการเชื่อมโยง ของเว็บเพจจะทำให้สามารถวิเคราะห์ได้ว่ามีเหตุการณ์สำคัญบางอย่างเกิดขึ้น ณ ช่วงเวลาหนึ่งๆได้

ในวิทยานิพนธ์เล่มนี้ได้นำเอาหลักการดังกล่าวไปประยุกต์ใช้ในการหาองค์ประกอบค่าแนวโน้ม (Trend Factor) ของเว็บเพจเพื่อใช้ในการพิจารณาค่าความสำคัญของเว็บเพจในช่วงเวลาต่างๆในลำดับต่อไป

อัลกอริทึม Age Based PageRank

เป็นอัลกอริทึมหนึ่งที่นำเสนอวิธีในการจัดลำดับเว็บเพจ โดยมีพื้นฐานของความคิดที่ว่าเว็บเพจที่เกิดขึ้นมาใหม่อาจมีค่าความสำคัญบางอย่าง หากคิดในแง่ของคุณค่าของข้อมูลที่เกิดขึ้น ซึ่งแน่นอนว่าการมีข้อมูลที่สดใหม่จะมีคุณค่ามากกว่าข้อมูลที่เก่า ยกตัวอย่างกรณีเช่น ปัจจุบันปี 2007 เว็บเพจที่แสดงถึงข่าวการจัดกีฬาโอลิมปิกที่กำลังจะจัดขึ้นในปี 2008 ย่อมต้องมีความสำคัญมากกว่าโอลิมปิกที่จัดขึ้นไปแล้วเมื่อปี 2004 ซึ่งจากการพิจารณาฐานข้อมูลเว็บเพจพบว่าเส้นการเชื่อมโยงสำหรับเว็บเพจใหม่นั้นจะมีจำนวนน้อยแต่ก็จะได้รับเส้นการเชื่อมโยงเพิ่มขึ้นจากเว็บเพจต่างๆเมื่อเวลาผ่านไป และในขณะเดียวกันเว็บเพจเก่าซึ่งเคยเป็นเว็บเพจมีคุณค่าแต่เนื่องจากเจ้าของเว็บเพจหยุดการปรับปรุงเปลี่ยนแปลงข้อมูลทำให้เกิดความล้าสมัยของข้อมูล ดังนั้นการจัดลำดับความสำคัญจึงควรที่จะทำให้เว็บเพจเก่ามีค่าความสำคัญลดลงตามอายุของเว็บเพจที่เพิ่มขึ้น ซึ่งจากกระบวนการวิจัยพบว่า การคำนวณเพจแร็งค์ที่ให้ความสำคัญกับเส้นการเชื่อมโยงเพียงอย่างเดียวมีข้อบกพร่องที่จะให้ค่าคะแนนแก่เว็บเพจเก่ามากกว่าเนื่องจากได้รับเส้นการเชื่อมโยงเพิ่มขึ้นตามเวลา ในขณะที่เว็บเพจใหม่จะได้รับค่าเฉลี่ยเพจแร็งค์น้อยเนื่องจากยังไม่เป็นที่รู้จักโดยทั่วไป

จากการพิจาณาถึงอายุของเว็บเพจ Baeza-Yates *et al.* (2002) เสนอฟังก์ชั่นการจัดความสำคัญของเว็บเพจใหม่เก่าตามอายุแสดงได้ดังนี้

กำหนดให้

age_i แทนค่าเวลาจากจุดอ้างอิงถึงเวลาการแก้ไขล่าสุดของเว็บเพจ i

A และ B เป็นค่าคงที่ปรับแต่งได้เพื่อใช้ประกอบในการคำนวณเพจแร็งค์โดยปรับใช้ในการถ่วงน้ำหนักกับอายุของเว็บเพจโดยมีค่าอยู่ระหว่างศูนย์ถึงหนึ่ง

และฟังก์ชั่นค่าความสำคัญของเว็บเพจ i ตามอายุสามารถแสดงได้ดังสมการที่ (6)

$$f(age_i) = (1 + A * e^{-B*age_i}) \qquad (6)$$

จากฟังก์ชั่นของอายุเมื่อนำไปรวมกับเพจแร็งค์สมการที่ (3) สามารถเขียนได้ดังสมการที่ (7)

$$PR^{(k+1)}(y) = d \sum_{x \in B(y)} PR^{(k)}(x) \cdot f(age_y) + (1-d) \cdot s(y) \qquad (7)$$

จากสมการที่ (6) พิจารณาหากเป็นกรณีที่เป็นเว็บเพจใหม่ซึ่งมีค่าอายุ (age_i) เป็นศูนย์ ค่าความสำคัญของเว็บเพจ $f(age_i)$ จะมีค่าเป็น (1+*A*) แต่หากเป็นเว็บเพจเก่ามีค่าอายุมาก (age->∞) จะมีแนวโน้มได้ค่าความสำคัญเป็น 1 ซึ่งเมื่อนำไปปรับปรุงในส่วนพิจารณาการถ่ายทอดความสำคัญของเว็บเพจผ่านการคำนวณเพจแร็งค์แบบดั้งเดิมดังสมการที่ (7) ซึ่งจากการทดลองของ Baeza-Yates *et al.* (2002) กับฐานข้อมูลชิเลียนเว็บซึ่งมีประมาณ 670,000 เว็บเพจ สามารถเพิ่มค่าความสำคัญกับเว็บเพจใหม่ได้แต่อย่างไรก็ตามค่าใหม่ที่ได้ยังเปลี่ยนแปลงไม่มากนักมื่อเปรียบเทียบกับการคำนวณเพจแร็งค์แบบดั้งเดิม ดังนั้นผลลัพธ์ที่ได้ยังถือว่าไม่ค่อยดีนัก

พิจารณาการปรับปรุงฟังก์ชั่นอีกกรณีหนึ่งสำหรับการหาค่าถ่วงน้ำหนักสำหรับเส้นการเชื่อมโยงกันของเว็บเพจโดยใช้หลักการประมาณอัตราการเปลี่ยนแปลงของเว็บเพจคือ พิจารณาเส้นการเชื่อมโยงกรณีจากเว็บเพจ x ไปยังเว็บเพจ y โดยมีจุดเวลาการแก้ไขล่าสุดเขียนแทนด้วย t_x และ t_y ตามลำดับ ซึ่งหากเป็นกรณีปกติเว็บเพจ x มีค่าจุดเวลาการแก้ไขล่าสุดมากกว่าเว็บเพจ y หรือเว็บเพจ x สร้างทีหลังแล้วสร้างเส้นการเชื่อมโยงมายังเว็บเพจ y ซึ่งในกรณีนี้ค่าการถ่วงน้ำหนัก w ของเส้นการเชื่อมโยงควรมีค่าเป็น 1 ในทางกลับกันกรณีที่เว็บเพจ y มีการปรับปรุงเว็บเพจตัวเอง หลังจากได้ถูกผู้อื่นคือเว็บเพจ x เชื่อมโยงมาแล้วค่าการถ่วงน้ำหนักของเส้นการเชื่อมโยงซึ่งได้จากฟังก์ชัน $f(age_i)$ ตามสมการที่ (6)

กำหนดให้

$w(t_x, t_y)$ คือ ฟังก์ชั่นการถ่วงน้ำหนักเส้นเชื่อมโยง จากเว็บเพจที่มีเวลาแก้ไข t_x ไปยังเว็บเพจที่มีเวลาแก้ไข t_y ซึ่งเราจะได้ดังสมการที่ (8)

$$w(t_x, t_y) = \begin{cases} 1 & t_x \geq t_y \\ f(t_y - t_x) & otherwise \end{cases} \qquad (8)$$

โดยที่ f แสดงถึงฟังก์ชั่นการลดทอนตามสมการที่ (6)

จากอัลกอริทึมเพจแร็งค์เดิมตามสมการที่ (3) สำหรับ $t(x,y)$ และ $s(y)$ คำนวณได้ดังนี้

$$t(x,y) = \frac{w(t_x, t_y)}{\sum_{\forall z: x \to z} w(t_x, t_z)} \quad (9)$$

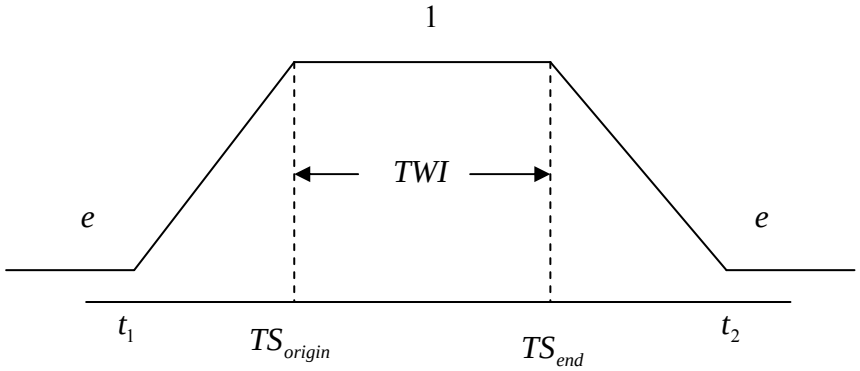
$$s(y) = \frac{f(t_y)}{\sum_{\forall z} f(t_z)} \quad (10)$$

แต่อย่างไรก็ตามจากฟังก์ชั่นการถ่วงน้ำหนักในสมการ (8) ที่ใช้หลักการพิจารณาจุดวันเวลาแก้ไขล่าสุดของเว็บเพจต้นทาง (เว็บเพจที่เป็นผู้ชี้) กับเว็บเพจปลายทาง (เว็บเพจที่ถูกชี้) โดยหากเว็บเพจต้นทางเป็นเว็บเพจใหม่กล่าวคือมีค่าจุดวันเวลาแก้ไขล่าสุดมากกว่าค่าวันเวลาแก้ไขล่าสุดของเว็บเพจปลายทาง ค่าความสำคัญของเส้นการเชื่อมโยงจะถูกกำหนดให้มีค่าเป็น 1 และในทางกลับกันหากเกิดกรณีการเปลี่ยนแปลงแก้ไขปรับปรุงกับเว็บเพจปลายทางทำให้จุดวันเวลาแก้ไขล่าสุดมากกว่าเว็บเพจต้นทางในกรณีนี้ค่าความสำคัญของเส้นการเชื่อมโยงจะมีค่ามากกว่า 1 ซึ่งวิธีการนี้ยังมีข้อบกพร่องอยู่เนื่องจากเราไม่สามารถรู้วันเวลาแก้ไขที่แท้จริงของเว็บเพจต้นทางและเว็บเพจปลายทางของเส้นการเชื่อมโยงแท้จริงได้

อัลกอริทึม T-Rank

ในปี ค.ศ. 2004 Berberich *et al.* ได้ศึกษาถึงผลกระทบของเวลาที่มีต่อการค้นคืนเว็บเพจของระบบสืบค้นข้อมูล โดยได้ยกตัวอย่างผลการค้นคืนด้วยคำถาม (query) ว่า “Olympics opening” ในช่วงฤดูร้อนของปี 2004 โดยผลลัพธ์ที่ค้นคืนจากระบบสืบค้นข้อมูลจะได้เว็บเพจ ที่มีเนื้อหาหรือข้อมูลเกี่ยวกับโอลิมปิคในฤดูหนาวปี 2000 ซึ่งจัดขึ้นที่เมืองเซาท์เลคซิตี้ (South Lake City มลรัฐหนึ่งในอเมริกา) ซึ่งในความเป็นจริงแล้ว ผู้ใช้มักสนใจข้อมูลใหม่ที่กำลังจะมาถึงมากกว่า กล่าวคือโอลิมปิคที่กำลังจะจัดขึ้นที่เอเทน (Athens) ในปี 2004 ณ ประเทศกรีก สาเหตุที่เป็นเช่นนี้เนื่องจากระบบสืบค้นข้อมูลได้จัดเรียงผลลัพธ์การค้นคืนตามค่าเพจแร็งค์และเว็บเพจที่มีข้อมูลโอลิมปิคในปี 2000 เป็นเว็บเพจที่เกิดขึ้นมาก่อนแล้วและมีเว็บเพจอื่นสร้างเส้นเชื่อมโยงมาอ้างอิงมากจึงมีค่าเพจแร็งค์มากกว่าข้อมูลในปี 2004 ในขณะที่เว็บเพจโอลิมปิคเอเทน 2004 มีค่าเพจแร็งค์ที่ต่ำเพราะเป็นเว็บเพจใหม่ มีอายุน้อยและยังไม่เป็นที่รู้จักโดยทั่วไปในระบบอินเตอร์เน็ตนั่นเอง Berberich *et al.*

ค่าความสด สำหรับในงานวิจัยนี้ กำหนดนิยามคือค่าที่แสดงถึงความสดใหม่ของเว็บเพจหากเว็บเพจนั้นๆมีการเปลี่ยนแปลงภายในกรอบช่วงเวลาของการสนใจ ก็จะมีค่าความสดมาก แสดงดังภาพที่ 9 ค่าความสดของเว็บเพจจะมีค่าสูงสุดเป็น 1 หากจุดเวลาของการแก้ไขครั้งล่าสุดของเว็บเพจอยู่ภายในกรอบเวลาการพิจารณาเวลาจุดเริ่มต้น (TS_{origin}) และจุดสิ้นสุดความสนใจ (TS_{end}) โดยช่วงเวลาของการสนใจนี้เรียกว่า "Temporal Window of Interest" (TWI)

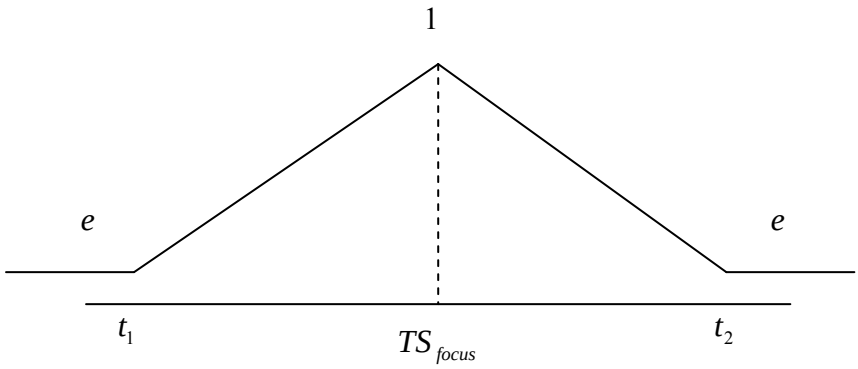


ภาพที่ 9 ค่าความสดเมื่อมีการเปลี่ยนแปลง ณ จุดเวลาต่างๆพิจารณาบนกรอบ temporal window of interest ณ จุดเวลา (timestamp)

จากกราฟดังภาพที่ 9 แกนนอน คือ เวลาซึ่งมีค่าเพิ่มขึ้นจากซ้ายไปขวา และแกนตั้ง แสดงค่าความสด ณ จุดเวลา (timestamp) ต่างๆ ซึ่งค่าความสด ณ จุดเวลาระหว่าง TS_{origin} กับ TS_{end} หรือภายในช่วงของ TWI กำหนดให้มีค่าความสดสูงสุดเท่ากับ 1 โดยที่ค่าความสดนี้จะมีค่าน้อยกว่า 1 ในช่วงก่อนและหลัง TWI ที่สนใจ โดยจะกำหนดให้มีค่าความสดต่ำสุดเท่ากับ e ($=10^{-7}$) ณ จุดเวลาก่อน t_1 และหลัง t_2 ซึ่งสมการคำนวณค่าความสด ณ จุดเวลา ts_x เมื่อ x คือเว็บเพจใดๆ สามารถแสดงเป็นสมการได้ดังนี้

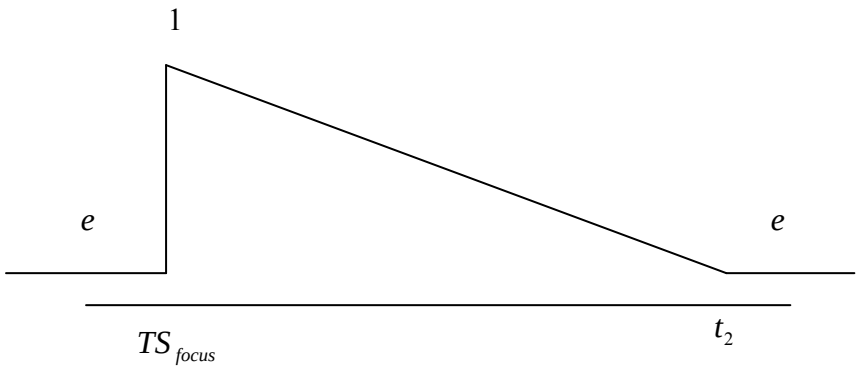
$$f(ts_x) = \begin{cases} 1 & TS_{origin} \le ts_x \le TS_{end} \\ \dfrac{1}{(TS_{origin} - ts_x) + 1} & t_1 \le ts_x < TS_{origin} \\ \dfrac{1}{(ts_x - TS_{end}) + 1} & TS_{end} < ts_x \le t_2 \\ e & otherwise \end{cases} \qquad (11)$$

จาก TWI ข้างต้น หากเป็นกรณีที่ผู้ใช้สนใจถึงการเกิดเหตุการณ์ ณ จุดเวลาหนึ่งๆเท่านั้น กรอบเวลา TWI นี้สามารถนำมาแก้ไขใหม่ และเรียกว่า “temporal focus of attention” ซึ่งในกรณีนี้จุดเวลา TS_{origin} จะมีค่าเท่ากับ TS_{end} ซึ่งเขียนแทนด้วย TS_{focus} โดยค่าความสดของเว็บเพจ ณ จุดเวลาดังกล่าวให้มีค่าสูงสุดเท่ากับ 1 ดังภาพที่ 10



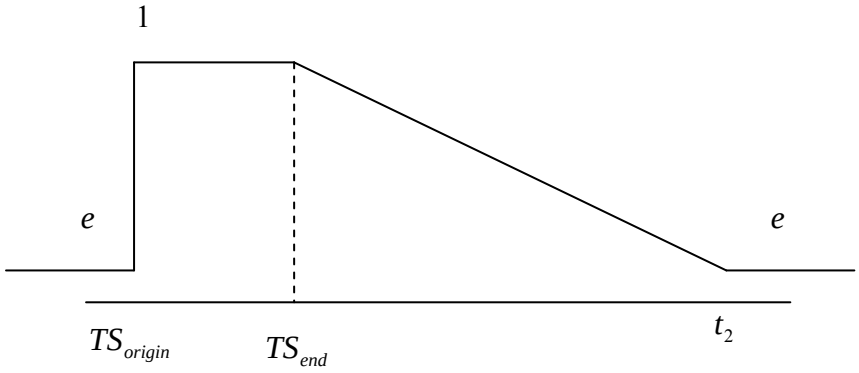
ภาพที่ 10 แสดง temporal window of interest กรณี $TS_{origin} = TS_{end}$

ตัวอย่างเช่น เหตุการณ์ก่อการร้ายในเมืองเมดริด (Madrid) ณ วันที่ 11 มีนาคม ค.ศ. 2004 ซึ่งในกรณีนี้จุดเวลา $TS_{origin} = TS_{end} = TS_{focus}$ = “11 มีนาคม ค.ศ. 2004” และถ้าหากผู้ใช้ไม่สนใจเว็บเพจที่เกิดขึ้นก่อนหน้าเหตุการณ์ก่อการร้ายในเมดริด กล่าวคือเว็บเพจที่มีการเปลี่ยนแปลง ณ จุดเวลาก่อนหน้าที่จะเกิดเหตุการณ์ถูกกำหนดให้ไม่มีความสำคัญ ในขณะหากเปลี่ยนแปลง ณ จุดเวลาตรงกับเหตุการณ์จะมีความค่าความสดสูงสุดเป็น 1 และมีค่าความสดน้อยลงเมื่อผ่านจุดเวลาดังกล่าวนั้นไปแล้ว ในกรณีนี้สามารถปรับค่าจุดเวลาเริ่มต้นสนใจโดยกำหนดให้ $t_1 = TS_{focus}$ = “11 มีนาคม ค.ศ. 2004” ซึ่งสามารถแสดงได้ดังภาพที่ 11



ภาพที่ 11 แสดง temporal window of interest กรณี $t_1 = TS_{focus}$

และหากเป็นกรณีที่ผู้ใช้สนใจเฉพาะเวลา ณ ช่วงเวลาขณะหนึ่ง (period) โดยไม่สนใจถึงช่วงของเวลาก่อนหน้า (ก่อน TS_{origin}) กล่าวคือเว็บเพจที่ถูกสร้างขึ้นจะยังไม่ได้รับความสนใจจนกระทั่งถึงจุดเวลาหนึ่ง (TS_{origin}) จึงได้รับความสนใจและมีค่าความสดสูงสุดเป็น 1 จนกว่าจะหมดช่วงเวลาของเหตุการณ์นั้นๆ ยกตัวอย่างเช่น เหตุการณ์ผู้ใช้สนใจข้อมูลการแข่งขันปั่นจักรยานนานาชาติที่ประเทศฝรั่งเศส (Tour de France) ในปี 2003 ซึ่งจะได้ว่า $t_1 = TS_{origin}$ และ $TS_{end} = TS_{origin} + 3$ เดือน เป็นต้น สามารถแสดงได้ดังภาพที่ 12



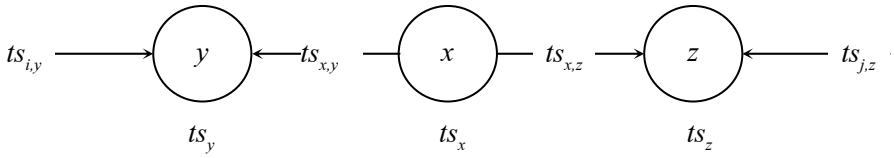
ภาพที่ 12 แสดง temporal window of interest กรณี $t_1 = TS_{Origin}$

สำหรับ *ค่ากิจกรรม* ของเว็บเพจใดๆ พิจารณาจากความถี่ที่เว็บเพจนั้นมีการเปลี่ยนแปลง (Update rate) ซึ่งคำนวณได้จากผลรวมของค่าความสด ณ จุดเวลา ($TS_{modified}$) ซึ่งมาจากการที่เครื่องมือเก็บข้อมูลตรวจพบการเปลี่ยนแปลงในช่วงระหว่าง t_1 และ t_2 แสดงดังสมการที่ (12) ต่อไปนี้

$$a(x) = \begin{cases} \sum_{t_1}^{t_2} f(ts) & ts \in TS_{modified}, TS_{modified} \neq \phi \\ e' & Otherwise \end{cases} \qquad (12)$$

ซึ่งจะเห็นว่าเว็บเพจใดมีการเปลี่ยนแปลงบ่อยครั้ง ก็จะยิ่งมีค่ากิจกรรมสูงตามไปด้วย ในทางกลับกันหากเว็บเพจไม่มีการเปลี่ยนแปลงเลย นั่นหมายความว่าค่ากิจกรรมจะต้องน้อยมาก ซึ่งในทางปฏิบัติกำหนดให้มีค่าต่ำสุดเท่ากับ e' ($=10^{-7}$)

พิจารณาตัวอย่างเว็บกราฟแสดงในภาพที่ 13 ซึ่งประกอบไปด้วยเว็บเพจ x, y และ z กำหนดให้เว็บเพจมีจุดเวลาแก้ไขล่าสุดเท่ากับ ts_x, ts_y และ ts_z และจุดเวลาแก้ไขล่าสุดของเส้นการเชื่อมโยงเท่ากับ $ts_{x,y}$, $ts_{x,z}$, $ts_{i,y}$ และ $ts_{j,z}$



ภาพที่ 13 ตัวอย่างเว็บกราฟอย่างง่ายซึ่งแสดงถึงจุดเวลาแก้ไขล่าสุดสำหรับเว็บเพจและเส้นการเชื่อมโยง

การประยุกต์ใช้ค่าความสดและค่ากิจกรรมในอัลกอริทึม T-Rank มีแนวทางโดยแก้ไขปรับปรุงสมการอัลกอริทึมเพจแร็งค์เดิมจากสมการที่ (3) ในส่วนที่เป็น $t(x,y)$ และ $s(y)$ เป็นดังสมการที่ (13) และ (14) ตามลำดับ

กำหนดให้

- $f(x)$ แทน ค่าฟังก์ชั่นความสดของเว็บเพจ x ณ จุดเวลา ts
- $f(x,y)$ แทน ค่าฟังก์ชั่นความสดของเส้นการเชื่อมโยงจากเว็บเพจ x ไปยัง y ณ จุดเวลา ts
- $a(x)$ แทน ค่ากิจกรรมของเว็บเพจ x
- $a(x,y)$ แทน ค่ากิจกรรมของเส้นการเชื่อมโยงจากเว็บเพจ x ไปยัง y
- w_{ti} และ w_{si} แทน ค่าถ่วงน้ำหนักของค่าในแต่ละค่าในฟังก์ชั่น t และ s ตามลำดับเมื่อ $i = 1,2,3,\ldots$

จากอัลกอริทึมเพจแร็งค์เดิมตามสมการที่ (3) สำหรับ $t(x,y)$ และ $s(y)$ คำนวณได้ดังนี้

$$t(x,y) = w_{t1}\frac{f(y)}{\sum_{\forall z:x\to z} f(z)} + w_{t2}\frac{f(x,y)}{\sum_{\forall z:x\to z} f(x,z)} + w_{t3}\frac{avg\{f(v,y) \mid \forall v : v \to y\}}{\sum_{\forall v:v\to w} avg\{f(v,w) \mid \forall v \forall w : v \to w\}} \quad (13)$$

$$s(y) = w_{s1}\frac{f(y)}{\sum_{\forall z} f(z)} + w_{s2}\frac{a(y)}{\sum_{\forall z} a(z)} + w_{s3}\frac{avg\{f(v,y) \mid \forall v : v \to y\}}{\sum_{\forall z} avg\{f(w,z) \mid \forall w : w \to z\}} + w_{s4}\frac{avg\{a(v,y) \mid \forall v : v \to y\}}{\sum_{\forall z} avg\{a(w,z) \mid \forall w : w \to z\}} \quad (14)$$

จากสมการที่ (13) และ (14) จะเห็นว่าอัลกอริทึม T-Rank ที่ Berberich *et al.* (2004) นำเสนอนั้นเป็นการช่วยจัดลำดับเว็บเพจในลักษณะกว้างๆกล่าวคืออัลกอริทึมอาศัยผู้ใช้เป็นผู้กำหนดค่าพารามิเตอร์ (parameter) พื้นฐานตามความสนใจ หากต้องการเพิ่มค่าความสำคัญให้แก่เว็บเพจช่วงใดก็กำหนดค่าพารามิเตอร์จุดเวลาเริ่มต้น TS_{origin} และจุดสิ้นสุด TS_{end} ของความสนใจ โดยค่าความสดของเว็บเพจที่อยู่ในช่วงที่สนใจจะถูกกำหนดให้มีค่าสูงสุดเป็น 1 ในขณะที่เว็บเพจที่อยู่นอกช่วงเวลาดังกล่าวก็จะถูกลดทอนค่าความสดลงแบบเส้นตรง (linear) ซึ่งสุดท้ายค่าของผลลัพธ์ที่ได้จากการคำนวณจะสามารถตอบคำถามการสืบค้น ได้เพียงคำถามใดคำถามหนึ่งได้เท่านั้น เช่น "Olympic opening" หรือไม่ว่าจะเป็น "Y2K problem" ซึ่งหากต้องการผลลัพธ์อื่นๆจำเป็นที่จะต้องกำหนดค่าพารามิเตอร์ดังกล่าวใหม่และคำนวณเพจแร็งค์อีกครั้ง ทำให้ต้องคำนวณเพจแร็งค์หลายครั้ง และผลลัพธ์อาจมีโอกาสเกิดข้อผิดพลาดขึ้นได้ หากได้กำหนดค่าพารามิเตอร์ที่ไม่เหมาะสม

สำหรับค่ากิจกรรมของเว็บเพจ อาจถือได้ว่าเป็นปัจจัยที่ดีตัวหนึ่งในการกำหนดค่าความสำคัญของเว็บเพจเนื่องจากโดยความหมายแล้วเว็บเพจที่ดีควรจะเป็นเว็บเพจที่มีการปรับปรุงอยู่เสมอ แต่ก็มีโอกาสสูงที่จะก่อให้เกิดข้อผิดพลาดบางอย่างขึ้นได้เนื่องจากระบบอินเตอร์เน็ตในปัจจุบันมีขนาดใหญ่ การใช้เครื่องมือที่เรียกว่า "เว็บคราวเลอร์" (web crawler) เพื่อเก็บรวบรวมข้อมูลเว็บเพจจำเป็นต้องมีรอบ (period) การเก็บข้อมูลหลายครั้ง ซึ่งโดยปกติเราจะมีทรัพยากร (resource) อยู่อย่างจำกัด เช่น ช่องทางการสื่อสารของเครือข่ายข้อมูล (bandwidth) ทำให้การคำนวณค่ากิจกรรมของเว็บเพจหนึ่งๆจึงเป็นไปได้ลำบาก นอกจากนี้ยังพบว่าอัลกอริทึมดังกล่าวต้องมีการปรับแต่งค่าถ่วงน้ำหนัก w_{ti} และ w_{si} ให้เหมาะสม ซึ่งค่านี้ขึ้นอยู่กับลักษณะของเว็บกราฟอีกด้วย

อัลกอริทึม TimedPageRank

จากการศึกษาโครงสร้างเว็บกราฟของ Yu *et al.* (2005) พบข้อบกพร่องในการคำนวณด้วยอัลกอริทึมเพจแร็งค์แบบดั้งเดิมเช่นกัน โดยในงานวิจัยพบว่าเว็บเพจที่มีคุณภาพในอดีตอาจไม่มีคุณภาพแล้วในปัจจุบันทั้งนี้เนื่องจากอัลกอริทึมเพจแร็งค์ดังกล่าวเป็นเทคนิคที่เอื้อให้ค่าความสำคัญต่อเว็บเพจเก่า ซึ่งมักมีจำนวนการอ้างอิงมาก ซึ่งต่างกับเว็บเพจใหม่ซึ่งยังไม่เป็นที่รู้จักจึงมีจำนวนการอ้างอิงน้อย นอกจากนี้ Yu *et al.* (2005) ยังได้จำแนกคุณลักษณะโดยทั่วไปของเว็บเพจ ดังรายละเอียดต่อไปนี้

1 **เว็บเพจเก่า (Old Pages)** เป็นเว็บเพจที่ปรากฏขึ้นบนอินเตอร์เน็ตเป็นเวลานานแล้วซึ่งสามารถแบ่งออกได้เป็นเว็บเพจที่มีคุณภาพ (quality pages) มีเส้นการเชื่อมโยงเข้าหาเป็น

1.1 **เว็บเพจเก่าที่มีคุณภาพและมีการปรับปรุงข้อมูล (Old quality pages that are up-to-date)** แสดงถึงเว็บเพจที่ปรากฏขึ้นบนอินเตอร์เน็ตมาเป็นเวลานานซึ่งมีจำนวนจุดเชื่องโยงเข้าหามาก และมีการเปลี่ยนแปลงแก้ไขข้อมูลให้ทันสมัยอยู่ตลอดเวลา ดังนั้นเว็บเพจนี้จึงถือว่าเป็นเว็บเพจเก่าที่มีคุณภาพ

1.2 **เว็บเพจเก่าที่มีคุณภาพแต่ไม่มีการปรับปรุงข้อมูล (Old quality pages that are not up-to-date)** แสดงถึงเว็บเพจที่ปรากฏขึ้นบนอินเตอร์เน็ตมาเป็นเวลานานแล้ว แต่ผู้สร้างได้หยุดเปลี่ยนแปลงแก้ไขข้อมูลบนหน้าเว็บเพจนั้น ดังนั้นทำให้ข้อมูลที่มีอยู่อาจเกิดความล้าสมัย

1.3 **เว็บเพจเก่าทั่วไปที่ยังคงเป็นเว็บเพจทั่วไป (Old common pages that remain common pages)** แสดงถึงเว็บเพจที่ปรากฏขึ้นบนระบบอินเตอร์เน็ตมาเป็นเวลานานแล้ว แต่มีกลุ่มเว็บเพจอื่นอ้างอิงมาหาไม่มากและเมื่อเวลาผ่านไปก็ยังคงมีกลุ่มเว็บเพจจำนวนไม่มากอ้างอิงมาหา

1.4 **เว็บเพจเก่าทั่วไปที่กลายเป็นเว็บเพจที่มีคุณภาพ (Old common pages that have become quality pages)** แสดงถึงเว็บเพจที่ปรากฏขึ้นบนอินเตอร์เน็ตมาเป็นเวลานานแล้ว และเมื่อเวลาผ่านไปจุดเชื่อมโยงที่ได้รับจากเว็บเพจอื่นมีจำนวนมากขึ้นซึ่งอาจเกิดจากมีการเพิ่มข้อมูลที่สำคัญในช่วงเวลาเก่านานมาแล้ว

2 **เว็บเพจใหม่ (New Pages)** เป็นเว็บเพจที่เพิ่งจะปรากฏขึ้นบนอินเตอร์เน็ต ซึ่งสามารถแบ่งออกเป็นกลุ่มได้ดังนี้

2.1 **เว็บเพจใหม่ที่มีคุณภาพ (New quality pages)** แสดงถึงเว็บเพจที่เพิ่งปรากฏขึ้นบนอินเตอร์เน็ตและมีข้อมูลเนื้อหาที่มีคุณภาพ แต่ด้วยเหตุว่าเป็นเว็บเพจที่ใหม่ดังนั้นจึงได้รับการอ้างอิงจากเว็บเพจอื่นน้อย เนื่องจากยังไม่เป็นที่รู้จักเท่าที่ควร อย่างไรก็ตามเมื่อเวลาผ่านไป เว็บเพจดังกล่าวนี้มักได้รับการอ้างอิงเพิ่มขึ้นในอัตราที่รวดเร็ว

2.2 **เว็บเพจใหม่ทั่วไป (New common pages)** แสดงถึงเว็บเพจที่เพิ่งปรากฏขึ้นบนอินเตอร์เน็ตซึ่งได้รับความสนใจและอ้างอิงถึงไม่มากนัก และเนื่องด้วยเป็นเว็บเพจที่มีข้อมูลเนื้อหาทั่วไปดังนั้นเมื่อเวลาผ่านไป ก็ยังคงมีเว็บเพจจำนวนไม่มากนักอ้างอิงมาหา

จากการแบ่งชนิดของเว็บเพจข้างต้น ในงานวิจัยนี้ Yu *et al.* (2005) นำเสนอแนวทางในการปรับปรุงอัลกอริทึมเพจแร็งค์เดิม โดยการมีจุดมุ่งหมายในการแก้ข้อบกพร่อง 2 ประการคือ

1. แก้ปัญหาการกำหนดค่าความสำคัญน้อยสำหรับเว็บเพจเก่าที่มีคุณภาพแต่ไม่มีการปรับปรุงข้อมูล (Old quality pages that are not up-to-date)
2. แก้ปัญหาสำหรับเว็บเพจใหม่โดยกำหนดค่าความสำคัญให้มากขึ้นสำหรับเว็บเพจใหม่ที่มีคุณภาพ (New quality pages)

และเพื่อให้สะดวกต่อการประเมินความถูกต้องเหมาะสมของอัลกอริทึม ในงานวิจัยของ Yu *et al.* (2005) จึงได้นำเสนอ และใช้ฐานข้อมูลบทความงานวิจัยแทนโครงสร้างของเว็บกราฟในการทดสอบ เนื่องจากโครงสร้างความสัมพันธ์การเชื่อมโยงของบทความวิจัยมีลักษณะใกล้เคียงกับเว็บกราฟ กล่าวคือ กำหนดให้แต่ละบทความคือโหนดในกราฟ และการอ้างอิงของบทความคือความสัมพันธ์การเชื่อมโยงของแต่ละโหนด ดังนั้น เราจึงสามารถคำนวณค่าความสำคัญของแต่ละบทความได้ นอกจากนี้ข้อดีอีกประการหนึ่งก็คือ เราสามารถทราบจุดเวลาของการตีพิมพ์ที่ปรากฏบนข้อความนั้นๆได้อย่างค่อนข้างชัดเจน

Yu *et al.* (2005) ได้นำอัลกอริทึมเพจแร็งค์พื้นฐานมาประยุกต์เพื่อคำนวณค่าความสำคัญของแต่ละบทความวิจัย โดยใช้ชื่อว่า “TimedPageRank” (TPR) อัลกอริทึมนี้นำเสนอวิธีการปรับลดค่าความสำคัญลงสำหรับบทความเก่า ด้วยอัตราความเสื่อมสภาพของบทความตามกาลเวลาที่เปลี่ยนไป

กำหนดให้

$C(x)$ แทน จำนวนการอ้างอิง (Citation count) ของบทความ x

w_x คือ ค่าถ่วงน้ำหนัก ซึ่งเป็นอัตราการเสื่อมสภาพของบทความ x

N คือ จำนวนบทความทั้งหมด

จากอัลกอริทึมเพจแร็งค์สมการที่ (3) สำหรับ $t(x,y)$ และ $s(y)$ คำนวณได้ดังนี้

$$t(x,y) = w_x \cdot \frac{1}{C(x)} \tag{15}$$

$$s(y) = \frac{1}{N} \tag{16}$$

โดยที่ w_x สามารถคำนวณได้จาก

$$w_x = DecayRate^{\frac{T-t_x}{12}} \tag{17}$$

เมื่อ $DecayRate$ เป็นค่าคงที่ค่าหนึ่ง (ในงานวิจัยกำหนดให้มีค่าเท่ากับ 0.5) และ T คือ จุดเริ่มต้นเวลาที่เราสนใจ (ในทางปฏิบัติกำหนดให้เป็น ณ จุดเวลาปัจจุบัน) และ t_x คือ จุดเวลาที่บทความ x

$T - t_x$ x

จากงานวิจัยนี้จะเห็นว่าค่าพารามิเตอร์การเสื่อมสภาพนั้นเป็นการคิดจากการทดลองซึ่งในงานวิจัย Yu *et al.* (2005) ใช้เท่ากับ 0.5 เพื่อให้ง่ายต่อการทดลองแต่ในความเป็นจริงแล้วค่าดังกล่าวนี้จะต้องมีการทดลองปรับอีกครั้งเพื่อหาอัตราที่เหมาะสมของอัตราการเสื่อมสภาพของบทความวิจัย ซึ่งจากคุณลักษณะเด่นของบทความวิจัยคือ มีการอ้างอิงระหว่างบทความแบบถาวร (permanent) กล่าวคือ เมื่อบทความวิจัยได้รับการตีพิมพ์แล้วจะไม่สามารถเปลี่ยนแปลงแก้ไขข้อมูลทั้งเนื้อหาและการอ้างอิงได้ ดังนั้นการหาค่าการเสื่อมสภาพนั้นสามารถทำได้ง่ายโดยดูจากอัตราการอ้างอิงของบทความวิจัยซึ่งจะลดลงตามเวลา ซึ่งแตกต่างกับเว็บเพจที่อาจมีการเปลี่ยนแปลงแก้ไขจุดเชื่อมโยง กล่าวคือเว็บเพจเก่าที่ปรากฏบนอินเตอร์เน็ตมีโอกาสที่จะเปลี่ยนแปลงแก้ไขข้อมูลหรือจุดเชื่อมโยงได้อีก ขณะเดียวกันค่าเวลาสำหรับการตีพิมพ์บทความวิจัยนั้นสามารถหาได้ แต่หากเปรียบเทียบกับระบบเว็บเพจที่ค่าเวลาการปรากฏของเว็บเพจนั้น ไม่สามารถตรวจสอบได้ ทำให้การกำหนดค่าพารามิเตอร์การเสื่อมสภาพให้เหมาะสมกับฐานข้อมูลที่มีอยู่ทำได้ยาก เพราะมีโอกาสที่ค่านี้จะเปลี่ยนไปในแต่ละช่วงเวลาเช่นกัน ซึ่งหากการประเมินค่าการเสื่อมสภาพไม่ถูกต้อง จะส่งผลให้ได้ผลลัพธ์ที่ผิดพลาด แต่อย่างไรก็ตามเราได้นำเอาแนวความคิดของอัลกอริทึมนี้ไปใช้เป็นแนวทางในการปรับปรุงวิธีการคำนวณเพจแร็งค์ใหม่ในวิทยานิพนธ์ฉบับนี้ ซึ่งจะได้กล่าวต่อไป

จากการศึกษางานวิจัยต่างๆดังที่ได้กล่าวมาแล้วข้างต้น สามารถนำมาสรุปเปรียบเทียบฟังก์ชั่นการถ่วงน้ำหนักได้ดังตารางที่ 3

ตารางที่ 3 เปรียบเทียบคุณลักษณะด้านต่างๆของอัลกอริทึม Age Based PageRank, T-Rank และ TimedPageRank

	Age Based PageRank	T-Rank	TimedPageRank
ฟังก์ชันการลดทอนค่าความสำคัญของเว็บเพจ	$f(age_i) = (1 + A * e^{-B*age_i})$	$f(ts) = \begin{cases} 1 & TS_{origin} \le ts \le TS_{end} \\ \frac{1}{(TS_{origin} - ts)+1} & t_1 \le ts < TS_{origin} \\ \frac{1}{(ts - TS_{end})+1} & TS_{end} < ts \le t_2 \\ e & otherwise \end{cases}$	$w_x = DecayRate^{\frac{T-t_x}{12}}$
พารามิเตอร์เริ่มต้นของระบบ	A และ B ซึ่งใช้พิจารณาค่าอายุของเว็บเพจจำเป็นต้องมีการปรับแต่งค่า	TS_{origin} , TS_{end} , $TS_{modified}$ ผู้ใช้จำเป็นต้องกำหนดจุดเวลาของเหตุการณ์ต่างๆที่สนใจ	$DecayRate$ ต้องมีการปรับแต่งค่าโดยอิงจากผลการทดลอง
ความรวดเร็วในการเตรียมข้อมูลก่อนการทดลอง	เตรียมได้รวดเร็ว	เตรียมได้ช้าเนื่องจากต้องคำนวณค่ากิจกรรมของเว็บเพจ ก่อน	เตรียมได้รวดเร็ว
การใช้ทรัพยากรของระบบในการคำนวณ	ใช้ทรัพยากรการสื่อสารบนเครือข่ายไม่มากนักเนื่องจากเก็บรวบรวมข้อมูลเพียงจุดเวลาที่มีการแก้ไขล่าสุดเท่านั้น	ใช้ทรัพยากรของระบบมาก เช่น ทรัพยากรการสื่อสารบนเครือข่าย (bandwidth) เนื่องจากต้องเก็บรวบรวมข้อมูลการเปลี่ยนแปลงทั้งหมดของทุกๆเว็บเพจ	ใช้ทรัพยากรการสื่อสารบนเครือข่ายไม่มากนักเนื่องจากเก็บรวบรวมข้อมูลเพียงจุดเวลาที่มีการแก้ไขล่าสุดเท่านั้น

อุปกรณ์และวิธีการ

อุปกรณ์

ฮาร์ดแวร์

1. เครื่องคอมพิวเตอร์ Radiant cluster 1 เครื่อง ประกอบด้วยอุปกรณ์ดังต่อไปนี้
   - ซีพียู AMD Opteron 240
   - หน่วยความจำหลัก 6 GB
   - ฮาร์ดดิกส์ขนาด 80 GB
   - การ์ดแลน

ซอฟต์แวร์

1. ระบบปฏิบัติการ Windows XP Professional
2. ระบบปฏิบัติการ Linux
3. Java1.5 คอมไพเลอร์
4. Eclipse (Java IDE)

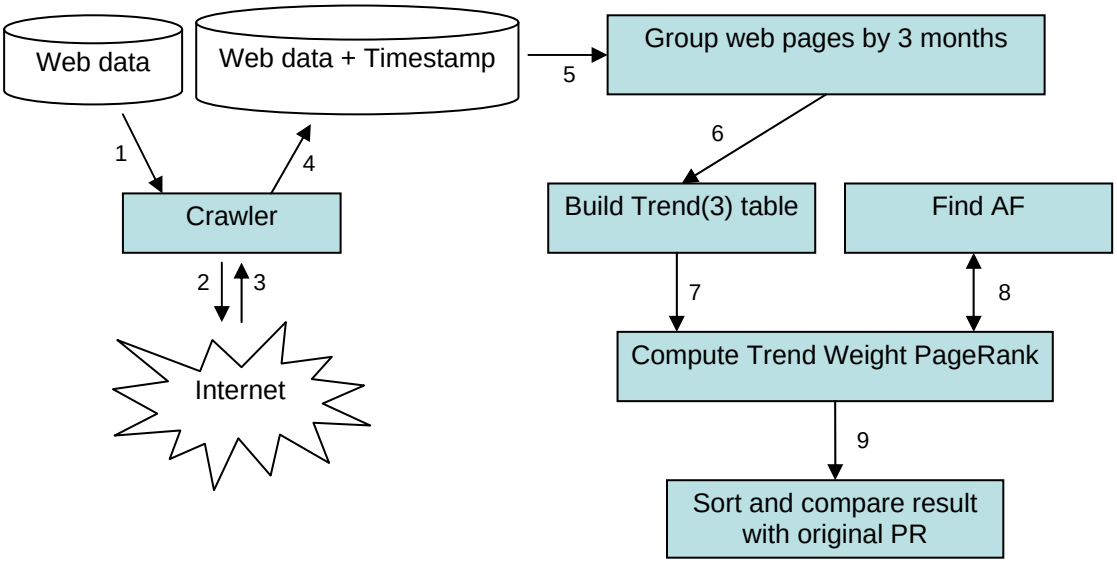
ชุดข้อมูลทดสอบ

ทดสอบบนฐานข้อมูลปีค.ศ. 2007 (โดเมนต่างประเทศ)แต่เนื่องด้วยฐานข้อมูลเว็บเพจปัจจุบันมีขนาดใหญ่ดังนั้นจึงทำการเก็บข้อมูลเฉพาะบางหัวเรื่อง (Topic) โดยหัวข้อที่เลือกคือคือ “Harry Potter”, “NCSEC”, “Olympic”, “Thailand tsunami”, “Tsunami”, “World trade center” จากนั้นใช้เครื่องมือ Google Web API ค้นหาเว็บเพจ ที่เป็นชุดตั้งต้น (Seed) โดยใช้หัวเรื่องดังที่กล่าว จากนั้นนำเว็บเพจชุดตั้งต้นมาขยายผลต่อไปโดยทำการเก็บเว็บเพจที่ชุดเว็บเพจตั้งต้นอ้างอิงถึง (Forward Link) และที่ถูกได้รับการอ้างอิง (Back Link) หลังจากที่ได้ฐานข้อมูลแล้วการใช้เครื่องมือสำหรับหาวันเวลาการแก้ไขล่าสุด (last modified) ของเว็บเพจแต่ละเว็บเพจนั้นต้องมีการพัฒนาให้มีการทำงานแบบขนาน และมีช่องการสื่อสารข้อมูล (bandwidth) ที่เพียงพอ ซึ่งสำหรับในวิทยานิพนธ์เล่มนี้พัฒนาเครื่องมือเก็บข้อมูลบนเครื่องที่มีการทำงานแบบขนานแต่ช่องทางการสื่อสารมีจำกัด ทำให้ทำการเก็บข้อมูลได้ช้า ดังนั้นผู้วิจัยจึงใช้วิธีการตัดกลุ่มของเว็บเพจชุดทดลองดังกล่าวเพื่อทำการศึกษาเบื้องต้นจำนวนประมาณ 380,458 เว็บเพจ ซึ่งประกอบไปด้วยจำนวนการอ้างอิง (citation) เท่ากับ 11,970,507 หรือโดยเฉลี่ยประมาณ 31.5 จำนวนการอ้างอิงต่อเว็บเพจ

วิธีการ

ภาพรวมของระบบ

ตามที่ได้กล่าวรายละเอียดในหัวข้อความรู้เบื้องต้นว่า เว็บเพจต่างๆบนอินเตอร์เน็ตจะประกอบไปด้วยเว็บเพจเก่า (old pages) และเว็บเพจใหม่ (new pages) ซึ่งในวิทยานิพนธ์ฉบับนี้เราสนใจถึงกรณีการแก้จุดด้อยของการคำนวณค่าความสำคัญให้กับเว็บเพจใหม่ ซึ่งตามอัลกอริทึมเพจแร็งค์แบบดั้งเดิม (Page *et al.*, 1998) พบว่าค่าความสำคัญของเว็บเพจใหม่เหล่านี้ ไม่ว่าจะเป็นเว็บเพจใหม่ที่มีคุณค่า (new quality page) หรือเว็บเพจทั่วไป (new common page) จะมีค่าน้อย เนื่องจากยังไม่เป็นที่รู้จักและมีเว็บเพจอื่นอ้างอิงมาหาไม่มากนัก หรืออาจกล่าวได้ว่า เว็บเพจใหม่เหล่านี้มักถูกจัดอยู่ในลำดับท้ายๆของผลลัพธ์การค้นคืน แต่อย่างไรก็ตามบางเว็บเพจใหม่อาจเป็นเว็บเพจที่มีคุณภาพ (ข้อมูลมีความสด) ก็ควรจะได้รับการกำหนดหรือคำนวณค่าความสำคัญใหม่ให้มีค่าเพิ่มขึ้นจากเดิม สำหรับแนวทางในการพัฒนาจะประยุกต์ใช้เทคนิคของ Yu *et al.* (2005) โดยกำหนดฟังก์ชันถ่วงน้ำหนักแนวโน้มตามเวลา (trend weight factor) ให้เป็นฟังก์ชันที่แสดงถึงแนวโน้มของจำนวนเว็บเพจที่มีการเปลี่ยนแปลง ณ ช่วงเวลาต่างๆเมื่อเทียบกับเว็บเพจทั้งหมดในฐานข้อมูลและจะถูกลดทอนลงตามเวลา กำหนดด้วยอายุ (aging factor) ของเว็บเพจนั้นๆ สำหรับภาพรวมการทำงานของระบบได้ออกแบบไว้ดังภาพที่ 14



ภาพที่ 14 แสดงภาพรวมการทำงานของอัลกอริทึมเพจแรงค์แบบถ่วงน้ำหนักตามแนวโน้ม

จากภาพที่ 14 ขั้นตอนการทำงานของระบบจะเป็นไปตามทิศทางของลูกศร โดยมีขั้นตอนต่างๆตามหมายเลขที่กำกับ ขั้นตอนที่ 1 ระบบจะอ่านฐานข้อมูลเว็บเพจที่มีอยู่ในฐานข้อมูลชุดทดสอบเพื่อตรวจสอบเวลาการแก้ไขล่าสุดจากเครื่องแม่ข่ายในอินเตอร์เน็ตดังขั้นตอนที่ 2 และ 3 ซึ่งรายละเอียดวิธีการหาเวลาการแก้ไขล่าสุดนั้นจะได้กล่าวถึงในลำดับต่อไป จากนั้นเมื่อระบบได้ค่าเวลาวันเวลาแก้ไขล่าสุดจากเครื่องเว็บเพจแม่ข่ายแล้ว ขั้นตอนที่ 4 ระบบบันทึกข้อมูลเว็บเพจพร้อมทั้งวันเวลาแก้ไขล่าสุดที่ได้ลงฐานข้อมูลใหม่ ขั้นตอนที่ 5 นำข้อมูลเว็บเพจและวันเวลาแก้ไขล่าสุดที่ได้ไปจัดกลุ่มเว็บเพจตามช่วงระยะเวลาการแก้ไขข้อมูลเว็บเพจส่วนใหญ่ (รายละเอียดประกอบดังภาพที่ 6) ซึ่งในกรณีนี้ใช้ระยะเวลา 3 เดือน ขั้นตอนที่ 6 สร้างตารางแนวโน้มจากเว็บเพจและเส้นการเชื่อมโยง (รายละเอียดประกอบดังตารางที่ 4 และ 5 ซึ่งหาได้จากจะกล่าวถึงถัดไป) ขั้นตอนที่ 7 และ 8 นำข้อมูลค่าแนวโน้มและค่าของอายุมาคำนวณตามอัลกอริทึมเพจแรงค์ถ่วงน้ำหนักตามแนวโน้ม ขั้นตอนที่ 9 จัดเรียงผลลัพธ์ที่ได้เปรียบเทียบกับอัลกอริทึมเพจแร็งค์มาตรฐาน

จากค่าที่ได้จากเงื่อนไขดังกล่าว(ค่าแนวโน้มกับค่าอายุ) จะสามารถเพิ่มค่าความสำคัญให้กับเว็บเพจใหม่ที่เกิดขึ้นมาใหม่ให้ได้อยู่ในอันดับต้นๆของผลการจัดลำดับข้อมูลเว็บเพจโดยผู้วิจัยได้เสนอให้ใช้ค่า ตัวแปรอันได้แก่ แนวโน้ม (Trend Factor) และค่าอายุของเว็บเพจ (Aging Factor) ที่วิเคราะห์ได้จากฐานข้อมูลเว็บเพจเนื่องจากเว็บเพจส่วนใหญ่มีอัตราการเสื่อมสภาพที่ค่อนข้างเร็ว ซึ่งสำหรับผลการทดลองกับฐานข้อมูลเว็บเพจซึ่งเป็นฐานข้อมูลที่เป็นแบบอิสระกล่าวคือตัวข้อมูลมีความหลากหลายไม่จำเพาะสำหรับเรื่องใดเรื่องหนึ่งดังนั้นการประเมินค่าเราจะทำการวัดจากความใหม่ล่าสุดจากผลลัพธ์การจัดเรียงอันดับเว็บเพจ การวัดค่าความนิยมของเว็บเพจและผลลัพธ์จากการสืบค้น 10 อันดับว่าให้เว็บเพจที่เกี่ยวข้องกับคำถามหรือไม่ ดูตารางประกอบได้จากตารางที่ 10 ในลำดับถัดไป ซึ่งผลลัพธ์สำหรับเว็บเพจที่ได้เปรียบเทียบกับอัลกอริทึมอื่นอยู่ในระดับปานกลางกล่าวคือให้ค่าผลลัพธ์การสืบค้นเป็นเว็บเพจใหม่ 80 เปอร์เซ็นต์ในขณะที่อัลกอริทึม T-Rank ให้ค่า 90 เปอร์เซ็นต์และเพจแร็งค์ให้ค่าเพียง 30 เปอร์เซ็นต์เท่านั้น ซึ่งในส่วนของรายละเอียดและวิธีการของอัลกอริทึมที่นำเสนอจะกล่าวถึงในลำดับถัดไป

อัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม (Trend Weight PageRank Algorithm - TWPR)

สืบเนื่องจากอัลกอริทึมการคำนวณเพจแร็งค์แบบดั้งเดิม (Page *et al.*,1998) มีจุดอ่อนอันเกิดจากการพิจารณาความสำคัญของเว็บเพจผ่านเส้นการเชื่อมโยงกันเพียงอย่างเดียวซึ่งจากหลักวิธีการดังกล่าว ทำให้เว็บเพจเก่าที่อายุมากมีค่าคะแนนเพจแร็งค์ค่อนข้างสูงในขณะกรณีที่เว็บเพจใหม่จะมีค่าเพจแร็งค์ค่อนข้างต่ำ ซึ่งบางกรณีเว็บเพจใหม่บางเว็บเพจอาจเป็นเว็บเพจที่สำคัญ จากสมมุติฐานที่ว่าข้อมูลที่สดใหม่กว่าย่อมเป็นข้อมูลที่สำคัญสำหรับการสืบค้นข้อมูลกว่า ดังนั้นในวิทยานิพนธ์นี้ จึงนำเสนอหลักวิธีการการถ่ายทอดความสำคัญของเว็บเพจแบบใหม่ ซึ่งปรับปรุงมาจากอัลกอริทึมพื้นฐานเพจแร็งค์เดิม โดยสนใจในเงื่อนไขปรับปรุงเพิ่ม 3 ส่วนด้วยกัน คือ การหาค่าอายุจากเว็บเพจ การหาค่าอายุของเส้นการเชื่อมโยง การหาค่าแนวโน้มทั้งของเว็บเพจและเส้นการเชื่อมโยง ซึ่งหลักการหาค่าแนวโน้มนั้นสามารถทำได้จากการจัดกลุ่มของข้อมูล ตามกรอบระยะเวลาการพิจารณาทีละ 3 เดือน(เป็นจำนวนเดือนที่เหมาะสมที่สุดจากการทดลองซ้ำหลายๆ ครั้ง) โดยรายละเอียดสามารถศึกษาต่อตามลำดับได้ดังนี้

ส่วนที่ 1 พิจารณา *อายุของเว็บเพจ* หมายถึงค่าจำนวนวันเวลาการแก้ไขล่าสุดของเว็บเพจแต่ละเว็บเพจ (ts_x) เมื่อเทียบกับเวลาปัจจุบัน ($TS_{current}$) ซึ่งสำหรับเว็บเพจที่เกิดขึ้นใหม่นี้จะมีค่าอายุน้อยมากหรือมีอายุเกือบเป็นศูนย์และจะกำหนดให้เว็บเพจเหล่านี้มีค่าความสำคัญสูงขึ้นจากปกติ ตรงกับสมมุติฐานที่ผู้ใช้งานมักต้องการข้อมูลข่าวสารที่มีความใหม่และทันต่อเหตุการณ์เสมอ ดังนั้นจากอัลกอริทึมที่นำเสนอจึงนำเกณฑ์อายุไปพิจารณาเพื่อเพิ่มค่าเพจแร็งค์ให้กับเว็บเพจใหม่และในทางตรงกันข้ามก็จะพยามยามลดค่าเพจแร็งค์กับเว็บเพจที่มีอายุมากหรือเป็นเว็บเพจที่เก่าและล้าสมัยลงเนื่องจากข้อมูลขาดความสดนั่นเอง อย่างไรก็ตาม ข้อมูลรายละเอียดของแต่ละเว็บเพจที่บ่งบอกถึงจุดเวลาที่เว็บเพจนั้นถูกสร้างขึ้น หรือถูกเปลี่ยนแปลงครั้งล่าสุดเมื่อใดนั้น เป็นเรื่องที่ทราบได้ยาก เนื่องจากผู้สร้างหรือเปลี่ยนแปลงเว็บเพจโดยมากมักไม่ระบุไว้ ดังนั้นเราจะประยุกต์ใช้เทคนิคพื้นฐานของ Amitay *et al.*, (2004) ในการหาค่าเวลาแก้ไขล่าสุดของเว็บเพจจาก HTTP header field แต่เนื่องจากวิธีการดังกล่าว สามารถใช้ได้กับเว็บเพจที่เป็นอพลวัตเว็บเพจ (static web page) กล่าวคือเป็นเว็บเพจประเภท .htm หรือ .html แต่หากเป็นเว็บเพจพลวัต (dynamic web page) เครื่องแม่ข่ายจะไม่คืนค่าเวลาการแก้ไขล่าสุด (last modified) ให้ เนื่องจากเว็บเพจมีการเปลี่ยนข้อมูลบนหน้าเว็บเพจตามเวลาเข้าชม ณ ขณะนั้นซึ่งได้แก่เว็บเพจประเภท .php .asp หรือ .jsp ดังนั้นเราจะใช้วิธีการหาค่าเวลาการแก้ไขล่าสุด จากส่วนประกอบที่ปรากฏบนหน้าพลวัตเว็บเพจนั้นๆแทน เช่น รูปภาพ และจะถือว่าเวลาที่ได้นี้แทนถึงค่าเวลาการแก้ไขล่าสุดของเว็บเพจนั้น

ส่วนที่ 2 พิจารณา *อายุของเส้นการเชื่อมโยง* เนื่องด้วยค่าเวลาการแก้ไขล่าสุดของเส้นการเชื่อมโยงไม่สามารถหาได้อย่างในส่วนแรกดังนั้นเราจะแบ่งพิจารณาให้อายุกับเส้นการเชื่อมโยงที่ถูกสร้างขึ้นออกเป็น 2 กรณี โดยมีสมมติฐานของการพิจารณาคือ การเปลี่ยนแปลงแก้ไขโครงสร้างของเส้นการเชื่อมโยงนี้เป็นเรื่องเดิม และช่วยทำให้เว็บเพจมีคุณภาพดีขึ้น

<u>กรณีที่ 1</u> เส้นการเชื่อมโยงเกิดจากเว็บเพจต้นทางที่มีค่าเวลาการแก้ไขล่าสุดเก่ากว่าเว็บเพจปลายทางที่มีค่าเวลาการแก้ไขล่าสุดใหม่กว่าซึ่งในกรณีนี้เราจะพิจารณาให้ค่าเวลาการแก้ไขล่าสุดกับเส้นการเชื่อมโยงเท่ากับเว็บเพจปลายทาง เพราะเหตุว่า ข้อมูลที่เว็บเพจปลายทางนั้นมีความสดใหม่กว่าดังนั้นเส้นการเชื่อมโยงจึงควรได้รับค่าความสำคัญมากด้วยนั้นเอง

<u>กรณีที่ 2</u> เส้นการเชื่อมโยงเกิดจากเว็บเพจต้นทางที่มีค่าเวลาการแก้ไขล่าสุดใหม่กว่าเว็บเพจปลายทางที่มีค่าเวลาการแก้ไขล่าสุดเก่ากว่าซึ่งในกรณีนี้เราพิจารณาให้ค่าเวลาการแก้ไขล่าสุดกับเส้นการเชื่อมโยงเท่ากับเว็บเพจเพจต้นทาง เพราะเหตุว่า เว็บเพจต้นทางนั้นมีการปรับปรุงเปลี่ยนแปลงแก้ไขข้อมูลให้มีความสดใหม่ โดยข้อมูลสดใหม่ที่ปรับปรุงนั้นอาจมีเนื้อหาพื้นฐานที่เกี่ยวข้องกันกับเว็บเพจเก่าจึงยังคงเส้นการเชื่อมโยงไว้กับเว็บเพจเก่าดังนั้นข้อมูลนี้จึงควรพิจารณาให้ค่าความสำคัญเช่นกัน

จากอายุของเว็บเพจและอายุของเส้นการเชื่อมโยง เราจะคำนวณค่าองค์ประกอบอายุ (Aging factor) ของเว็บเพจเสียก่อน พิจารณาจากช่วงเวลาการสนใจหรือ TWI โมเดลซึ่งเรานำมาประยุกต์ใช้กับเว็บเพจ โดยจะถือว่าเว็บเพจใดๆ ที่มีจุดเวลา (timestamp) ของการเปลี่ยนแปลงครั้งล่าสุดอยู่ในช่วงที่เราสนใจกำหนดให้มีค่าองค์ประกอบของอายุเป็นไปตามระยะเวลาของจุดเวลาที่มีการเปลี่ยนแปลงครั้งล่าสุดนั้นเทียบกับจุดเวลาปัจจุบัน ($TS_{current}$) แสดงดังในสมการที่ (18)

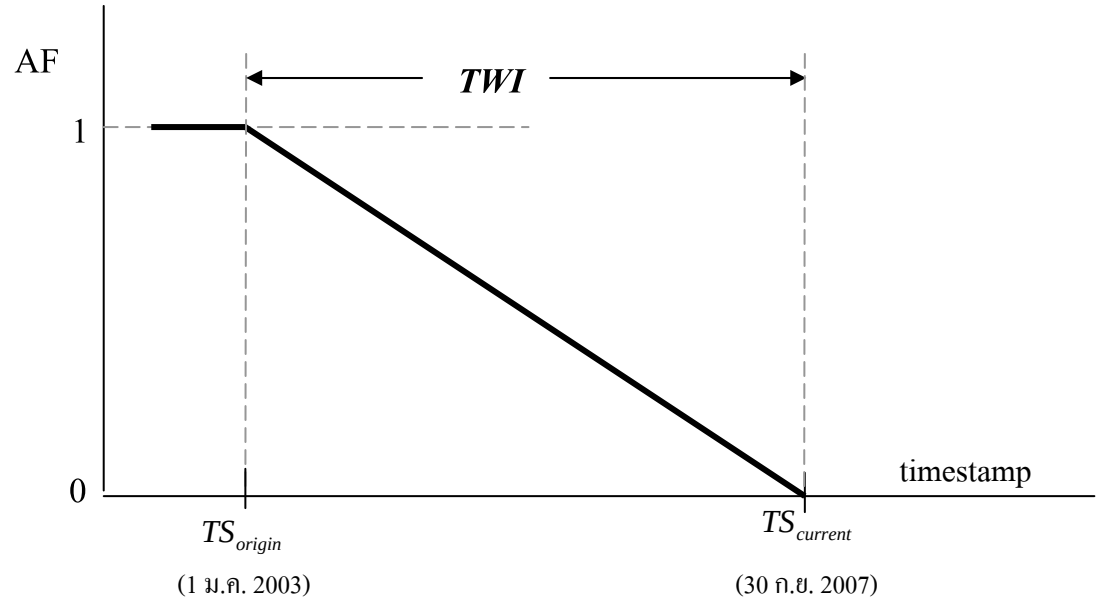
กำหนดให้

- $AF(x)$ แทน ค่าองค์ประกอบอายุของเว็บเพจ x
- ts_x แทน จุดเวลาของการเปลี่ยนแปลงครั้งล่าสุดของเว็บเพจ x
- $TS_{current}$ เป็นค่าจุดเวลาปัจจุบัน ณ ตอนที่เริ่มทำการคำนวณ
- TS_{origin} เป็นค่าจุดเวลาอ้างอิงตั้งต้น

$$AF(x) = \begin{cases} \dfrac{TS_{current} - ts_x}{TS_{current} - TS_{origin}} & TS_{origin} \le ts_x \le TS_{current} \\ 1 & ts_x < TS_{origin} \\ 0 & otherwise \end{cases} \qquad (18)$$

จากสมการที่ 18 เราเสนอให้ใช้ $AF(x)$ ตัวนี้เพื่อบ่งบอกความสดของข้อมูลโดยถ้าข้อมูลสดมากจะมีค่าอายุเป็น 0 และจะมีค่าอายุเพิ่มมากขึ้นตามลำดับและเมื่ออยู่นอกช่วงของการพิจารณาจะมีค่าเป็น 1 สำหรับในการทดลองเรากำหนดจุดเวลาอ้างอิงตั้งต้น (TS_{origin}) กับเว็บเพจที่มีวันเวลาการแก้ไขล่าสุดตั้งแต่วันที่ 1 มกราคม ปี 2003 ถึง ($TS_{current}$) วันที่ 30 กันยายน ปี 2007 รวมเป็น 57 เดือน ซึ่งสาเหตุของช่วงการศึกษาอยู่ในช่วงนี้เนื่องจากข้อมูลในฐานข้อมูลส่วนใหญ่อยู่ในช่วงนี้ซึ่งจากสมการ (18) เราสามารถแทนค่า $TS_{origin} = TS_{last57months}$

และจากสมการที่ 18 ค่าอายุเราสามารถเขียนแสดงความสัมพันธ์ของอายุกับช่วงเวลาที่ผู้ใช้สนใจหรือกราฟ TWI ได้ดังภาพที่ 15 โดยค่าองค์ประกอบอายุของเว็บเพจนี้จะมีค่าสูดสุดเป็น 1 สำหรับเว็บเพจที่มีวันเวลาการแก้ไขล่าสุดก่อนและตรงกับค่าจุดเวลาอ้างอิงตั้งต้น สำหรับเว็บเพจใหม่ที่มีวันเวลาการแก้ไขล่าสุดหลังจุดเวลาอ้างอิงตั้งต้นจะมีค่าองค์ประกอบอายุลดลงตามสมการเส้นตรงและมีค่าองค์ประกอบอายุต่ำสุดเป็น 0 สำหรับเว็บเพจที่มีค่าวันเวลาแก้ไขล่าสุดตรงกับวันที่ปัจจุบัน ณ ตอนที่เริ่มทำการคำนวณ



ภาพที่ 15 แสดงค่าของ AF

ส่วนที่ 3 พิจารณา *ค่าแนวโน้ม (Trend Factor)* ซึ่งค่านี้จะหมายถึงค่าทางสถิติที่ได้จากการนับจำนวนกลุ่มของเว็บเพจ ณ ช่วงเวลาที่กำหนดเทียบกับจำนวนของเว็บเพจทั้งหมดที่ศึกษาและจากการศึกษาถึงพฤติกรรมการเปลี่ยนแปลงข้อมูลของเว็บเพจดังภาพที่ 6 ซึ่งเว็บเพจส่วนใหญ่มีการเปลี่ยนแปลงที่เวลามากกว่า 4 เดือน แต่ช่วงดังกล่าวนี้ยาวเกินไปซึ่งอาจมีหลายเหตุการณ์เกิดขึ้นเมื่อนำมาหาค่าแนวโน้มอาจไม่สมบูรณ์ดังนั้นในการทดลองเราได้กำหนดแบ่งช่วงเวลาให้น้อยกว่าเวลาดังกล่าวโดยมีค่าเป็นครั้งละ 3 เดือน ในเวลา 1 ปีเราสามารถแบ่งกลุ่มเวลาออกได้เป็น 4 ควอเตอร์ (quarter) ซึ่งเราจะทำการหาสัดส่วนจำนวนของเว็บเพจที่มีค่าเวลาแก้ไขล่าสุดตามแต่ละควอเตอร์ โดยค่าที่ได้นี้จะทำให้เราสามารถทราบได้ว่า ณ ช่วงเวลาใดของเว็บเพจที่ค่าแนวโน้มมีค่าสูงแสดงว่าช่วงนั้นมีเหตุการณ์บางอย่างเกิดขึ้นทำให้มีจำนวนเว็บเพจเปลี่ยนแปลงช่วงนี้เป็นจำนวนมากและจากหลักการที่ข้อมูลใหม่นั้นถือเป็นข้อมูลที่มีน่าสนใจดังนั้นจึงแบ่งช่วงการหาค่าแนวโน้มออกเป็น 2 ช่วง โดยช่วงแรกพิจารณาถึงเว็บเพจที่มีจุดเวลาการแก้ไขล่าสุดอยู่ในช่วงเวลาที่ทำการศึกษาคือ 57 เดือนก่อนปัจจุบันซึ่งในช่วงนี้เว็บเพจจะต้องมีการเสื่อมสภาพเป็นไปตามอัตราส่วนค่าแนวโน้มยกกำลังด้วยค่าอายุโดยหากเป็นเว็บเพจใหม่ค่าอายุใกล้เคียงศูนย์จะทำให้ได้ค่าถ่วงน้ำหนักสูงสุดในขณะเดียวกันเว็บเพจเก่าก็จะได้ค่าถ่วงน้ำหนักน้อยลงตามค่าแนวโน้มและช่วงที่สองซึ่งเว็บเพจมีจุดเวลาการแก้ไขล่าสุดที่อยู่นอกช่วงกรอบการพิจารณาจะกำหนดให้มีค่าถ่วงน้ำหนักต่ำสุดเป็น $e\,(10^{-7})$ สามารถเขียนความสัมพันธ์ได้ดังสมการที่ (19)

สมการกำหนดให้

- $TREND(x)$ แทน ค่าแนวโน้มของเว็บเพจ x
- $PAGE_{q(i)}$ แทน จำนวนของเว็บเพจทั้งหมดที่อยู่ในควอเตอร์ $q(i)$
- N แทน จำนวนควอเตอร์ทั้งหมดที่พิจารณาถึงปัจจุบัน
- $TS_{last51months}$ แทน เวลาที่เริ่มทำการศึกษา

$$TREND(x) = \begin{cases} \dfrac{PAGE_{q(i)}}{\sum_{i=1}^{N} PAGE_{q(i)}} & TS_{last57month} \le ts_x \le TS_{current} \\ e & otherwise \end{cases} \qquad (19)$$

จากสมการข้างต้น พิจารณากรณีที่ค่าแนวโน้มจะต้องมีการปรับลดลงไปตามเวลาซึ่งเราจะประยุกต์ใช้ความสัมพันธ์จากสมการที่ (19) เขียนสมการการถ่วงน้ำหนักได้ดังสมการที่ (20)

กำหนดให้

- $AF(x)$ แทน ค่าองค์ประกอบอายุของเว็บเพจ x

- $W(x)$ แทน ค่าถ่วงน้ำหนักตามแนวโน้มของเว็บเพจ x

$$W(x) = TREND(x)^{AF(x)} \qquad (20)$$

จากสมการแบบทั่วไปของเพจแร็งค์สมการที่ (3) เราทำการปรับอัลกอริทึมใหม่โดยพิจารณาการให้ค่าความสำคัญ 2 ส่วนคือ

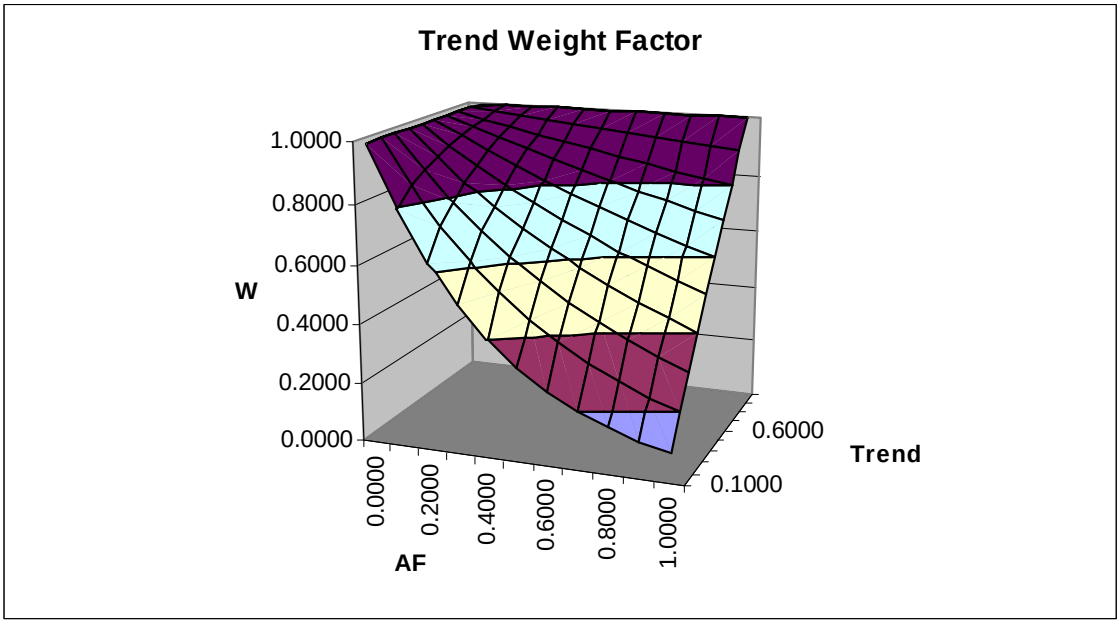
1. ส่วนที่มีการถ่ายทอดค่าความสำคัญให้แก่เว็บเพจอื่น คือ $t(x,y)$ ซึ่งแทน สัดส่วนของค่าความสำคัญที่ x จะส่งมอบให้กับ y ซึ่งแทนด้วยค่าถ่วงน้ำหนักของเว็บเพจ y เขียนแทนด้วย $w(y)$ และค่าถ่วงน้ำหนักของเส้นการเชื่อมโยงจาก $x \to y$ เขียนแทนด้วย $w(x,y)$ และค่าคงที่ A มีค่าเท่ากับ 0.5 เนื่องจากเราพิจารณาให้ค่าความสำคัญระหว่างเว็บเพจและเส้นเชื่อมโยงเท่ากัน คำนวณได้จากสมการที่ (21)

$$t(x,y) = \left(A\frac{W(y)}{\sum_{\forall z: x \to z} W(z)} + (1-A)\frac{W(x,y)}{\sum_{\forall z: x \to z} W(x,z)} \right) \qquad (21)$$

2. ส่วนที่เป็น random jump หรือ $s(y)$ คือค่าโอกาสที่เว็บเพจอื่นๆจะกระโดดมาหาเว็บเพจ y ซึ่งคำนวณได้จากสมการที่ (22)

$$s(y) = \frac{W(y)}{\sum_{\forall z} W(z)} \qquad (22)$$

จากสมการที่ (20) สามารถแสดงกราฟความสัมพันธ์ระหว่างค่า W, AF และ TREND โดยแกน AF แทนอายุ แกน Trend แทนค่าแนวโน้ม ส่วนแกน W แสดงค่าถ่วงน้ำหนักตามแนวโน้ม ซึ่งจะเห็นว่าหากค่าแนวโน้มมีค่าสูงสุด 1 ค่าอายุที่เปลี่ยนไปจะไม่ส่งผลกระทบต่อค่าถ่วงน้ำหนักตามแนวโน้ม ในขณะเดียวกันหากอายุมีค่าต่ำสุด 0 (หมายถึงสดที่สุด) ค่าแนวโน้มที่เปลี่ยนไปจะไม่ส่งผลกระทบต่อค่าถ่วงน้ำหนักตามแนวโน้มโดยจะยังคงมีค่าสูงสุดเป็น 1 และสุดท้ายค่าแนวโน้มที่มีค่าสูงจะมีค่าถ่วงน้ำหนักตามแนวโน้มลดลงช้ากว่าค่าแนวโน้มที่มีค่าต่ำกว่าเมื่อเว็บเพจมีอายุมากขึ้นเรื่อยๆ สามารถดูรูปประกอบได้ดังรูปที่ 16



ภาพที่ 16 แสดงค่าถ่วงน้ำหนักตามแนวโน้ม

ผลและวิจารณ์

ผล

แนวทางในการวัดผล

เรามีวิธีการวัดผล 2 ประการ

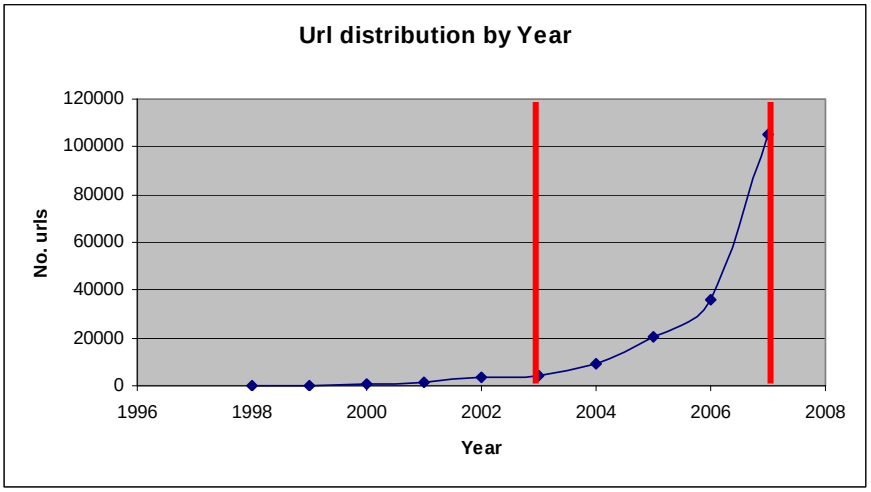
1) *เว็บเพจที่เป็นที่นิยม(Page Popularity)* ซึ่งเป็นวิธีการที่นำเสนอโดย Yu *et al.* (2005) โดยใช้หลักการดั้งเดิมของเพจแร็งค์ที่ใช้วิธีการนับจำนวนจุดเชื่อมโยงที่ชี้เข้าหาเว็บเพจนั้นๆ (in-degree) เนื่องจากจำนวนของจุดเชื่อมโยงที่ชี้เข้าหาเว็บเพจมากแสดงได้ถึงการเป็นเว็บเพจที่น่าสนใจ

2) *วัดความใหม่ล่าสุด(Page up-to datedness)* เป็นวิธีที่วิทยานิพนธ์เล่มนี้นำเสนอใหม่ เนื่องจากเราสนใจถึงวิธีในการเพิ่มค่าความสำคัญให้แก่เว็บเพจที่มีค่าวันเวลาการแก้ไขล่าสุดใกล้เคียงปัจจุบันมากที่สุด เพราะในช่วงเวลาดังกล่าวข้อมูลมีค่าความสำคัญมากดังที่กล่าวมาแล้วดังข้างต้น

ซึ่งจากวิธีการวัดผลทั้ง 2 จะใช้ข้อมูลผลลัพธ์จากการจัดเรียงลำดับเว็บเพจที่ได้จากการคำนวณเพจแร็งค์ที่มีค่าคะแนนสูงสุด 20 และ 50 อันดับแรก ตามลำดับ เพื่อแสดงให้เห็นถึงประสิทธิภาพในการคัดแยกเว็บเพจที่สำคัญจากทั้งระบบและเป็นวิธีที่นิยมใช้ในงานวิจัย Yu *et al.* (2005) สำหรับทำการวัดประสิทธิภาพรวมของระบบ ซึ่งวิธีการดังกล่าวสามารถขยายผลโดยทำการเก็บรวบรวมข้อมูลเว็บเพจให้เพิ่มมากขึ้นเป็นเว็บกราฟขนาดใหญ่และนำไปเชื่อมต่อกับระบบสืบค้นข้อมูลได้

ขั้นตอนดำเนินการทดลอง

จากฐานข้อมูลเว็บเพจที่กล่าวไว้ในวิธีการทดลองเมื่อทำการหา จุดวันเวลาการแก้ไขล่าสุด (lastmodified) ของแต่ละเว็บเพจ จะกำหนดจุดเวลาเริ่มต้นสนใจ TS_{origin} โดยดูจากอัตราการกระจายตัวของเว็บเพจโดยใช้วิธีการนับจำนวนประชากรของเว็บเพจที่มีค่าวันเวลาแก้ไขล่าสุดในแต่ละกลุ่มช่วงเวลา โดยกำหนดช่วงละ 3 เดือน แสดงดังภาพที่ 17 แกนตั้งแสดงถึงจำนวนยูอาร์แอลหรือเว็บเพจส่วนแกนแนวนอนแสดงข้อมูลรายปี



ภาพที่ 17 แสดงการกระจายตัวของค่าเวลาแก้ไขล่าสุดของเว็บเพจต่างๆตามปี

จากข้อมูลดังภาพที่ 17 จำนวนเว็บเพจที่เราสนใจจะกำหนดช่วงเวลาเริ่มต้นเป็นวันที่ 1 มกราคม ค.ศ. 2003 ถึงปัจจุบันซึ่งคือวันที่ 31 กันยายน ค.ศ. 2007 โดยเราทำการหาค่าแนวโน้มได้จากสมการที่ (19) ซึ่งค่าแนวโน้มทั้งจากข้อมูลเว็บเพจและเส้นการเชื่อมโยงแบ่งออกเป็นช่วงละ 3 เดือนซึ่งสามารถแสดงผลได้ดังตารางที่ 4 และ 5 ตามลำดับ

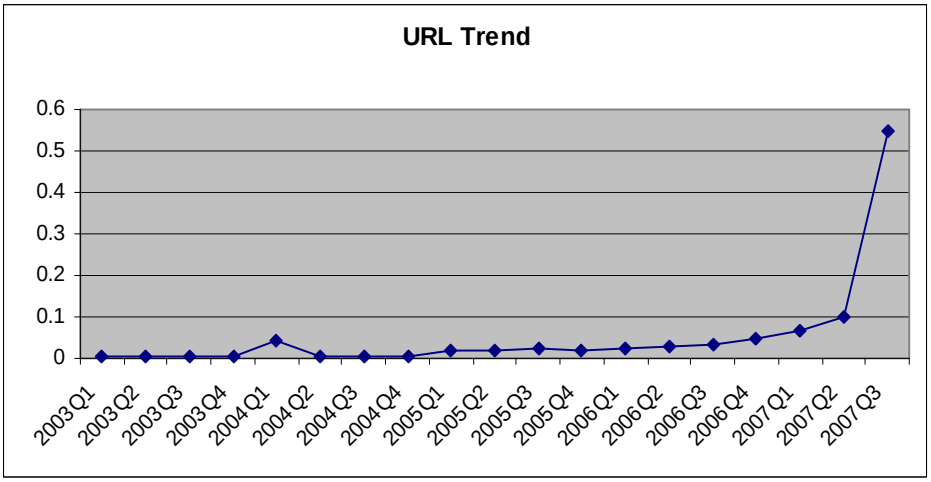
ตารางที่ 4 แสดงค่าแนวโน้มของเว็บเพจในแต่ละควอเตอร์

ช่วงเวลา	ผลรวมค่าแนวโน้ม(เว็บเพจ)	ช่วงเวลา	ผลรวมค่าแนวโน้ม (เว็บเพจ)	ช่วงเวลา	ผลรวมค่าแนวโน้ม(เว็บเพจ)
2003Q1	0.003029	2005Q1	0.016731	2007Q1	0.067402
2003Q2	0.003701	2005Q2	0.020331	2007Q2	0.100795
2003Q3	0.004911	2005Q3	0.021486	2007Q3	0.546509
2003Q4	0.003474	2005Q4	0.018028		
2004Q1	0.045135	2006Q1	0.021455		
2004Q2	0.006168	2006Q2	0.027139		
2004Q3	0.006347	2006Q3	0.031816		
2004Q4	0.007081	2006Q4	0.048461		

ตารางที่ 5 แสดงค่าแนวโน้มของเส้นการเชื่อมโยงในแต่ละควอเตอร์

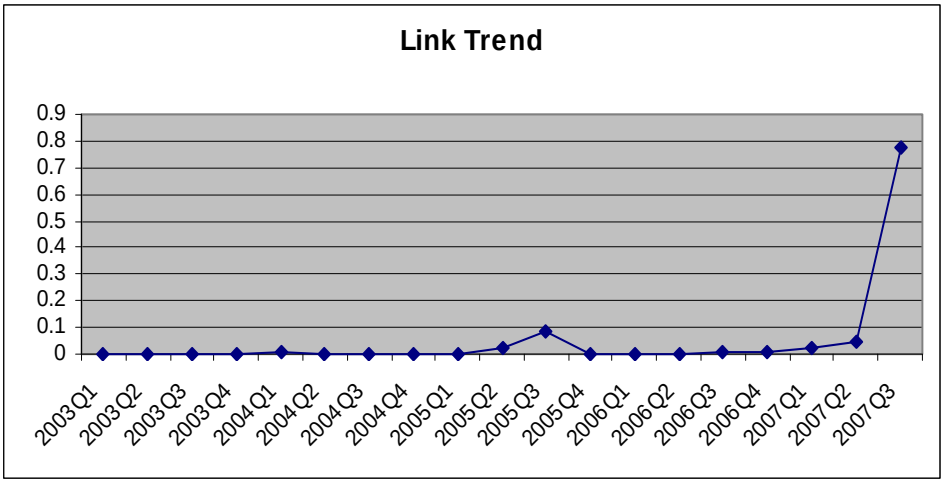
ช่วงเวลา	ผลรวมค่าแนวโน้ม (เส้นการเชื่อมโยง)	ช่วงเวลา	ผลรวมค่าแนวโน้ม(เส้นการเชื่อมโยง)	ช่วงเวลา	ผลรวมค่าแนวโน้ม(เส้นการเชื่อมโยง)
2003Q1	0.000221	2005Q1	0.002663	2007Q1	0.019532
2003Q2	0.001172	2005Q2	0.023856	2007Q2	0.048438
2003Q3	0.000483	2005Q3	0.087829	2007Q3	0.774507
2003Q4	0.000245	2005Q4	0.003035		
2004Q1	0.011179	2006Q1	0.003424		
2004Q2	0.00085	2006Q2	0.003077		
2004Q3	0.001324	2006Q3	0.005611		
2004Q4	0.0022	2006Q4	0.010353		

จากค่าแนวโน้มที่ได้จากตารางที่ 4 และ 5 พบว่าช่วงเวลาที่ค่าแนวโน้มขึ้นถึงจุดสูงสุดทั้งในส่วนของเว็บเพจและเส้นการเชื่อมโยงนั้นเหมือนกันคือ 2007Q3 ซึ่งค่าแนวโน้มที่มากนี้จะทำให้เมื่อคำนวณค่าความสำคัญของเว็บเพจโดยอัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มจะให้ผลลัพธ์เป็นเว็บเพจใหม่ที่ค่อนข้างชัดเจนดูจากภาพประกอบได้ดังภาพที่ 18 และ 19



ภาพที่ 18 แสดงค่าแนวโน้มในแต่ละควอเตอร์ของเว็บเพจ x

จากฐานข้อมูลซึ่งเว็บเพจชุดตั้งต้นมีฐานข้อมูลเกี่ยวกับเรื่อง Harry Potter, NCSEC, Olympic และอื่นๆ ซึ่งสอดคล้องกับภาพที่ 18 ซึ่งพบว่าระหว่างเวลาเริ่มต้นปี ค.ศ. 2004 มีค่าแนวโน้มที่สูงขึ้นก่อนที่จะเกิดเหตุการณ์เรื่อง Olympic ซึ่งกำลังจะจัดขึ้น (Athen 2004) เนื่องจากเว็บเพจโดยทั่วไปทราบเหตุการณ์ล่วงหน้าแล้ว ซึ่งหลังจากนั้น ภาพที่ 19 จะพบว่าค่าแนวโน้มจากเส้นการเชื่อมโยงนั้นมีค่าที่สูงขึ้นช้ากว่าเหตุการณ์ที่เกิดขึ้นจริงคือ เกิดเหตุการณ์ ซึนามิ(ธันวาคม ค.ศ. 2004) และการก่อการร้ายตึก World Trade Center (กันยายน ค.ศ. 2004) ณ ประเทศสหรัฐอเมริกา โดยค่าแนวโน้มที่ได้จะอยู่ช่วงประมาณ มีนาคม ค.ศ. 2005 ถึง ธันวาคม ค.ศ. 2005 เนื่องจากเว็บเพจโดยทั่วไปไม่ทราบเหตุการณ์ดังกล่าวล่วงหน้า แต่หลังจากเกิดเหตุการณ์ขึ้นแล้วก็ทำให้เกิดการเปลี่ยนแปลงแก้ไขจุดเชื่อมโยงจำนวนมาก และพิจารณาช่วงระหว่างปี ค.ศ. 2006 ถึง ค.ศ. 2007 กราฟในภาพที่ 18 จะสูงขึ้นเรื่อยเนื่องจากเว็บเพจส่วนใหญ่จะมีค่าเวลาล่าสุดอยู่ในช่วงนี้ ประกอบกับภาพที่ 19 ที่จะพุ่งสูงขึ้นที่ 2007Q3 เนื่องจากกลุ่มของเว็บเพจที่เป็นพอร์ทัล (portal) มีการปรับปรุงแก้ไขข้อมูลเส้นการเชื่อมโยง



ภาพที่ 19 แสดงค่าแนวโน้มในแต่ละควอเตอร์ของเส้นการเชื่อมโยง

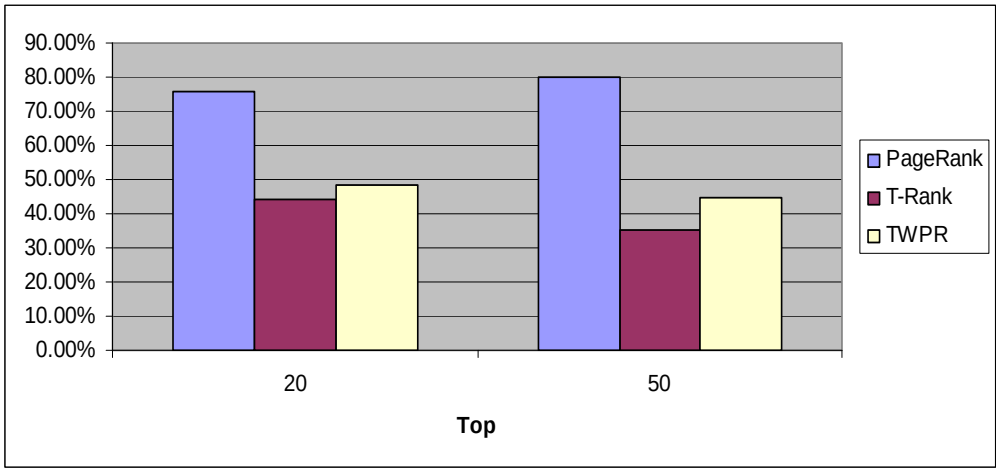
ดังที่กล่าวมาแล้วข้างต้นว่าเราจะใช้วิธีการวัดผลเป็น 2 ประการ

ประการที่ 1 วัดค่าความนิยม ใช้วิธีการวัดจากเส้นการเชื่อมโยงที่ชี้เข้าหาเว็บเพจนั้นซึ่งเป็นพื้นฐานจากอัลกอริทึมเพจแร็งค์แบบดั้งเดิมซึ่งเว็บเพจใดเป็นเว็บเพจที่น่าสนใจเว็บเพจนั้นย่อมมีจำนวนเส้นการเชื่อมโยงเข้าหาเว็บเพจเป็นจำนวนมาก โดยวิธีการวัดจะใช้ชุดข้อมูลการเรียงลำดับเว็บเพจสูงสุด 20 และ 50 อันดับแรกตามลำดับ

ตารางที่ 6 แสดงเปรียบเทียบจำนวนจุดเชื่อมโยงเว็บเพจที่ได้รับสูงสุด 20 อันดับและ 50 อันดับแรก โดยใช้อัลกอริทึมเพจแร็งค์, T-Rank, TWPR ตามลำดับ

1	2	3	4	5	6	7	8
Top Urls	PageRank		T-Rank		TWPR		Best Citation Count
20	204891	75.87%	118990	44.06%	130151	48.20%	270049
50	344646	79.96%	152599	35.40%	193490	44.89%	431030

จากตารางที่ 6 ข้อมูลในหลักที่ 2, 4, 6 และ 8 แสดงถึงจำนวนจุดเชื่อมโยงที่ชี้เข้าหาเว็บเพจ Top 20 และ 50 โดยจำนวนจุดเชื่อมโยงที่ดีที่สุดในเว็บเพจหรือมากที่สุด (Best Citation Count) 20 อันดับ และ 50 อันดับแรกแสดงในหลักที่ 8 ซึ่งมีค่าเท่ากับ 270049 และ 431030 ตามลำดับ ส่วนข้อมูลในหลักที่ 3, 5 และ 7 แสดงถึงเปอร์เซ็นต์ของจำนวนเส้นการเชื่อมโยงเว็บเพจในแต่ละอัลกอริทึมเมื่อเทียบกับจำนวนเส้นการเชื่อมโยงที่ดีที่สุด ซึ่งจากตารางจะเห็นว่าสำหรับชุดข้อมูลเรียงลำดับสูงสุด 20 อันดับแรก อัลกอริทึมแบบเพจแร็งค์สามารถให้เว็บเพจที่มีจำนวนจุดเชื่อมโยงสูงสุดถึง 75.87 เปอร์เซ็นต์ และอัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มให้ค่ามากอันดับสองโดยมีค่าเป็น 48.2 เปอร์เซ็นต์ในขณะที่อัลกอริทึม T-Rank จะให้ค่าต่ำสุดคือ 44.06 เปอร์เซ็นต์ โดยหากพิจารณาข้อมูลในแถวที่ 2 ข้อมูลสูงสุด 50 อันดับแรก เพจแร็งค์ยังคงให้ค่าสูงสุดคือ 79.96 เปอร์เซ็นต์ ในขณะที่ อัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มให้ค่ารองลงมาคือ 44.89 เปอร์เซ็นต์และอัลกอริทึม T-Rank ให้ค่าน้อยสุดคือ 35.4 เปอร์เซ็นต์ตามลำดับ จากตารางที่ 6 สามารถแสดงกราฟความสัมพันธ์ได้ดังภาพที่ 20 โดยแกนตั้งแสดงจำนวนเปอร์เซ็นต์ของจุดเชื่อมโยงเมื่อเทียบกับจำนวนจุดเชื่อมโยงที่ดีที่สุดในเว็บเพจ แกนนอนแสดงชุดจำนวนเว็บเพจเรียงลำดับสูงสุด 20 และ 50 อันดับแรก จากการวัดค่าความนิยมซึ่งแสดงถึงความสำคัญของเว็บเพจตามวิธีของ Yu *et al.*(2005) ในลำดับถัดไปจะได้อธิบายถึงวิธีการวัดความใหม่ล่าสุดของเว็บเพจเนื่องจากผู้ใช้ส่วนใหญ่สนใจข้อมูลที่มีความสด



ภาพที่ 20 แสดงค่าความนิยมของเว็บเพจของแต่ละอัลกอริทึม

ประการที่ 2 วัดความใหม่ล่าสุด โดยวัดจากจำนวนของเว็บเพจที่อยู่ในช่วงเวลาล่าสุด (2007Q3) เทียบกับที่เวลาต่างๆ (200XQY) จากผลลัพธ์การจัดเรียงอันดับเว็บเพจสูงสุด 20 และ 50 อันดับแรกตามลำดับ

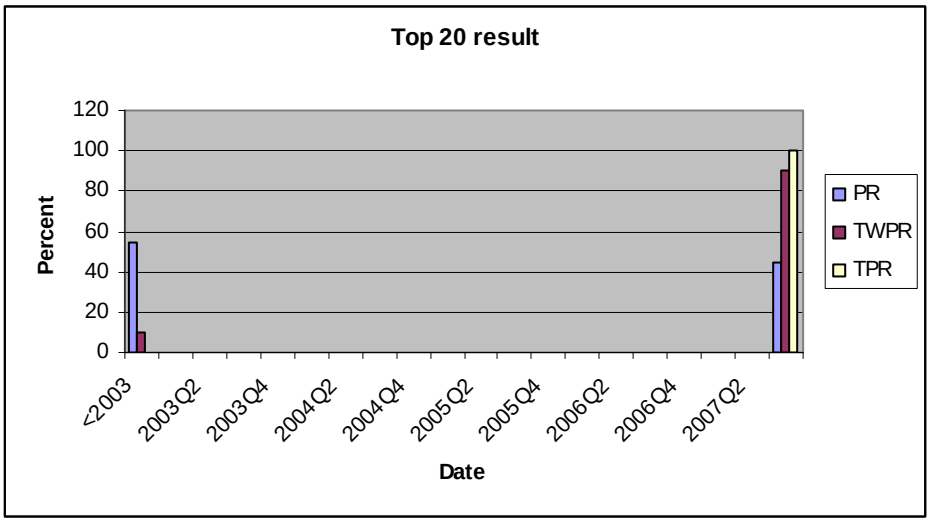
ผลลัพธ์การจัดเรียงลำดับเว็บกราฟสูงสุด 20 อันดับแรก

จากผลลัพธ์การเรียงลำดับเว็บกราฟสูงสุด 20 อันดับแรกพบว่าถ้าใช้อัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มจะได้เว็บเพจที่เป็นผลลัพธ์อยู่ในกลุ่มช่วงระหว่างปี 2007Q3 มากถึง 18 เว็บเพจ ในขณะที่เพจแร็งค์จะให้ค่าในช่วงดังกล่าว 9 เว็บเพจ และผลลัพธ์จาก อัลกอริทึม T-Rank ให้ 20 เว็บเพจดังแสดงในตารางที่ 7

ตารางที่ 7 แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 20 อันดับแรก

Top20	<	2003				2004				2005				2006				2007		
	2003	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3
PR	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
TWPR	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18
T-Rank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20

และจากตารางที่ 7 เขียนแสดงค่าความแม่นยำจากการจัดเรียงผลลัพธ์เปรียบเทียบเป็นเปอร์เซ็นต์ ได้ดังกราฟในภาพที่ 21 ซึ่งจะเห็นว่าผลลัพธ์จากการคำนวณด้วยอัลกอริทึม T-Rank จะมีเว็บเพจที่มีวันเวลาแก้ไขล่าสุดอยู่ในช่วง 2007Q3 สูงสุดถึง 100 เปอร์เซ็นต์ ในขณะที่อัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มนั้นจะมีเว็บเพจที่อยู่ในช่วงเวลาดังกล่าว 90 เปอร์เซ็นต์ และเพจแร็งค์ได้ 45 เปอร์เซ็นต์สามารถดูภาพประกอบได้ดังภาพที่ 21



ภาพที่ 21 แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 20 อันดับแรก

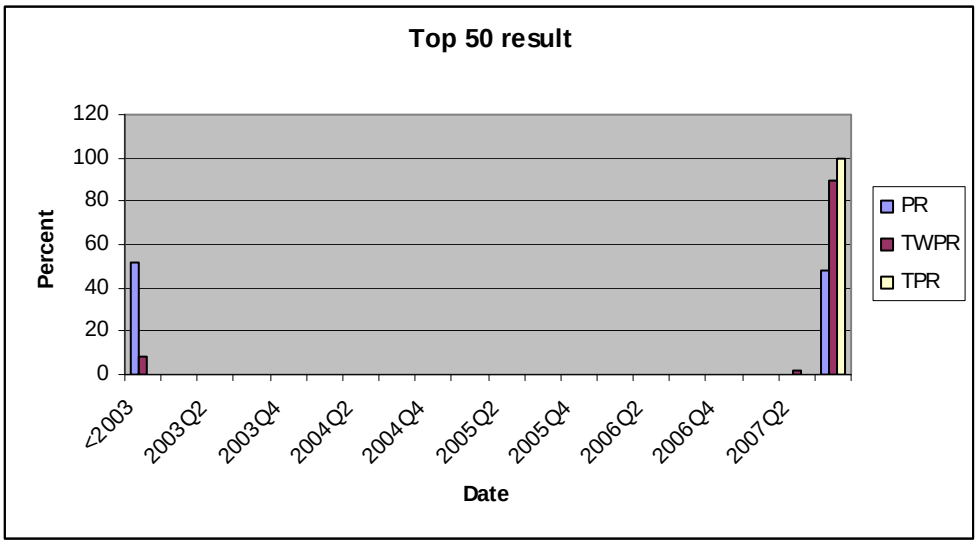
ผลลัพธ์การจัดเรียงลำดับเว็บกราฟสูงสุด 50 อันดับแรก

ในผลลัพธ์การจัดเรียงลำดับเว็บกราฟสูงสุด 50 อันดับแรกพบว่าถ้าใช้อัลกอริทึม T-Rank ให้จะให้เว็บเพจที่อยู่ในช่วงเวลา 2007Q3 มากสุดถึง 50 เว็บเพจ ในขณะอัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มจะได้เป็นเว็บเพจที่มีวันเวลาแก้ไขล่าสุดอยู่ในช่วงเวลาดังกล่าวรองลงมาคือ 45 เว็บเพจ และอัลกอริทึมเพจแร็งค์จะให้ค่าน้อยที่สุดคือ 24 เว็บเพจ สามารถดูตารางประกอบได้ดังตารางที่ 8

ตารางที่ 8 แสดงเปรียบเทียบผลลัพธ์การจัดอันดับเว็บเพจ 50 อันดับแรก

Top50	< 2003	2003				2004				2005				2006				2007		
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3
PR	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	24
TWPR	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	45
T-Rank	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50

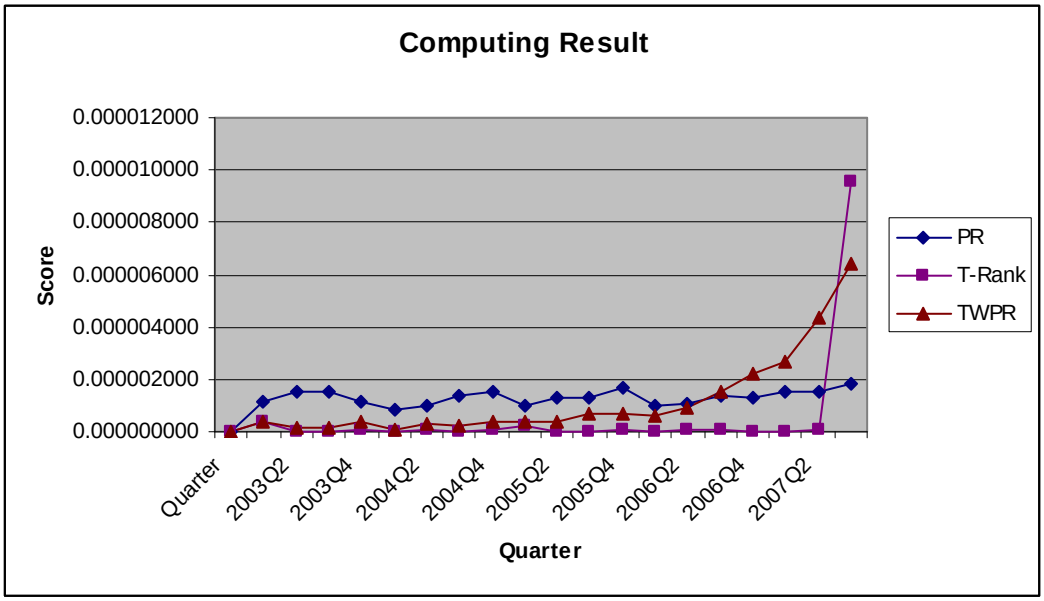
และจากตารางที่ 8 เขียนแสดงค่าความแม่นยำจากการจัดเรียงผลลัพธ์เปรียบเทียบเป็นเปอร์เซ็นต์ ได้ดังกราฟในภาพที่ 22 ซึ่งจะเห็นว่าผลลัพธ์จากการคำนวณด้วยอัลกอริทึม T-Rank จะมีเว็บเพจที่มีวันเวลาแก้ไขล่าสุดอยู่ในช่วง 2007Q3 สูงสุดถึง 100 เปอร์เซ็นต์ ในขณะที่อัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มนั้นจะมีเว็บเพจที่อยู่ในช่วงเวลาดังกล่าว 90 เปอร์เซ็นต์และเพจแร็งค์แบบดั้งเดิมจะให้ค่าน้อยที่สุดคือ 48 เปอร์เซ็นต์



ภาพที่ 22 แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับเว็บเพจ 50 อันดับแรก

จากการเปรียบเทียบการจัดเรียงลำดับทั้ง 20 และ 50 อันดับแรกซึ่งเป็นอันดับที่ผู้ค้นหาข้อมูลส่วนใหญ่สนใจรวมถึง 30 และ 40 อันดับแรกด้วยแต่เพื่อให้เห็นความแตกต่าง ดังนั้นจะแสดงการเปรียบเทียบดังที่กล่าวในข้างต้นซึ่งแสดงให้เห็นว่าอัลกอริทึม T-Rank สามารถเพิ่มค่าความสำคัญให้แก่เว็บเพจใหม่ได้ดีกว่าเนื่องจากค่าฟังก์ชั่นการถ่วงน้ำหนักเว็บเพจที่ อยู่ในช่วงที่สนใจจะถูกถ่วงให้มีค่าสูงสุดเป็น 1 ในขณะที่อัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มนั้นจะให้ค่าการถ่วงน้ำหนักสำหรับทุกๆเว็บเพจไม่เท่ากัน โดยค่าที่ได้จะขึ้นอยู่กับค่าแนวโน้มยกกำลังด้วยค่าอายุของแต่ละเว็บเพจ ซึ่งจากผลลัพธ์ที่ได้ในช่วง 2007Q3 ที่ได้นี้หากทำการตรวจสอบค่าความนิยมของเว็บเพจดังประการที่ 1 ซึ่งได้แสดงมาแล้วในภาพที่ 20 จะพบว่าผลลัพธ์ที่ได้จากอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มมีคุณภาพมากกว่า

พิจารณาความสัมพันธ์อีกรูปแบบหนึ่งระหว่างเวลาและค่าคะแนนเพจแร็งค์ของเว็บเพจที่คำนวณได้จากอัลกอริทึมต่างๆ โดยจากกราฟดังภาพที่ 23 เราจะใช้มาตราค่าเฉลี่ยของคะแนนเพจแร็งค์ที่กลุ่มของเว็บเพจที่อยู่ในช่วงเวลาต่างๆ (ค่าคะแนนรวมเพจแร็งค์ของทุกเว็บเพจในแต่ละควอเตอร์หารด้วยจำนวนเว็บเพจทั้งหมด) ซึ่งในที่นี้จะเห็นว่าอัลกอริทึม PageRank ให้ค่าคะแนนเว็บเพจสูงสุดอยู่ในช่วงเวลาเก่า (ปี 2003 และ ปี 2006) ในขณะที่ T-Rank ให้เว็บเพจในช่วงล่าสุด (ปี 2007Q3) สูงสุดในขณะที่ TWPR ให้ค่าผลลัพธ์เป็นเว็บเพจในช่วงล่าสุดสูงสุดและมีค่าลดลงตามลำดับอายุของเว็บเพจ



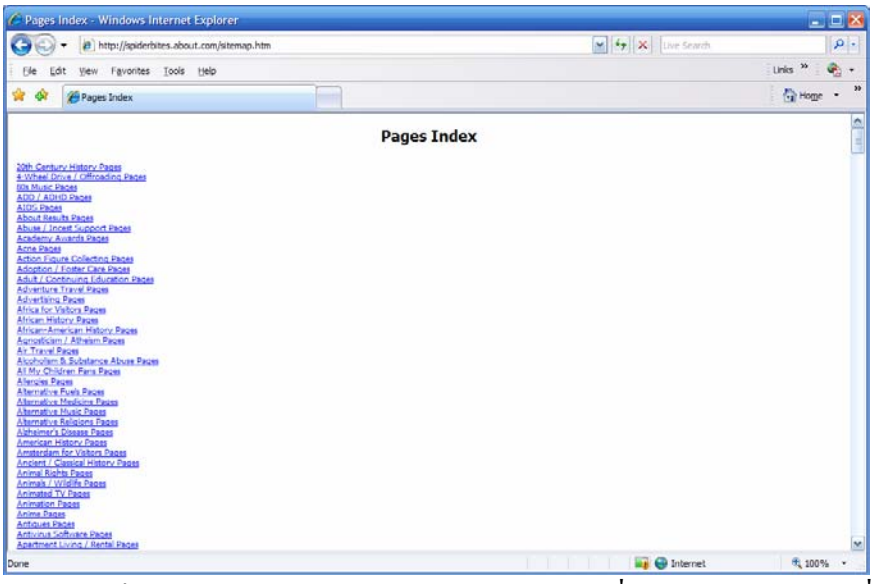
ภาพที่ 23 แสดงค่าคะแนนของเพจแร็งค์ของเว็บเพจตามเวลาต่างๆ

และจากการสุ่มตรวจผลลัพธ์จากการจัดเรียงอันดับโดยดูเนื้อหาที่ปรากฏจากยูอาร์แอลระหว่างอัลกอริทึมเพจแร็งค์กับอัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มสามารถแสดงได้ดังตารางที่ 9

ตารางที่ 9 แสดงเปรียบเทียบผลลัพธ์การจัดเรียงอันดับ 20 อันดับแรกระหว่าง PR และ TWPR

Rank	***PR***	***TWPR***
1	http://www.nytco.com/	http://www.christiananswers.net/dictionary/home.html
2	http://www.about.com/	http://www.howstuffworks.com/jobs.htm
3	http://www.heavygames.com/	http://www.howstuffworks.com/faq.htm
4	**http://spiderbites.about.com/sitemap.htm**	http://www.howstuffworks.com/hsw-contact.htm
5	http://www.about.com/gi/pages/ethics.htm	http://www.howstuffworks.com/terms-and-conditions.htm
6	**http://www.about.com/gi/pages/patent.htm**	http://mobiltravelguide.howstuffworks.com/
7	http://caloriecount.about.com/	http://consumerguideauto.howstuffworks.com/
8	http://www.google.com/	**http://auto.howstuffworks.com/**
9	http://www.google.co.th/	http://communication.howstuffworks.com/
10	http://www.usa.gov/	http://science.howstuffworks.com/
11	http://www.christiananswers.net/dictionary/home.html	http://computer.howstuffworks.com/
12	http://www.howstuffworks.com/jobs.htm	http://electronics.howstuffworks.com/
13	http://www.howstuffworks.com/faq.htm	http://entertainment.howstuffworks.com/
14	http://www.howstuffworks.com/hsw-contact.htm	http://recipes.howstuffworks.com/
15	http://www.howstuffworks.com/terms-and-conditions.htm	http://health.howstuffworks.com/
16	http://www.job.onru.ru/	http://people.howstuffworks.com/
17	http://www.anekdoty.onru.ru/	http://money.howstuffworks.com/
18	http://www.hon.ch/HONcode/Conduct.html	http://travel.howstuffworks.com/
19	http://www.google.co.th/intl/th/about.html	http://www.google.com/
20	http://www.google.co.th/en	http://www.google.co.th/

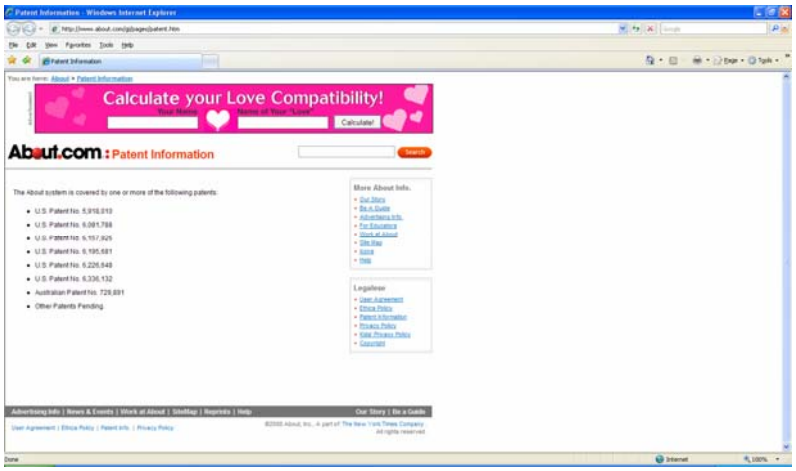
หากทำการตรวจสอบยูอาร์แอลโดยพิจารณาในแง่เนื้อหาของข้อมูลจะเห็นว่า ผลลัพธ์การจัดลำดับจากการคำนวณเพจแร็งค์แบบดั้งเดิมในลำดับที่ 4 คือ http://spiderbites.about.com/sitemap.htm ซึ่งสามารถแสดงหน้าเว็บเพจได้ดังภาพที่ 24 เป็นเว็บเพจที่เป็นพอร์ทัลคือได้รับเส้นการเชื่อมโยงสูง ในขณะที่เมื่อเทียบกับการคำนวณแบบถ่วงน้ำหนักตามแนวโน้มในลำดับที่ 8 ซึ่งแสดงได้ดังภาพที่ 25 ซึ่งแสดงถึงเนื้อหาที่ค่อนข้างใหม่กว่าและเนื่องจากชุดข้อมูลเว็บเพจ howstuffwork นั้นมีค่าแนวโน้มที่อยู่ในช่วง 2007Q3 ซึ่งมีอายุต่ำดังนั้นค่าอันดับจากกลุ่มของเว็บเพจ howstuffwork จึงปรากฏอยู่เป็นจำนวนมาก



ภาพที่ 24 แสดงเว็บเพจ http://spiderbites.about.com/sitemap.htm ซึ่งแสดงเพียงเส้นการเชื่อมโยง

ภาพที่ 25 แสดงเว็บเพจ http://auto.howstuffworks.com/ ซึ่งแสดงเนื้อหาที่ค่อนข้างละเอียด

พิจารณาเว็บเพจจากตารางที่ 9 จากผลลัพธ์การจัดเรียงเว็บเพจแบบ PR อันดับที่ 6 ดังรูปที่ 26 ซึ่งแสดงให้เห็นว่าให้ข้อมูลที่น่าสนใจน้อยกว่าเมื่อเทียบกับอันดับที่ 8 ของ TWPR ในรูปที่ 25



ภาพที่ 26 แสดงเว็บเพจ http://www.about.com/gi/pages/patent.htm ซึ่งแสดงข้อมูล patent ที่คลอบคลุมทั้งหมดของเว็บเพจ about

อย่างไรก็ตามการการจัดเรียงลำดับเว็บเพจจากการคำนวณค่าเพจแร็งค์เพียงอย่างเดียว ซึ่งแสดงให้เห็นว่าอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มนั้นให้ข้อมูลที่น่าสนใจในอันดับแรกของการจัดเรียงผลลัพธ์เมื่อเทียบกับเพจแร็งค์แบบดั้งเดิมซึ่งผู้พัฒนาก็ได้พัฒนานำค่าคะแนนที่ได้รวมเข้ากับระบบสืบค้นข้อมูลซึ่งในเบื้องต้นเมื่อทดลองกับคำถามที่ผู้ใช้สนใจคือ “harry potter” ดังตารางที่ 10 ก็ได้ผลเป็นที่น่าพอใจโดยอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มให้ผลลัพธ์ที่เกี่ยวข้องมากถึง 14 เว็บเพจ ในขณะที่ T-Rank ให้มากกว่าเพียงเว็บเพจเดียวคือเป็น 15 เว็บเพจและแบบเพจแร็งค์ดั้งเดิมให้ 9 เว็บเพจ

ตารางที่ 10 แสดงเปรียบเทียบผลลัพธ์การค้นคืน 20 อันดับแรกระหว่าง TWPR และ PR และ T-Rank

Order	**TWPR**	**PR**	**T-Rank**
1	http://www.crossroad.to/text/articles.html	http://www.heavygames.com/	**http://www.harry-potter-games.com/Harry_Potter_Fan_Fiction.htm**
2	**http://www.harry-potter-games.com/index.html**	**http://lumos.mugglenet.com/**	**http://www.the-leaky-cauldron.org/**

ตารางที่ 10 แสดงเปรียบเทียบผลลัพธ์การค้นคืน 20 อันดับแรกระหว่าง TWPR และ PR และ T-Rank (ต่อ)

Order	TWPR	PR	T-Rank
3	**http://www.harry-potter-games.com/es/index.html**	http://www.blogcatalog.com/search	**http://www.harry-potter-games.com/index.html**
4	**http://www.harry-potter-games.com/pt/index.html**	http://www.kottke.org/	**http://www.harry-potter-games.com/J.K._Rowling.htm**
5	**http://www.the-leaky-cauldron.org/**	**http://www.crossroad.to/text/articles.html**	**http://www.harry-potter-games.com/Pictures_of_Emma_Watson.htm**
6	http://www.activityvillage.co.uk/index.htm	**http://www.the-leaky-cauldron.org/**	**http://www.harry-potter-games.com/Pictures_of_Tom_Felton.htm**
7	**http://www.harry-potter-games.com/Harry_Potter_Fan_Fiction.htm**	http://www.factmonster.com/	**http://www.harry-potter-games.com/Harry_Potter_Rumours.htm**
8	http://www.bettybowers.com/	**http://www.kidsreads.com/HP07/content/index.asp**	**http://www.harry-potter-games.com/Draco_Malfoy.htm**
9	**http://www.harrypotterfacts.com/_chapters.htm**	http://www.rottentomatoes.com/	**http://www.harry-potter-games.com/Harry_Potter_Characters.htm**
10	**http://www.harry-potter-games.com/Pictures_of_Emma_Watson.htm**	**http://www.harry-potter-games.com/index.html**	**http://www.activityvillage.co.uk/index.htm**
11	**http://www.harry-potter-games.com/Pictures_of_Tom_Felton.htm**	http://www.digital-digest.com/movies/index.php	**http://www.harry-potter-games.com/Emma_Watson.htm**
12	**http://www.harry-potter-games.com/J.K._Rowling.htm**	http://www.snitchseeker.com/	http://glosslip.com/
13	**http://www.harry-potter-games.com/Harry_Potter_Rumours.htm**	**http://www.harrypottertime.com/**	http://www.crossroad.to/text/articles.html

ตารางที่ 10 แสดงเปรียบเทียบผลลัพธ์การค้นคืน 20 อันดับแรกระหว่าง TWPR และ PR และ T-Rank (ต่อ)

Order	**TWPR**	**PR**	**T-Rank**
14	http://glosslip.com/	**http://shop.mugglenet.com/**	http://www.nlm.nih.gov/
15	**http://www.btinternet.com/~harrypotterguide/harrypotter.htm**	**http://www.harry-potter-games.com/es/index.html**	**http://www.btinternet.com/~harrypotterguide/harrypotter.htm**
16	**http://www.harry-potter-games.com/Draco_Malfoy.htm**	**http://www.harry-potter-games.com/pt/index.html**	http://travel.howstuffworks.com/
17	**http://www.harry-potter-games.com/Harry_Potter_Characters.htm**	http://www.factmonster.com/funfacts.html	**http://www.christnhp.org/Recs.htm**
18	**http://www.harry-potter-games.com/Emma_Watson.htm**	http://au.movies.yahoo.com/	http://www.ugoplayer.com/
19	http://www.nlm.nih.gov/	http://m.technorati.com/	**http://www.crossroad.to/articles2/HP-Movie.htm**
20	http://www.landoverbaptist.org/	http://www.factmonster.com/games.html	**http://blogs.raincoast.com/weblog/C14/**

จากตารางที่ 10 พิจารณาผลลัพธ์ระหว่าง PR และ TWPR ถึงแม้ว่า แบบเพจแร็งค์จะให้เว็บเพจได้หลากหลายมากกว่าแต่ จากการสุ่มตรวจสอบเนื้อหาเว็บเพจที่ปรากฏอยู่ในอันดับต้นๆ เช่น ยูอาร์แอล http://lumos.mugglenet.com/ ดังภาพที่ 27 แสดงข้อมูลแก้ไขล่าสุด 20 มกราคม ค.ศ. 2008 ของเพจแร็งค์นั้นมีข้อมูลอัพเดพที่เก่ากว่าเว็บเพจ http://www.the-leaky-cauldron.org/ ภาพที่ 28 แสดงข้อมูลแก้ไขล่าสุด 22 กุมภาพันธ์ 2008 ซึ่งอยู่ในลำดับที่ 5 ของ TWPR และอยู่ในอันดับที่ 2 ของการคำนวณ T-Rank

ภาพที่ 27 แสดงเว็บเพจ http://lumos.mugglenet.com/

ภาพที่ 28 แสดงเว็บเพจ http://www.the-leaky-cauldron.org/

และจากตารางที่ 10 พิจารณาผลลัพธ์ระหว่าง TWPR และ T-Rank ถึงแม้ว่าอัลกอริทึม T-Rank จะให้เว็บเพจใหม่ได้จำนวนมากกว่าแต่เมื่อตรวจยูอาร์เอลพบว่า เว็บเพจที่อยู่ในอันดับแรกๆนั้นมาจากเครื่องแม่ข่ายเดียวกันในขณะที่ TWPR จะให้ความหลากหลายกว่า ดังนั้นอัลกอริทึม TWPR จึงน่าสนใจมากกว่า

วิจารณ์

ในวิทยานิพนธ์เล่มนี้ศึกษาและประยุกต์ ใช้องค์ประกอบทางด้านเวลาปรับเข้ากับอัลกอริทึมเพจแร็งค์เดิม เพื่อเพิ่มประสิทธิภาพของผลลัพธ์ให้แสดงเว็บเพจที่มีความสดมากขึ้น ซึ่งจากผลการทดลองเบื้องต้น กับชุดข้อมูลเว็บเพจที่กล่าวมาแล้วนั้น แสดงให้เห็นว่า อัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้ม “TWPR” ที่นำเสนอ ให้ประสิทธิภาพที่ดีขึ้นเมื่อเทียบกับอัลกอริทึมเพจแร็งค์ดั้งเดิม โดยจากประการที่ 1 ในการวัดผลซึ่งนับจากจำนวนจุดเชื่อมโยงที่ได้รับโดย Yu *et al.* (2005) ผลลัพธ์ที่ได้อยู่ในระดับปานกลางเมื่อเทียบกับอัลกอริทึม PR และ T-Rank และจากประการที่ 2 ซึ่งนับเว็บเพจที่มีค่าวันเวลาแก้ไขล่าสุดใกล้เคียงกับค่าเวลาปัจจุบัน (2007Q3) จากลำดับแรก 20 และ 50 จากการคำนวณซึ่ง TWPR ให้ค่าอยู่ในระดับปานกลางอีกเช่นกัน อย่างไรก็ตาม ผลลัพธ์ที่ได้อยู่ในระดับปานกลางเมื่อเทียบกับอัลกอริทึม T-Rank ในประการทางที่ 2 ซึ่งจากการคิดค่าแนวโน้มนั้นเป็นค่าที่ได้มาจากกลุ่มของข้อมูล ซึ่งเพิ่มค่าความสำคัญให้แก่เว็บเพจดังกล่าวน้อยกว่าแบบอัลกอริทึม T-Rank ที่พิจารณาให้ค่าความสดของเว็บเพจในช่วงเวลาดังกล่าวสูงสุดเป็น 1 ซึ่งทำให้เว็บเพจในช่วงเวลาดังกล่าวได้รับค่าคะแนนที่สูงกว่าอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มมาก ดังนั้นจากการคำนวณจัดเรียงผลลัพธ์ที่ได้จากอัลกอริทึม T-Rank เว็บเพจใหม่จะมีมากที่สุด อย่างไรก็ตามหากพิจารณาในแง่คุณภาพของเว็บเพจโดยนับจำนวนเส้นการเชื่อมโยงอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มจะให้ประสิทธิภาพที่มากกว่ากว่าอัลกอริทึม T-Rank จากการทดลองกับระบบสืบค้นข้อมูล โดยระบบสืบค้นให้ข้อมูลเป็นเว็บเพจที่หลากหลายกว่าดังนั้นหากพิจารณาโดยรวมอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มนี้จะให้ผลลัพธ์โดยรวมที่ดีกว่าเนื่องจากให้ผลเป็นเว็บเพจใหม่ในขณะเดียวกันเว็บเพจที่ได้ยังคงเว็บเพจที่มีคุณภาพด้วย

สรุปและข้อเสนอแนะ

สรุป

จากผลการทดลองกับฐานข้อมูลที่ทำการเก็บจากระบบอินเตอร์เน็ตแสดงให้เห็นว่าอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้ม สามารถให้ผลลัพธ์การจัดเรียงลำดับ 20 ลำดับแรกให้เว็บเพจที่มีความสดใหม่มากกล่าวคืออยู่ในช่วง 2007Q3 ถึง 90 เปอร์เซ็นต์ในขณะที่อัลกอริทึมเพจแร็งค์ให้เว็บเพจที่อยู่ในช่วงดังกล่าว 45 เปอร์เซ็นต์(ข้อมูลตารางที่ 7) กล่าวคือมีประสิทธิภาพมากกว่าเพจแร็งค์โดยรวมเฉลี่ย 45 เปอร์เซ็นต์ ในด้านการเพิ่มค่าความสำคัญให้กับเว็บเพจใหม่ ซึ่งถือว่าสามารถทำให้เว็บเพจใหม่มีโอกาสอยู่ในลำดับแรกๆของผลลัพธ์การค้นคืนข้อมูล ซึ่งหากเป็นเว็บเพจใหม่ที่มีคุณภาพและมีการเปลี่ยนแปลงแก้ไขเว็บเพจให้มีความสดใหม่อยู่เสมอก็จะได้รับจำนวนจุดเชื่อมโยงเพิ่มขึ้นไปตามเวลา สามารถทำให้คงอยู่ในลำดับแรกๆได้เป็นอย่างดี ในขณะเดียวกันเว็บเพจใหม่ที่ไม่มีการเปลี่ยนแปลงแก้ไขข้อมูลเว็บเพจก็จะมีอายุมากขึ้นเรื่อยๆและอาจได้รับจำนวนเส้นการเชื่อมโยงน้อยลงค่าแนวโน้มก็จะลดลง ซึ่งเมื่อนำมาจัดลำดับด้วยอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มก็จะทำให้มีค่าความสำคัญลดลง และพิจารณาอีกแง่หนึ่งโดยนับจากจำนวนของเส้นการเชื่อมโยงที่ชี้เข้าหาเว็บเพจซึ่งเป็นการวัดจากความนิยมของเว็บเพจ (Page popularity) นั่นเอง ซึ่งจากผลการทดลองอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มให้เว็บเพจใน 20 และ 50 อันดับแรกมีเปอร์เซ็นต์จุดเชื่อมโยงชี้เข้าหาเว็บเพจสูง 48.2 และ 44.89 เปอร์เซ็นต์ตามลำดับซึ่งมีค่าที่สูงกว่าเมื่อเทียบกับอัลกอริทึม T-Rank ซึ่งแสดงว่า อัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มนี้สามารถให้เว็บเพจที่มีคุณภาพมากกว่าอัลกอริทึม T-Rank นั่นเอง

เพราะฉะนั้น สามารถที่จะสรุปได้ว่าวิธีการของอัลกอริทึมเพจแร็งค์แบบถ่วงน้ำหนักตามแนวโน้มที่นำเสนอนี้ จะให้ผลลัพธ์เป็นเว็บเพจใหม่ในขณะเดียวกันก็ยังคงเว็บเพจที่มีคุณภาพและด้วยคุณลักษณะของอัลกอริทึมนี้ ผู้วิจัยคาดว่าเหมาะสมสำหรับนำไปใช้ในการจัดลำดับกลุ่มเว็บเพจที่มีการเปลี่ยนแปลงข้อมูลบ่อยครั้งและข้อมูลบนเว็บเพจมีอัตราการเสื่อมสภาพที่ค่อนข้างเร็วซึ่งลักษณะของเว็บเพจ ยกตัวอย่าง เช่น เว็บเพจฐานข้อมูลด้านการข่าว เว็บเพจฐานข้อมูลด้านเทคโนโลยี เนื่องจากฐานข้อมูลเหล่านี้ผู้ใช้มีความต้องการที่จะค้นหาเว็บเพจที่มีความสดใหม่อยู่เสมอซึ่งจะได้นำเสนอต่อไปในข้อเสนอแนะ

ข้อเสนอแนะ

เพื่อผลลัพธ์ที่ดีขึ้นของการคำนวณเพจแร็งค์โดยใช้เวลาเป็นองค์ประกอบ ในส่วนของการกำหนดค่าแนวโน้ม (Trend) นั้นเป็นการหาจากสัดส่วนจำนวนของเว็บเพจที่อยู่ในช่วงเวลาที่กำหนดเมื่อเทียบกับจำนวนของเว็บเพจทั้งหมด โดยในงานวิจัยนี้ได้เลือกทดลองศึกษากับเว็บเพจในช่วงระยะเวลา 3 เดือน ซึ่งจากวิธีการดังกล่าวเป็นการประมาณค่าจากช่วงเวลาอัตราเฉลี่ยของการเปลี่ยนแปลงแก้ไขข้อมูล ซึ่งค่านี้อาจจะสามารถปรับจูนขยายช่วงเวลาเพิ่มขึ้นเป็น 4 เดือนหากคุณลักษณะการเปลี่ยนแปลงข้อมูลเว็บเพจนั้นเกิดขึ้นช้าแล้วปรับลดลงน้อยกว่า 3 เดือนในกรณีที่การเปลี่ยนแปลงนั้นเกิดขึ้นเร็วซึ่งข้อมูลเหล่านี้จำเป็นที่จะต้องมีการศึกษาเพิ่มเติมเพื่อให้การปรับค่านี้เหมาะสมที่สุด และจากการที่อัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มนั้นใช้เวลาเป็นหลัก ดังนั้นเครื่องมือที่ใช้เก็บข้อมูลเวลาการแก้ไขล่าสุดของเว็บเพจจำเป็นจะต้องใช้เวลาในการเก็บข้อมูลให้เร็วที่สุดเพื่อให้ข้อมูลมีความสดใหม่มากที่สุด ซึ่งในความเป็นจริงหากเราใช้ฐานข้อมูลแบบทั่วไปซึ่งมีขนาดใหญ่จะทำให้การเก็บข้อมูลเป็นไปได้ยากเพราะทรัพยากรที่ใช้ในการค้นหามีจำนวนจำกัดเช่น ช่องทางสื่อสารข้อมูลของเครือข่ายอินเตอร์เน็ต ดังนั้นเราอาจสามารถแบ่งฐานข้อมูลเฉพาะที่มีความต้องการเฉพาะความสดใหม่ ดังเช่น ฐานข้อมูลการข่าวหรือเทคโนโลยี มาแยกคำนวณด้วยอัลกอริทึมแบบถ่วงน้ำหนักตามแนวโน้มซึ่งจะช่วยเสริมให้การจัดลำดับเว็บเพจสามารถแสดงข้อมูลที่สดใหม่กว่าเมื่อเทียบกับอัลกอริทึมเพจแร็งค์แบบดั้งเดิม

ในส่วนของการพิจารณาเว็บเพจใหม่นั้นในวิทยานิพนธ์เล่มนี้เป็นการพิจารณาเปรียบเทียบโดยใช้เวลาการแก้ไขล่าสุดของเว็บเพจเป็นเกณฑ์หลักในการพิจารณา ซึ่งค่านี้หาได้จากเครื่องมือเก็บข้อมูล (Crawler) ทำการเก็บข้อมูล ณ เครื่องแม่ข่าย โดยหากช่วงเวลาของการทำการเก็บข้อมูลไม่เหมาะสม กล่าวคือระหว่างที่ระบบทำการเก็บข้อมูล เว็บเพจต้นทางยังไม่ถึงกำหนดช่วงเวลาจำเป็นสำหรับการเปลี่ยนแปลงแก้ไขข้อมูลก็อาจส่งผลให้ข้อมูลที่ได้เป็นเวลาเก่าทำให้ลำดับของเว็บเพจลดลง ซึ่งจำเป็นอย่างยิ่งที่จะต้องพัฒนาเครื่องมือเก็บข้อมูลให้ดี

เอกสารและสิ่งอ้างอิง

D. Fetterly, M. Manasse, M. Najork, and J. Wiener. 2003. A Large-Scale Study of the Evolution of Web Pages. **International World Wide Web Conference.** 12: 669-678

E. Amitay, D. Carmel, M. Herscovici, R. Lempel, and A. Soffer . 2004. Trend detection through temporal link analysis. **Journal of the American Society for Information Science and Technology.** 14(55): 1270-1281

E. Garfield. 1998. The use of journal impact factors and citation analysis for evaluation of science. **Unpublished presentation at Cell Separation, Hematology and Journal Citation Analysis Mini Symposium in tribute to Arne Bøyum, Rikshospitalet**. 69 (3): 224-229

J. Cho and H. Garcia-Molina. 2003. Estimating frequency of change. **ACM Transactions on Internet Technology.** 3(3): 256-290

J. Cho and H. Garcia-Molina. 2000. Synchronizing a database to improve freshness. **ACM International Conference on Management of Data (SIGMOD).** 26: 117-128

J. Cho and H. Garcia-Molina. 2000. The evolution of the web and implications for an incremental crawler. **The VLDB Journal.** 26: 200-209

J. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. **Proceedings of the ACM-SIAM Symposium on Discrete Algorithms.** 9: 604-632

K. Berberich, M. Vazirgiannis and G. Weikum. 2004. T-Rank: Time-aware Authority Ranking. **Proceedings of the International Workshop on Algorithms and Models for the Web-**

Graph. 31(3243): 131-142

L. Egghe. 2001. A noninformetric analysis of the relationship between citation age and journal productivity. **Journal of the American Society for Information Science and Technology.** 52(5): 371-377.

L. Adamic, B.A. Huberman. 2001. The Web's hidden order. **Communications of the ACM**. 9(44): 55-60

L. Page, S. Brin, R. Motwani and T. Winograd. 1998. **The PageRank Citation Ranking: Bringing Order to the Web**. Technical Report, Stanford Univ.

P.S. Yu, X. Li and B. Liu. 2005. Adding the Temporal Dimension to Search - A Case Study in Publication Search. **Proceeding of the IEEE/WIC/ACM International Conference on Web Intelligence.** 10: 543-549

R. Baeza-Yates, F. Saint-Jean, and C. Castillo. 2002. Web Structure, Dynamics and Page Quality. **Proceedings of the 9th edition of the Symposium on String Processing and Information Retrieval, LNCS**, 9: 117-130

T.H. Haveliwala. 1999. **Efficient Computation of PageRank**. Technical Report, Stanford Univ.

W.E. Donath, A.J. Hoffman. 1972. Algorithm for partitioning of graphs and computer logic based on eigenvectors of connections matrices. **IBM Technical Disclosure Bulletin**. 15: 938-944

ประวัติการศึกษา และการทำงาน

ชื่อ –นามสกุล	นายศุภกร กาญจนการุณ
วัน เดือน ปี ที่เกิด	วันที่ 16 พฤศจิกายน 2522
สถานที่เกิด	บุรีรัมย์
ประวัติการศึกษา	วศ.บ. (วิศวกรรมคอมพิวเตอร์) คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น (พ.ศ.2549)
ตำแหน่งหน้าที่การงานปัจจุบัน	พนักงานบริษัท
สถานที่ทำงานปัจจุบัน	เอส เอส ซี โซลูชั่น
ผลงานดีเด่นและรางวัลทางวิชาการ	
ทุนการศึกษาที่ได้รับ	