กรณีศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอว์เลอร์

สุทิน อุทธบูรณ์

งานนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาการสารสนเทศ มหาวิทยาลัยบูรพา
ตุลาคม 2560

A CASE STUDY OF WEB RESEARCH DATA SCRAPING BY WEB CRAWLERS.

SUTIN UTTABOON

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT

FOR THE MASTER DEGREE OF SCIENCE IN INFORMATION TECHNOLOGY

FACULTY OF INFORMATICS BURAPHA UNIVERSITY

OCTOBER 2017

คณะกรรมการควบคุมงานนิพนธ์และคณะกรรมการสอบงานนิพนธ์ได้พิจารณางานนิพนธ์ของ นายสุทิน อุทธบูรณ์ ฉบับนี้แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพาได้

คณะกรรมการควบคุมงานนิพนธ์

ดร.ทัศนีย์ เจริญพร อาจารย์ที่ปรึกษา

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุรางคนา ธรรมลิขิต)

..... กรรมการ
(ดร.คนึงนิจ กุโบลา)

..... กรรมการ
(ดร.ทัศนีย์ เจริญพร)

คณะวิทยาการสารสนเทศ อนุมัติให้รับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยบูรพา

..... คณบดีคณะวิทยาการสารสนเทศ
(ผู้ช่วยศาสตราจารย์ ดร.กฤษณะ ชินสาร)
วันที่ 25 เดือน ธันวาคม พ.ศ. 2560

กิตติกรรมประกาศ

งานนิพนธ์นี้สำเร็จได้โดยได้รับความกรุณาและความช่วยเหลือจากอาจารย์ ดร.ทัศนีย์ เจริญพร อาจารย์ผู้ควบคุมงานนิพนธ์ ตลอดระยะเวลาที่จัดทำงานนิพนธ์ฉบับนี้ อาจารย์ได้ให้การช่วยเหลือการทำงานทุกด้าน ทำให้งานนิพนธ์นี้มีความคืบหน้าในการทำงานที่รวดเร็ว แม้ในการทำงานนิพนธ์ ครั้งนี้จะมีอุปสรรคและผลลัพธ์ของการทำงานที่ไม่เป็นไปตามเป้าหมายหลายครั้ง แต่ด้วยเพราะกำลังใจและความเอาใจใส่ที่อาจารย์มอบให้ ทำให้ผู้จัดทำงานนิพนธ์มีกำลังใจในการที่จะดำเนินงานนิพนธ์นี้

ขอขอบพระคุณ ผศ.ดร.สุรางคนา ธรรมลิขิต ที่คอยสอนและให้คำแนะนำที่ดี เพื่อให้งานนิพนธ์เกิดความสำเร็จได้ตามระยะเวลาที่กำหนด

ขอขอบพระคุณ ผศ.ดร.ณัฐนนท์ ลีลาตระกูล ที่คอยให้คำปรึกษา ติดตามความคืบหน้าในการทำงาน รวมทั้งให้คำแนะนำที่เป็นประโยชน์ต่อการศึกษา ทำให้ผู้จัดทำงานนิพนธ์มีความเข้าใจจุดมุ่งหมายของการทำงานนิพนธ์มากขึ้น ทั้งยังคอยกระตุ้นให้ทำงานนิพนธ์นี้จนสำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณ ผศ.ดร.กฤษณะ ชินสาร ที่คอยให้คำปรึกษาแนะนำแนวทาง ในการเรียนหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ในครั้งนี้

ขอขอบพระคุณ คุณพ่อ คุณแม่ ที่คอยเติมกำลังแรงใจตลอดการทำงานนิพนธ์ในครั้งนี้และเป็นแบบอย่างในการทำงาน ทำให้ผู้จัดทำงานนิพนธ์ไม่ย่อท้อต่ออุปสรรคและมีความตั้งใจในการทำงานนิพนธ์นี้ให้สำเร็จ

ขอขอบคุณเพื่อนๆ ร่วมรุ่น ป.โท เทคโนโลยีสารสนเทศรุ่น 10 ทุกคน สำหรับการดูแลเอาใจใส่ให้ความช่วยเหลือและกำลังใจที่มอบให้ตลอดระยะเวลาของการศึกษา

สุทิน อุทธบูรณ์

57920646: สาขาวิชา: เทคโนโลยีสารสนเทศ; วท.ม. (เทคโนโลยีสารสนเทศ)
คำสำคัญ: เว็บครอเลอร์/ โปรแกรมค้นหา/ งานวิจัย

สุทิน อุทธบูรณ์: กรณีศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอเลอร์
(A CASE STUDY OF WEB RESEARCH DATA SCRAPING BY WEB CRAWLERS.)
อาจารย์ผู้ควบคุมงานนิพนธ์: ทัศนีย์ เจริญพร, Ph.D., 47 หน้า. ปี พ.ศ. 2560.

งานนิพนธ์นี้ นำเสนอผลการศึกษาและประยุกต์ใช้วิธีการสกัดข้อมูลบนเว็บเพจด้วยเว็บครอเลอร์จากเว็บไซต์ที่รวบรวมงานวิจัยต่างๆ มาแสดงไว้บนเว็บไซต์เดียวกัน เพื่อให้สามารถค้นหาได้สะดวกและรวดเร็วขึ้น โดยใช้ภาษาและเครื่องมือที่ปรากฏอยู่ในปัจจุบัน ได้แก่ Nodejs และ Cheerio ซึ่งเป็นเครื่องมือที่สามารถดึงข้อมูลจากเว็บไซต์ด้วยวิธีการเข้าถึงโครงสร้าง HTML ของเว็บไซต์นั้น ๆ เพื่อสกัดข้อมูลที่ต้องการ และจัดเก็บข้อมูลที่ได้ลงในฐานข้อมูล สำหรับนำไปสร้างส่วนแสดงผลลัพธ์ต่อไป โดยได้ทดลองสกัดข้อมูลงานวิจัยจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา เว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย และเว็บไซต์คลังข้อมูลงานวิจัยไทย ผลการทดลองสกัดข้อมูลงานวิจัยทั้ง 3 เว็บไซต์ ที่มีจำนวนงานวิจัยรวมทั้งหมด 543,695 รายการนั้น ใช้เวลาในการครอว์ทั้งสิ้น 1,434 นาที เมื่อนำข้อมูลทั้งหมดมาหาค่า Precision, Recall และ F-Measure เพื่อหาประสิทธิภาพของการสกัดข้อมูล พบว่า ค่า Precision เท่ากับร้อยละ 99 ของสัดส่วนของจำนวนข้อมูล (Records) ที่สกัดได้ตรงตามความต้องการต่อจำนวนข้อมูลงานวิจัยทั้งหมด ค่า Recall เท่ากับร้อยละ 99 ของสัดส่วนของจำนวนข้อมูล(Records) ที่สืบค้นได้ตรงตามความต้องการต่อจำนวนข้อมูลที่ตรงตามความต้องการ และเมื่อวัดค่า F-measure เพื่อหาประสิทธิภาพของการนำเครื่องมือดังกล่าวมาใช้ในการครอว์และสืบค้น พบว่ามีค่าความถูกต้องร้อยละ 99 จึงแสดงให้เห็นว่าการประยุกต์ใช้วิธีการสกัดข้อมูลบนเว็บเพจด้วยเว็บครอเลอร์นี้มีประสิทธิภาพ ผลของการศึกษาสามารถนำมาใช้เป็นส่วนหนึ่งของระบบบริหารจัดการประวัติการทำงานของบุคลากรต่อไปได้

57920646: MAJOR; INFORMATION TECHNOLOGY; M.Sc.
(INFORMATION TECHNOLOGY)
KEYWORDS: WEB CRAWLERS / SEARCH ENGINE / RESEARCH

SUTIN UTTABOON: A CASE STUDY OF WEB RESEARCH DATA SCRAPING BY WEB CRAWLERS. THESIS ADVISOR: THATSANEE CHAROENPORN, D., 44 P. 2017.

This thesis presents the study and application of data extraction method on Web page by WebCrawler in order to facilitate the data searching. The application has been conducted by applying the current existing programming language and tool including Nodejs and Cheerio. They are able to extract the required information from the websites by accessing the HTML structure and store it in the local database for further searching and retrieving. The experiment has been done on the 543,695 research information records from 3 main research Websites of Thailand including Burapha University's research Website, Thailand Library Network Project Website, and Thai National Research Repository Website. The result presents that all research information records can be extracted within 1,434 minutes. Precision-Recall and F-measure are employed to evaluate the accuracy of extracting and search result. The result value of 0.99 can be illustrated the high accuracy of applying the proposed method. The consequence of the application can be used as part of the approaching personnel management system development.

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญภาพ

สารบัญภาพ (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันงานวิจัยทั้งในประเทศและต่างประเทศมีจำนวนเพิ่มมากขึ้น และด้วยความก้าวหน้าทางเทคโนโลยีทำให้มีการรวบรวมผลงานวิจัยไว้ในเว็บไซต์ต่าง ๆ เพื่อให้บริการแก่ผู้ที่สนใจแต่เนื่องจากมีเว็บไซต์ให้บริการข้อมูลงานวิจัยเป็นจำนวนมาก จึงทำให้ผู้ใช้งานต้องค้นหางานวิจัยที่สนใจจากหลาย ๆ เว็บไซต์เพื่อให้ได้ข้อมูลงานวิจัยที่ครอบคลุมและตรงตามความต้องการ ทำให้นักวิจัยใช้เวลานานในการรวบรวมข้อมูลงานวิจัยที่สนใจ รวมไปถึงการได้ข้อมูลงานวิจัยที่เป็นข้อมูลเดียวกันจากเว็บไซต์หลายแหล่ง เพื่อเป็นการอำนวยความสะดวกให้แก่นักวิจัย ผู้จัดทำงานนิพนธ์จึงได้มีการพัฒนาระบบค้นหาข้อมูลงานวิจัย ที่รวบรวมข้อมูลงานวิจัยจากแหล่งต่าง ๆ ด้วยเทคนิคการดึงข้อมูลจากเว็บไซต์ที่เรียกว่า “Web Scraping” หรือ “Web Crawler”

งานนิพนธ์นี้จะกล่าวถึง การค้นหาข้อมูลแบบเฉพาะเจาะจงเกี่ยวกับข้อมูลงานวิจัยโดยจะนำข้อมูลงานวิจัยจากเว็บไซต์ที่ให้บริการ มารวมกันในแหล่งเดียวหรือเว็บไซต์เดียว เพื่อให้ง่ายต่อการค้นหาและการตรวจสอบ ดังนั้น การทำให้เว็บไซต์มีความสามารถในการเก็บรวบรวมข้อมูลได้ต้องอาศัยหลักการสกัดข้อมูลที่เรียกว่า Web Crawler ดังที่ได้กล่าวมาแล้วข้างต้น ด้วยการสกัดข้อมูลงานวิจัยที่ต้องการจากเว็บไซต์ที่กำหนด เพื่อนำข้อมูลที่ได้มาจัดเก็บในฐานข้อมูล และนำข้อมูลไปพัฒนาระบบค้นหางานวิจัยต่อไป

อนึ่ง งานนิพนธ์นี้เป็นจุดเริ่มต้นของการพัฒนาระบบฐานข้อมูลนักวิจัยไทย กล่าวคือสามารถรวบรวมงานวิจัยและจำแนกเป็นรายการตามผู้สร้างระบบได้แล้ว จะสามารถต่อยอดเป็น Profile หรือประวัติงานวิจัยของนักวิจัยแต่ละท่านได้อย่างรวดเร็ว

1.2 วัตถุประสงค์การศึกษา

1. เพื่อศึกษาขั้นตอน วิธีการ และออกแบบการสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์
2. เพื่อรวบรวมข้อมูลงานวิจัยจากเว็บไซต์ต่าง ๆ มาจัดเก็บในฐานข้อมูลเดียว
3. เพื่อวัดประสิทธิภาพการสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์ที่ออกแบบไว้
4. เพื่อพัฒนาระบบค้นหางานวิจัยที่ได้จากการสกัดข้อมูลด้วยเว็บครอว์เลอร์

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการศึกษา

1. เป็นแนวทางในการศึกษาวิธีการสกัดข้อมูลได้ด้วยเว็บครอว์เลอร์ และนำไปประยุกต์ใช้ในด้านอื่น
2. เป็นแนวทางในการสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์ และนำไปใช้ในการพัฒนาระบบค้นหาข้อมูลงานวิจัยต่อไป
3. ระบบค้นหาข้อมูลงานวิจัยที่รวบรวมงานวิจัยจากเว็บไซต์ต่าง ๆ ได้รับการพัฒนาขึ้น
4. เป็นจุดเริ่มต้นของการพัฒนาเว็บไซต์รวบรวมงานวิจัยของนักวิจัยไทย

1.4 ขอบเขตของการศึกษา

งานนิพนธ์กรณีศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอว์เลอร์นี้มีความตั้งใจที่จะพัฒนาเว็บไซต์ เพื่อให้เป็นศูนย์กลางข้อมูลสารสนเทศด้านผลงานวิจัย ผู้จัดทำงานนิพนธ์ โดยได้ดำเนินการศึกษาโครงสร้างเว็บไซต์งานวิจัยต่าง ๆ เพื่อใช้เครื่องมือสกัดข้อมูลของเว็บไซต์งานวิจัยนั้น และจัดเก็บลงฐานข้อมูลสำหรับการค้นหาผลงานวิจัย ผ่านทางหน้าเว็บไซต์ที่จัดทำขึ้น โดยมีขอบเขตของงานนิพนธ์ดังต่อไปนี้

1. สกัดข้อมูลงานวิจัยจากเว็บไซต์ภายในประเทศไทยทั้งหมด 3 เว็บไซต์ โดยการวิเคราะห์โครงสร้าง HTML ของแต่ละเว็บไซต์เพื่อการสกัดข้อมูลงานวิจัยที่ต้องการ

1.1 เว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา ประกอบไปด้วย 5 ส่วนหลัก ได้แก่

ส่วนที่ 1 คือ ส่วนหัวของเว็บไซต์

ส่วนที่ 2 คือ ส่วนของการค้นหางานวิจัย

ส่วนที่ 3 คือ ส่วนของลิ้งก์ที่เกี่ยวข้อง

ส่วนที่ 4 คือ ส่วนของการเข้าสู่ระบบเข้าใช้งานและการสมัครลงทะเบียน

ส่วนที่ 5 คือ ส่วนแสดงผลลัพธ์

ซึ่งข้อมูลงานวิจัยที่ต้องการสกัดข้อมูลนั้น อยู่ในส่วนแสดงผลลัพธ์ โดยใช้ Cheerio สกัดข้อมูล ตารางรายละเอียดงานวิจัย

1.2 เว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย งานวิจัยที่ต้องการถูกจัดเก็บในตารางแสดงผลลัพธ์ ซึ่งงานนิพนธ์นี้จะออกแบบฟังก์ชันสกัดข้อมูลเพื่อรองรับการสกัดข้อมูลในตารางแสดงผลลัพธ์

1.3 เว็บไซต์คลังข้อมูลงานวิจัยไทย เว็บไซต์งานวิจัยนี้ข้อมูลงานวิจัยที่ต้องการมีโครงสร้าง HTML ในรูปแบบชื่อ Class ซึ่งจะต้องออกแบบฟังก์ชันสกัดข้อมูลงานวิจัยที่สามารถเข้าถึงโครงสร้าง Class ที่ต้องการ

2. สร้างฟังก์ชันสกัดข้อมูลงานวิจัยที่รองรับทั้ง 3 เว็บไซต์ในข้อที่ 1 อาศัยเทคนิคการสกัดข้อมูลด้วย Cheerio เป็นภาษา JavaScript ที่ทำงานในฝั่งเครื่องคอมพิวเตอร์แม่ข่าย และสกัดข้อมูลตามลำดับคิว และตามเวลาที่กำหนด

3. นำผลลัพธ์งานวิจัยที่ได้จากการสกัดข้อมูลมาจัดเก็บในฐานข้อมูล เพื่อนำข้อมูลที่ได้ ไปพัฒนาเว็บไซต์สำหรับการค้นหางานวิจัย

4. พัฒนาเว็บไซต์ค้นหางานวิจัย ที่ผู้ใช้งานสามารถค้นหางานวิจัยได้จากชื่อผลงานวิจัย และชื่อผู้เขียนงานวิจัย

5. วัดประสิทธิภาพการสกัดข้อมูลงานวิจัย ด้านระยะเวลาในการสกัดข้อมูล และด้านความถูกต้องในการสกัดข้อมูล

1.5 ระยะเวลาในการดำเนินงาน

สำหรับการศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอว์เลอร์ และพัฒนาระบบค้นหางานวิจัยในครั้งนี้ ได้มีการกำหนดระยะเวลาในการดำเนินงาน เพื่อให้บรรลุตามวัตถุประสงค์ที่กำหนดไว้ โดยมีรายละเอียดดังตารางที่ 1-1

ตารางที่ 1-1 ตารางระยะเวลาในการดำเนินงาน

รายละเอียดการศึกษา	ปี พ.ศ. 2559			ปี พ.ศ. 2560						
	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.
1. ศึกษาข้อมูลของความรู้พื้นฐานเกี่ยวกับการทำงานเว็บครอว์เลอร์การค้นหางานวิจัย										
2. ศึกษาวิธีการการทำงานของเว็บครอว์เลอร์การค้นหางานวิจัย										
3. ศึกษาและรวบรวมงานวิจัยที่เกี่ยวข้อง										
4. ออกแบบโปรแกรม และฐานข้อมูลสำหรับ เว็บครอว์เลอร์การค้นหางานวิจัย										
5. พัฒนาเว็บครอว์เลอร์การสกัดข้อมูลงานวิจัย และพัฒนาระบบสืบค้นข้อมูลงานวิจัยไทย										
6. ทดสอบการทำงานทั้งระบบ										
7. ปรับปรุงระบบสืบค้นข้อมูลงานวิจัยไทย										
8. สรุปผลการทดลอง และสรุปปัญหา										
9. จัดทำเอกสารและส่งรายงานฉบับสมบูรณ์										

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานนิพนธ์นี้ได้นำวิธีการสกัดข้อมูลด้วยเว็บครอว์เลอร์มาใช้ในการสกัดข้อมูลงานวิจัย เพื่อนำข้อมูลงานวิจัยที่ได้มาพัฒนาระบบค้นหางานวิจัย ซึ่งผู้ใช้งานสามารถค้นหางานวิจัยที่ผู้จัดทำงานนิพนธ์รวบรวมมาจากเว็บไซต์งานวิจัยในประเทศไทยและเว็บไซต์งานวิจัยต่างประเทศได้ ในบทนี้จะนำเสนอทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการพัฒนา ประกอบด้วยโปรแกรมค้นหา หลักการทำงานเว็บครอว์เลอร์และงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดดังต่อไปนี้

2.1 โปรแกรมค้นหา (Search Engine)

โปรแกรมที่ช่วยในการสืบค้นหาข้อมูลบนอินเทอร์เน็ต ค้นหาข้อมูลต่าง ๆ เช่น ข้อความ รูปภาพ ภาพเคลื่อนไหว เพลง ซอฟต์แวร์ แผนที่ ข้อมูลบุคคล กลุ่มข่าว เป็นต้น ซึ่งแตกต่างกันไปขึ้นอยู่กับผู้ให้บริการแต่ละราย เสิร์ชเอนจินส่วนใหญ่จะค้นหาข้อมูลจากคำสำคัญ (Keyword) ที่ผู้ใช้ป้อนเข้าไป จากนั้นก็จะแสดงรายการผลลัพธ์ที่มันคิดว่าผู้ใช้น่าจะต้องการขึ้นมา

ปัจจุบันเสิร์ชเอนจินบางตัว เช่น กูเกิล จะบันทึกประวัติการค้นหาและการเลือกผลลัพธ์ของผู้ใช้ไว้ด้วย และจะนำประวัติที่บันทึกไว้นั้นเพื่อช่วยกรองผลลัพธ์ในการค้นหาครั้งต่อ ๆ ไป หลักการของโปรแกรมค้นหาเริ่มจากทำการเก็บรวบรวมข้อมูลต่าง ๆ จากเว็บไซต์ บนอินเทอร์เน็ต มาทำดัชนีให้อยู่ในรูปแบบที่ง่ายต่อการสืบค้นข้อมูล เมื่อผู้ใช้งานระบุคำค้นหา ที่ต้องการระบบจะทำการแสดงข้อมูลที่คาดว่าตรงตามความต้องการให้แก่ผู้ใช้งาน ส่วนประกอบของโปรแกรมค้นหาประกอบไปด้วย 3 ส่วน ดังนี้

1. เว็บครอว์เลอร์ คือ โปรแกรมเก็บรวบรวมข้อมูลจากเว็บไซต์ต่าง ๆ ตามลำดับที่กำหนดไว้ จากนั้นเก็บข้อมูลที่ได้ลงฐานข้อมูล

2. ดัชนี คือ การสร้างดัชนีให้กับข้อมูลเพื่อช่วยให้การค้นหาข้อมูลรวดเร็วยิ่งขึ้น

2.1 การกรองข้อมูล คือ กระบวนการเพื่อตรวจสอบว่าข้อมูลที่พบนี้สามารถนำไปทำดัชนีได้หรือไม่

2.2 แยกคำ คือ การรับข้อมูลจากขั้นตอนการกรองในรูปแบบสายอักขระ แล้วทำการตัดสายอักขระออกเป็นคำ ๆ เพื่อตรวจสอบว่าสามารถนำไปทำ ดัชนีได้หรือไม่

2.3 สร้างดัชนี ขั้นตอนนี้จะทำหน้าที่ตรวจสอบคำศัพท์แต่ละคำที่ได้มาจากการแยกคำ แล้วพิจารณาว่าคำศัพท์คำนั้นสมควรที่จะนำมาทำดัชนีหรือไม่

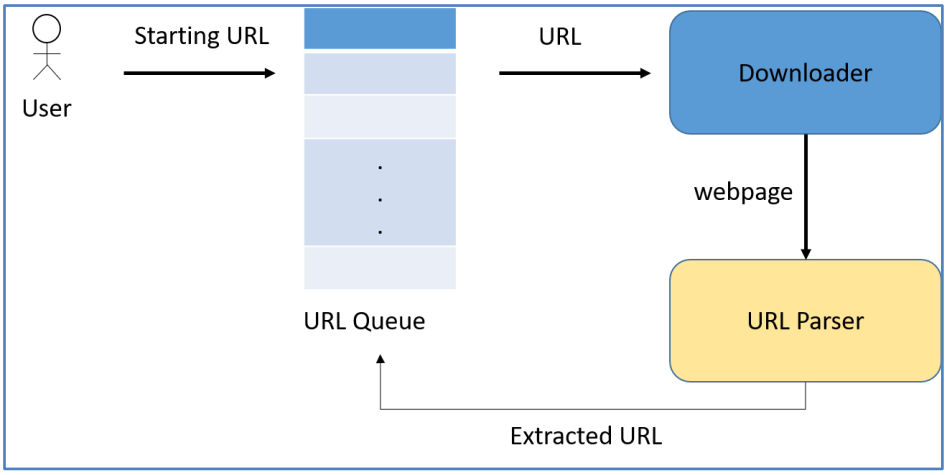
3. หน้าเว็บค้นหา (Searcher) คือ โปรแกรมสำหรับรับคำศัพท์ที่ต้องการค้นหาและเปรียบเทียบคำศัพท์ และดัชนีที่สร้างไว้จากนั้น แสดงผลลัพธ์ ตามลำดับความสำคัญ ที่มา: กลยุทธ บพิตร. (2555) ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับ เว็บครอว์เลอร์ที่ใช้ในโปรแกรมค้นหา

2.2 ความรู้เบื้องต้นเกี่ยวกับเว็บครอว์เลอร์

เว็บครอว์เลอร์เป็นโปรแกรมที่ถูกพัฒนาขึ้นเพื่อใช้ประโยชน์ในการรวบรวมยูอาร์-แอลจากอินเทอร์เน็ต ซึ่งนักพัฒนาโปรแกรมพยายามออกแบบและสร้างเว็บครอว์เลอร์ให้สามารถทำงานให้มีประสิทธิภาพสูงสุด เพื่อให้ได้ข้อมูลที่ต้องการ อย่างไรก็ตามการออกแบบและลักษณะกระบวนการทำงานยังคงมีพื้นฐานเดียวกัน ซึ่งในหัวข้อนี้กล่าวถึงแบบจำลองพื้นฐานสำหรับการออกแบบและพัฒนาเว็บครอว์เลอร์ ซึ่งประกอบด้วยส่วนต่าง ๆ ดังนี้

1. แบบจำลองพื้นฐานของเว็บครอว์เลอร์

เว็บครอว์เลอร์เป็นโปรแกรมที่สามารถเก็บรวบรวมเว็บเพจ โดยอัตโนมัติ เริ่มจากการสร้างยูอาร์แอลคิว หรือการจัดลำดับคิวสำหรับการครอว์เลอร์ การเก็บยูอาร์แอลโดยเริ่มจากการเก็บยูอาร์แอลเริ่มต้น ซึ่งยูอาร์แอลใหม่ที่อยู่ในเว็บเพจนั้นจะถูกแยกออกมาและนำไปตรวจสอบว่ามีการเก็บยูอาร์แอลนี้ก่อนหรือไม่ หากตรวจสอบพบว่าเคยมีการเก็บมาก่อนจะไม่เก็บซ้ำอีกรอบ แต่หากพบว่าไม่เคยเก็บยูอาร์แอลนี้มาก่อนเว็บครอว์เลอร์จะต้องเก็บเว็บเพจยูอาร์แอลนั้น ขั้นตอนนี้จะถูกทำงานวนลูปจนสิ้นสุดการทำงานเมื่อเก็บเว็บเพจได้ครบตามจำนวนที่กำหนดหรือไม่พบยูอาร์แอลที่สามารถเก็บต่อไป ดังภาพที่ 2-1 แสดงให้เห็นถึงการทำงานแบบพื้นฐานของเว็บครอว์เลอร์

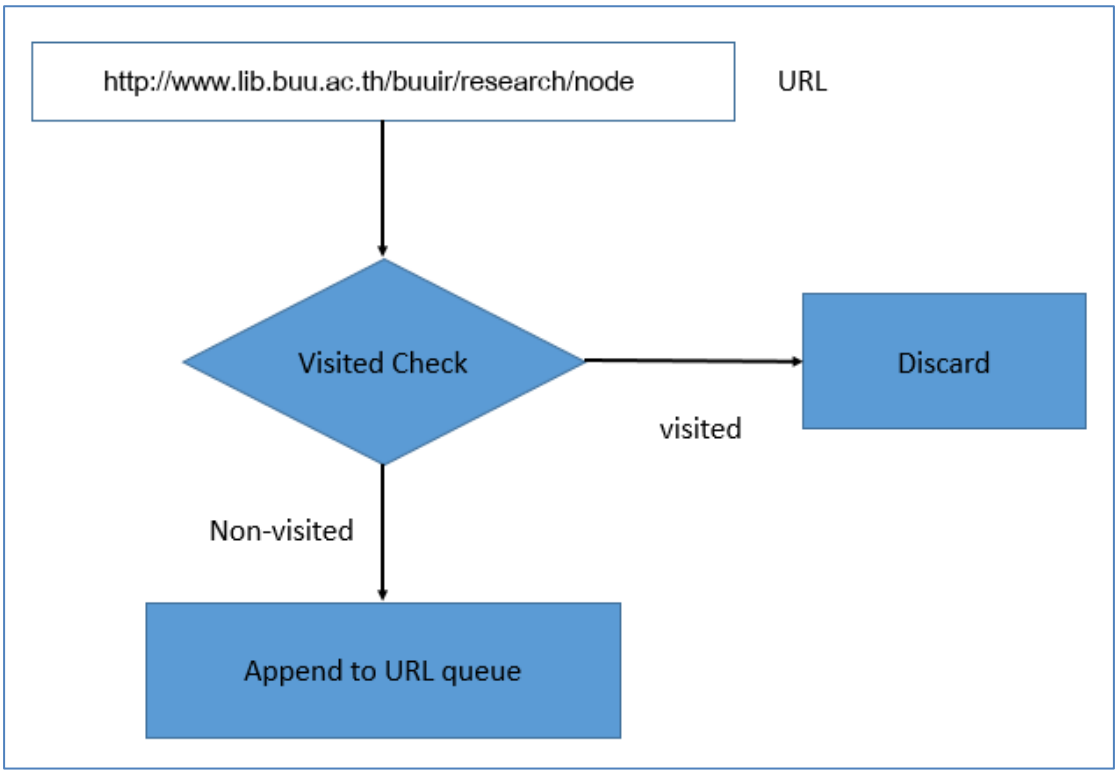


ภาพที่ 2-1 แบบจำลองพื้นฐานของเว็บครอว์เลอร์ ที่มา : กลยุทธ บพิตร. (2555) ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับ เว็บครอว์เลอร์ที่ใช้ในโปรแกรมค้นหา

2.3 แบบจำลองพื้นฐานของเว็บครอว์เลอร์

แบบจำลองพื้นฐานของเว็บครอว์เลอร์ประกอบด้วย 3 ส่วนหลัก ได้แก่ ยูอาร์แอลคิว (URL Queue) ตัวดาวน์โหลด (Downloader) และตัวพาร์สยูอาร์แอล (URL Parser) ดังภาพที่ 2-1 โดยแต่ละส่วนประกอบจะมีลักษณะดังนี้ ที่มา : กลยุทธ บพิตร. (2555) ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับ เว็บครอว์เลอร์ที่ใช้ในโปรแกรมค้นหา

1. ยูอาร์แอลคิว ทำหน้าที่เก็บยูอาร์แอลที่เว็บครอว์เลอร์พบในเว็บเพจ ซึ่งเป็นการจัดลำดับยูอาร์แอลในลักษณะเข้าก่อนออกก่อน ซึ่งยูอาร์แอลคิวมีกระบวนการทำงาน และมีการตรวจสอบว่า ยูอาร์แอลที่เข้ามาเป็นยูอาร์แอลที่เคยเก็บมาก่อนหรือไม่ หากเคยเก็บมาแล้วยูอาร์แอลนั้นจะไม่ถูกนำมาเก็บในยูอาร์แอลคิว เพื่อป้องกันความซ้ำซ้อนในการรวบรวมเว็บเพจ ดังภาพที่ 2-2



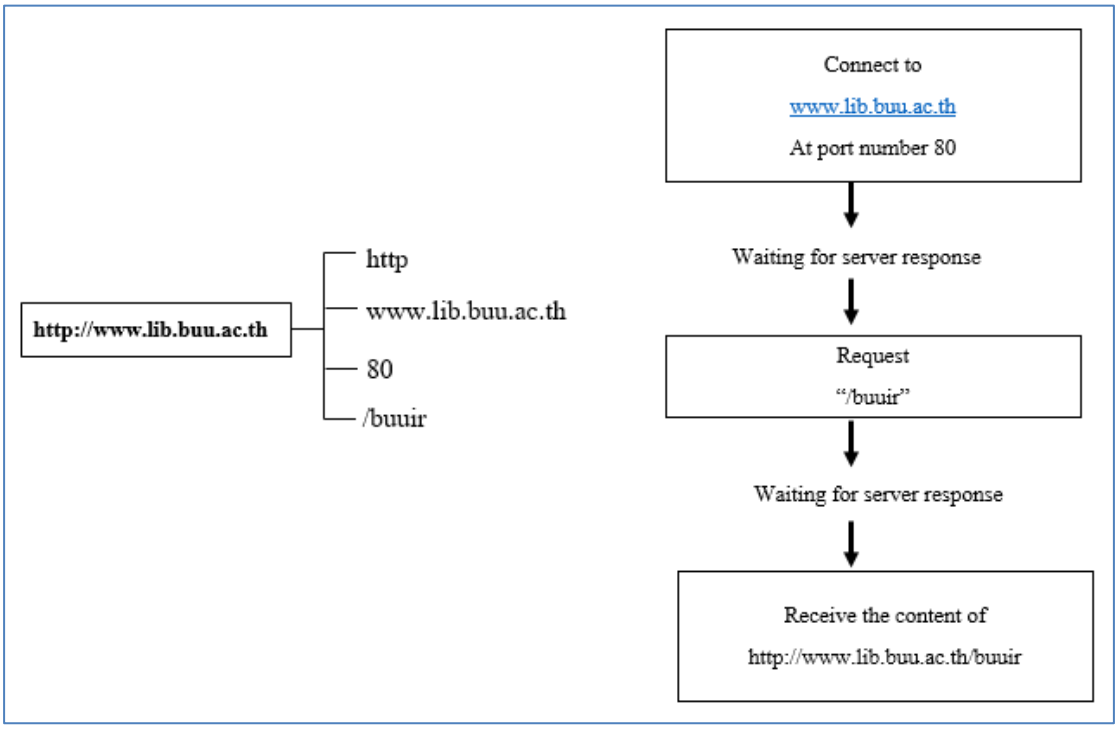
ภาพที่ 2-2 กระบวนการตรวจสอบยูอาร์แอลก่อนจัดเก็บในยูอาร์แอลคิว ที่มา : กลยุทธ บพิตร. (2555) ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับเว็บครอว์เลอร์ที่ใช้ในโปรแกรมค้นหา

2. ตัวดาวน์โหลด ทำหน้าที่เก็บเว็บเพจจากอินเทอร์เน็ต โดยตัวดาวน์โหลดจะดึงส่วนของยูอาร์แอลจากยูอาร์แอลคิว ซึ่งยูอาร์แอลจะถูกแบ่งออกเป็น 4 ส่วน ได้แก่ โปรโตคอล (Protocol) เซิร์ฟเวอร์ (Server) พอร์ต (Port) และพาธ (Path) แสดงดังตารางที่ 2-1

ตารางที่ 2-1 ตารางตัวอย่างการแยกส่วนประกอบของยูอาร์แอล

ยูอาร์แอล	โปรโตคอล	เซิร์ฟเวอร์	พอร์ต	พาธ
http://www.lib.buu.ac.th	http	www.lib.buu.ac.th	80	/buuir
http://www.tnrr.in.th	http	www.tnrr.in.th	80	/index
http://tdc.thailis.or.th	http	tdc.thailis.or.th	80	/tdc

การแยกส่วนประกอบของยูอาร์แอลจากตารางที่ 1 โปรโตคอลจะช่วยแยก และเลือกประเภทของข้อมูลได้ เช่น โปรโตคอล http จะเป็นข้อมูลเว็บเพจ เป็นต้น ส่วนของเซิร์ฟเวอร์ และพอร์ตทำให้เว็บครอว์เลอร์ทราบว่าจะร้องขอเว็บเพจนี้จากเว็บเซิร์ฟเวอร์ใด และที่พอร์ตหมายเลขใด และส่วนพาธเป็นการอ้างอิงถึงเว็บเพจที่ต้องการ แสดงดังภาพที่ 2-3



ภาพที่ 2-3 การร้องขอเว็บเพจจากเซิร์ฟเวอร์

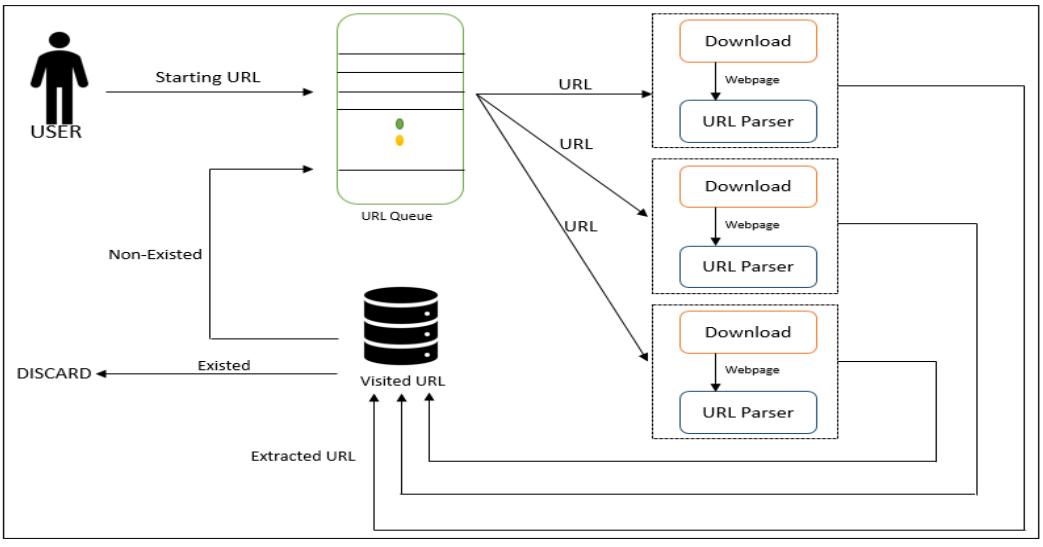
2. ตัวพาร์สเซอร์ยูอาร์แอล ทำหน้าที่แยกส่วนของยูอาร์แอลที่พบในเว็บเพจออกมาโดย ยูอาร์แอลที่ปรากฏในเว็บเพจจะถูกกำกับด้วยแท็ก HTML ดังแสดงตารางที่ 2-2

ตารางที่ 2-2 ตารางแท็กในภาษา HTML ที่กำกับยูอาร์แอล

แท็ก	**ตัวอย่างการใช้แท็ก**
a	BUU
img	<img src="http://www.lib.buu.ac.th/buuir/research/themes/bartik/logo.png">
h2	User login

การทำงานของเว็บครอว์เลอร์เบื้องต้นเพื่อให้เข้าใจง่ายยิ่งขึ้น ซึ่งผู้จัดทำงานนิพนธ์ขอเรียกว่า การทำงานแบบเดี่ยว (Sequential Processing) เนื่องจากปัจจุบันจำนวนเว็บเพจมีปริมาณเพิ่มมากขึ้น และมีแนวโน้มเพิ่มสูงขึ้นเป็นจำนวนพันล้านต่อปี จึงทำให้มีการใช้เว็บครอว์เลอร์แบบขนานเพื่อลดเวลาในการเก็บรวบรวมเว็บเพจจากอินเทอร์เน็ต

การทำงานเว็บครอว์เลอร์แบบขนานช่วยเพิ่มประสิทธิภาพในการเก็บรวบรวมเว็บเพจซึ่งส่วนของตัวดาวน์โหลด และตัวพาร์สเซอร์จะถูกกำหนดให้มีมากกว่าหนึ่งหน่วย โดยที่ตัวดาวน์-โหลดและตัวยูอาร์แอลพาร์สเซอร์จะทำงานขนานกันไปอย่างเป็นอิสระต่อกัน แต่ยังคงใช้ยูอาร์แอลคิวเดียวกัน เพื่อลดความซ้ำซ้อนของการทำงาน ยูอาร์แอลคิวจะทำหน้าที่แจกจ่ายงานและควบคุมทิศทางของงาน ดังภาพที่ 2-4



ภาพที่ 2-4 การทำงานเว็บครอว์เลอร์แบบขนาน

2.4 การควบคุมลำดับของการเก็บเว็บเพจด้วยยูอาร์แอลคิว

ยูอาร์แอลคิวมีส่วนสำคัญในการควบคุมลำดับการเข้าออกของยูอาร์แอล ซึ่งหากมีการปรับปรุงยูอาร์แอลคิวให้เหมาะสมจะช่วยให้เว็บครอเลอร์ทำงานได้อย่างมีประสิทธิภาพยูอาร์แอลคิวที่ใช้ในเว็บครอว์เลอร์มี 2 แบบ ได้แก่

1. ยูอาร์แอลคิวแบบเข้าก่อนออกก่อน (FIFO Queue)

เป็นคิวปกติ ซึ่งยูอาร์แอลที่เข้ามาก่อนได้ออกก่อนตามลำดับ ในการสร้างยูอาร์แอลคิวแบบนี้สามารถใช้โครงสร้างข้อมูลแบบอาร์เรย์ (Array) และแบบลิงค์ลิสต์ (Linked-list)

2. ยูอาร์แอลคิวแบบเข้าออกตามค่าความสำคัญ (Priority Queue)

เป็นการจัดลำดับความสำคัญของยูอาร์แอล โดยควบคุมให้ยูอาร์แอลที่มีความสำคัญมากที่สุดไปก่อน ซึ่งยูอาร์แอลประเภทนี้มักใช้ในเว็บครอเลอร์ที่จัดลำดับความสำคัญในการเก็บเว็บเพจ ยกตัวอย่างเช่น จัดลำดับความสำคัญตามความเร็วในการเก็บข้อมูล จัดลำดับความสำคัญตามเว็บเพจที่มีการเปลี่ยนแปลงบ่อย หรือ ความสำคัญจากคุณภาพเนื้อหาของเว็บเพจ เป็นต้น

2.5 การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่อง

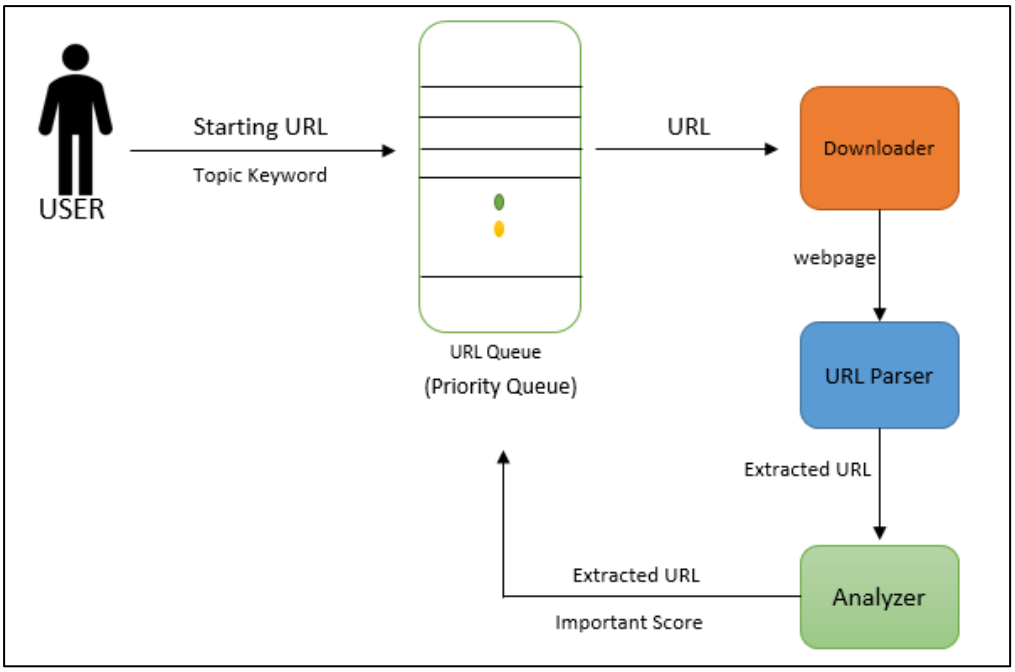
เสิร์จเอ็นจินที่มีชื่อเสียง เช่น Yahoo, Google และ Bing เป็นเสิร์จเอ็นจินที่รวบรวมเว็บเพจจากอินเทอร์เน็ตเพื่อให้บริการสืบค้นแก่ผู้ใช้ ซึ่งเก็บรวบรวมเว็บเพจให้ได้ปริมาณมากที่สุดเพื่อให้ได้ผลลัพธ์ที่ใกล้เคียงกับการสืบค้น และครอบคลุมทั่วทั้งอินเทอร์เน็ต อย่างไรก็ตามหากจะคาดการณ์จำนวนเว็บเพจที่มีอยู่ในอินเทอร์เน็ต และระยะเวลาที่ใช้ในการเก็บรวบรวมเว็บเพจเหล่านั้น เป็นสิ่งที่ทำได้ยาก ซึ่งมีเว็บเพจจำนวนมากมายที่รวบรวมมาได้ไม่มีประโยชน์ในการทำดัชนี เนื่องจากเว็บเพจมีเนื้อความไม่สมบูรณ์ ซึ่งปัญหาเหล่านี้ทำให้เกิดแนวคิดการเก็บรวบรวมเว็บเพจให้ตรงกับเป้าหมายที่ต้องการ โดยกำหนดหัวเรื่องที่ต้องการ และเก็บรวบรวมเฉพาะเว็บเพจที่กล่าวถึงเรื่องนั้น เว็บเพจที่รวบรวมได้จะมีประโยชน์ในการทำดัชนีหรือการสืบค้นมากยิ่งขึ้น ซึ่งเรียกว่า การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่อง

แนวคิดนี้ยังคงใช้เว็บครอว์เลอร์เป็นเครื่องมือในการเก็บรวบรวม ซึ่งเว็บครอว์เลอร์นี้ต้องมีความสามารถในการเลือกเก็บเฉพาะเว็บเพจที่มีเนื้อหาเกี่ยวข้องกับหัวเรื่องเท่านั้น ดังนั้นเว็บเพจที่สามารถเก็บได้อาจมีจำนวนไม่มากนักแต่ในส่วนของเนื้อหาภายในเว็บเพจกลับมีคุณภาพสูง เว็บครอว์เลอร์ประเภทนี้เรียกว่า การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่อง (Topic-specific web crawler) ซึ่งในหัวข้อนี้จะกล่าวถึงรูปแบบการทำงานของเว็บครอว์เลอร์ประเภทนี้ รวมไปถึงส่วนประกอบที่สำคัญ และทฤษฎีที่นำมาประยุกต์ใช้เพื่อทำให้เว็บครอว์เลอร์มีความสามารถในการเลือกเก็บเว็บเพจ

2.5.1 หลักการทำงานการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่อง

เว็บครอว์เลอร์แบบเฉพาะเจาะจงหัวเรื่องมีหลักการทำงานคล้ายคลึงกับ เว็บครอว์เลอร์แบบธรรมดาทั่วไป เนื่องจากการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องเพื่อต้องการได้เว็บเพจที่ตรงกับคำค้นที่เราต้องการ เพราะฉะนั้นผู้ใช้เว็บครอว์เลอร์ไม่เพียงระบุยูอาร์แอลเริ่มต้นให้กับเว็บครอว์เลอร์เท่านั้น แต่จำเป็นต้องกำหนดคำสำคัญของหัวเรื่อง (Topic keyword) ที่เกี่ยวข้องกับหัวเรื่องนั้นด้วย ซึ่งหัวเรื่อง และคำสำคัญจำเป็นต่อเทคนิคนี้อย่างมาก เนื่องจากเว็บครอว์เลอร์จะอาศัยเทคนิคการเปรียบเทียบคำสำคัญกับเนื้อหาภายในเว็บเพจ และวิเคราะห์ว่าเว็บเพจที่พบมีความเป็นไปได้มากน้อยเพียงใดที่จะเกี่ยวข้องหรือคล้ายคลึงกับหัวเรื่องที่กำหนด ซึ่งความแม่นยำในการวิเคราะห์ส่งผลต่อยูอาร์แอลใหม่ที่พบในเว็บเพจนั้นด้วย หากเนื้อหาในเว็บเพจนั้นมีความเกี่ยวข้องสูง จึงมีความเป็นไปได้ที่ยูอาร์แอลใหม่ในเว็บเพจนั้น จะมีความเกี่ยวข้องเช่นกัน ในทางกลับกันเนื้อหาที่ไม่เกี่ยวข้องกับหัวเรื่อง ยูอาร์แอลใหม่ภายในเว็บเพจจะถูกลดความน่าเชื่อถือที่จะเป็นเว็บเพจที่เกี่ยวข้อง

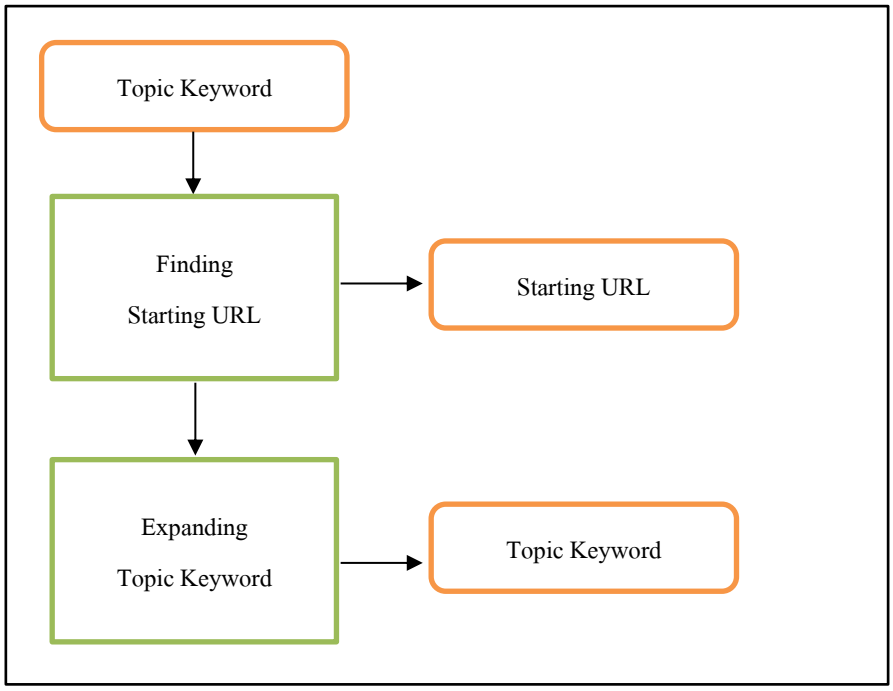
เว็บครอว์เลอร์แบบเฉพาะเจาะหัวเรื่องมีส่วนการทำงานที่เพิ่มเติมขึ้นมา ได้แก่ ตัววิเคราะห์ (Analyzer) ทำหน้าที่วิเคราะห์ความเกี่ยวข้องของเว็บเพจที่เก็บได้กับหัวเรื่องที่ต้องการมีมากน้อยเพียงใด และวิเคราะห์ว่ายูอาร์แอลที่พบในเว็บเพจนั้นมีความเกี่ยวข้องหรือไม่ ซึ่งผลลัพธ์จากการวิเคราะห์ได้จะเป็นตัวเลขที่เรียกว่า ค่าความสำคัญ (Important Score) และยูอาร์แอลคิวที่ใช้ในเว็บครอว์เลอร์แบบเฉพาะเจาะจงหัวเรื่องเป็นการเข้าออกตามลำดับความสำคัญ เมื่อตัววิเคราะห์ส่งให้ยูอาร์แอลที่พบใหม่มาพร้อมกับค่าความสำคัญ ยูอาร์แอลคิวจะเรียงลำดับการเข้าออกของยูอาร์แอลตามค่าความสำคัญ เนื่องจากค่าความสำคัญบ่งบอกถึงความเกี่ยวข้องกับหัวเรื่อง ดังนั้นยูอาร์แอลที่เกี่ยวข้องมากที่สุดจะถูกจัดให้อยู่ลำดับก่อนยูอาร์แอลที่มีค่าความสำคัญน้อยกว่า ดังแสดงในภาพที่ 2-5



ภาพที่ 2-5 หลักการทำงานเบื้องต้นเว็บครอว์เลอร์แบบเฉพาะเจาะจงหัวเรื่อง

หลักการทำงานการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องแบ่งขั้นตอนย่อยได้ ดังนี้

1. ยูอาร์แอลเริ่มต้นและคำสำคัญของหัวเรื่อง ยูอาร์แอลเริ่มต้นและคำสำคัญของหัวเรื่องเป็นสิ่งที่ช่วยให้เว็บครอว์เลอร์สามารถเก็บรวบรวมเว็บเพจแบบเฉพาะเจาะจงได้ ซึ่งคำสำคัญของหัวเรื่องเป็นคำที่บ่งบอกหัวเรื่องและมักพบคำเหล่านั้นในหัวเรื่องนั้น ยูอาร์แอลเริ่มต้นที่เกี่ยวข้องกับหัวเรื่องจะนำทางให้เว็บครอว์เลอร์ไปพบกับเว็บครอว์เลอร์ที่ตรงกับหัวเรื่องที่กำหนด ในทางกลับกันหากยูอาร์แอลเริ่มต้นไม่มีความเกี่ยวข้องกับหัวเรื่องที่กำหนดเลย เว็บครอว์เลอร์อาจไม่สามารถเก็บเว็บเพจที่เกี่ยวข้องกับหัวเรื่องได้ คำสำคัญของหัวเรื่องที่ถูกกล่าวถึงมากที่สุดในเว็บเพจนั้นย่อมมีประโยชน์ในการค้นหาเช่นกัน ซึ่งขั้นตอนในการหายูอาร์แอลเริ่มต้นและคำสำคัญของหัวเรื่อง แสดงได้ดังภาพที่ 2-6



ภาพที่ 2-6 ขั้นตอนการหายูอาร์แอลเริ่มต้นและคำสำคัญของหัวเรื่อง

จากภาพที่ 2-6 จะเห็นว่ามีขั้นตอนหลัก 2 ขั้นตอน โดยเริ่มจากการกำหนดคำสำคัญของหัวเรื่อง จากนั้นหายูอาร์แอลเริ่มต้น หลังจากนั้นจึงหาคำสำคัญของหัวเรื่องเพิ่มเติมจากคำสำคัญของหัวเรื่องที่กำหนดไว้ก่อนหน้า ซึ่งมีขั้นตอนย่อย ดังต่อไปนี้

1.1 การหายูอาร์แอลเริ่มต้น มีเทคนิคทั้งหมด 3 วิธี ได้แก่

1.1.1 ผู้ใช้กำหนดเอง (User Defined) วิธีนี้มักพบในเว็บครอว์เลอร์ที่มีระบบใหญ่ ซึ่งผู้ใช้มีอิสระในการกำหนดยูอาร์แอลเริ่มต้นเอง การให้ผู้ใช้สามารถกำหนดยูอาร์แอลเริ่มต้นเองเป็นสิ่งดี เนื่องจาก ผู้ใช้ที่มีความเชี่ยวชาญในหัวเรื่องนั้น อาจเป็นประโยชน์ต่อการได้ยูอาร์แอลเริ่มต้นที่ดีที่สุด

1.1.2 การใช้เสิร์จเอ็นจิน (Search Engine) วิธีนี้เป็นการหายูอาร์แอลเริ่มต้นแบบอัตโนมัติโดยการนำคำสำคัญของหัวเรื่องที่กำหนดให้ไปสืบค้นในเสิร์จเอ็นจิน จากนั้นนำผลลัพธ์ที่ได้มาเป็นยูอาร์แอลเริ่มต้น

1.1.3 ไดเรกทอรีสร้างโดยผู้เชี่ยวชาญ (Expert human-edited Directory) ไดเรกทอรี คือ กลุ่มของเว็บเพจที่ถูกจัดหมวดหมู่ตามหัวเรื่องที่กำหนด ภายในหัวเรื่องหนึ่งประกอบไปด้วยเว็บเพจที่กล่าวถึงเรื่องนั้น วิธีนี้เป็นการใช้เว็บเพจที่ถูกจัดหมวดหมู่แล้วมาเป็น

ยูอาร์แอลเริ่มต้นให้แก่เว็บครอว์เลอร์ ซึ่งสามารถสร้างได้ 2 แบบ ได้แก่ แบบระบบอัตโนมัติและแบบผู้เชี่ยวชาญ (มนุษย์)

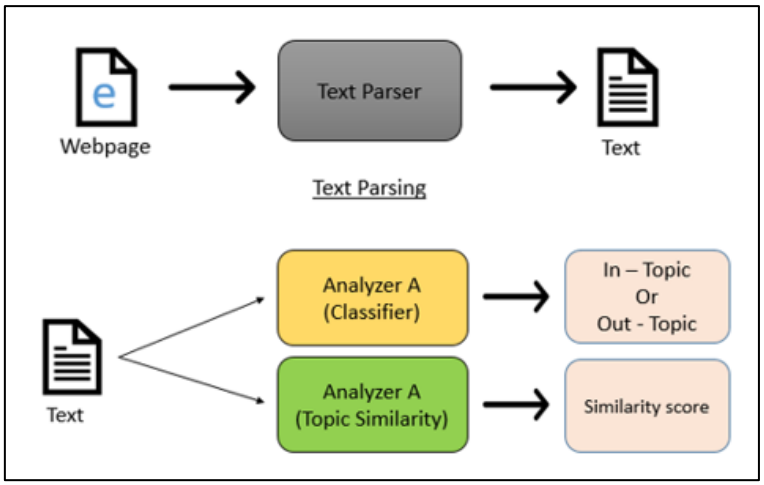
2. การหาคำสำคัญของหัวเรื่องเพิ่มเติม แบ่งออกเป็น 2 วิธี ดังนี้

2.1 เพิ่มเติมจากผลลัพธ์การสืบค้น (Expanding from search result) หลังจากการหายูอาร์แอลเริ่มต้นด้วยวิธีสืบค้นจากเสิร์จเอ็นจินแล้ว นำคำสำคัญที่กำกับแต่ละยูอาร์แอลที่ได้จากสืบค้นมาเป็นคำสำคัญของหัวเรื่องเพิ่มเติม เนื่องจากคำสำคัญเพิ่มเติมเหล่านี้เป็นคำที่อยู่ในตำแหน่งใกล้เคียงกับคำสำคัญที่กำหนดไว้เริ่มแรก ซึ่งเป็นไปได้มากที่จะเป็นคำสำคัญของหัวเรื่องนั้นเช่นกัน

2.2 เพิ่มเติมจากไดเรกทอรี (Expanding from directory) วิธีนี้เกิดขึ้นหลังจากการหายูอาร์เริ่มต้นจากไดเรกทอรีที่สร้างโดยมนุษย์ผู้เชี่ยวชาญ โดยคำสำคัญที่ใกล้เคียงยูอาร์แอลเริ่มต้นในไดเรกทอรีเหล่านั้น จะถูกนำมาใช้เป็นคำสำคัญของหัวเรื่องเพิ่มเติม เนื่องจากคำสำคัญเหล่านี้ถูกเขียนโดยผู้เชี่ยวชาญในหัวเรื่องนั้น ๆ ซึ่งเป็นไปได้ที่คำสำคัญเหล่านี้มีความเกี่ยวข้องกับเรื่องที่สนใจ

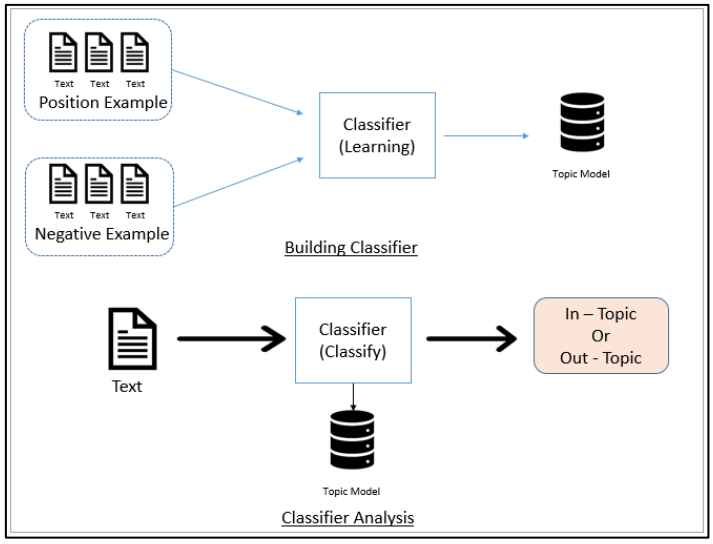
3. การวิเคราะห์เว็บเพจ ช่วยให้ทราบว่าเว็บเพจนี้มีความเกี่ยวข้องกับหัวเรื่องที่กำหนดมากน้อยเพียงใด และยังมีประโยชน์ในการกำหนดความสำคัญให้กับยูอาร์แอลภายในเว็บเพจด้วย หากเว็บเพจมีความเกี่ยวข้องจะส่งผลให้ยูอาร์แอลภายในเว็บเพจมีความเป็นไปได้สูงที่จะเป็นยูอาร์แอลที่เกี่ยวข้องเช่นเดียวกัน หรือในทางกลับกัน ยูอาร์แอลที่พบในเว็บเพจที่ไม่เกี่ยวข้องย่อมมีความน่าจะเป็นน้อยกว่าที่จะมีความเกี่ยวข้องกับหัวเรื่อง

การวิเคราะห์เว็บเพจเริ่มต้นจากการนำเว็บเพจส่งให้ตัวพาร์สเซอร์ข้อความ (Text Parser) ซึ่งมีหน้าที่แยกส่วนเฉพาะเนื้อความที่อยู่ในเว็บเพจออกมา จากนั้นจึงนำส่วนเนื้อความมาวิเคราะห์ ดังภาพที่ 2-7 ซึ่งวิธีการวิเคราะห์แบ่งออกเป็น 2 วิธี คือการใช้ตัวจัดหมวดหมู่ (Classifier) และการวิเคราะห์ความคล้ายคลึงของหัวเรื่อง (Topic Similarity)



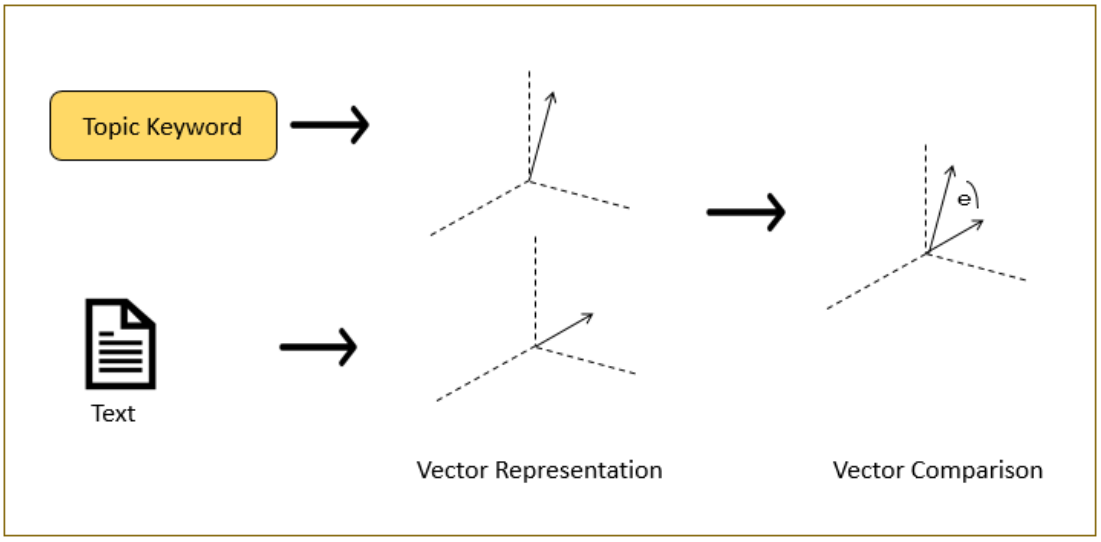
ภาพที่ 2-7 ขั้นตอนการวิเคราะห์เว็บเพจ

3.1 การวิเคราะห์ด้วยตัวจัดหมวดหมู่ (Classifier) วิธีนี้เป็นการค้นหาคำตอบว่าเว็บเพจนั้นถูกจัดอยู่ในหัวเรื่องนั้นหรือไม่ ซึ่งมีหลักการในการค้นหาลักษณะพิเศษในแต่ละหมวดหมู่เพื่อให้การจัดหมวดหมู่มีความถูกต้องสูง โดยตัวจัดหมวดหมู่จะเรียนรู้เว็บเพจตัวอย่างเกี่ยวกับหัวเรื่องที่ต้องการจัดหมวดหมู่ 2 แบบ คือ ตัวอย่างที่ถูก (Positive example) และตัวอย่างที่ผิด (Negative example) เพื่อให้ได้แบบจำลองของหัวเรื่อง (Topic model) เมื่อพบเว็บเพจใด ๆ จะนำเว็บเพจตรวจสอบกับแบบจำลองของหัวเรื่อง ผลลัพธ์ที่ได้คือ เว็บเพจจัดอยู่ในหมวดหมู่นี้หรือไม่ ดังแสดงในภาพที่ 2-8



ภาพที่ 2-8 การวิเคราะห์เว็บเพจด้วยตัวจัดหมวดหมู่

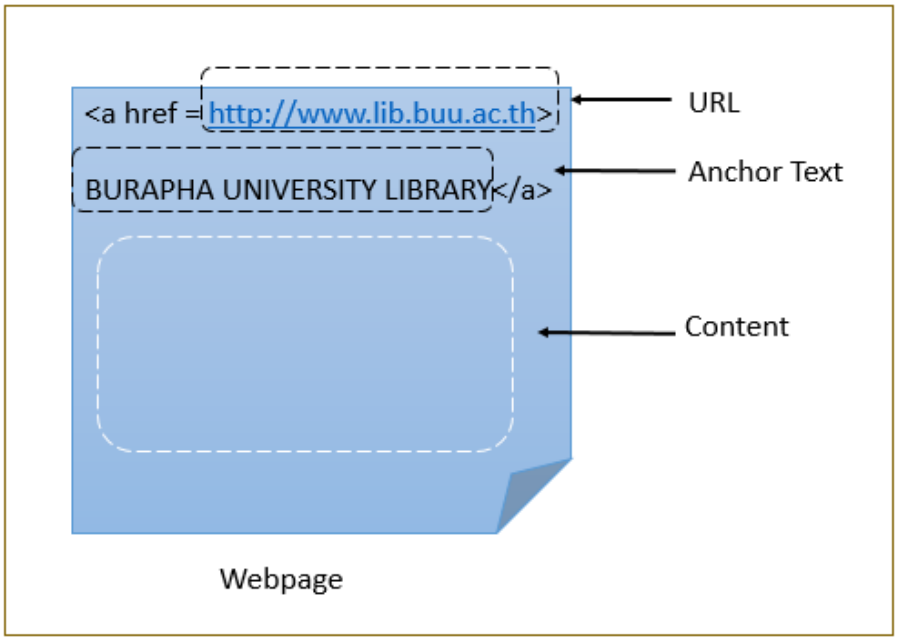
3.2 การวิเคราะห์ความคล้ายคลึงของหัวเรื่อง (Topic Similarity) เป็นวิธีการเปรียบเทียบความคล้ายคลึงกันระหว่างคำสำคัญของหัวเรื่องและเนื้อหาของเว็บเพจ โดยใช้ทฤษฎีแบบจำลองเวคเตอร์สเปซ (Vector Space Model) หลักการคือ ปรับคำสำคัญของหัวเรื่องและเนื้อหาของเว็บเพจให้อยู่ในรูปแบบเวคเตอร์ และนำเวคเตอร์ทั้งสองมาเปรียบเทียบทิศทาง หากมีทิศทางเดียวกันหรือใกล้เคียงกัน จะแสดงว่า คำสำคัญของหัวเรื่องและเนื้อหาของเว็บเพจมีความคล้ายคลึงกัน ดังภาพที่ 2-9



ภาพที่ 2-9 การวิเคราะห์เว็บเพจด้วยความคล้ายคลึงของหัวเรื่อง

4. ค่าความสำคัญของยูอาร์แอล

เมื่อเว็บครอเลอร์ได้เว็บเพจมาหนึ่งเว็บเพจ ยูอาร์แอลที่ปรากฏอยู่ในเว็บเพจนั้นจะถูกแยกออกมาเพื่อเก็บต่อไป ภาพที่ 2-10 แสดงส่วนประกอบของเว็บเพจ ยูอาร์แอลซึ่งจะถูกแยกและเก็บข้อมูลดังตารางที่ 2-1 ในแต่ละยูอาร์แอลจะมีข้อความอธิบายเนื้อความยูอาร์แอลเรียกว่า แองเคอร์เท็กซ์ (Anchor text) และข้อความอื่น ๆ ที่ไม่อยู่ภายในแท็กเรียกว่า เนื้อความ (Content)



ภาพที่ 2-10 ส่วนประกอบภายในเว็บเพจ

การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องไม่ใช่วิธีการเรียงลำดับยูอาร์แอลที่ถูกพบก่อน-หลัง เนื่องจากยูอาร์แอลที่พบก่อนอาจจะไม่มีความเกี่ยวข้องกับหัวเรื่องที่กำหนด ดังนั้นจึงต้องมีกระบวนการในการวิเคราะห์ความเกี่ยวข้องของเว็บเพจที่ถูกเก็บมา โดยพิจารณาจากส่วนประกอบภายในเว็บเพจ ดังภาพที่ 2-10 ซึ่งการวิเคราะห์จะได้ผลลัพธ์คือ ค่าความสำคัญของยูอาร์แอล ยูอาร์แอลที่มีค่าความสำคัญมากจะถูกคาดการณ์ว่ามีความเกี่ยวข้องสูงมากและจะถูกจัดเก็บก่อน วิธีการวิเคราะห์ค่าความสำคัญของยูอาร์แอล มีวิธีต่าง ๆ ดังนี้

4.1 ค่าความสำคัญจากเนื้อความของเว็บเพจ เนื้อความสามารถบ่งบอกถึงหัวเรื่องของเว็บเพจได้ จึงคาดการณ์ได้ว่า ยูอาร์แอลที่ปรากฏภายในเว็บเพจที่มีความเกี่ยวข้อง มีความเป็นไปได้สูงมากที่จะเป็นยูอาร์แอลของเว็บเพจที่มีความเกี่ยวข้องกัน

4.2 ค่าความสำคัญจากแองเคอร์เท็กซ์ แองเคอร์เท็กซ์เป็นข้อความที่อธิบายถึงเว็บเพจของยูอาร์แอล การคาดการณ์ความเกี่ยวข้องของยูอาร์แอลด้วยแองเคอร์เท็กซ์ช่วยให้เกิดความแม่นยำมายิ่งขึ้น เนื่องจากบางครั้งการคาดการณ์ด้วยเนื้อความอาจพบกับกรณีพิเศษ เช่น เนื้อความของเว็บเพจไม่เกี่ยวข้องกับหัวเรื่องที่กำหนด แต่แองเคอร์เท็กซ์ของยูอาร์แอลมีความเกี่ยวข้อง หรือเว็บเพจที่ไม่มีเนื้อความ ปรากฏเพียงยูอาร์แอลและแองเคอร์แท็กซ์ของยูอาร์แอล ดังนั้นการใช้แองเคอร์แท็กซ์จึงช่วยลดความผิดพลาดจากกรณีที่เนื้อความไม่สามารถบ่งบอกความเกี่ยวข้องได้

4.3 ค่าความสำคัญจากส่วนประกอบของยูอาร์แอล การแยกส่วนของยูอาร์แอลหรือแยกคำออกมาจากยูอาร์แอลเพื่อนำมาคำนวณค่าความสำคัญ ตัวอย่างเช่น หัวเรื่องที่กำหนดคือ Research และ http://www.lib.buu.ac.th/buuir/research/node?page=0 คือ ยูอาร์แอลที่พบเมื่อแยกส่วนของยูอาร์แอลออกมา ในส่วนพาธที่พบกับคำว่า Research ซึ่งให้ความหมายที่เกี่ยวข้องอย่างมากกับหัวเรื่องที่กำหนด

4.4 ค่าความสำคัญจากการเชื่อมโยงของยูอาร์แอล วิธีการนี้จะกำนดระยะของการเชื่อมโยง หากเกินระยะที่กำหนดไว้จะถูกลดความสำคัญลงไปเรื่อย ๆ

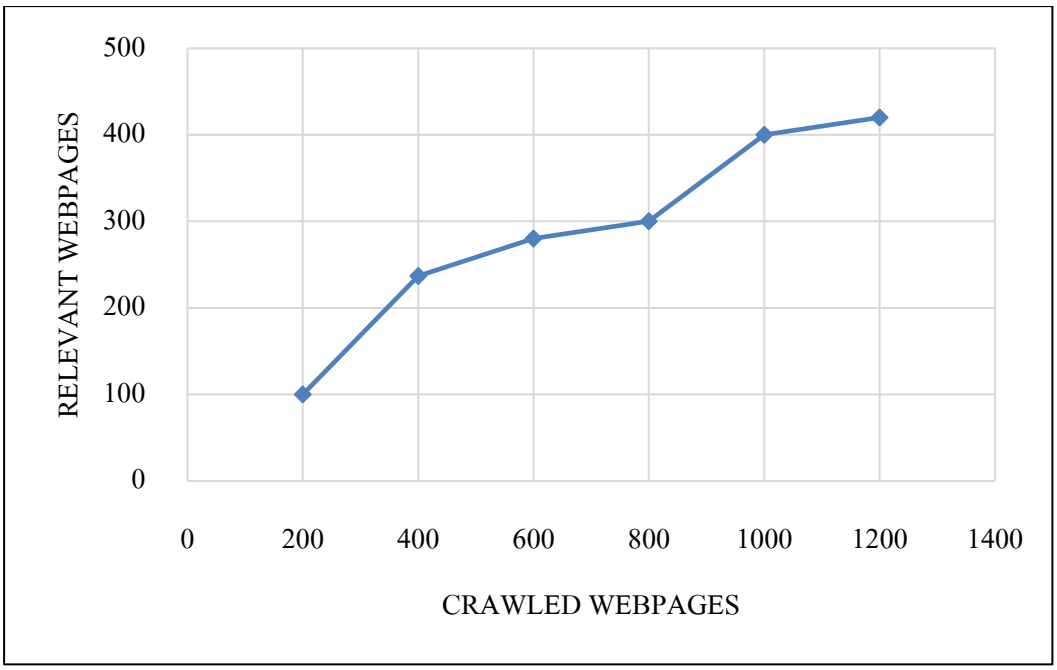
4.5 ค่าความสำคัญจากยูอาร์แอลที่มาจากเว็บเพจเดียวกัน หากยูอาร์แอลส่วนใหญ่ที่มาจากเว็บเพจเดียวกันถูกคาดการณ์ว่ามีความเกี่ยวข้องสูงกับหัวเรื่องที่กำหนด ยูอาร์แอลที่เหลือจะถูกให้ความสำคัญเพิ่มขึ้น

4.6 ค่าความสำคัญจากตำแหน่งที่ยูอาร์แอลปรากฏ การพิจารณาจากตำแหน่งการปรากฏของยูอาร์แอลช่วยให้คาดการณ์ได้ว่า ยูอาร์แอลนั้นมีความสำคัญมากน้อยแค่ไหนนักออกแบบและพัฒนาเว็บเพจจึงคำนึงถึงการวางตำแหน่งของส่วนประกอบต่าง ๆ ในเว็บเพจยูอาร์แอลที่อยู่ในตำแหน่งที่ดีจะได้รับความสำคัญสูง

4.7 ค่าความสำคัญจากการเชื่อมโยงเข้าหา_เว็บเพจต่าง ๆ ที่มีการเชื่อมโยงกันโดยมียูอาร์แอลเป็นตัวอ้างอิง ซึ่งเว็บเพจที่ถูกเว็บเพจอื่น ๆ อ้างอิงถึงเป็นจำนวนมากเป็นเว็บเพจที่ได้รับความนิยม ความสำคัญของยูอาร์แอลสำหรับวิธีนี้คือ การให้ค่าความสำคัญสูงแก่ยูอาร์แอลที่ได้รับความนิยมสูง

5. การวัดผลของการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่อง

การวัดผลของการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องมีเป้าหมายเพื่อเก็บรวบรวมเว็บเพจที่เกี่ยวข้องกับหัวเรื่องที่กำหนดให้ได้มากที่สุด โดยเฉพาะช่วงเริ่มต้นควรเก็บเว็บเพจที่เกี่ยวข้องให้ได้ปริมาณมาก ดังภาพที่ 2-11



ภาพที่ 2-11 ตัวอย่างกราฟวัดผลการเก็บเว็บเพจแบบเฉพาะเจาะจง

จากภาพที่ 2-11 แกน X คือจำนวนเว็บเพจที่เว็บครอว์เลอร์สามารถเก็บรวบรวมได้ตามช่วงเวลา แกน Y คือจำนวนเว็บเพจที่เกี่ยวข้องกับหัวเรื่องที่กำหนดตามช่วงเวลาที่เก็บเว็บเพจได้สามารถอ่านได้ว่า เมื่อเว็บครอว์เลอร์เก็บเว็บเพจได้ 800 เว็บเพจ ภายใน 800 เว็บเพจนั้นมีเว็บเพจที่เกี่ยวข้องกับหัวเรื่องที่กำหนดอยู่ 300 เว็บเพจ คิดเปอร์เซ็นต์ความสำเร็จในการพบเว็บเพจที่เกี่ยวข้องกับหัวเรื่อง 50 เปอร์เซนต์ เป็นต้น ซึ่งเว็บครอว์เลอร์แบบเฉพาะเจาะจงหัวเรื่องที่ดีควรมีแนวโน้มของกราฟดังภาพที่ 2-11 เพราะช่วงแรกเว็บครอว์เลอร์สามารถเก็บเว็บเพจที่เกี่ยวข้องได้จำนวนมาก

ซึ่งเป็นข้อดีกรณีที่มีพื้นที่การเก็บข้อมูลมีจำนวนจำกัด แต่เว็บครอว์เลอร์ สามารถเก็บเว็บเพจที่เกี่ยวข้องได้จำนวนเพียงพอแล้วในช่วงเริ่มต้น ดังนั้นจึงทำให้ใช้ทรัพยากรในการจัดเก็บอย่างมีประสิทธิภาพ

อย่างไรก็ตามการทำให้เว็บครอเลอร์สามารถเก็บรวบรวมเว็บเพจได้อย่างมีประสิทธิภาพขึ้นนั้นอยู่กับกระบวนการของเว็บครอว์เลอร์ ซึ่งในหัวข้อนี้จะกล่าวถึงการนำผลลัพธ์จากการเก็บเว็บเพจมาหาค่าความแม่นยำในการเก็บเว็บเพจที่เกี่ยวข้อง ซึ่งสามารถแบ่งวิธีการวัดความเกี่ยวข้องของเว็บเพจได้ 2 วิธี ดังต่อไปนี้

1. การวัดด้วยเทคนิคทางคอมพิวเตอร์แบ่งออกเป็น 3 วิธีการ ดังต่อไปนี้

1.1 เทคนิคการจัดหมวดหมู่ (Classifier Technique) การนำเอาระบบการจัดหมวดหมู่มาคำนวณหาความเกี่ยวข้องของเว็บเพจที่รวบรวมมาได้ ซึ่งเว็บเพจที่ถูกจัดอยู่ในหัวเรื่องที่กำหนดจะถือว่าเป็นเว็บเพจที่เกี่ยวข้อง ซึ่งต้องมีการสอนระบบจัดหมวดหมู่ด้วยการให้ตัวอย่างที่ถูก (Positive Example) และตัวอย่างที่ผิด (Negative Example) ให้ระบบจัดหมวดหมู่เรียนรู้เพื่อให้การจัดหมวดหมู่ได้ผลลัพธ์ที่แม่นยำมากที่สุด

1.2 เทคนิคระบบสืบค้น (Retrieval System Technique) เป็นการนำเอาระบบสืบค้นมาช่วยในการวัดผล เช่นระบบสืบค้นชื่อว่า SMART มาใช้ โดยนำเว็บเพจที่เก็บได้ไปทดสอบกับระบบ SMART ซึ่งจะทำหน้าที่สืบค้นหัวเรื่องที่กำหนดจากเว็บเพจเหล่านั้น และจัดลำดับผลลัพธ์ออกมา ซึ่งวัดผลด้วยการคำนวณจากอันดับเปรียบเทียบกับจำนวนเว็บเพจที่เก็บได้ตามช่วงเวลานั้น การวัดผลด้วยระบบสืบค้นจะบอกถึงอันดับความสำคัญของเว็บเพจที่เก็บได้นั้นมากน้อยเพียงใด

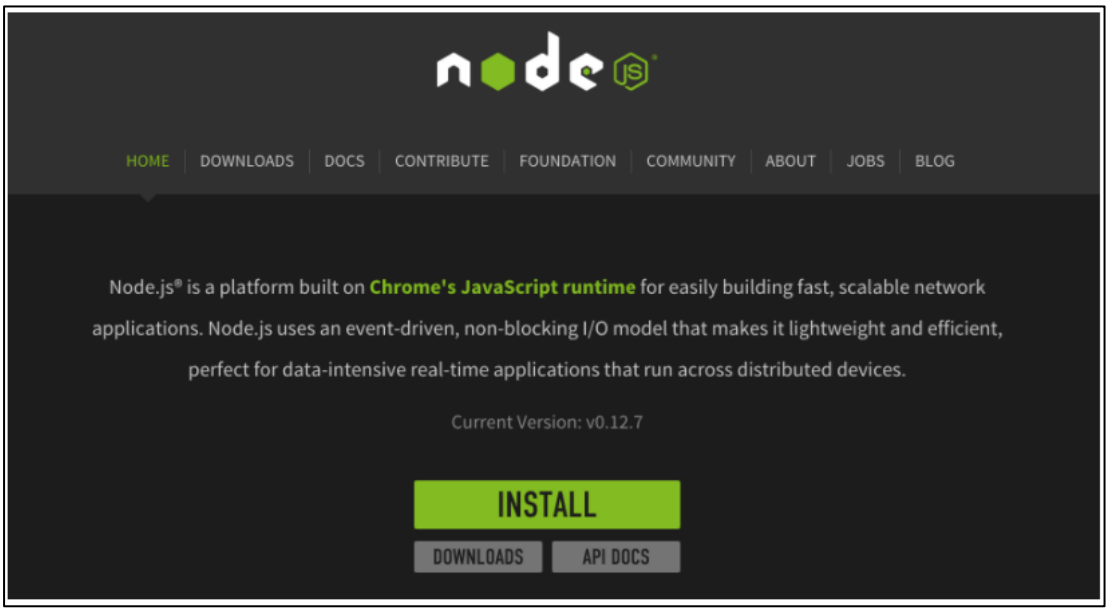
1.3 เทคนิคความคล้ายคลึงของหัวเรื่อง (Topic Similarity Technique) เป็นการคำนวณค่าความคล้ายคลึงระหว่างหัวเรื่องและเว็บเพจที่เก็บมาได้ โดยใช้ทฤษฎีของแบบจำลองเวคเตอร์สเปซ ผลลัพธ์ที่ได้จะแสดงให้เห็นถึงเว็บเพจที่เก็บได้ในแต่ละช่วงเวลามีความเกี่ยวข้องกับหัวเรื่องที่กำหนดหรือไม่

2. การวัดด้วยมนุษย์ผู้เชี่ยวชาญ วิธีนี้จะทำให้ได้คำตอบที่แม่นยำมากที่สุด ผู้เชี่ยวชาญที่กล่าวถึงต้องเป็นผู้ที่สมัครใจที่จะช่วยวัดผลด้วย โดยเว็บครอว์เลอร์เก็บรวบรวมเอกสารงานวิจัยและให้ผู้เชี่ยวชาญตรวจสอบว่าข้อมูลที่เก็บมาเป็นเอกสารงานวิจัยหรือไม่ จะเห็นได้ว่าการใช้ผู้เชี่ยวชาญทำให้การตรวจสอบที่แม่นยำ

2.6 ความรู้เบื้องต้นเกี่ยวกับ Node.js

Node.js คือ Cross Platform Runtime Environment เขียนด้วย JavaScript สำหรับการทำงานฝั่ง Server หรือใช้สำหรับเป็น Web Server และมีลักษณะเป็น Open Source

Node.js มีความเร็วของการประมวลผล จึงทำให้ Application ที่เขียนด้วย Node.js นั้นมีจำนวนเพิ่มขึ้นอย่างรวดเร็ว ซึ่งรวมไปถึง Application ที่จะช่วยให้การพัฒนาเว็บไซต์เป็นไปอย่างราบรื่นมากขึ้นด้วย

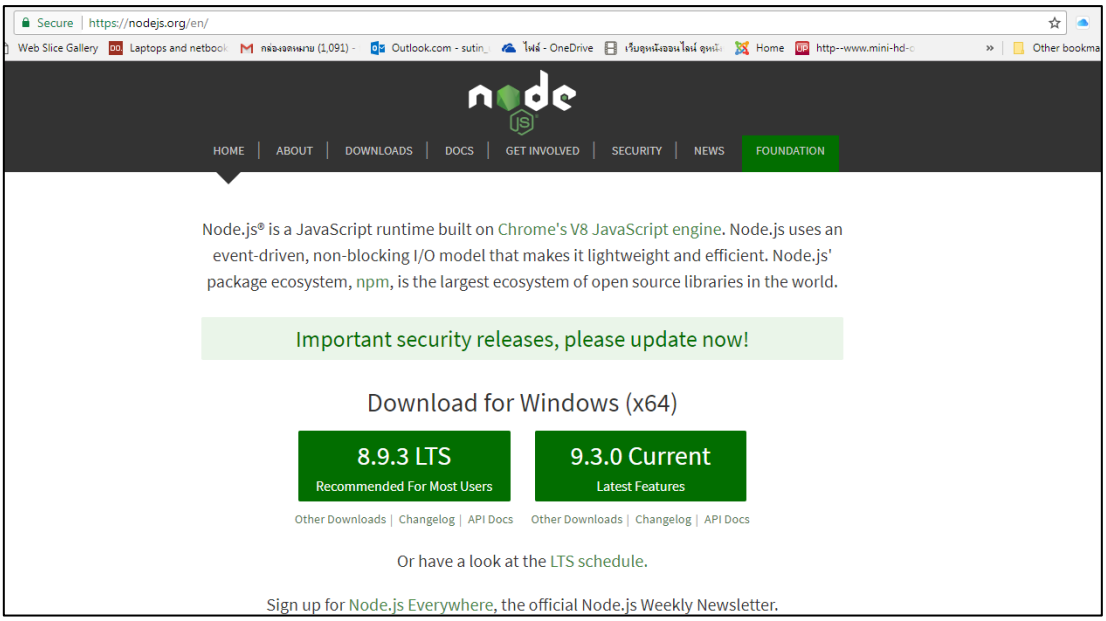


ภาพที่ 2-12 ตัวอย่างเว็บไซต์ของ Node.js ที่มา : http://www.siamhtml.com/introduction-to-node-js/

2.6.1 ขั้นตอนการติดตั้ง Node.js และการใช้คำสั่ง npm

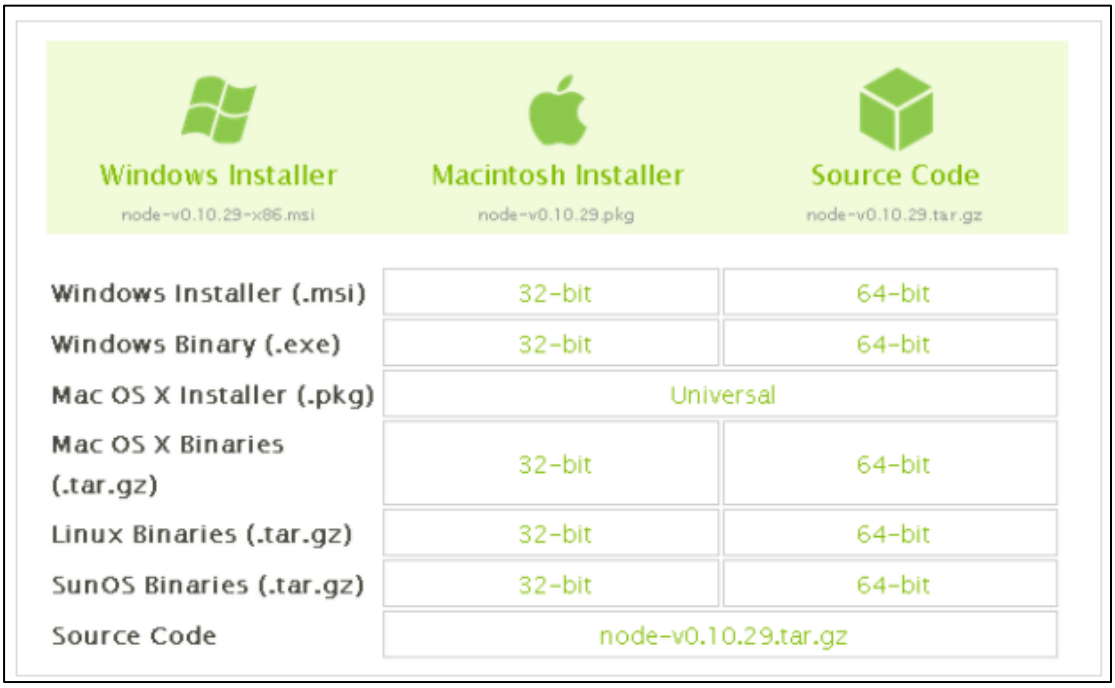
1. เข้าเว็บไซต์ https://nodejs.org/en/

2. ดาวน์โหลดตัว Installer จะต้องทำการเลือกตัว Installer ที่เหมาะสมกับระบบปฏิบัติการ ที่จะใช้ติดตั้งโปรแกรม Node.js โดยในตัวอย่างภาพที่ 2-13 เลือกใช้ระบบปฏิบัติการแบบ 64 bit

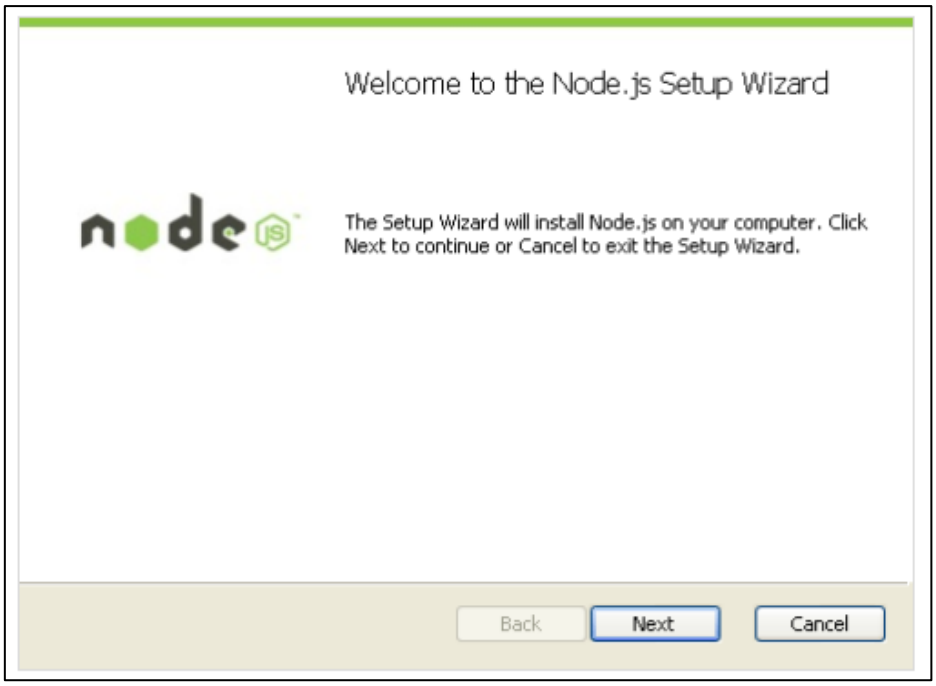


ภาพที่ 2-13 หน้าเว็บไซต์ดาวน์โหลดโปรแกรม Node.js ที่มา : http://www.siamhtml.com/ introduction- to-node-js/

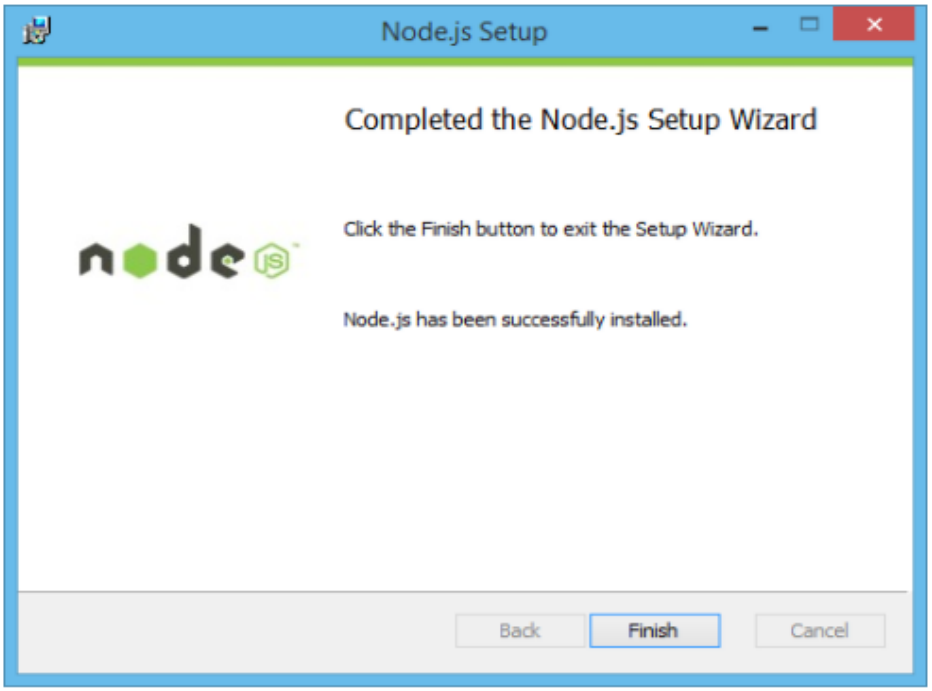
ทำการเลือกระบบปฏิบัติการที่ใช้งาน และทำการติดตั้งโปรแกรม ตัวอย่างดังภาพที่ 2-14



ภาพที่ 2-14 เลือกระบบปฏิบัติการที่ต้องการใช้ ที่มา : http://www.siamhtml.com/introduction-to-node-js/



ภาพที่ 2-15 ดาวน์โหลดและติดตั้งโปรแกรม Node.js ที่มา http://www.siamhtml.com/introduction-to-node-js/



ภาพที่ 2-16 ติดตั้งโปรแกรม Node.js สำเร็จ ที่มา : http://www.siamhtml.com/introduction-to-node-js/

3. การตรวจสอบเลขเวอร์ชันของ Node.js

เมื่อติดตั้งเสร็จแล้ว ให้เปิด Command-line Interface (Command Prompt, Terminal) ขึ้นมา แล้วพิมพ์คำสั่ง node –v เลขเวอร์ชันของ Node.js จะแสดงขึ้นมา ตัวอย่างดังภาพที่ 2-17

```
node -v
```

ภาพที่ 2-17 ตรวจสอบเลขเวอร์ชันของ Node.js ที่มา : http://www.siamhtml.com/introduction-to-node-js/

4. ขั้นตอนการติดตั้ง package npm

เนื่องจากการใช้ Application ใน Node.js นั้นไม่สามารถเขียนหรือสร้างขึ้นมาใช้เองได้ จึงมีวิธีเลือกเอา Application นั้นมาใช้งานเรียกว่าการใช้คำสั่ง “npm”

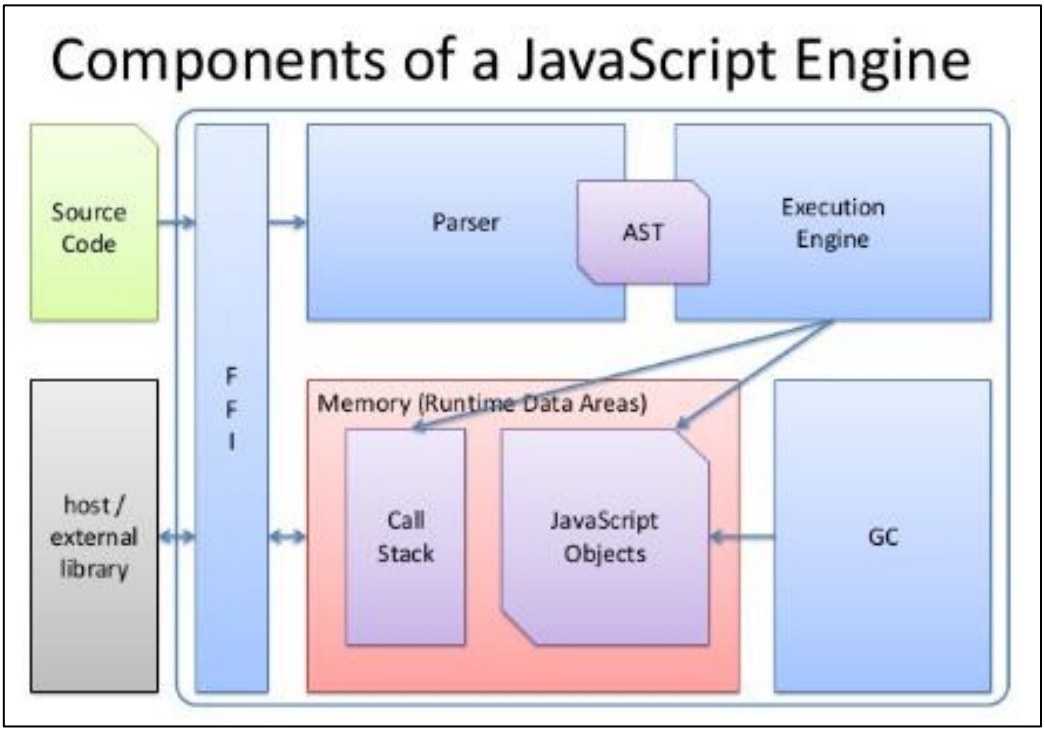
npm นั้นจะถูกติดตั้งมาพร้อมกับ Node.js เพื่อทำหน้าที่จัดการ Package เสริมต่าง ๆ การติดตั้ง Application หรือการติดตั้ง module ต่าง ๆ ที่เป็น dependency ของ Application โดยระบุชื่อ Package ที่ต้องการ “npm” ก็จะไปดำเนินการตรวจสอบชื่อ Package นั้นใน Registry เมื่อพบแล้วตัวโปรแกรมจะดาวน์โหลด Package นั้น ๆ มา นอกจากนั้นการนำ Application ที่เขียนไปเพิ่มไว้ใน Registry ของ npm ก็สามารถทำผ่าน npm ได้เช่นกัน สำหรับวิธีใช้ npm นั้นจะทำได้โดยการทำงานผ่าน Command-line Interface แล้วเข้าไปยัง Path ที่ต้องการจะติดตั้งหลังจากนั้นให้พิมพ์คำสั่งดังภาพที่ 2-18 Package ที่ระบุก็จะถูกติดตั้งเรียบร้อย

```
npm install ชื่อแพคเกจ
```

ภาพที่ 2-18 ติดตั้ง Package โดยเรียกผ่านคำสั่ง npm ที่มา : http://www.siamhtml.com/introduction-to-node-js/

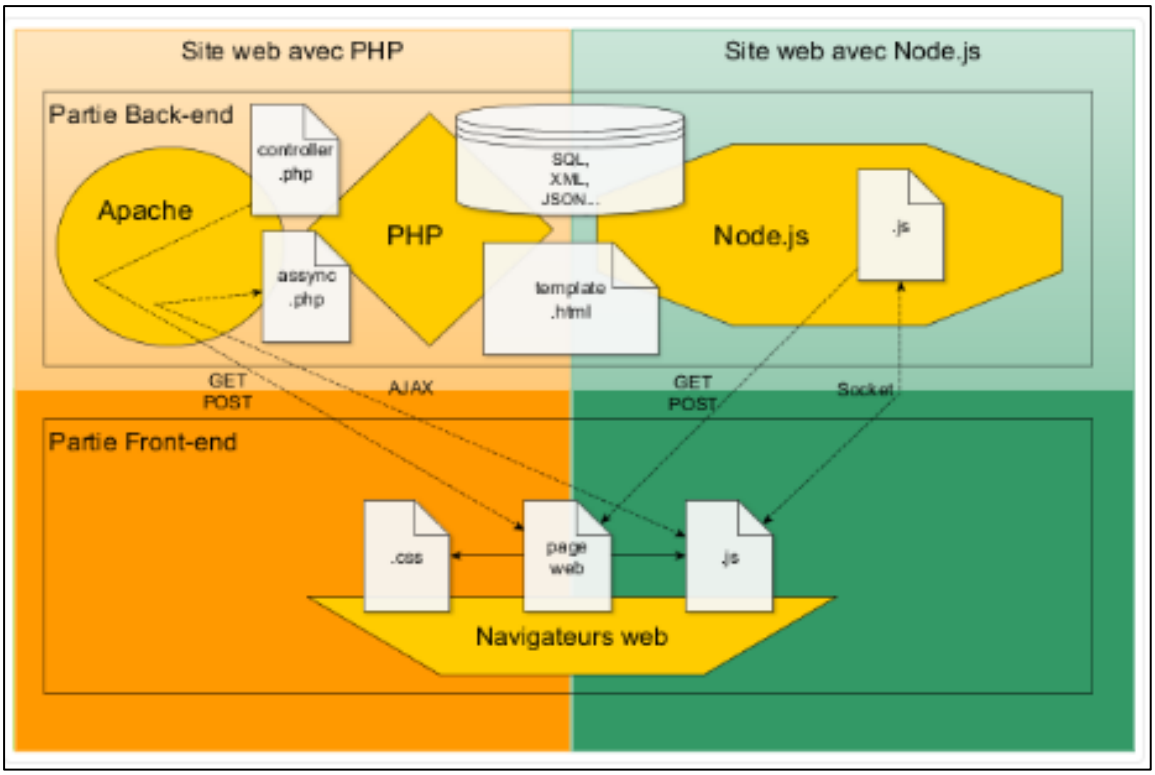
Package ที่น่าสนใจได้จาก official website ของ npm ได้เลย ที่ https://www.npmjs.com/ โดยทางเว็บไซต์จะมีหน้าการจัดอันดับ Package ที่มียอดดาวน์โหลดสูงสุดเอาไว้ด้วย ทำให้เราทราบว่าในขณะนี้ มีคนกำลังนิยมใช้ Package อะไรกันอยู่บ้างสำหรับวิธีใช้งานของแต่ละแพคเกจนั้นจะ

แตกต่างกันออกไป โดยเราสามารถอ่านวิธีใช้งานพื้นฐานของ Package นั้น ๆ ได้ที่หน้ารายละเอียดของ Package ที่ทาง npmjs.org ได้จัดทำไว้ แต่ถ้ายังไม่ละเอียดพอ ก็สามารถเข้าไปอ่านคู่มือการใช้งานได้ที่เว็บไซต์หลักของ Package นั้น ๆ Node.js ไม่ใช่ภาษาใหม่ และไม่ใช่ตัว Compiler ใหม่ ถ้าเอาไปเทียบกับ PHP ก็จะเข้าใจตรงกันว่า PHP คือ ภาษา Computer แต่ Node.js เวลาจะเขียน syntax ที่ใช้มันคือ ภาษา JavaScript เหมือนที่ใช้เขียนหน้าเว็บทั่วไป ผู้นิพนธ์ทราบว่าไม่ใช่ภาษาใหม่ ส่วนที่บอกว่าไม่ใช่ Compiler ตัวใหม่ ก็เพราะว่า Node.js ใช้ Compiler ของ Google JavaScript Engine V8



ภาพที่ 2-19 ส่วนประกอบของ JavaScript Engine ที่มา : http://www.dekcomcr.com/blog/?p=90

สรุป Node.js คือ environment หรือ สภาพแวดล้อม ที่ช่วยให้การทำงานได้ง่าย และเร็ว พร้อมตัวช่วยต่าง ๆ เพื่อความเหมาะสม ส่วนความแตกต่าง ระหว่าง PHP กับ Node.js ก็คือ วิธีการทำงานของ PHP กับ Node.js ทำงานแตกต่างกันโดยสิ้นเชิง ดังภาพที่ 2-20



ภาพที่ 2-20 ภาพแสดงความแตกต่างระหว่าง PHP กับ Node.js ที่มา : http://www.dekcomcr.com/blog/?p=90

กล่าวคือ PHP เวลาทำงาน จะเริ่มทำงานตั้งแต่บรรทัดแรกของไฟล์ .php ไปทีละบรรทัดไปเรื่อย ๆ จนถึงบรรทัดสุดท้าย ตัวอย่างเช่น บรรทัดส่วนเริ่มต้นของไฟล์ .php ทำงาน นาน 1 วินาที เช่น การเปิดไฟล์ที่มีขนาดใหญ่ และส่วนอื่นทำงาน 0.5 วินาที ผลของการทำงานก็คือ หน้า PHP นั้นจะใช้เวลาประมวลผล 1.5 วินาทีจึงจะเสร็จ แต่ส่วนของ Node.js ถ้าทำการใช้คำสั่งให้ทำงานเหมือนกันผลที่ได้ ก็คือ ผลการประมวลผลบรรทัดแรก ไปทีละบรรทัดเหมือนกัน แต่ว่า ถ้าการทำงานบางอย่างที่ต้องใช้เวลา เช่น การเปิดอ่านไฟล์ขนาดใหญ่ ก็จะเริ่มทำงานสิ่งนั้น แล้วข้ามไปทำสิ่งใหม่ต่อ โดยที่สิ่งที่เพิ่งทำผ่านไปยังไม่เสร็จ ซึ่งในส่วนนี้คือ key ของ Node.js

ยกตัวอย่าง เช่น การเปิดไฟล์ text ขนาดใหญ่ และบรรทัดต่อไป ใช้ไฟล์ text ที่ได้มาแสดงผลลัพธ์ ผลลัพธ์จะเป็น text แบบขาดหายไม่สมบูรณ์ทั้งไฟล์ เพราะการเปิดอ่านไฟล์ไม่สำเร็จการทำงานจะข้ามไปแสดงผลต่อไปโดยไม่สนใจความสมบูรณ์ของไฟล์ นี้คือการทำงานของ Node.js ส่วนการทำงานของ PHP ผู้นิพนธ์ขอยกตัวอย่างเช่น การเปิดไฟล์ทำงานไปทีละบรรทัดโดยมีเงื่อนไขบังคับว่า ถ้างานที่ทำอยู่ยังไม่เสร็จ ห้ามทำงานในบรรทัดต่อไป PHP มีการทำงานลักษณะทีละบรรทัดจนเสร็จ จึงทำให้การทำงานของ Node.js เร็วกว่าการทำงานของ PHP

การดึงข้อมูลเว็บไซต์ด้วย Node.js และ Cheerio

การดึงข้อมูลเว็บไซต์ด้วยการใช้ Node.js และ Cheerio ซึ่งเทคนิคการดึงข้อมูลเว็บไซต์ต่าง ๆ นี้ เรียกว่า “Web Scraping” หรือ “Web Crawler”

ขั้นตอนที่ 1 Create project การสร้างโครงการ

การสร้างโปรเจ็คด้วย npm init หรือ ใช้ไฟล์ package.json ตัวอย่างดังภาพที่ 2-21

```
{
  "name": "nodejs-google-play-information",
  "version": "1.0.0",
  "scripts": {
    "start": "node index.js",
    "dev": "nodemon index.js"
  },
  "engines": {
    "node": "^4.2.0"
  },
  "dependencies": {
    "cheerio": "^0.20.0",
    "hapi": "^13.0.0",
    "request": "^2.69.0"
  }
}
```

ภาพที่ 2-21 ภาพการเริ่มสร้างโครงการ ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

โดยตัวอย่างในที่นี่คือการดึงข้อมูลรายละเอียดของ App บน Google Play โดยใช้ชื่อของ application Id หรือก็คือชื่อ Package Name การทำ App Android ต้องมีชื่อที่ไม่ซ้ำกัน ทำให้สามารถใช้ชื่อนี้ในการเข้าดูรายละเอียด application แต่ละหน้าได้

ขั้นตอนที่ 2 Create Server with Hapi.js การสร้าง Server ด้วย Hapi.js

หลังจากได้ไฟล์ index.js แล้วต่อไปจะสร้าง Server ขึ้นมาใช้งานด้วย Hapi.js ตัวอย่างดังภาพที่ 2-22

```
'use strict';

const Hapi = require('hapi');
const server = new Hapi.Server();

server.connection({
  host: 'localhost',
  port: 8088
});

server.route({
  method: 'GET',
  path: '/{appId}',
  handler: (req, reply) => {
    reply({message: 'Hello World'});
  }
});

server.start(err => {
  console.log(`Server running at ${server.info.uri}`);
});
```

ภาพที่ 2-22 การสร้าง Server ขึ้นมาใช้งานด้วย Hapi.js ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

จากรายละเอียดดังภาพที่ 2-22 เขียนด้วย ES6 ซึ่งมีใน Node v4.2.4 โดยต้องกำหนด use strict โดยไม่ต้องใช้ Label ในการ compile เป็น ES5 และส่วนโค้ดอื่น ๆ ก็เป็นการเริ่มกำหนด route โดย path /appId ทดสอบสั่งรัน Server

```
node index.js
```

ภาพที่ 2-23 ทดสอบการรัน Server ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

และเมื่อเข้า http://localhost:8088/appId ก็จะได้ข้อความแสดงผล ดังภาพที่ 2-24

```
{
    "message": "Hello World"
}
```

ภาพที่ 2-24 ผลการทดสอบ Server ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

ขั้นตอนต่อไปคือการรับค่า appId หลังจากนั้นก็ใช้ request module เพื่อเปิดหน้าเว็บของ Google Play ด้วย appId ตัวอย่างโค้ดการเขียน รายละเอียดดังภาพที่ 2-25

```
const URL = 'https://play.google.com/store/apps/details?id=';

server.route({
  method: 'GET',
  path: '/{appId}',
  handler: (req, reply) => {
    let appId = req.params.appId;
    let lang = req.query.lang || 'en';
    let url = `${URL}${appId}&hl=${lang}`;

    reply({
      url: url
    });
  }
});
```

ภาพที่ 2-25 การรับค่า appId เพื่อเปิดหน้าเว็บ Google Play ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

ขั้นตอนที่ 3 Use Request การใช้งานการร้องขอ

Module ที่จะทำให้สามารถ call HTTP request ได้คือ Request ซึ่งการใช้งาน Request โดยมี syntax ดังนี้

Request (URL, callback): URL คือ URL ที่ต้องการ call ส่วน callback เป็น callback function ซึ่งมี (err, response และ body) 3 ตัว

1. err: หากการ call HTTP มี error เกิดขึ้น
2. response: เป็นค่า response ที่ตอบกลับมาจาก server เช่น header, status Code
3. body: เป็นข้อมูล body ที่ server ส่งกลับมา เหมือนหน้า HTML ทั่วไปตามหน้าเว็บไซต์ตัวอย่าง ดังภาพที่ 2-26

```
const request = require('request');

request('http://devahoy.com', (err, response, body) {

  if (!err && response.statusCode === 200) {
      console.log('body : ' +  body);
  }
});
```

ภาพที่ 2-26 ภาพโค้ดของการ Request ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

ขั้นตอนที่ 4 Cheerio

เมื่อได้ค่าของ body จากการ call HTTP แล้วลำดับต่อมาเรื่องของการใช้ cheerio เพื่อหา DOM element ในหน้า HTML นั้น โดยมี syntax เหมือนกับ jQuery มี html ดังภาพที่ 2-27

```
<ul id="fruits">
  <li class="apple">Apple</li>
  <li class="orange">Orange</li>
  <li class="pear">Pear</li>
</ul>
```

ภาพที่ 2-27 ภาพ syntax ของภาษา HTML ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

การใช้ Cheerio และการ Selector ตัวอย่างดังภาพที่ 2-28

```
const cheerio = require('cheerio');

let $ = cheerio.load(html);

$('.apple', '#fruits').text()
//=> Apple

$('ul .pear').attr('class')
//=> pear

$('li[class=orange]').html()
//=> Orange
```

ภาพที่ 2-28 ภาพตัวอย่างการใช้ Cheerio และการ Selector ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

ขั้นตอนที่ 5 การใช้ Cheerio Selector

ตัวอย่างที่เราต้องการดึงข้อมูลมา โดยเข้าเว็บ Facebook on Google Play ขั้นตอนนี้ใช้ Chrome Developer Tools เข้ามาช่วย ทำได้โดยการเลือก More Tools => Developer Tool ลำดับแรกกำหนดสิ่งต้องการคือ Title ของแอพพลิเคชัน ด้านบน พบว่า .document-title selector จะเป็น $('.document-title').text() ดังภาพ 2-29

ภาพที่ 2-29 ภาพแสดง Chrome Developer Tools ที่มา : https://devahoy.com/posts/scraping-web-with-nodejs/

2.7 งานวิจัยที่เกี่ยวข้อง

1. นายกลยุทธ บพิตร, 2554 ได้ศึกษาและวิจัยเกี่ยวกับการสกัดข้อมูลสินค้าบนเว็บเพจด้วยเว็บครอว์เลอร์ที่นำไปใช้ในโปแกรมค้นหาสินค้า ซึ่งเน้นการวิเคราะห์โครงสร้างเอกสาร HTML และการวิเคราะห์คำสำคัญ โดยเริ่มจากวิเคราะห์โครงสร้างหน้ารวมของเว็บขายสินค้า จากนั้นค้นหาโหนดของสินค้าโดยการวิเคราะห์รูปและราคา และนำโหนดของสินค้าที่ได้ไปทำการสกัดรายละเอียดข้อมูลสินค้า สำหรับการประเมินประสิทธิภาพของการทำงานของขั้นตอนการสกัดข้อมูลสินค้า ผลลัพธ์ที่ได้จากการทำงานขั้นตอนวิธีการสกัดข้อมูลสินค้าจะถูกเปรียบเทียบกับผลลัพธ์การสกัดข้อมูลสินค้าโดยมนุษย์ ซึ่งผู้จัดทำสุ่มเลือกยูอาร์แอลจากเว็บไซต์พาณิชย์อิเล็กทรอนิกส์ จำนวนทั้งสิ้น 60 ยูอาร์แอล ผลการประเมินพบว่าขั้นตอนและวิธีการสกัดข้อมูลสินค้ามีความแม่นยำในการสกัดข้อมูลสินค้าร้อยละ 88.4 สำหรับเว็บร้านค้าออกไลน์และเว็บแคตตาล็อกสินค้าออนไลน์ ในขณะที่มีความแม่นยำในการสกัดข้อมูลสินค้าร้อยละ 77.3 สำหรับตลาดกลางอิเล็กทรอนิกส์

งานวิจัยนี้มีหลักการคล้ายคลึงกับงานนิพนธ์ที่จัดทำขึ้น โดยการวิเคราะห์โครงสร้าง HTML เว็บไซต์ที่ต้องการสกัดข้อมูล จากการวิเคราะห์รูป ราคาและรายละเอียดสินค้าแต่ในงานนิพนธ์ที่จัดทำขึ้นจะวิเคราะห์โครงสร้าง HTML ในส่วนข้อมูลที่ต้องการเท่านั้น ตามโครงสร้างแต่ละเว็บไซต์ที่ต้องการสกัดข้อมูล

2. วิริยะ แก้วมรินทร์, 2551 เป็นตัวอย่างงานวิจัยเกี่ยวกับการค้นหาบริการของเว็บเซอร์วิสแบบสื่อความหมายโดยใช้ตัวค้นหาบนเว็บ งานวิจัยนี้นำเสนอ สถาปัตยกรรมแบบระดับชั้นของระบบสืบค้นเว็บเซอร์วิสอย่างมีความหมาย โดยการใช้ Crawler เป็นองค์ประกอบหลักในการค้นหาบริการที่อยู่ในเว็บไซต์ต่าง ๆ Crawler สามารถทำงานในสภาพแวดล้อมที่เป็นมัลติเทรดเพื่อเพิ่มความสามารถในการค้นหาเว็บเซอร์วิสที่กระจ่ายอยู่ได้อย่างพร้อมกันรวมไปถึงการสืบค้นจากไดเรกทอรีกลางที่หลากหลายด้วย ซึ่งคำอธิบายเว็บเซอร์วิสที่ถูกสืบค้นได้จะถูกแปลงให้เป็นภาษาที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ เช่น ภาษา OWL-S ระบบงานที่ได้จากการวิจัยนี้มีความยืดหยุ่น

และง่ายต่อการสืบค้นข้อมูลเว็บเซอร์วิสที่มีความซับซ้อนที่สอดคล้องกับความต้องการของผู้ใช้อย่างแท้จริง

งานวิจัยนี้แตกต่างจากงานนิพนธ์ที่จัดทำขึ้นในด้านภาษาที่นำมาพัฒนา Web Crawler ซึ่งในงานนิพนธ์นี้เลือกใช้ Node.js และ Cheerio เนื่องจากเป็นเทคโนโลยีที่ใหม่กว่าแล้วทำงานบนเครื่องแม่ข่ายซึ่งสามารถเรียกใช้งานได้ตลอดและมีประสิทธิภาพในการประมวลผลที่ดี

3. นิรันดร์ อังควัฒนวิทย์, 2545 ได้ทำการวิจัยเกี่ยวกับการเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องด้วยเว็บครอว์เลอร์แบบเรียนรู้ได้ เว็บครอว์เลอร์แบบเฉพาะเจาะจงหัวเรื่องใช้สำหรับเลือกเก็บเว็บเพจที่มี หัวเรื่องตรงกับความต้องการ มีงานวิจัยที่ผ่านมาค้นคว้าอัลกอริทึมในการเลือกเก็บเว็บเพจ มีจุดประสงค์เพื่อให้เว็บครอว์เลอร์เก็บเว็บเพจที่ตรงกับหัวเรื่องที่ต้องการ ให้ได้มากที่สุดเมื่อเทียบเป็นสัดส่วนกับจำนวนเว็บเพจที่เก็บมาทั้งหมด งานที่ผ่านมา ได้นำเสนอการเก็บเว็บเพจเพียงครั้งเดียวเท่านั้น คำถามคือเว็บเพจคราวเลอร์ทำอย่างไร ในการเก็บเว็บเพจครั้งต่อไปเพื่อเพิ่มเติมหรือติดตามการเปลี่ยนแปลงของเว็บเพจ บทความนี้เสนออัลกอริทึมการเก็บเว็บเพจครั้งแรกและครั้งต่อไปของเว็บครอว์เลอร์ โดยนำข้อมูลการเก็บเว็บเพจครั้งก่อนมาสร้างฐานความรู้สำหรับการเก็บเว็บเพจครั้งต่อไป ได้แก่ ยูอาร์แอลเริ่มต้น คำสำคัญของหัวเรื่อง การทำนายยูอาร์แอล ฐานความรู้นี้ เปรียบเสมือนความรู้ของเว็บครอว์เลอร์ที่รวบรวมจากประสบการณ์ในการเก็บเว็บเพจ ครั้งก่อนหน้า เมื่อนำความรู้มาใช้ในการเก็บเว็บเพจครั้งต่อไป ประสิทธิภาพการเก็บเว็บเพจควรดีขึ้น

งานวิจัยที่สร้างเว็บคลอเลอร์สำหรับเก็บหัวเรื่องที่ต้องการ โดยอัลกอริทึมสามารถเรียนรู้ได้เองในการเก็บข้อมูลครั้งต่อไป จากการสร้างฐานข้อมูลให้เว็บคลอเลอร์ได้เรียนรู้ ซึ่งแตกต่างจากงานนิพนธ์ครั้งนี้ที่จะเน้นการเก็บรายละเอียดข้อมูลงานวิจัยจากเว็บไซต์ที่ต้องการ โดยอัลกอริทึมจะเฉพาะเจาะจงในแต่ละเว็บไซต์

4. ณรงค์ ล่ำดี, 2552 ได้ศึกษาค้นคว้าทักษะการใช้โปรแกรมค้นหาของนักศึกษาวิทยาลัยราชพฤกษ์ ในครั้งนี้มีวัตถุประสงค์ ศึกษาปัญหาในการใช้งานโปรแกรมค้นหา ทราบถึงทักษะและความเข้าใจในการใช้งานโปรแกรมค้นหา เพื่อเปรียบเทียบผลหลัง จากมีความเข้าใจที่ถูกต้องในการใช้งานโปรแกรมค้นหาโดยการศึกษาดังกล่าวใช้นักศึกษาวิทยาลัยราชพฤกษ์เป็นกลุ่มตัวอย่างจำนวน 345 คนในการดำเนินการศึกษาจะเก็บ รวบรวมข้อมูลในเรื่องการใช้งานอินเทอร์เน็ต, การใช้งานโปรแกรมค้นหา และการใช้งานโปรแกรมค้นหาขั้นสูง ซึ่งจะนำผลลัพธ์จากการทำแบบทดสอบมาเปรียบก่อนและหลังได้รับคู่มือแนะนำการใช้โปรแกรมค้นหา เพื่อหาประสิทธิภาพการเรียนรู้ในการใช้โปรแกรมค้นหาผลการศึกษาพบว่า ทักษะการใช้โปรแกรมค้นหาของนักศึกษา

วิทยาลัยราชพฤกษ์อยู่ในระดับพอใช้-น้อย โดยส่วนใหญ่ยังมีความเข้าใจผิดถึงความหมายของโปรแกรมค้นหา และประเภทของโปรแกรมค้นหาที่ใช้ รวมไปถึงความเข้าใจในการใช้งานโปรแกรมค้นหาขั้นสูงที่นักศึกษาส่วนใหญ่ไม่เคยผ่านการใช้งาน จากการทดสอบพบว่าทักษะในส่วนการใช้งานขั้นสูง ในเรื่องตรรกะบูลีน (X =1.65,S.D. =0.32) อยู่ในระดับน้อย และตัวดำเนินการขั้นสูง (X =1.69, S.D. =0.23) อยู่ในระดับน้อย หลังจากที่ได้รับคู่มือความเข้าใจ ในเรื่องตรรกะบูลีน (X =3.99, S.D.=0.58) อยู่ในระดับมาก และตัวดำเนินการขั้นสูง (X =4.04, S.D.=0.39) อยู่ในระดับมาก ในการศึกษาครั้งนี้ได้เสนอแนวทางในการแก้ไขปัญหาการใช้งานโปรแกรมค้นหาด้วยการเพิ่มทักษะและความเข้าใจที่ถูกต้องในการใช้งานโปรแกรมค้นหา โดยจัดทำคู่มือการใช้งานแก่นักศึกษาทำให้ความเข้าใจและทักษะในการใช้งานโปรแกรมค้นหาเพิ่มขึ้นจากระดับน้อยเป็นระดับมาก งานวิจัยนี้เป็นการศึกษาการใช้งานโปรแกรมค้นหาของกลุ่มนักศึกษาที่นำมาเป็นข้อมูลตัวอย่าง

ด้วยการวัดผลจากแบบสอบถามการใช้งานโปรแกรมค้นหาและหลังจากใช้คู่มือแนะนำ ซึ่งแตกต่างจากงานนิพนธ์ครั้งนี้ที่ใช้พัฒนาโปรแกรมสำหรับสกัดข้อมูลเพื่อนำข้อมูลที่ได้มารวบรวมสร้างระบบสารสนเทศสำหรับค้นหางานวิจัย

บทที่ 3
วิธีดำเนินงานนิพนธ์และเครื่องมือ

บทนี้จะกล่าวถึงวิธีดำเนินงานนิพนธ์และเครื่องมือที่ใช้ในการดำเนินงานการจัดทำนิพนธ์กรณีศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอว์เลอร์ โดยเริ่มจากการศึกษาค้นคว้าข้อมูลที่เกี่ยวข้อง การออกแบบขั้นตอนและวิธีการสกัดข้อมูลงานวิจัย การกำหนดแบบแผนการวัดประสิทธิผล และการวัดประสิทธิผล โดยมีรายละเอียดดังต่อไปนี้

3.1 การศึกษาค้นคว้าข้อมูลที่เกี่ยวข้อง

ขั้นตอนนี้ผู้จัดทำงานนิพนธ์ได้ทำการศึกษาเพื่อหาขั้นตอนและวิธีการสกัดข้อมูลงานวิจัยบนเว็บเพจเพื่อนำไปใช้ในระบบค้นหางานวิจัย ซึ่งมีหัวข้อดังต่อไปนี้

3.1.1 ศึกษาเกี่ยวกับโครงสร้างเว็บไซต์งานวิจัย

งานนิพนธ์นี้ได้ศึกษาโครงสร้างเว็บไซต์งานวิจัยในประเทศไทย ที่ได้เก็บรวมรวมผลของงานวิจัยต่าง ๆ เช่น เว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา เว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย เว็บไซต์คลังข้อมูลงานวิจัยไทย และเว็บไซต์ IEEE Xplore Digital Library เป็นต้น ซึ่งข้อมูลที่สำคัญและจำเป็นในเว็บไซต์งานวิจัยที่พบได้แก่ ชื่องานวิจัย ผู้เขียนงานวิจัย ปีที่เขียนงานวิจัย ยูอาร์แอลของงานวิจัย เอกสารงานวิจัยและบทคัดย่อ

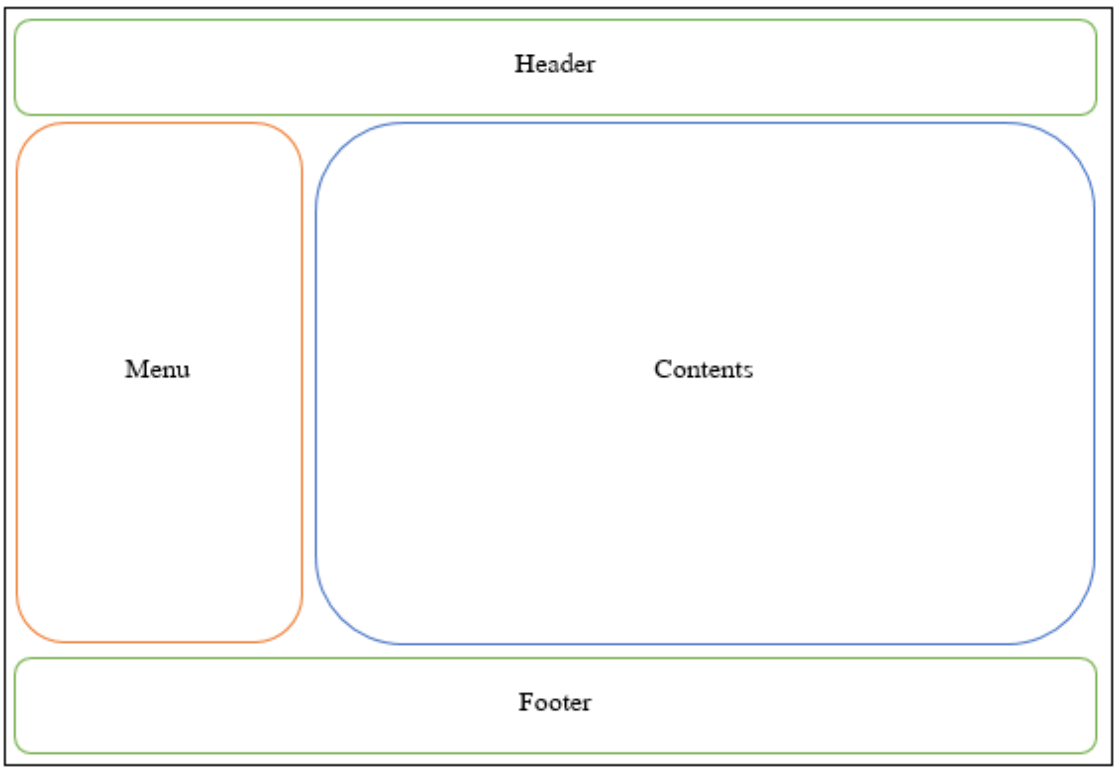
จากการศึกษาเว็บไซต์ทั้งหมดพบว่า โครงสร้างเว็บไซต์ประกอบไปด้วยส่วนหลักทั้งหมด 4 ส่วน ได้แก่

ส่วนที่ 1 ส่วนหัว คือ ส่วนแสดงชื่อเว็บไซต์ Logo และ Title

ส่วนที่ 2 ส่วนเมนู คือ ส่วนที่เป็นจุด ลิ้งค์ไปยังหน้าเว็บเพจอื่น โดยแบ่งเป็นหมวดหมู่

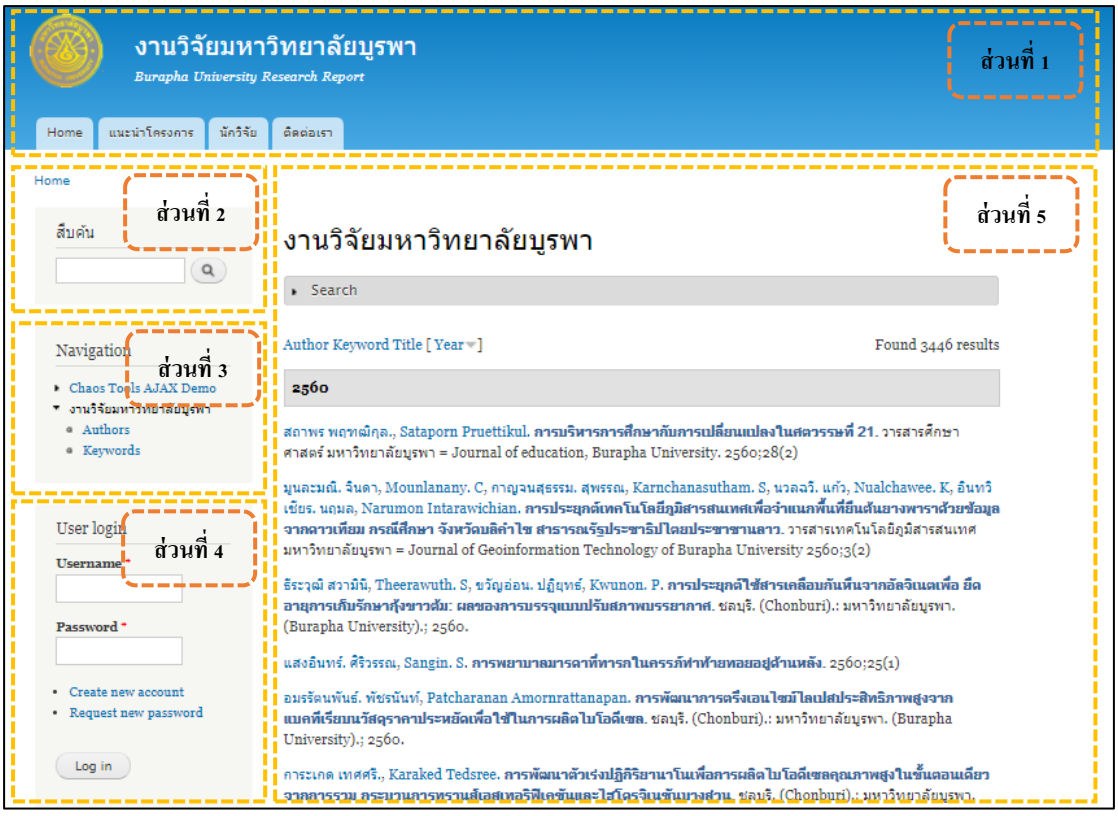
ส่วนที่ 3 ส่วนของเนื้อหา คือ ส่วนของการนำเสนอข้อมูลจากผลลัพธ์ที่ค้นหา ส่วนนี้เป็นส่วนที่สำคัญที่สุด

ส่วนที่ 4 ส่วนท้าย คือ ส่วนของ e-mail หรือส่วนของการอ้างอิง หรือส่วนของลิขสิทธ์



ภาพที่ 3-1 ภาพโครงสร้างพื้นฐานของเว็บไซต์ ที่มา : เนณุภา สุภเวชย์. (2551). การสร้างเว็บเพจด้วยภาษา HTML

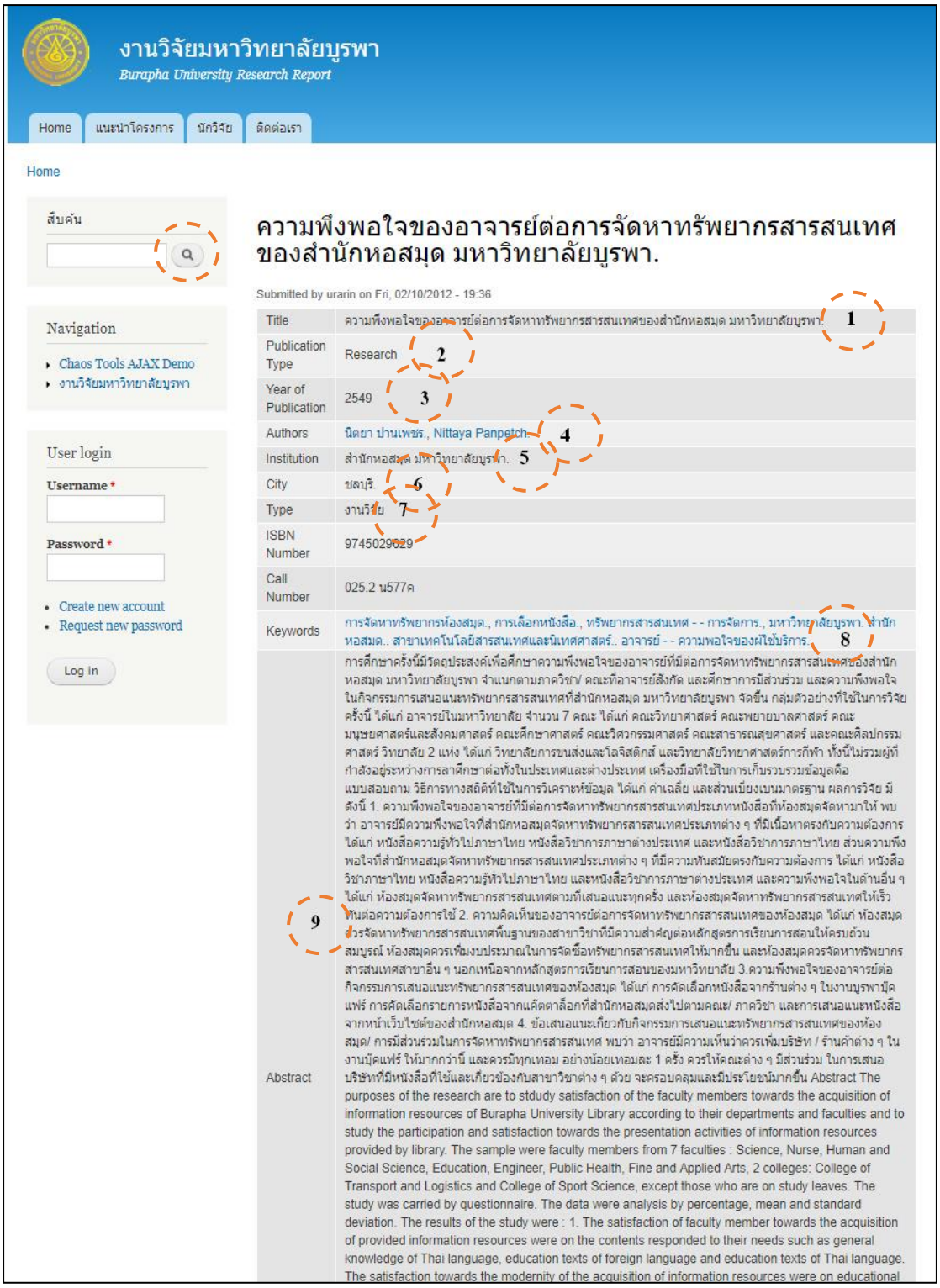
โครงสร้างเว็บไซต์งานวิจัยเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา เว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย เว็บไซต์คลังข้อมูลงานวิจัยไทยทั้ง 3 เว็บไซต์นี้มีโครงสร้างที่คล้ายคลึงกัน ซึ่งสามารถใช้เป็นประโยชน์ต่อการวิเคราะห์ เพื่อสกัดข้อมูลต่อไป ในบทนี้ผู้จัดทำงานนิพนธ์ยกตัวอย่างเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพาในการศึกษาโครงสร้างเว็บไซต์เพื่อออกแบบฟังก์ชันสกัดข้อมูลที่สามารถดึงข้อมูลงานวิจัยที่ต้องการออกมาได้เพื่อนำไปจัดเก็บในฐานข้อมูลต่อไป ซึ่งมีรูปแบบโครงสร้างดังต่อไปนี้



ภาพที่ 3-2 ตัวอย่างผลการค้นหางานวิจัยจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา

ภาพที่ 3-2 เป็นตัวอย่างเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา สามารถวิเคราะห์ส่วนประกอบของเว็บไซต์ ซึ่งประกอบด้วยส่วนต่าง ๆ ดังต่อไปนี้

1. ส่วนที่ 1 ส่วนหัวเป็นส่วนที่แสดงรายละเอียดเว็บไซต์ ประกอบไปด้วย ชื่อระบบงาน และเมนูหลักได้แก่ เมนูHome เมนูแนะนำโครงการ เมนูนักวิจัยและเมนูติดต่อเรา ส่วนใหญ่นักพัฒนาระบบจะใช้รูปภาพเป็นองค์ประกอบหลัก
2. ส่วนที่ 2 ส่วนการค้นหางานวิจัย ซึ่งส่วนนี้จะแสดงรายละเอียดเกี่ยวกับลิ้งก์งานวิจัยที่เกี่ยวข้องและค้นหาข้อมูล
3. ส่วนที่ 3 ส่วนการแสดงลิ้งก์ที่เกี่ยวข้อง
4. ส่วนที่ 4 ส่วนแสดงการเข้าสู่ระบบเข้าใช้งานและสมัครลงทะเบียนเป็นสมาชิก
5. ส่วนที่ 5 ส่วนแสดงผลลัพธ์ ส่วนนี้จะแสดงผลลัพธ์งานวิจัยจากการค้นหาของผู้ใช้งานซึ่งเป็นส่วนสำคัญที่ผู้จัดทำงานนิพนธ์สนใจในการนำข้อมูลไปใช้ในการพัฒนาระบบค้นหา เมื่อทำการใช้คำเพื่อเข้าสืบค้นงานวิจัยในส่วนของการค้นหา แต่ละงานวิจัยจะได้ผลลัพธ์ดังภาพที่ 3-2 ต่อไปนี้



ภาพที่ 3-3 รายละเอียดงานวิจัยของเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา

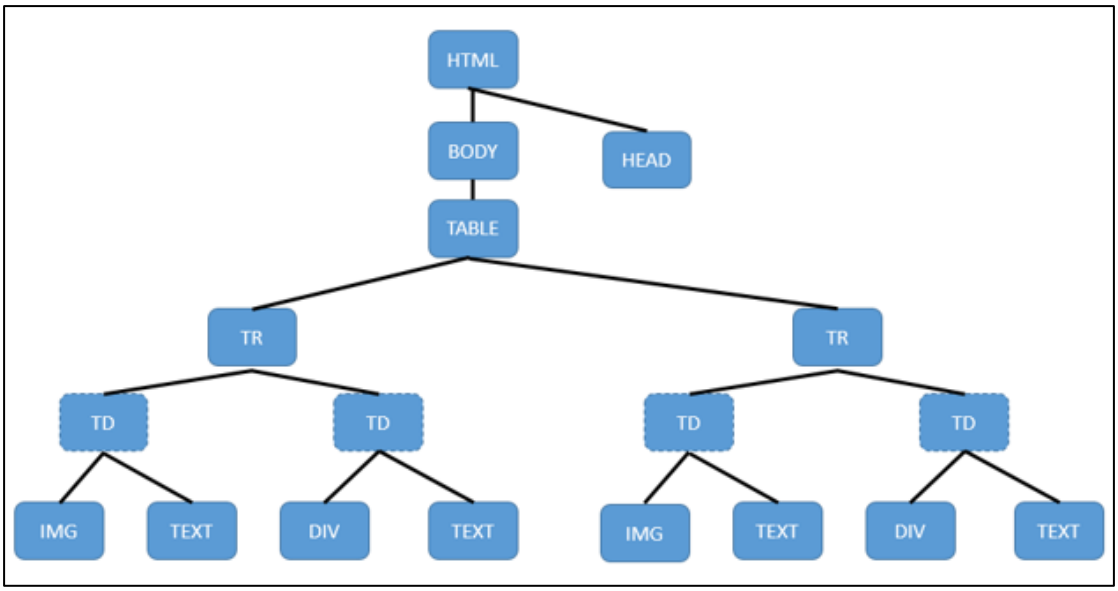
ภาพที่ 3-3 แสดงรายละเอียดงานวิจัยของเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา ซึ่งเป็นผลลัพธ์จากการสืบค้น มีข้อมูลครบตามที่ผู้จัดทำงานนิพนธ์ต้องการ ดังต่อไปนี้

1. Title คือ หัวข้องานวิจัย
2. Publication Type คือ ชนิดที่เผยแพร่
3. Year of Publication คือ ปีที่เผยแพร่งานวิจัย
4. Author คือ ผู้เขียนงานวิจัย
5. Institution คือ สถาบันที่จัดทำงานวิจัย
6. City คือ เมือง
7. Type คือ ชนิดงานวิจัย
8. Keywords คือ คำสำคัญในงานวิจัย
9. Abstract คือ บทคัดย่องานวิจัย

จากข้อมูลข้างต้นเป็นรายละเอียดงานวิจัยที่ต้องการ ซึ่งวิธีการสกัดข้อมูลจากส่วนต่าง ๆ จะอธิบายในหัวข้อต่อไป

3.1.2 เว็บครอว์เลอร์ที่ใช้ในการสกัดข้อมูลงานวิจัย

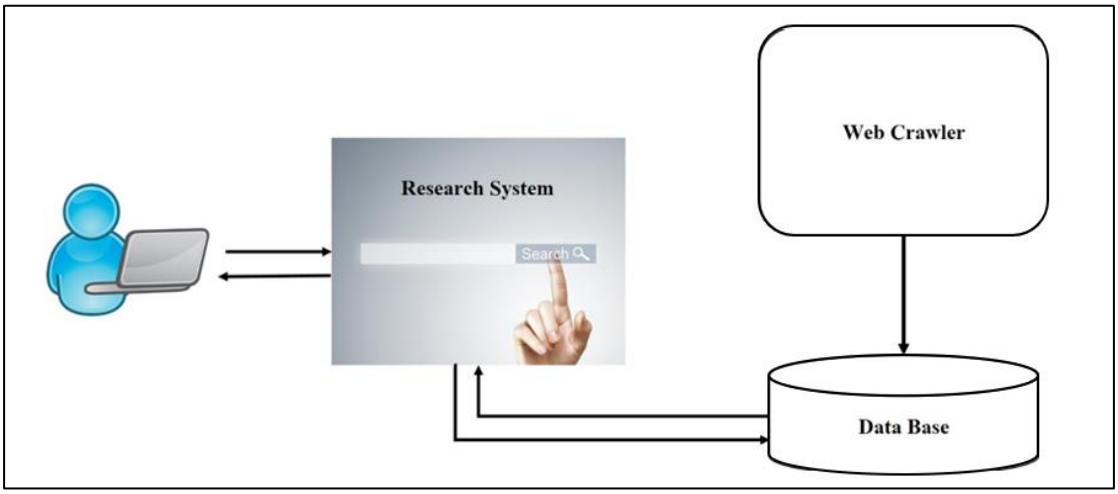
จากที่ได้กล่าวถึงในบทที่ 2 หัวข้อที่ 2.4 พบว่าวิธีการสกัดข้อมูลสามารถทำได้หลายวิธี ซึ่งไม่สามารถใช้เทคโนโลยีเว็บเซอร์วิสได้ ดังนั้นผู้จัดทำงานนิพนธ์จึงจำเป็นต้องใช้การสกัดข้อมูลจากโครงสร้าง HTML ที่เรียกว่า Node.js และ Cheerio แทน ซึ่งเป็นเทคนิคการดึงข้อมูลจากเว็บไซต์ที่เรียกว่า “Web Scraping” หรือ “Web Crawler” เป็นการโหลดข้อมูลโหนดของ HTML มา ซึ่งข้อมูล HTML ที่ได้ มีโครงสร้างลักษณะต้นไม้ หรือ DOM Tree ดังภาพที่ 3-3 ลักษณะลำดับจะแสดงชั้นของโหนด เริ่มต้นจาก Root หรือ โหนดแม่ ไปยังโหนดที่ต่ำกว่า หรือ โหนดลูก ซึ่งโหนดแม่สามารถมีโหนดลูกได้หลายโหนด แต่โหนดลูกสามารถมีโหนดแม่ได้เพียงโหนดเดียว ตัวอย่างเช่น โหนด TR มีโหนด TD ที่เรียกว่าโหนดลูก (ในกรอบเส้นประ) จำนวน 2 โหนดหรือเท่าไหร่ก็ได้ เป็นต้น



ภาพที่ 3-4 โครงสร้างโหนดในภาษา HTML

3.1.3 ออกแบบและวิเคราะห์ระบบค้นหางานวิจัย

ขั้นตอนการออกแบบระบบค้นหางานวิจัย ซึ่งประกอบด้วย 3 ส่วนหลักได้แก่ เว็บคลอว์เลอร์สำหรับสกัดข้อมูลงานวิจัย ฐานข้อมูลสำหรับเก็บข้อมูลงานวิจัย และเว็บไซต์สำหรับแสดงผลลัพธ์งานวิจัยจากการค้นหางานวิจัย ซึ่งแสดงดังภาพที่ 3-4



ภาพที่ 3-5 ไดอะแกรมแสดงภาพรวมระบบค้นหางานวิจัย

จากภาพที่ 3-5 แสดงภาพรวมระบบค้นหางานวิจัย ซึ่งเมื่อผู้ใช้งานระบุคำค้น ระบบจะนำคำค้นหาดังกล่าว ค้นหาข้อมูลงานวิจัยในฐานข้อมูลที่ได้จากการสกัดข้อมูลงานวิจัยจากเว็บไซต์ที่ต้องการและนำผลลัพธ์ที่ได้มาแสดงให้แก่ผู้ใช้งาน

3.2 การออกแบบขั้นตอนและวิธีการสกัดข้อมูลงานวิจัย

การออกแบบขั้นตอนและวิธีการสกัดข้อมูลงานวิจัยนี้ จะใช้วิธีการวิเคราะห์โครงสร้าง HTML เพื่อหาโหนดของข้อมูลงานวิจัยที่ต้องการ ซึ่งแบ่งขั้นตอนได้ดังต่อไปนี้

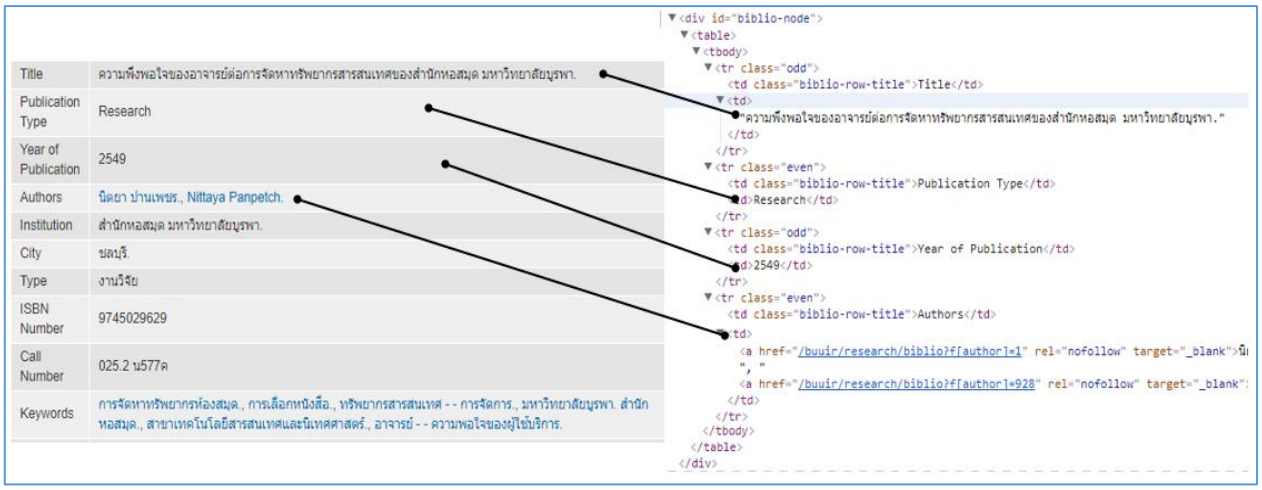
3.2.1 วิเคราะห์โครงสร้าง HTML ในหน้าเว็บไซต์งานวิจัย

จากการศึกษางานวิจัยที่เกี่ยวข้อง พบว่า Cheerio เป็นเครื่องมือที่ใช้สกัดข้อมูลที่ใช้กันอย่างแพร่หลาย ในการสกัดข้อมูล งานวิจัยนี้ Cheerio จะช่วยโหลดข้อมูล HTML ทั้งหมดเพื่อหา DOM Element โดย Cheerio จะท่องหน้าเว็บและตรวจสอบจนพบข้อความที่ต้องการ ที่อยู่ใน Tag ที่ต้องการด้วย

การทำงานของ Cheerio ใช้วิธีแยกโหนดแต่ละอันที่ต้องการออกจากกัน และเข้าถึงข้อมูลเพื่อค้นหาคำที่ต้องการและทำการสกัดออกมา โดยเรียกโหนดแต่ละอันว่า โหนดข้อมูลงานวิจัย ตัวอย่างข้อมูลงานวิจัยจากเว็บไซต์ งานวิจัยมหาวิทยาลัยบูรพาประกอบด้วยโหนดต่าง ๆ ดังนี้

- ไอดี “biblio-node” คือโหนดแม่ที่ประกอบด้วยโหนดลูกที่เก็บข้อมูลงานวิจัยที่ต้องการ
- Tag “Table” คือ ตารางรายละเอียดงานวิจัย
- Tag “TR” ลำดับที่ 1 คือ โหนดลูกที่เก็บชื่องานวิจัย
- Tag “TR” ลำดับที่ 2 คือ โหนดลูกที่เก็บชนิดที่เผยแพร่งานวิจัย
- Tag “TR” ลำดับที่ 3 คือ โหนดลูกที่เก็บปีที่เผยแพร่งานวิจัย
- Tag “TR” ลำดับที่ 4 คือ โหนดลูกที่เก็บชื่อผู้เขียนงานวิจัยทั้งหมด
- Tag “a” คือ โหนดลูกที่เก็บชื่อผู้เขียนงานวิจัยและยูอาร์แอลประวัติผู้เขียนงานวิจัย

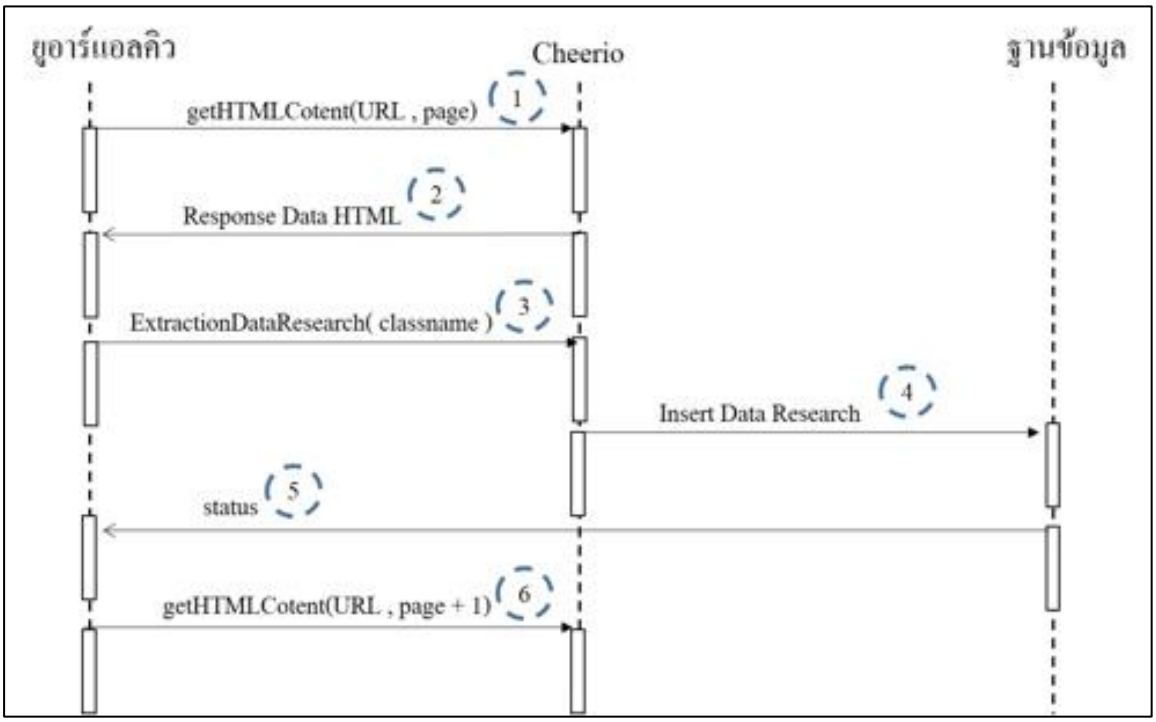
จากตัวอย่างจะเห็นได้ว่าไอดี “biblio-node” คือ ส่วนของข้อมูลงานวิจัยที่ต้องการ ผู้จัดทำงานนิพนธ์จึงโหลดข้อมูลโหนดต่าง ๆ ที่อยู่ภายใต้ไอดี “biblio-node” ใน แท็ก “table” เพื่อนำมาสกัดข้อมูลงานวิจัยที่ต้องการได้แก่ ดังภาพที่ 3-6



ภาพที่ 3-6 โครงสร้าง HTML ข้อมูลงานวิจัยจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา

3.2.2 วิเคราะห์กระบวนการทำงานของเว็บครอว์เลอร์

การวิเคราะห์กระบวนการทำงานของเว็บครอว์เลอร์ เพื่อวิเคราะห์ภาพรวมการทำงานเว็บครอว์เลอร์สำหรับสกัดข้อมูลงานวิจัย ซึ่งผู้จัดทำงานนิพนธ์ได้กำหนดลำดับการสกัดข้อมูลงานวิจัยของเว็บครอว์เลอร์ ดังภาพที่ 3-7



ภาพที่ 3-7 ลำดับการสกัดข้อมูลงานวิจัยของเว็บครอว์เลอร์

ภาพที่ 3-7 สามารถอธิบายลำดับการสกัดข้อมูลงานวิจัยของเว็บครอว์เลอร์ได้ ดังต่อไปนี้

1. โหลดเว็บเพจงานวิจัยที่กำหนด ขั้นตอนนี้เป็นการโหลดข้อมูล HTML เว็บเพจที่ต้องการโดยการกำหนดยูอาร์แอลและหน้าที่ต้องการโหลด

2. เมื่อคำสั่งในข้อที่ 1 ทำเสร็จสิ้นจะส่งข้อมูล HTML ที่โหลดได้คืนให้

3. สกัดข้อมูลงานวิจัยจากโครงสร้าง HTML ที่โหลดมาได้ ซึ่งกำหนด Class ที่จัดเก็บข้อมูลงานวิจัยที่ต้องการสกัดข้อมูล

4. เมื่อสกัดข้อมูลงานวิจัยตามข้อที่ 3 ขั้นตอนนี้จะจัดเก็บข้อมูลที่ได้ในฐานข้อมูล

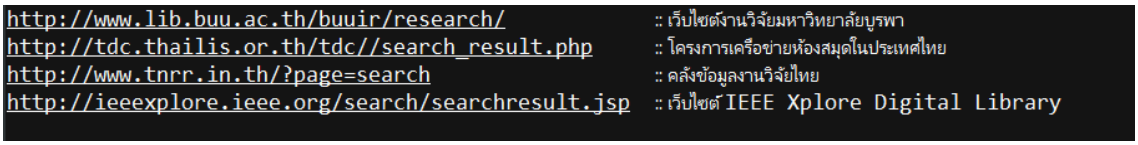
5. ตอบกลับสถานะการจัดเก็บข้อมูลงานวิจัย ซึ่งกรณีที่มีข้อผิดพลาดจะระบุข้อผิดพลาดกลับมาด้วย

6. โหลดเว็บเพจงานวิจัยที่กำหนดในหน้าต่อไป ซึ่งทำตามขั้นตอนที่ 1 – 5 จนครบทุกหน้าเว็บเพจของ URL

3.2.3 การออกแบบขั้นตอนการทำงานของเว็บครอว์เลอร์

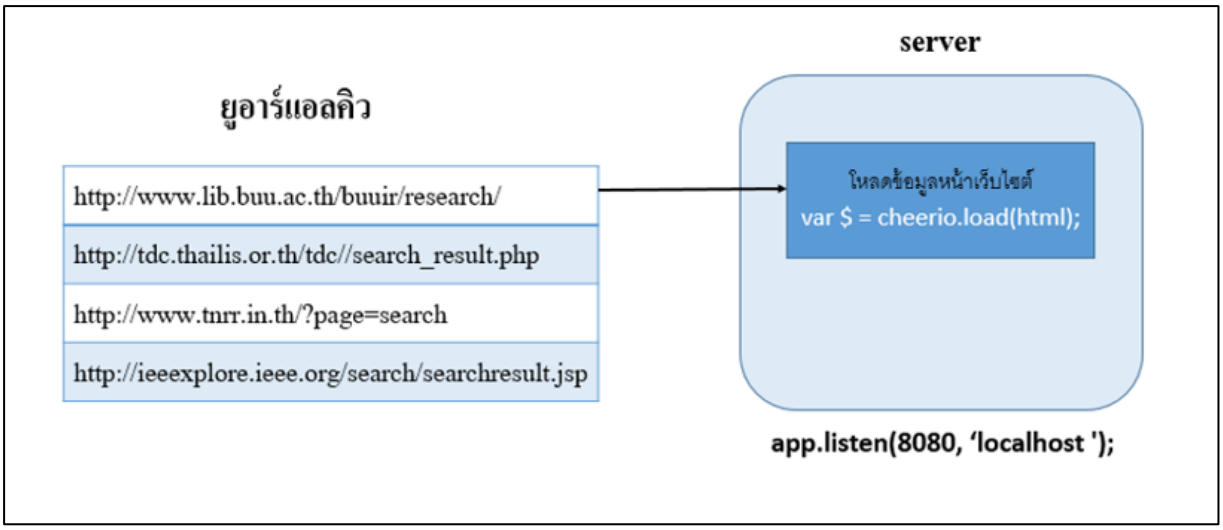
การออกแบบขั้นตอนกระบวนการทำงานของเว็บครอเลอร์ เพื่อการสกัดข้อมูลงานวิจัยสามารถแบ่งขั้นตอนได้ ดังต่อไปนี้

1. จัดการยูอาร์แอลคิว คือ การเก็บรวบรวมยูอาร์แอลเว็บไซต์ฐานข้อมูลงานวิจัยที่ต้องการสกัดข้อมูลและจัดลำดับยูอาร์แอลทั้งหมด ดังภาพที่ 3-8



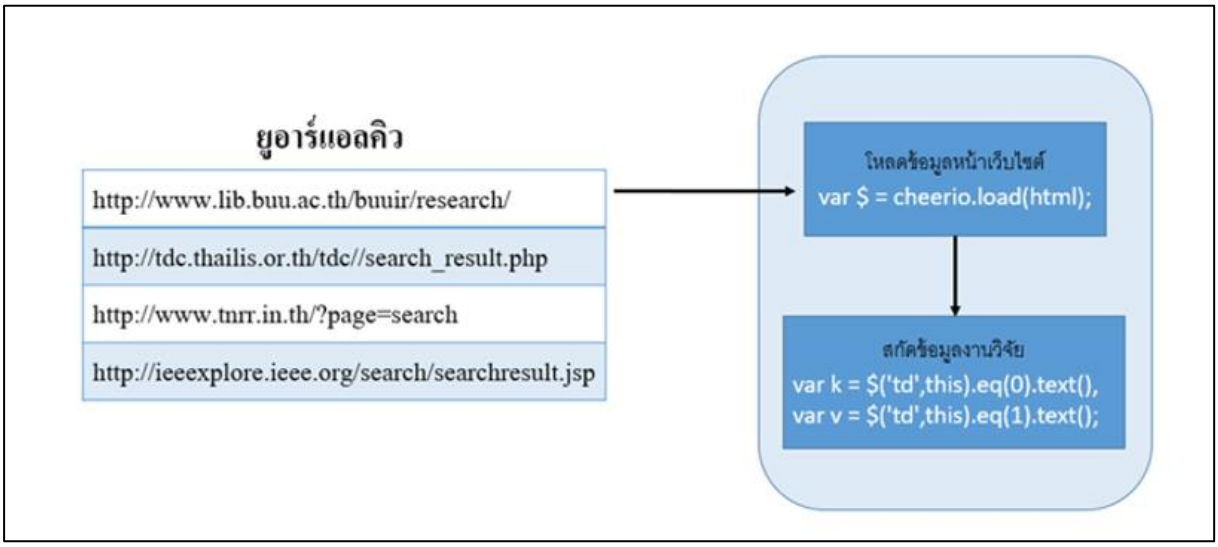
ภาพที่ 3-8 ยูอาร์แอลคิวเว็บไซต์งานวิจัย

2. โหลดข้อมูลหน้าเว็บไซต์ เป็นกระบวนการดึงข้อมูล HTML เว็บไซต์งานวิจัยจากยูอาร์แอล ซึ่งในงานวิจัยนี้ใช้ Cheerio ในการโหลดข้อมูล HTML เพื่อหา DOM Element หรือในหน้า HTML นั้น ๆ ดังภาพที่ 3-9



ภาพที่ 3-9 การโหลดข้อมูล HTML จากยูอาร์แอลคิว

3. สกัดข้อมูลงานวิจัย เป็นกระบวนการกำหนดส่วนของข้อมูลงานวิจัยที่ต้องการจากการวิเคราะห์โครงสร้าง HTML และสกัดข้อมูลตามคิวของยูอาร์แอลโดยใช้คำสั่ง var v = $('td',this).eq(1).text() ; ทำการวนลูปเก็บข้อมูลที่ต้องการในทุก ๆ แถวในตาราง ดังภาพที่ 3-10



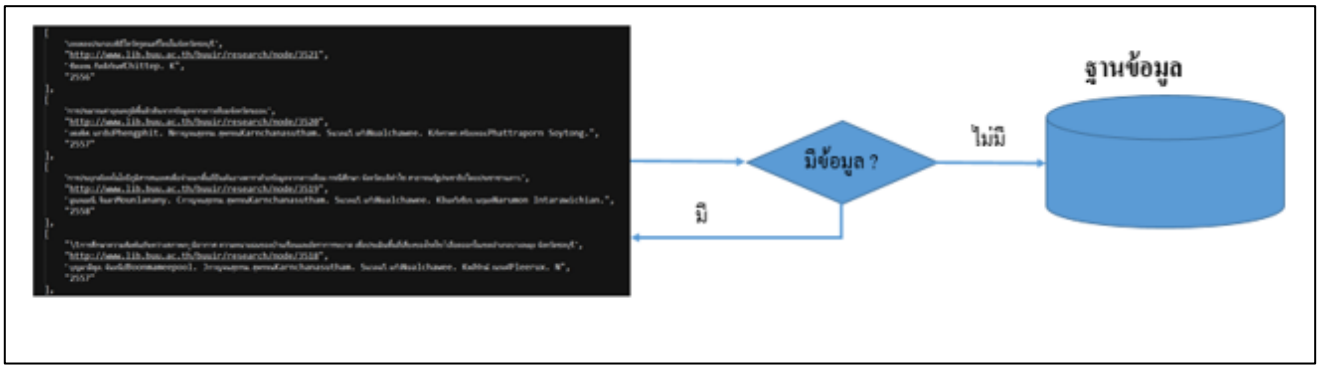
ภาพที่ 3-10 การสกัดข้อมูลงานวิจัยจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา

4. Output ข้อมูลงานวิจัย ผลลัพธ์ที่ได้จากขั้นตอนการสกัดข้อมูล ได้แก่ ชื่องานวิจัย ยูอาร์แอลงานวิจัย ชื่อผู้เขียน ปีที่เขียน จะถูกจัดเก็บในรูปแบบ Json ซึ่งมีโครงสร้าง ดังภาพที่ 3-11


```
{
    "Title": "ความพึงพอใจของอาจารย์ต่อการจัดหาทรัพยากรสารสนเทศของสำนักหอสมุด มหาวิทยาลัยบูรพา.",
    "Publication Type": "Research",
    "Year of Publication": "2549",
    "Authors": {
        "นิตยา ปานเพชร.": "http://www.lib.buu.ac.th//buuir/research/biblio?f[author]=1",
        "Nittaya Panpetch.": "http://www.lib.buu.ac.th//buuir/research/biblio?f[author]=928"
    },
    "Institution": "สำนักหอสมุด มหาวิทยาลัยบูรพา.",
    "City": "ชลบุรี.",
    "Type": "งานวิจัย",
    "ISBN Number": "9745029629",
    "Call Number": "025.2 น577ค",
    "Keywords": "การจัดหาทรัพยากรห้องสมุด., การเลือกหนังสือ., ทรัพยากรสารสนเทศ - - การจัดการ., มหาวิทยาลัยบูรพา. สำนักหอสมุด., สาขาเทคโนโลยีสารสนเทศและนิเทศศาสตร์., อาจารย์ - - ความพอใจของ
    "Abstract": "การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาความพึงพอใจของอาจารย์ที่มีต่อการจัดหาทรัพยากรสารสนเทศของสำนักหอสมุด มหาวิทยาลัยบูรพา จำแนกตามภาควิชา/ คณะที่อาจารย์สังกัด และคุ
    "Alternate Title": "Satisfaction of faculty members of Burapha University Library.",
    "link": "http://www.lib.buu.ac.th/buuir/research/node/1"
},
```

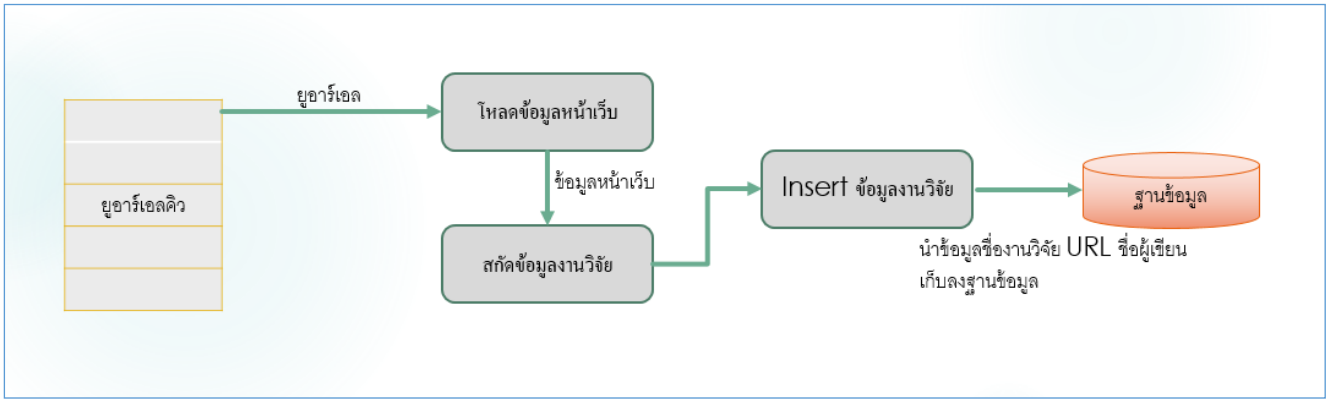

ภาพที่ 3-11 ผลลัพธ์งานวิจัยที่ได้จากการสกัดข้อมูล

5. จัดเก็บข้อมูลงานวิจัย เป็นกระบวนการบันทึกข้อมูลที่ได้จากขั้นตอนการสกัดข้อมูลงานวิจัย ซึ่งก่อนที่จะเพิ่มข้อมูลงานวิจัยในฐานข้อมูลจะทำการตรวจสอบว่าข้อมูลงานวิจัยนั้น ๆ มีในฐานข้อมูลหรือไม่ กรณีที่ไม่มีในฐานข้อมูลงานวิจัยเท่านั้น จึงจะถูกจัดเก็บ เพื่อป้องกันความซ้ำซ้อนของข้อมูล ดังภาพที่ 3-12



ภาพที่ 3-12 การจัดเก็บข้อมูลงานวิจัยในฐานข้อมูล

จากขั้นตอนข้างต้นเป็นการทำงานของเว็บครอว์เลอร์สำหรับสกัดข้อมูลงานวิจัย ซึ่งภาพรวมขั้นตอนทั้งหมดแสดงดังภาพที่ 3-13 หลังจากทำทุกขั้นตอนทั้งหมดจะได้ฐานข้อมูลงานวิจัยเพื่อนำข้อมูลที่ได้ ไปพัฒนาเว็บไซต์ระบบฐานข้อมูลงานวิจัยต่อไป



ภาพที่ 3-13 ภาพรวมกระบวนการทำงานเว็บครอว์เลอร์

3.2.4 การออกแบบฐานข้อมูลงานวิจัย

ฐานข้อมูลงานวิจัยสำหรับจัดเก็บข้อมูลที่ได้จากขั้นตอนการสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์ ประกอบด้วย 2 ตารางหลัก ดังต่อไปนี้

1. ตารางรายละเอียดงานวิจัย

ตารางรายละเอียดงานวิจัยจัดเก็บข้อมูลพื้นฐานงานวิจัยที่จำเป็นสำหรับนำไปพัฒนาเว็บไซต์ค้นหางานวิจัย มีรายละเอียดดังนี้

ตารางที่ 3-1 ตารางรายละเอียดงานวิจัย (Research)

ลำดับที่	ชื่อฟิวล์	คำอธิบาย
1.	Research_id	คีย์หลักของงานวิจัย
2.	Title	หัวข้อของงานวิจัย
3.	Link	ยูอาร์แอลของงานวิจัย
4.	Year	ปีที่เขียนของงานวิจัย
5.	Status	สถานะของงานวิจัย
6.	CreateBy	สร้างโดย(ผู้วิจัย)
7.	CreateDt	วันที่สร้างงานวิจัย

2. ตารางชื่อผู้เขียนงานวิจัย (Author)

ตารางชื่อผู้เขียนงานวิจัยจัดเก็บชื่อผู้เขียนงานวิจัยนั้น ๆ ซึ่งจัดเก็บแยกจากรายละเอียดงานวิจัย เนื่องจากงานวิจัยหนึ่งสามารถมีผู้เขียนได้หลายท่าน มีรายละเอียดดังนี้

ตารางที่ 3-2 ตารางผู้เขียนงานวิจัย

ลำดับที่	ชื่อฟิวล์	คำอธิบาย
1.	Author_id	คีย์หลักผู้เขียนของงานวิจัย
2.	Research_id	คีย์หลักของงานวิจัย
3.	FullName	ชื่อของผู้เขียนงานวิจัย
4.	CreateBy	สร้างโดย(ผู้วิจัย)
5.	CreateDt	วันที่สร้างงานวิจัย

3.3 การกำหนดแบบแผนการวัดประสิทธิผล

การกำหนดแบบแผนการวัดประสิทธิผลประกอบด้วยขั้นตอน ดังต่อไปนี้

3.3.1 การสร้างเว็บครอว์เลอร์เพื่อวัดประสิทธิผล

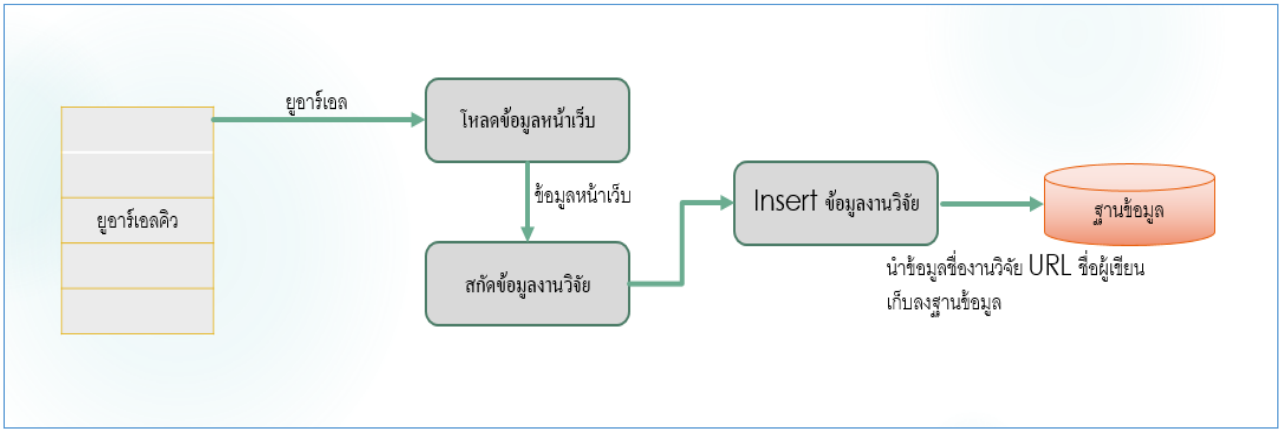
ส่วนประกอบของเว็บครอว์เลอร์ดังภาพที่ 3-14 สามารถอธิบายส่วนประกอบได้ ดังนี้

1. ยูอาร์แอลคิว คือ ยูอาร์แอลทั้งหมดที่ต้องการให้เว็บครอว์เลอร์ดาวน์โหลด ซึ่งถูกจัดเรียงตามลำดับที่ผู้จัดทำงานวิจัยกำหนด

2. ยูอาร์แอลเริ่มต้น คือ ยูอาร์แอลที่เว็บครอว์เลอร์กำลังเข้าไปดาวน์โหลดข้อมูล

3. สกัดข้อมูล คือ กระบวนการคัดเลือกข้อมูลส่วนที่ต้องการ เพื่อนำไปเก็บในฐานข้อมูล

4. ฐานข้อมูล คือ ที่เก็บรวบรวมข้อมูลงานวิจัยทั้งหมดที่ได้จากกระบวนการสกัดข้อมูล

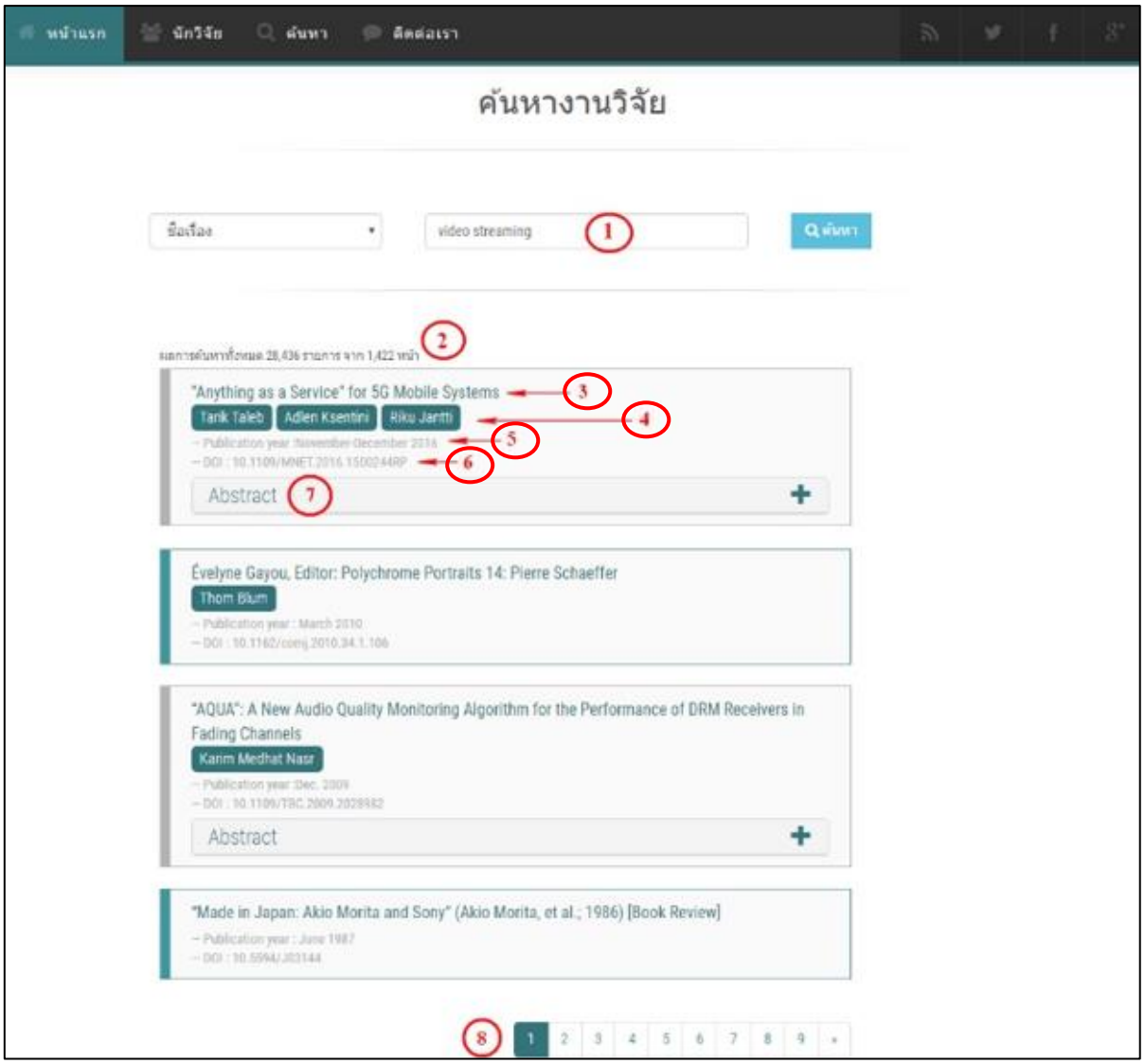


ภาพที่ 3-14 องค์ประกอบเว็บครอว์เลอร์

ดังนั้น จากภาพที่ 3-14 งานนิพนธ์นี้จะดำเนินตามกระบวนการทั้งหมดและนำข้อมูลงานวิจัยที่ได้แสดงผลที่หน้าเว็บไซต์ที่ผู้จัดทำงานนิพนธ์พัฒนาขึ้น เพื่อใช้ในการวัดประสิทธิผลของขั้นตอนการทำงานของเว็บครอว์เลอร์

3.3.2 การสร้างเว็บแสดงผลที่ได้จากการสกัดข้อมูลงานวิจัยของเว็บครอว์เลอร์

จากกระบวนการสกัดข้อมูลงานวิจัยที่พัฒนาขึ้นมานั้น สามารถสกัดข้อมูลงานวิจัยซึ่งประกอบด้วย ชื่องานวิจัย ยูอาร์แอลงานวิจัย ชื่อผู้เขียน ยูอาร์แอลผู้เขียน ปีที่เขียนงานวิจัย และบทคัดย่อ เพื่อให้ง่ายต่อการวัดประสิทธิผลและตรวจสอบความถูกต้องของข้อมูลที่สกัดมา ผู้จัดทำงานนิพนธ์จึงได้พัฒนาเว็บไซต์สำหรับแสดงข้อมูลงานวิจัยที่ได้จากการสกัดข้อมูล ดังภาพที่ 3-15



ภาพที่ 3-15 หน้าจอเว็บไซต์ค้นหางานวิจัย

จากภาพที่ 3-15 หมายเลขที่กำหนด สามารถอธิบายส่วนต่าง ๆ ของหน้าเว็บไซต์ได้ดังต่อไปนี้

1. หมายเลข 1 คือ ช่องสำหรับใส่คำค้นหางานวิจัย ซึ่งสามารถค้นหาได้จากชื่องานวิจัยและชื่อผู้เขียน
2. หมายเลข 2 แสดงจำนวนผลลัพธ์งานวิจัยที่ค้นหาได้ทั้งหมด จากคำค้นที่ผู้ใช้งานระบุ
3. หมายเลข 3 แสดงชื่อของงานวิจัย
4. หมายเลข 4 แสดงชื่อผู้เขียนงานวิจัยนี้ โดยแสดงผู้เขียนทั้งหมด
5. หมายเลข 5 แสดงปีที่เขียนงานวิจัย

6. หมายเลข 6 แสดงเลข DOI กรณีงานวิจัยนั้นมาจากเว็บไซต์ IEEE
7. หมายเลข 7 แสดงบทคัดย่อ โดยผู้ใช้งานจะต้องกดที่ปุ่ม “Abstract”
8. หมายเลข 8 คือ จำนวนหน้าผลลัพธ์ข้อมูลงานวิจัย

3.4 การวัดประสิทธิผล

การสกัดข้อมูลงานวิจัยโดยเว็บครอว์เลอร์ ด้วยการดึงข้อมูลงานวิจัยที่มีทั้งหมดในเว็บไซต์นั้น ๆ มาจัดเก็บในฐานข้อมูลนี้ สามารถวัดประสิทธิผลการสกัดข้อมูลได้ โดยใช้วิธีการวัดค่าประสิทธิภาพพื้นฐานในการค้นหาข้อมูลและความครบถ้วนของข้อมูลที่ต้องการ มีขั้นตอนดังต่อไปนี้

การคำนวณผลการประเมิน

ขั้นตอนนี้เป็นการวัดประสิทธิผลการสกัดข้อมูล โดยใช้ค่าวัดประสิทธิภาพพื้นฐานในการค้นหาข้อมูล (F-Measure) โดยการคำนวณจะใช้สูตร ดังสมการที่ 1

$$F = 2\left(\frac{Precision\ x\ Recall}{Precision+Recall}\right) \qquad (3.1)$$

โดยที่

Precision คือ ค่าที่บ่งบอกถึงอัตราผลลัพธ์ที่ไม่ถูกต้องจากการค้นหา

Recall คือ ร้อยละของสัดส่วนของจำนวนข้อมูล (records) ที่สืบค้นได้ตรงตามความต้องการต่อจำนวนข้อมูลที่ตรงตามความต้องการ

ที่มา : ดร.เอกสิทธิ์ พัชรวงศ์ศักดา. (2014). AN INTRODUCTION TO DATA MINING TECHNIQUES (THAI VERSION)

บทที่ 4
ผลการดำเนินงาน

จากการศึกษาและดำเนินงานนิพนธ์ กรณีศึกษาการสกัดข้อมูลงานวิจัยบนเว็บเพจด้วยเว็บครอว์เลอร์ จากการสกัดข้อมูลหน้าเว็บไซต์ที่มีโครงสร้างแบบ HTML การกำหนดจัดลำดับคิวของยูอาร์แอล การบันทึกไฟล์จากการสกัดข้อมูล และจัดเก็บไว้ในฐานข้อมูล สำหรับการพัฒนาเว็บไซต์งานวิจัย มีผลการดำเนินงานดังต่อไปนี้

4.1 การกำหนดลำดับยูอาร์แอลคิว

กระบวนการทำงานของระบบมีขั้นตอนแรก คือ การกำหนดยูอาร์แอลคิว เนื่องจากตัวอย่างเว็บไซต์งานวิจัยที่นำมาสกัดข้อมูลงานวิจัยทั้ง 3 เว็บไซต์มีจำนวนงานวิจัยทั้งหมดประมาณ 4.5 ล้านรายการ ทำให้การทำงานของเว็บครอว์เลอร์ในการสกัดข้อมูลนั้นใช้เวลาในการทำงานนาน ผู้จัดทำงานนิพนธ์จึงกำหนดลำดับการทำงานและกำหนดเวลาการทำงานในแต่ละยูอาร์แอล โดยใช้คำสั่ง Crontab ในระบบปฏิบัติการ Ubuntu โดยมีรายละเอียดดังตารางที่ 4-1

ตารางที่ 4-1 การกำหนดลำดับยูอาร์แอลคิวและการกำหนดเวลาการสกัดข้อมูล

คำสั่ง	รายละเอียด
0 14 * * 0 /curl http://localhost:8080/crawlerBuuIr	เก็บข้อมูลงานวิจัยจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา ทุกวันอาทิตย์ เวลา 14.00 น.
0 15 * * 0 /curl http://localhost:8080/crawlerBuuTdc	เก็บข้อมูลงานวิจัยจากเว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย ทุกวันอาทิตย์ เวลา 15.00 น.
0 16 * * 0 /curl http://localhost:8080/crawlerTnrr	เก็บข้อมูลงานวิจัยจากเว็บไซต์คลังข้อมูลงานวิจัยไทย ทุกวันอาทิตย์ เวลา 16.00 น.

4.2 การกำหนดคำสั่ง Node.js

เนื่องจากงานนิพนธ์นี้ใช้ภาษา Node.js ในการทำงานฝั่งเครื่องคอมพิวเตอร์แม่ข่ายเพื่อรอรับคำสั่งร้องขอจากเครื่องคอมพิวเตอร์ลูกข่ายด้วยคำสั่งตามตารางที่ 4-1 เพื่อสกัดข้อมูลงานวิจัยและนำผลลัพธ์ที่ได้จัดเก็บในรูปแบบ Json ซึ่งมีขั้นตอนดังต่อไปนี้

1. ติดตั้ง Package คำสั่ง Screen ด้วยคำสั่ง ดังภาพที่ 4-1

```
- ระบบปฏิบัติการ ubuntu
        $ sudo aptitude install screen
 - ระบบปฏิบัติการ centos
        $ yum install screen
```

ภาพที่ 4-1 ตัวอย่างคำสั่งติดตั้ง Package คำสั่ง Screen

2. เรียกใช้คำสั่ง Screen ดังภาพที่ 4-2

```
$ screen -S ชื่อไฟล์
เช่น
$ screen -S Webcrawler.js
```

ภาพที่ 4-2 ตัวอย่างการเรียกใช้คำสั่ง Screen

4.3 ขั้นตอนการสกัดข้อมูลงานวิจัย

หลังจากขั้นตอนกำหนดยูอาร์แอลคิวตามตารางที่ 4-2 เมื่อมีคำสั่งร้องขอจากเครื่องคอมพิวเตอร์ลูกข่าย เครื่องคอมพิวเตอร์แม่ข่ายจะทำการโหลด HTML ตามคิวที่กำหนด ซึ่งมีขั้นตอน ดังต่อไปนี้

1. ขั้นตอนโหลดโหนด HTML ด้วยคำสั่ง Cheerio

ขั้นตอนนี้จะทำการสร้างออฟเจ็คใน Cheerio เพื่อโหลดข้อมูล HTML จากนั้นสามารถใช้คำสั่ง Method ในการจัดการเกี่ยวกับ Element ใน DOM HTML ยกตัวอย่างการสกัดข้อมูลงานวิจัยเว็บไซต์คลังข้อมูลงานวิจัยไทย ด้วยคำสั่ง var $ = cheerio.load(html); ได้ข้อมูล HTML ดังภาพที่ 4-3

```
<div class="tasks">
    <div class="task task">
        <h4><a href="?page=result_search&amp;record_id=10051018">การพัฒนาการผลิตถุงเพาะ ถุงปลูก และการประยุกต์ใช้สำหรับพื้นที่ปลูกทางการเกษตรในประเทศไทย</a></h4>
        <div class="tmeta"><i class="icon-pushpin"></i> ฐานข้อมูลโครงสร้างพื้นฐานภาครัฐด้านวิทยาศาสตร์และเทคโนโลยี <i class="icon-tag"></i> 2562</div>
    </div>
    <div class="task important">
        <h4><a href="?page=result_search&amp;record_id=10156166">แผนงานวิจัยเครื่องจักรกลอัตโนมัติสำหรับอ้อย</a></h4>
        <div class="tmeta"><i class="icon-pushpin"></i> กรมวิชาการเกษตร ศูนย์เทคโนโลยีสารสนเทศและการสื่อสาร <i class="icon-tag"></i> 2562</div>
    </div>
    <div class="task cool">
        <h4><a href="?page=result_search&amp;record_id=77107">การบูรณาการฐานข้อมูลและการถ่ายทอดเทคโนโลยีในการจัดการทรัพยากรน้ำแบบพอเพียง</a></h4>
        <div class="tmeta"><i class="icon-pushpin"></i> สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์ <i class="icon-tag"></i> 2561</div>
    </div>
    <div class="task task">
        <h4><a href="?page=result_search&amp;record_id=87684">การรณรงค์การใช้ทรัพยากรการเรียนรู้อิเล็กทรอนิกส์เพื่อป้องกันความรุนแรงสำหรับเด็กวัยรุ่น (ต่อเนื่อง)</a></h4>
        <div class="tmeta"><i class="icon-pushpin"></i> สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์ <i class="icon-tag"></i> 2561</div>
    </div>
    <div class="task important">
        <h4><a href="?page=result_search&amp;record_id=93224">การศึกษาและพัฒนาระบบโทรมาตรสมบูรณ์ในลุ่มแม่น้ำเจ้าพระยา</a></h4>
        <div class="tmeta"><i class="icon-pushpin"></i> สถาบันวิจัยและพัฒนาแห่งมหาวิทยาลัยเกษตรศาสตร์ <i class="icon-tag"></i> 2561</div>
    </div>
</div>
```

ภาพที่ 4-3 ตัวอย่างข้อมูล DOM HTML จากการใช้คำสั่ง Cheerio

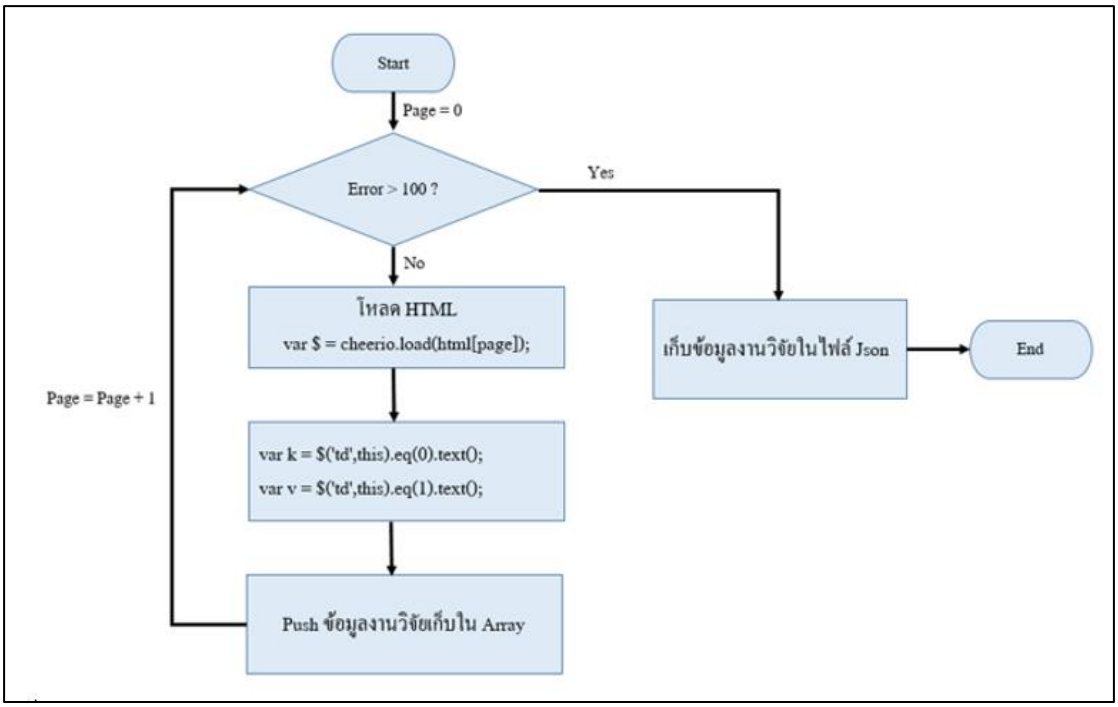
2. ขั้นตอนการเลือกโหนดหรือ Element ที่ใช้ในการสกัดข้อมูลงานวิจัย

จากตัวอย่างโครงสร้างข้อมูล DOM HTML ในภาพที่ 4-3 จะนำไปสู่ขั้นตอนต่อไป ขั้นตอนนี้จะเป็นการเลือก Element ที่ใช้ในการสกัดข้อมูลงานวิจัย ข้อมูลงานวิจัยที่ต้องการจากเว็บไซต์คลังข้อมูลงานวิจัยไทยจะอยู่ในคลาสแม่ที่ชื่อว่า “biblio-node” ซึ่งประกอบด้วยข้อมูลดังต่อไปนี้

- Tag “TR” ลำดับที่ 1 คือ โหนดลูกที่เก็บชื่องานวิจัย
- Tag “TR” ลำดับที่ 2 คือ โหนดลูกที่เก็บชนิดที่เผยแพร่งานวิจัย
- Tag “TR” ลำดับที่ 3 คือ โหนดลูกที่เก็บปีที่เผยแพร่งานวิจัย
- Tag “TR” ลำดับที่ 4 คือ โหนดลูกที่เก็บชื่อผู้เขียนงานวิจัยทั้งหมด
- Tag “a” คือ โหนดลูกที่เก็บชื่อผู้เขียนงานวิจัยและยูอาร์แอลประวัติผู้เขียนงานวิจัย

3. ขั้นตอนสกัดข้อมูลงานวิจัย

หลังจากทำขั้นตอนที่ 2 เสร็จสิ้น ขั้นตอนถัดไปคือ การสกัดข้อมูลการวิจัยซึ่งขั้นตอนนี้จะทำการวนลูปเพื่อสกัดข้อมูลงานวิจัยตาม Element ที่กำหนดในข้อที่ 2 ทุก ๆ แถวในตารางทำกระบวนการในข้อ 2 กับทุก ๆ ยูอาร์แอลที่เกี่ยวข้องจนกว่า หน้าเว็บเพจที่โหลดมาไม่มีข้อมูลที่ต้องการมากกว่า 100 เว็บเพจ จะสิ้นสุดการทำงาน จากนั้นนำข้อมูลงานวิจัยที่สกัดได้บันทึกลงในไฟล์ Json ซึ่งแสดงได้ดังภาพที่ 4-4



ภาพที่ 4-4 ขั้นตอนการสกัดข้อมูลจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา

4.4 ขั้นตอนการจัดเก็บข้อมูลงานวิจัยลงฐานข้อมูล

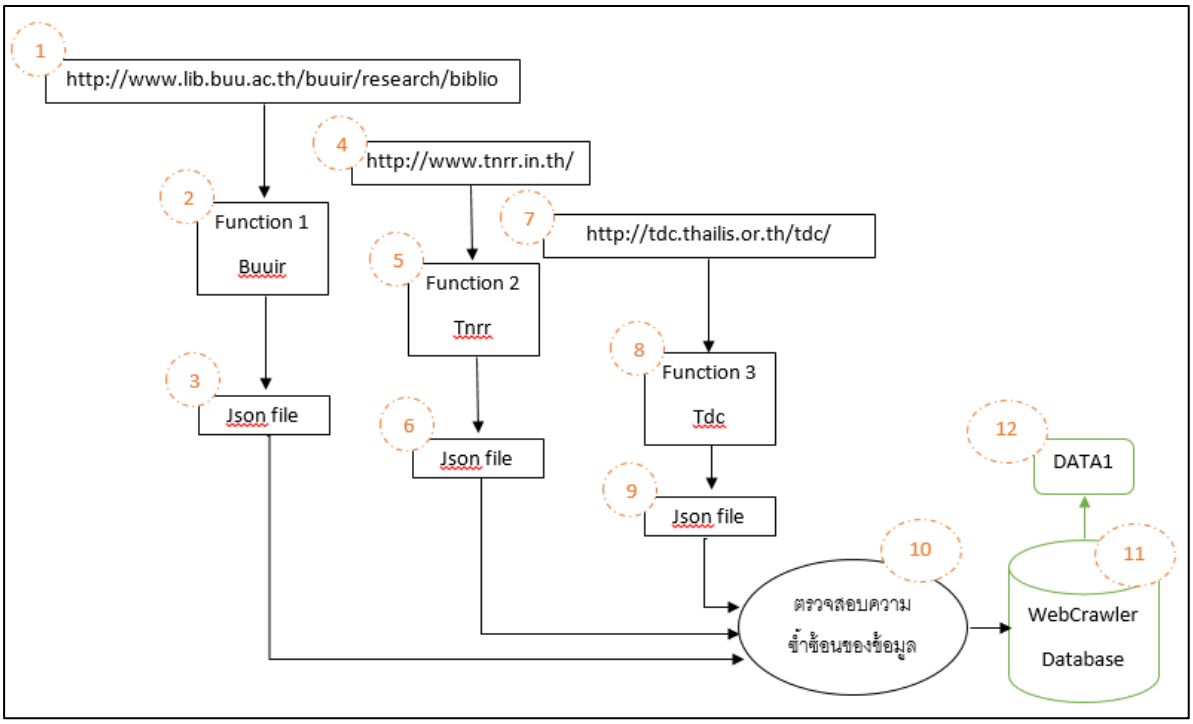
หลังจากดำเนินการสกัดข้อมูลงานวิจัยมาได้แล้ว ผลลัพธ์งานวิจัยที่ได้จะถูกจัดเก็บในไฟล์ Json ซึ่งในหัวข้อนี้ จะนำข้อมูลงานวิจัยที่ได้จากการสกัดข้อมูลมาจัดเก็บลงในฐานข้อมูลที่ได้ออกแบบไว้ในบทที่ 3 หัวข้อที่ 3.2.3 ก่อนจะบันทึกข้อมูลงานวิจัยลงในฐานข้อมูลนั้น จะต้องตรวจสอบความซ้ำซ้อนของข้อมูล ดังภาพในบทที่ 3 ภาพที่ 3-12 การสกัดข้อมูลจากเว็บไซต์นี้สามารถสรุปขั้นตอนการสกัด ได้ดังภาพที่ 4-4

4.5 สรุปขั้นตอนการดำเนินงาน

การดำเนินงานในงานนิพนธ์นี้สามารถแบ่งขั้นตอนหลักออกเป็น 2 ส่วนหลัก คือส่วนการสกัดข้อมูล และการค้นหาข้อมูลงานวิจัย โดยการทำงานของส่วนโปรแกรมการค้นหาข้อมูลงานวิจัยนั้น จะใช้ฐานข้อมูลของผู้จัดทำงานนิพนธ์ ร่วมกับฐานข้อมูลของเว็บไซต์ IEEE โดยมีภาพรวมการทำงานของระบบทั้งหมด และสามารถแสดงความสัมพันธ์ระหว่างส่วนของการสกัดข้อมูล และส่วนของการค้นหา ได้ดังนี้

1. ส่วนขั้นตอนการสกัดข้อมูล

ขั้นตอนนี้ดำเนินการตามกระบวนการในข้อที่ 4.1 การกำหนดลำดับยูอาร์แอลคิว ข้อที่ 4.2 การกำหนดคำสั่ง Node.js และ ข้อที่ 4.3 ขั้นตอนการสกัดข้อมูลงานวิจัย ซึ่งภาพรวมทั้งระบบทั้งหมดแสดงดังภาพที่ 4-5 และมีการทำงานระบุตามหมายเลขที่ 1-12 ดังต่อไปนี้



ภาพที่ 4-5 ขั้นตอนการสกัดข้อมูล

1. ขั้นตอนการสกัดข้อมูลจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา (หมายเลขที่ 1) ด้วยการเรียกใช้ฟังก์ชันที่ 1 (หมายเลขที่ 2) ซึ่งผลลัพธ์ที่ได้จะจัดเก็บในไฟล์ outputBuuir.json (หมายเลขที่ 3)

2. ขั้นตอนการสกัดข้อมูลจากเว็บไซต์คลังข้อมูลงานวิจัยไทย (หมายเลขที่ 4) ด้วยการเรียกใช้ฟังก์ชันที่ 2 (หมายเลขที่ 5) ซึ่งผลลัพธ์ที่ได้จะจัดเก็บในไฟล์ outputTnrr.json (หมายเลขที่ 6)

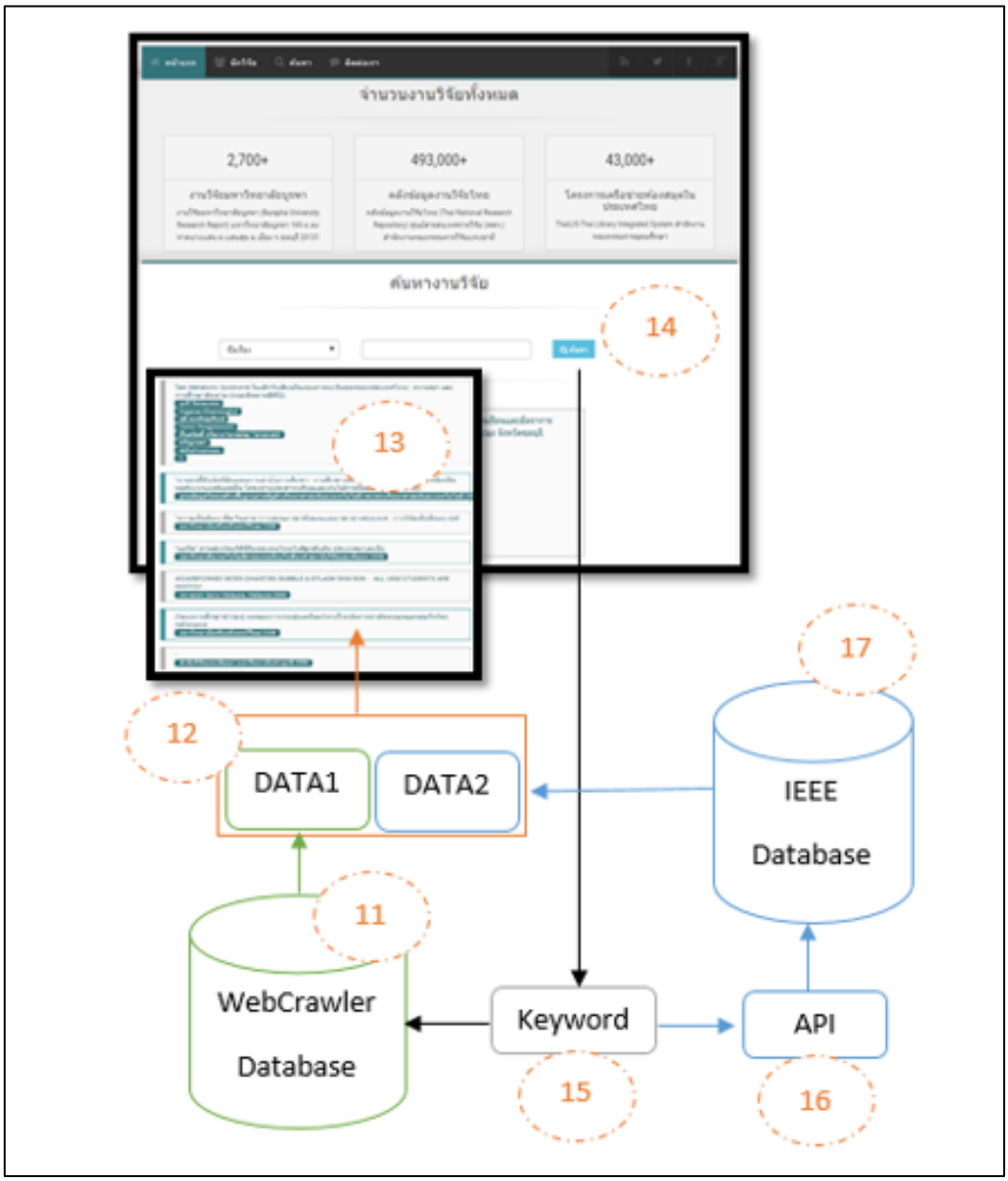
3. ขั้นตอนการสกัดข้อมูลจากเว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย (หมายเลขที่ 7) ด้วยการเรียกใช้ฟังก์ชันที่ 3 (หมายเลขที่ 8) ซึ่งผลลัพธ์ที่ได้จะจัดเก็บในไฟล์ outputTdc.json (หมายเลขที่ 9)

4. ขั้นตอนการจัดเก็บข้อมูลงานวิจัยที่ได้จากการสกัดข้อมูลทั้ง 3 เว็บไซต์ลงในฐานข้อมูล ซึ่งก่อนบันทึกลงฐานข้อมูล Web Crawler Database (หมายเลขที่ 11) จะต้องตรวจสอบความซ้ำซ้อนของข้อมูลก่อน (หมายเลขที่ 10)

5. ไฟล์ข้อมูลที่ได้จากฐานข้อมูลของผู้จัดทำงานนิพนธ์ (หมายเลขที่ 12)

2. ส่วนขั้นตอนการค้นหาข้อมูลงานวิจัย

หลังจากที่ข้อมูลได้ถูกสกัดจากกระบวนการครอว์เลอร์แล้ว ลำดับต่อไปข้อมูลถูกจัดเก็บลงฐานข้อมูล WebCrawler Database เรียบร้อยแล้ว เมื่อผู้ใช้งานต้องการใช้โปรแกรมค้นหางานวิจัย จะสามารถเข้าใช้ได้จาก Internet Web browsers ต่าง ๆ เช่น Internet Explorer Google Chrome Safari Mozilla Firefox เป็นต้น เมื่อเข้าสู่เว็บไซต์ค้นหางานวิจัยแล้ว ให้ผู้ใช้งานระบุคำค้นหาเพื่อค้นหางานวิจัยที่ต้องการ ระบบจะดำเนินการหางานวิจัยตามคำค้นหาที่ระบุ การทำงานระบุตามหมายเลขที่ 11-17 ซึ่งมีขั้นตอนแสดงตามภาพที่ 4-6 ดังต่อไปนี้



ภาพที่ 4-6 การค้นหาข้อมูลงานวิจัย

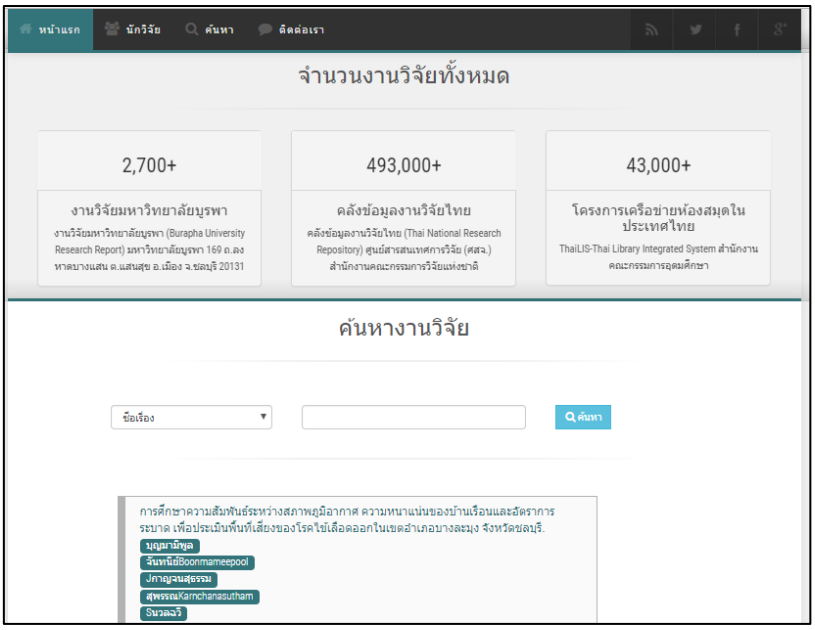
1. นำคำค้นที่ระบุ (Key word) กรอกที่ช่องค้นหางานวิจัยในเว็บไซต์ และกดปุ่มค้นหา (หมายเลข 14) ไปค้นหางานวิจัยที่ตรงกับคำค้นในฐานข้อมูล Web Crawler Database (หมายเลข 11) ผลลัพธ์ที่ได้ คือ Data 1

2. นำคำค้นที่ระบุ (Key word) กรอกที่ช่องค้นหางานวิจัยในเว็บไซต์ และกดปุ่มค้นหา (หมายเลข 14) ไปร้องขอข้อมูลจากเว็บไซต์ IEEE โดยผ่าน API (หมายเลข 16) ผลลัพธ์ที่ได้ คือ Data 2

3. นำข้อมูล Data 1 และ Data 2 มารวมกันและจัดเรียงข้อมูลตามชื่องานวิจัย (หมายเลข 12)

4. นำข้อมูลที่จัดเรียงเรียบร้อยแสดงผลลัพธ์บนหน้าเว็บเพจ (หมายเลข 13)

เมื่อนำทั้ง 2 ส่วนมารวมกัน คือ ส่วนขั้นตอนการสกัดข้อมูล และส่วนขั้นตอนการค้นหาข้อมูลงานวิจัย จะเกิดระบบใหม่ที่เรียกว่า “โปรแกรมค้นหางานวิจัย” ที่มีข้อมูลของงานวิจัยทั้งหมด 4 เว็บไซต์ใช้เป็นฐานข้อมูลสำหรับการสนับสนุน คือ การสกัดข้อมูลจากเว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา การสกัดข้อมูลจากเว็บไซต์คลังข้อมูลงานวิจัยไทย การสกัดข้อมูลจากเว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย สำหรับการสกัดข้อมูลจากเว็บไซต์ IEEE นั้นไม่สามารถทำด้วยเทคนิคการสกัดด้วย คำสั่ง Cheerio อันด้วยมาจากสาเหตุที่ โครงสร้างเว็บไซต์ ของเว็บไซต์ IEEE นี้ มีโครงสร้างที่ไม่ใช่แบบ HTML จะเป็นโครงสร้างที่มีการป้องกันการสกัดข้อมูลจากภายนอก ในการที่จะดึงข้อมูลของเว็บไซต์ IEEE นี้ ทางผู้ดำเนินงานนิพนธ์ได้ดำเนินการขออนุญาตไปทางผู้ให้บริการเว็บไซต์ IEEE เพื่อขอการสกัดข้อมูลและนำข้อมูลมาแสดงตามที่ผู้ใช้งานระบุตามคำค้นหา รายละเอียดของภาพรวมการทำงาน แสดงดังภาพที่ 4-7 ภาพตัวอย่างหน้าเว็บไซต์โปรแกรมค้นหางานวิจัย และภาพที่ 4-8 ภาพรวมกระบวนการสกัดข้อมูลและการแสดงผลการค้นหางานวิจัย



ภาพที่ 4-7 ภาพตัวอย่างหน้าเว็บไซต์โปรแกรมค้นหางานวิจัย

4.6 ผลการทดลอง

ผลลัพธ์จากการทดลองสกัดข้อมูลงานวิจัยที่ได้จากขั้นตอนการออกแบบและวิธีการสกัดข้อมูลงานวิจัย เมื่อทดสอบสกัดข้อมูลงานวิจัยทั้ง 3 เว็บไซต์สามารถสรุปเวลาที่ใช้ในการสกัดข้อมูลงานวิจัยได้ ดังตารางที่ 4-2

ตารางที่ 4-2 ตารางระยะเวลาที่ใช้ในการสกัดข้อมูล

เว็บไซต์งานวิจัย	จำนวนงานวิจัย (งานวิจัย)	เวลาที่ใช้ (นาที)
งานวิจัยมหาวิทยาลัยบูรพา	6,880	4
คลังข้อมูลงานวิจัยไทย	493,207	1,370
โครงการเครือข่ายห้องสมุดในประเทศไทย	43,608	60
รวมเว็บไซต์ด้านงานวิจัย	543,695	1,434

จากตารางที่ 4-2 เว็บไซต์ที่ใช้เวลาในการสกัดข้อมูลน้อยที่สุด คือ เว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา ใช้เวลาทั้งสิ้น 4 นาที และเว็บไซต์ที่ใช้ระยะเวลาในการสกัดข้อมูลนานที่สุด คือ เว็บไซต์คลังข้อมูลงานวิจัยไทย ใช้เวลาทั้งสิ้น 22 ชั่วโมง เนื่องจากมีจำนวนงานวิจัยมากที่สุดและโครงสร้าง HTML มีความซับซ้อน

หลังจากการทดลองเก็บข้อมูลงานวิจัยทั้ง 3 เว็บไซต์ที่มีจำนวนงานวิจัยทั้งหมด 543,695 รายการ มีจำนวนข้อมูลงานวิจัยที่มีอยู่ในเว็บไซต์แต่ไม่ถูกสกัดจำนวน 3,487 รายการ เนื่องจากเว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย มีโครงสร้าง HTML ที่ได้จากการโหลดข้อมูลไม่รองรับภาษาไทยจึงทำให้ข้อมูลไม่ครบถ้วน ดังตารางที่ 4-3

ตารางที่ 4-3 ตารางจำนวนการสกัดข้อมูลงานวิจัย

เว็บไซต์งานวิจัย	จำนวนงานวิจัย ที่สกัดได้	จำนวนงานวิจัย ที่สกัดไม่ได้
งานวิจัยมหาวิทยาลัยบูรพา	6,880	0
คลังข้อมูลงานวิจัยไทย	493,207	2,921
โครงการเครือข่ายห้องสมุดในประเทศไทย	43,608	566
รวมรวมเว็บไซต์ด้านงานวิจัย	543,695	3,487

เมื่อนำข้อมูลทั้งหมดมาหาค่า Precision, Recall และ F-Measure เพื่อหาประสิทธิภาพของการสกัดข้อมูล โดยให้

A แทน จำนวนข้อมูลที่ได้จากการสกัดข้อมูลและตรงตามความต้องการ

B แทน จำนวนข้อมูลที่ได้จากการสกัดข้อมูลแต่ไม่ตรงตามความต้องการ

C แทน จำนวนข้อมูลที่ตรงตามความต้องการในฐานข้อมูลแต่ไม่ได้สกัดมา

Precision คือ ร้อยละของสัดส่วนของจำนวนข้อมูล (records) ที่สกัดได้ตรงตามความต้องการต่อจำนวนข้อมูลงานวิจัยทั้งหมด

$$\text{Precision} = \frac{A}{A+B} \qquad (4.1)$$

$$\text{Precision} = \frac{542{,}189}{542{,}189+1{,}506}$$

$$\text{Precision} = 0.99$$

Recall คือ ร้อยละของสัดส่วนของจำนวนข้อมูล (records) ที่สืบค้นได้ตรงตามความต้องการต่อจำนวนข้อมูลที่ตรงตามความต้องการ

$$\text{Recall} = \frac{A}{A+C} \qquad (4.2)$$

$$\text{Recall} = \frac{542{,}189}{542{,}189+3{,}487}$$

$$\text{Recall} = 0.99$$

F-measure ค่าเอฟเมเชอร์ คือ ค่าวัดประสิทธิภาพพื้นฐานในการค้นหาข้อมูล โดยเป็นการนำเอาค่า Precision และ Recall มาใช้ในการคำนวณ ซึ่งค่า Recall คือ ค่าที่บ่งบอกถึงอัตราที่ถูกต้องจากการค้นหา ส่วนค่า Precision คือ ค่าที่บ่งบอกถึงอัตราผลลัพธ์ที่ไม่ถูกต้องจากการค้นหา

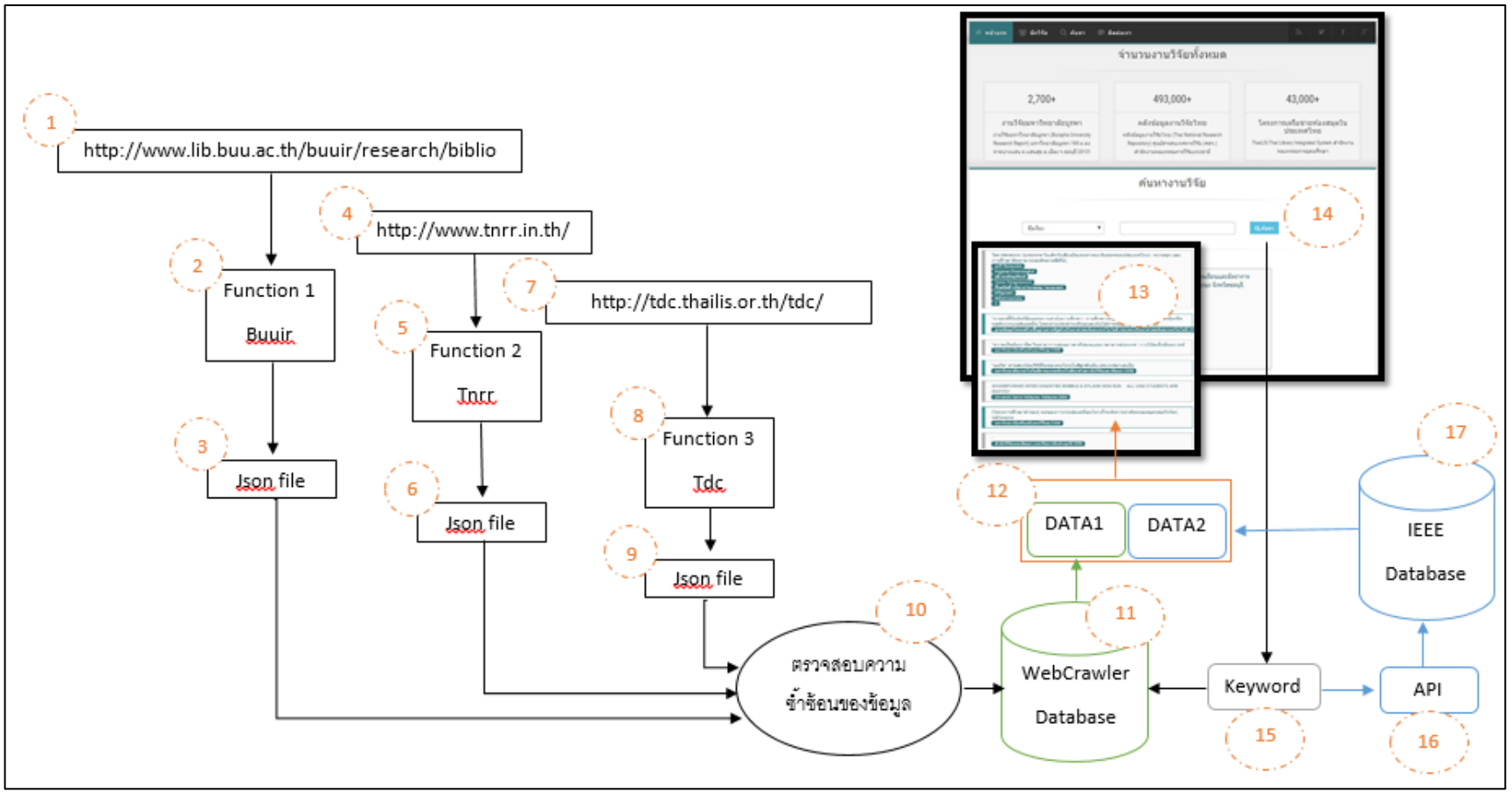
$$F = 2\left(\frac{Precision\ x\ Recall}{Precision+Recall}\right) \qquad (4.3)$$

$$F = 2\left(\frac{0.99x\ 0.99}{0.99+0.99}\right)$$

$$F = 0.99$$

แสดงถึงประสิทธิภาพการค้นหาข้อมูลงานวิจัย ด้วยเว็บครอเลอร์มีค่าเท่ากับ 99%

ที่มา: ดร.เอกสิทธิ์ พัชรวงศ์ศักด. (2014) .AN INTRODUCTION TO DATA MINING TECHNIQUES (THAI VERSION)



ภาพที่ 4-8 ภาพรวมกระบวนการสกัดข้อมูลและการแสดงผลการค้นหางานวิจัย

บทที่ 5
สรุปและอภิปรายผล

งานนิพนธ์นี้นำเสนอขั้นตอนการสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์ จากเว็บไซต์งานวิจัยในประเทศไทย 3 เว็บไซต์ ได้แก่ เว็บไซต์งานวิจัยมหาวิทยาลัยบูรพา เว็บไซต์โครงการเครือข่ายห้องสมุดในประเทศไทย เว็บไซต์คลังข้อมูลงานวิจัยไทย และเว็บไซต์ต่างประเทศ 1 เว็บไซต์ ได้แก่ เว็บไซต์ IEEE Xplore Digital Library ในบทนี้ได้สรุปผลและอภิปรายผลการดำเนินงานทั้งหมดที่นำเสนอไปและข้อเสนอแนะ การค้นคว้าวิจัยในอนาคต

5.1 สรุปผลการดำเนินงาน

การสกัดข้อมูลงานวิจัยด้วยเว็บครอว์เลอร์จากเว็บไซต์ทั้งหมด 3 เว็บไซต์ ผลลัพธ์ที่ได้จากการสกัดข้อมูล ได้แก่ ชื่องานวิจัย ยูอาร์แอลงานวิจัย ชื่อผู้เขียน ปีที่เขียนงานวิจัยและบทคัดย่อ โดยข้อมูลที่ได้จะถูกจัดเก็บในไฟล์ Json ซึ่งข้อมูลงานวิจัยที่ถูกจัดเก็บในไฟล์จะถูกอ่านเพื่อนำไปจัดเก็บในฐานข้อมูลที่ได้ออกแบบไว้ ประกอบด้วย 2 ตาราง ได้แก่ ตาราง Research และตาราง Author ซึ่งขั้นตอนการจัดเก็บข้อมูลงานวิจัยลงในฐานข้อมูลจะตรวจสอบงานวิจัยที่ถูกสกัดได้ว่ามีอยู่ในฐานข้อมูลหรือไม่ กรณีที่ไม่มีในฐานข้อมูลงานวิจัยนั้น ๆ จึงจะถูกจัดเก็บเพื่อป้องกันความซ้ำซ้อนของข้อมูล

หลังจากจัดเก็บงานวิจัยลงในฐานข้อมูลเรียบร้อยแล้ว ผู้ดำเนินงานนิพนธ์ได้ดำเนินการพัฒนาหน้าจอสำหรับค้นหาข้อมูลงานวิจัยที่สกัดได้ เพื่อให้ผู้ใช้งานสามารถค้นหางานวิจัย ด้วยการค้นหาจากชื่องานวิจัยและจากชื่อผู้เขียน เมื่อผู้ใช้กดปุ่มค้นหาจะแสดงผลลัพธ์งานวิจัยทั้งหมดที่ตรงตามเงื่อนไขที่ผู้ใช้งานระบุ โดยที่ผู้ใช้ไม่ต้องค้นหางานวิจัยจากทีละเว็บไซต์ดังเดิม

ผลการดำเนินงานนิพนธ์ วัดประสิทธิภาพได้ดังนี้

1. จำนวนข้อมูลที่สกัดมาได้ เท่ากับ 543,695 รายการ และระบบตรวจสอบการเพิ่มขึ้นแบบอัตโนมัติของข้อมูลทุกสัปดาห์
2. สัดส่วนของจำนวนข้อมูล (records) ที่สกัดได้ตรงตามความต้องการต่อจำนวนข้อมูลงานวิจัยทั้งหมดหรือค่า Precision เท่ากับ 0.99 คิดเป็น 99%
3. ค่าวัดประสิทธิภาพพื้นฐานในการค้นหาข้อมูลประสิทธิภาพการค้นหาข้อมูลงานวิจัยด้วยเว็บครอเลอร์หรือค่า F-measure เท่ากับ 0.99 คิดเป็น 99%

5.2 ข้อเสนอแนะ

จากการดำเนินงานนิพนธ์นี้พบว่า ยังมีประเด็นที่สามารถนำไปพัฒนาต่อหรือปรับปรุงประสิทธิภาพของขั้นตอนวิธีเพิ่มเติม เพื่อให้ได้วิธีการสกัดข้อมูลงานวิจัยที่มีประสิทธิภาพที่ดีขึ้น ซึ่งมีรายละเอียดดังนี้

1. การพัฒนาเว็บครอว์เลอร์ให้สามารถรองรับการสกัดข้อมูลงานวิจัยจากเว็บไซต์ที่มีโครงสร้าง HTML ที่แตกต่างกัน เนื่องจากในงานนิพนธ์นี้ พัฒนาเว็บครอว์เลอร์ขึ้นมาเพื่อรองรับการสกัดข้อมูลงานวิจัยในแต่ละเว็บไซต์นั้น ๆ หรือกล่าวได้ว่า 1 ฟังก์ชั่น ต่อ 1 เว็บไซต์งานวิจัย

2. การแบ่งการทำงานในการสกัดข้อมูลงานวิจัยเพื่อลดเวลาในการสกัดข้อมูล ยกตัวอย่างการสกัดข้อมูลงานวิจัยจากเว็บไซต์คลังข้อมูลงานวิจัยไทย ที่มีงานวิจัยทั้งหมด 493,683 งานวิจัย แบ่งเป็นทั้งหมด 41,141 หน้า ใช้เวลาในการสกัดข้อมูลทั้งหมด 22 ชั่วโมง ซึ่งหากมีการแบ่งการทำงานในการสกัดข้อมูลงานวิจัยออกเป็นส่วนย่อย ๆ จะทำให้ลดระยะเวลาในการสกัดข้อมูลลงได้

3. การพัฒนาเว็บไซต์ค้นหางานวิจัยให้สามารถค้นหาจากปีที่เขียนงานวิจัยหรือค้นหาจากคำค้นสำคัญ เพื่อเพิ่มทางเลือกในการค้นหางานวิจัยให้แก่ผู้ใช้

4. การพัฒนาต่อยอดจากฐานข้อมูลนี้ เช่น ระบบประวัติผลงานวิจัยของบุคคลต่างๆ

5. การพัฒนาเว็บครอว์เลอร์เพื่อให้สามารถสกัดข้อมูลที่ต้องการ จากเว็บไซต์ที่มีโครงสร้าง เว็บไซต์แบบ API เพื่อให้ได้ความครอบคลุมข้อมูลทั้งหมด

บรรณานุกรม

กลยุทธ บพิตร. (2555). *ขั้นตอนและวิธีการสกัดข้อมูลสินค้าบนเว็บเพจสำหรับเว็บครอเลอร์ที่ใช้ในโปรแกรมค้นหา.* วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต, เทคโนโลยีสารสนเทศ, วิศวกรรมเว็บ, มหาวิทยาลัยธุรกิจบัณฑิตย์.

ชาญชัย ศุภอรรถกร. (2554).จัดการฐานข้อมูลด้วย MySQL สำนักพิมพ์: ซิมพลิฟาย, บริษัท ซีเอ็ดยูเคชั่น จำกัด (มหาชน)

รศ. ชาญชัย ศุภอรรถกร. (2560). *สร้างเว็บแอพพลิเคชัน PHP MySQL+AJAX jQuery ฉบับสมบูรณ์* สำนักพิมพ์: ซิมพลิฟาย, บริษัท ซีเอ็ดยูเคชั่น จำกัด (มหาชน)

ณรงค์ ล่ำดี. (2552). *การศึกษาทักษะการใช้โปรแกรมค้นหาของนักศึกษาวิทยาลัยราชพฤกษ์* วิทยาลัยราชพฤกษ์.

เนณุภา สุภเวชย์.(2551). การสร้างเว็บเพจด้วยภาษา HTML กรุงเทพฯ: วิตตี้ กรุ๊ป.

นิรันดร์ อังควัฒนวิทย์. (2545). *การเก็บเว็บเพจแบบเฉพาะเจาะจงหัวเรื่องด้วยเว็บคราวเลอร์แบบเรียนรู้ได้มหาวิทยาลัยเกษตรศาสตร์.* วิทยานิพนธ์วิศวกรรมศาสตร์มหาบัณฑิต, วิศวกรรมศาสตร์, วิศวกรรมคอมพิวเตอร์, มหาวิทยาลัยเกษตรศาสตร์.

บัณฑิต จามรภูติ. (2553). *คัมภีร์ Ubuntu Linux Server เล่ม 1-2* สำนักพิมพ์: บัณฑิต จามรภูติ, บริษัท ซีเอ็ดยูเคชั่น จำกัด (มหาชน)

เอกบิณ ใจแก้วมา. (2559). *ANGULARJS 2 + NODEJS (API) + MONGODB ฉบับ BEGINNER:* ลำปาง: บริษัท ไดร์ซอฟต์ เทคโนโลยี จำกัด.

Chai Phonbopit. (2557). *แชร์เทคนิค ความรู้ และประสบการณ์เกี่ยวกับ Java, Android, JavaScript, Node.js, Angular.js, Phaser.js และ LibGDX รวมถึงเรื่องต่างๆที่เจ้าของบล็อคสนใจ.* วันที่ค้นข้อมูล 3 มิถุนายน 2560,เข้าถึงได้จาก https://devahoy.com/posts/scraping-web-with-nodejs/

Chai Phonbopit. (2014). *สร้าง API ง่ายๆ ด้วย Node.js และ Express.* วันที่ค้นข้อมูล 1 กรกฎาคม 2560, เข้าถึงได้จาก http://www.siamhtml.com/restful-api-with-node-js-and-express/

Eakasit Pacharwongsakda. (2014). *AN INTRODUCTION TO DATA MINING TECHNIQUES (THAI VERSION)*. กรุงเทพมหานคร: เอเชีย ดิจิตอลการพิมพ์.

RapidMiner 7 Operator Reference Manual, RapidMinerGmbH

บรรณานุกรม (ต่อ)

Michael Schrenk. (2015). *Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL.* United states of America : William Pollock.

Marc Wandschneider. (2013). *Learning Node.js a Hands-On Guide to Building Web Applications In JavaScript Addison-Wesley*. Boston: San Francisco: America.

Mike Chen. (2017).*Web Crawler/Spider for NodeJS + server-side jQuery ;-).* วันที่ค้นข้อมูล 1 กรกฎาคม 2560, เข้าถึงได้จาก https://github.com/bda-research/node-crawler

Richard Lawson. (2015). *Web Scraping with Python.* United Kingdom: Packt Publishing Ltd.

Ryan Mitchell. (2013). *Instant Web Scraping with Java Paperback.* United Kingdom: Packt Publishing Ltd.