รายงานการวิจัยฉบับสมบูรณ์

การสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์

Thai Herb Information Extraction from Multiple Websites

รศ.ดร. พรฤดี เนติโสภากุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินจากเงินรายได้ประจำปีงบประมาณ 2556
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

รายงานการวิจัยฉบับสมบูรณ์

การสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์

Thai Herb Information Extraction from Multiple Websites

รศ.ดร. พรฤดี เนติโสภากุล

ได้รับทุนสนับสนุนงานวิจัยจากเงินจากเงินรายได้ประจำปีงบประมาณ 2556

คณะเทคโนโลยีสารสนเทศ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ชื่อโครงการ (ภาษาไทย) การสกัดพืชสมุนไพรไทยจากหลายเว็บไซต์
ชื่อโครงการ (ภาษาอังกฤษ) Thai Herb Information Extraction from Multiple Websites
แหล่งเงิน คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ประจำปีงบประมาณ 2556 จำนวนเงินที่ได้รับการสนับสนุน 50,000 บาท
ระยะเวลาการทำวิจัย ตั้งแต่ 1 ตุลาคม พ.ศ. 2555 ถึง 1 สิงหาคม พ.ศ. 2556
ชื่อ-สกุล หัวหน้าโครงการ
รศ.ดร. พรฤดี เนติโสภากุล
คณะเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

บทคัดย่อ

เว็บไซต์เผยแพร่ข้อมูลพืชสมุนไพร อาจเป็นเว็บไซต์ที่มีโครงสร้างที่แน่นอน หรือเว็บไซต์ที่มีโครงสร้างไม่แน่นอนก็ได้ ดังนั้น กระบวนการสกัดข้อมูลพืชสมุนไพรจากหลายเว็บไซต์ จึงปรับตามลักษณะของเว็บไซต์ดังกล่าว โดยข้อมูลที่ต้องการสกัด ได้แก่ ชื่อพืชสมุนไพร ทั้งชื่อทางการ ชื่อภาษาอังกฤษ ชื่อวิทยาศาสตร์ ชื่อวงศ์ และชื่ออื่นๆ อาการต่างๆ ที่รักษาได้ ส่วนต่างๆ ที่ใช้รักษาอาการ และความเป็นพิษต่อร่างกาย ในการสกัดข้อมูลเบื้องต้น ได้ใช้แท็ก HTML และไฟล์เทมเพลตที่เก็บคำบ่งชี้หัวข้อของเว็บไซต์ในการดึงข้อมูลที่เกี่ยวข้อ จากนั้น ใช้ regular expression ในการสกัดชื่อพืชสมุนไพร และส่วนที่ใช้รักษา ส่วนการสกัดอาการใช้คำบ่งชี้อาการ และชื่ออาการที่ได้จากการเรียนรู้ ผลการสกัดพบว่า ระบบ มีความแม่นยำและความครบถ้วนในการสกัดชื่อพืชสมุนไพร 98-100% มีความแม่นยำในการสกัดส่วนที่ใช้รักษา 95% ความครบถ้วน 92% มีความแม่นยำในการสกัดชื่ออาการ 90% ความครบถ้วน 85% และมีความแม่นยำและความครบถ้วนในการสกัดข้อมูลความเป็นพิษต่อระบบร่างกาย 100%

คำสำคัญ การสกัดข้อมูลบนเว็บไซต์, นิพจน์ปรกติ, พืชสมุนไพรไทย

Research Title: Thai Herb Information Extraction from Multiple Websites

Researcher: Assoc. Prof. Ponrudee Netisopakul, Ph.D.

Faculty: Information Technology. **Department:** Information Technology.

ABSTRACT

Websites containing Thai herb information have different formats; some have certain structure but some have irregular structure. This research aims to extract Thai herb information from multiple websites. The process must be able to apply to each website accordingly. Information needed to extract include Thai herb names; those are an official name, an English name, a scientific name, its family, and other names, list of symptoms it can be used to treat, parts of used, and toxic information. The preparation process uses HTML tags and template files storing topic indicators for each website and each section of information. After extracting the related section, regular expressions are applied to extract names and parts of used. In addition, a symptom indicator word list is used to extract list of symptoms. The list is expanded using new symptom names learned during the process. The experimental result using totally 100 webpages achieves 98-100% precision and recall in extraction Thai herb names, 95% precision and 92% recall in extracting parts of used, 90% precision and 85% recall in extracting symptom names. Finally, the process achieves 100% in extracting toxic information.

Keywords : Thai herb information extraction, Regular expression, Thai herb

กิตติกรรมประกาศ

งานวิจัยเรื่อง การสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ ได้รับทุนสนับสนุนการวิจัยจากสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง จากแหล่งเงินรายได้คณะเทคโนโลยีสารสนเทศประจำปีงบประมาณ พ.ศ.2556 ผู้วิจัยขอขอบพระคุณคณะและสถาบันที่ให้การสนับสนุนทุนวิจัยมา ณ ที่นี้

รศ.ดร. พรฤดี เนติโสภากุล

สารบัญ

หน้า

บทคัดย่อภาษาไทย I

บทคัดย่อภาษาอังกฤษ II

กิตติกรรมประกาศ III

สารบัญ IV

สารบัญตาราง VI

สารบัญภาพ VII

บทที่ 1 บทนำ 1

1.1 ความเป็นมาและความสำคัญของปัญหา 1

1.2 วัตถุประสงค์ของการวิจัย 1

1.3 ขอบเขตของการวิจัย 2

1.4 วิธีดำเนินการวิจัย 2

1.5 ประโยชน์ที่คาดว่าจะได้รับ 2

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 3

2.1 ทฤษฎีที่เกี่ยวข้อง 3

2.2 งานวิจัยที่เกี่ยวข้อง 9

บทที่ 3 วิธีดำเนินการวิจัย 16

3.1 การสำรวจโครงสร้างแท็ก HTML ของเว็บไซต์ที่เกี่ยวข้อง 16

3.2 กระบวนการของการสกัดพืชสมุนไพรไทยจากหลายเว็บไซต์ 19

บทที่ 4 ผลการวิจัย 30

4.1 การวัดประสิทธิภาพการเปรียบเทียบของการสกัดชื่ออาการ
โดยใช้ไฟล์เรียนรู้ชื่ออาการและไม่ใช้ไฟล์เรียนรู้ชื่ออาการ 31

4.1 การวัดประสิทธิภาการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทยที่อยู่ในแต่ละเว็บไซต์ 33

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 36

4.1 สรุปผลการวิจัย 36

4.2 ข้อเสนอแนะ 38

บรรณานุกรม 39

ภาคผนวก 41

ภาคผนวก ก Regular expression **ที่ใช้สกัดข้อมูลพืชสมุนไพรไทย** 42

สารบัญ (ต่อ)

หน้า

ประวัตินักวิจัย 48

สารบัญตาราง

ตารางที่ หน้า

3.1 แท็กที่บ่งชี้ชื่อของหัวข้อพืชสมุนไพรไทย 18

3.2 ตัวอย่าง Regular expression ที่ใช้สกัดข้อมูลพืชสมุนไพรไทย 27

4.1 การสกัดข้อมูลชื่ออาการโดยไม่ใช้ไฟล์เรียนรู้ชื่ออาการ 32

4.2 การสกัดข้อมูลอาการโดยใช้ไฟล์เรียนรู้ชื่ออาการ 32

4.3 จำนวนข้อมูลดิบของข้อมูลพืชสมุนไพรไทยที่ระบบสามารถสกัดได้ 34

4.4 ค่า Precision และ Recall ของการสกัดข้อมูลพืชสมุนไพรไทย 35

ก.1 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของ
มหาวิทยาลัยราชภัฏธนบุรี 43

ก.2 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของ
สำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ
สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี 44

ก.3 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของ
ฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ 45

ก.4 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของ
ฐานข้อมูลพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม 46

ก.5 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของ
ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย 47

สารบัญภาพ

ภาพที่ หน้า

2.1 ตัวอย่างซอร์สโค้ดของเว็บเพจสภาพอากาศบนเว็บไซต์ 3

2.2 ตัวอย่างโครงสร้างต้นไม้ของแท็ก HTML 4

2.3 ตัวอย่างหน้าเว็บเพจของการประชุมวิชาการ 5

2.4 ตัวอย่างซอร์สโค้ดของหน้าเว็บเพจการประชุมวิชาการ 5

2.5 แพ็ตเทิร์นที่ใช้สกัดข้อมูลบนหน้าเว็บเพจประชุมวิชาการ 6

2.6 ผลลัพธ์ของการสกัดข้อมูลของงานประชุมวิชาการ 7

2.7 ตัวอย่างซอร์สโค้ดของเว็บเพจราคาห้องพักในโรงแรม 7

2.8 ตัวอย่างกฎที่ใช้สกัดข้อมูลบนหน้าเว็บเพจราคาห้องพักในโรงแรม 8

2.9 ตัวอย่างผลลัพธ์ของการสกัดข้อมูลบนหน้าเว็บเพจราคาห้องพักในโรงแรม 9

2.10 (a) การกำนดแท็กที่ผู้ใช้ต้องการแสดงผลลัพธ์และไม่ต้องการแสดงผลลัพธ์
(b) หน้าเว็บเพจก่อนการลบแท็กของรูปภาพ (c) หน้าเว็บเพจหลังลบแท็กของรูปภาพ 10

2.11 การสกัดลิงค์ที่เชื่อมโยงข่าวในเว็บไซต์ 11

2.12 ผลลัพธ์ของการสกัดเนื้อหาข่าวในแต่ละเว็บเพจ 12

2.13 ตัวอย่างขั้นตอนของการหาอนุประโยคและวลี 14

2.14 ตัวอย่างการกำกับแท็กเชิงความหมายโดยใช้ออนโทโลยี 15

3.1 โครงสร้างแท็ก HTML ที่แน่นอน 17

3.2 โครงสร้างแท็ก HTML ที่มีไม่แน่นอน 17

3.3 ขั้นตอนการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ 19

3.4 ขั้นตอนของการรวบรวมชื่ออาการ 20

3.5 ตัวอย่างข้อมูลของหัวข้อสรรพคุณ ที่อยู่ในโครงสร้าง HTML ที่แน่นอน 21

3.6 ตัวอย่างข้อมูลที่ถูกสกัดจากโครงสร้าง HTML ของหัวข้อสรรพคุณ 22

3.7 ตัวอย่างหัวข้อสรรพคุณที่อยู่ในโครงสร้าง HTML ที่ไม่แน่นอน 22

3.8 ตัวอย่างข้อมูลหลังจากการสัดข้อมูลที่อยู่ในโครงสร้าง HTML ที่ไม่แน่นอน 23

3.9 ผลลัพธ์ของการแยกประโยคและตัดคำของเนื้อหาสรรพคุณ 23

3.10 แสดงผลลัพธ์ของการสกัดข้อมูลโดยใช้คำบ่งชี้ 24

3.11 กระบวนการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย 25

3.12 ผลลัพธ์ของการกำหนดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย
ในเว็บเพจที่มีโครงสร้าง HTML ที่แน่นอน 26

สารบัญภาพ (ต่อ)

ภาพที่	หน้า
3.13 ผลลัพธ์ของการกำหนดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ในเว็บเพจที่มีโครงสร้าง HTML ที่ไม่แน่นอน	26
3.14 ผลลัพธ์ของการสกัดชื่อพืชสมุนไพรไทย	27
3.15 ผลลัพธ์ของการสกัดส่วนที่ใช้และชื่ออาการ	28
3.16 ผลลัพธ์ของการสกัดชื่ออาการโดยใช้คำบ่งชี้	29
3.17 ผลลัพธ์ของการสกัดชื่ออาการโดยใช้ไฟล์ "Learned symptom name"	29

บทที่ 1
บทนำ

1.1. ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันประชาชนส่วนใหญ่ให้ความสนใจพืชสมุนไพรไทยมาใช้บรรเทาอาการมากยิ่งขึ้น เนื่องจากพืชสมุนไพรไทยสามารถหาได้ง่าย และเพาะปลูกได้เองตามครัวเรือน อีกทั้งพืชสมุนไพรไทยบางชนิดยังนำมาประกอบเป็นอาหารได้ แต่ถึงอย่างไรก็ตามประชาชนส่วนใหญ่ ยังไม่สามารถเลือกพืชสมุนไพรไทยที่นำมาใช้บรรเทาอาการได้เอง อีกทั้งในปัจจุบันผู้เชี่ยวชาญพืชสมุนไพรไทยนั้นก็ลดน้อยลง ทำให้ประชาชนส่วนใหญ่ไม่สามารถเข้าถึงผู้เชี่ยวชาญได้ แต่ในปัจจุบันเทคโนโลยีอินเตอร์เนตพัฒนาไปอย่างรวดเร็ว ทำให้ประชาชนสามารถแลกเปลี่ยนและเข้าถึงข้อมูลต่างๆ ทำได้ง่ายมากยิ่งขึ้น ดังนั้นจึงทำให้เกิดเว็บไซต์ที่นำเสนอข้อมูลพืชสมุนไพรไทยขึ้นอย่างมากมาย [1][2][3][4][5] ทำให้ประชาชนที่สนใจพืชสมุนไพรไทยสามารถเข้าถึงข้อมูลพืชสมุนไพรไทยได้สะดวกยิ่งขึ้น

แต่ถึงอย่างไรก็ตามข้อมูลพืชสมุนไพรไทยที่นำเสนอในบางเว็บไซต์ ยังมีข้อมูลที่ไม่ครบถ้วน เช่น [1][2][3][4] ได้นำเสนอข้อมูลพืชสมุนไพรไทยที่สามารถรักษาอาการของผู้ใช้ได้ แต่ไม่ได้บอกความเป็นพิษของพืชสมุนไพรไทยแต่ละชนิด แต่ในทางกลับกัน [5] นำเสนอข้อมูลพืชสมุนไพรไทยที่เป็นพิษกับระบบร่างกายของพืชสมุนไพรไทย แต่ไม่ได้บอกถึงสรรพคุณของพืชสมุนไพรไทยแต่ละชนิด เป็นต้น นอกจากนี้บางเว็บเพจอาจนำเสนอสรรพคุณของพืชสมุนไพรไทยแต่ละชนิดไม่ครบถ้วน เช่น บางเว็บเพจนำเสนอสรรพคุณของขิงสามารถรักษาอาการปวดท้องได้ แต่บางเว็บเพจอาจไม่ได้นำเสนอสรรพคุณของขิงสามารถรักษาอาการดังกล่าวได้ เป็นต้น ทำให้ประชาชนส่วนใหญ่นั้นต้องเข้าไปหลายเว็บไซต์เพื่อได้ข้อมูลพืชสมุนไพรไทยต่างๆ ที่ครบถ้วนดังนั้นผู้วิจัยเกิดแนวคิดที่พัฒนากระบวนการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ เพื่อช่วยให้ประชาชนที่สนใจพืชสมุนไพรไทยได้รับข้อมูลพืชสมุนไพรไทยที่ครบถ้วนมากยิ่งขึ้น

1.2. วัตถุประสงค์ของงานวิจัย

1.2.1. เพื่อคิดค้นกระบวนการสกัดข้อมูลพืชสมุนไพรไทยจากเว็บไซต์

1.2.2. เพื่อทำให้ข้อมูลพืชสมุนไทยแต่ละชนิดมีความครบถ้วนมากยิ่งขึ้น

1.3. ขอบเขตการวิจัย

1.3.1. คิดค้นกระบวนการสกัดข้อมูลพืชสมุนไพรไทยจากเว็บไซต์ ที่สามารถสกัดข้อมูลได้ไม่น้อยกว่า 3 เว็บไซต์

1.3.2. ข้อมูลพืชสมุนไพรไทยของแต่ละชนิดที่จะสกัดในเว็บไซต์ ประกอบด้วย ชื่อวิทยาศาสตร์ ชื่อสามัญ ชื่อวงศ์ ส่วนประกอบของพืชสมุนไพรไทย และ ชื่ออาการ ที่พืชสมุนไพรไทยแต่ละชนิดรักษาได้

1.3.3. วัดประสิทธิภาพการสกัดข้อมูลพืชสมุนไพรไทยจากเว็บไซต์ โดยวัดจากความถูกต้อง (precision) และ ความครบถ้วน (recall) ของการสกัดข้อมูลในแต่ละเว็บไซต์

1.4. วิธีดำเนินการวิจัย

1.4.1. รวบรวมและศึกษางานวิจัยที่เกี่ยวข้องกับการสกัดข้อมูลบนเว็บไซต์

1.4.2. คิดค้นกระบวนการที่สามารถสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์

1.4.3. พัฒนาโปรแกรมตามการสกัดข้อมูลพืชสมุนไพรไทยตามที่ได้คิดค้นไว้

1.4.4. วัดประสิทธิภาพระบบตามที่พัฒนา

1.4.5. จัดทำรายงานผลการดำเนินการ

1.5. ประโยชน์ที่คาดว่าจะได้รับ

1.5.1. ได้กระบวนการที่สามารถสกัดพืชสมุนไพรไทยจากหลายเว็บไซต์

1.5.2. ข้อมูลพืชสมุนไพรไทยแต่ละชนิดมีความครบถ้วนสมบูรณ์มากยิ่งขึ้น

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

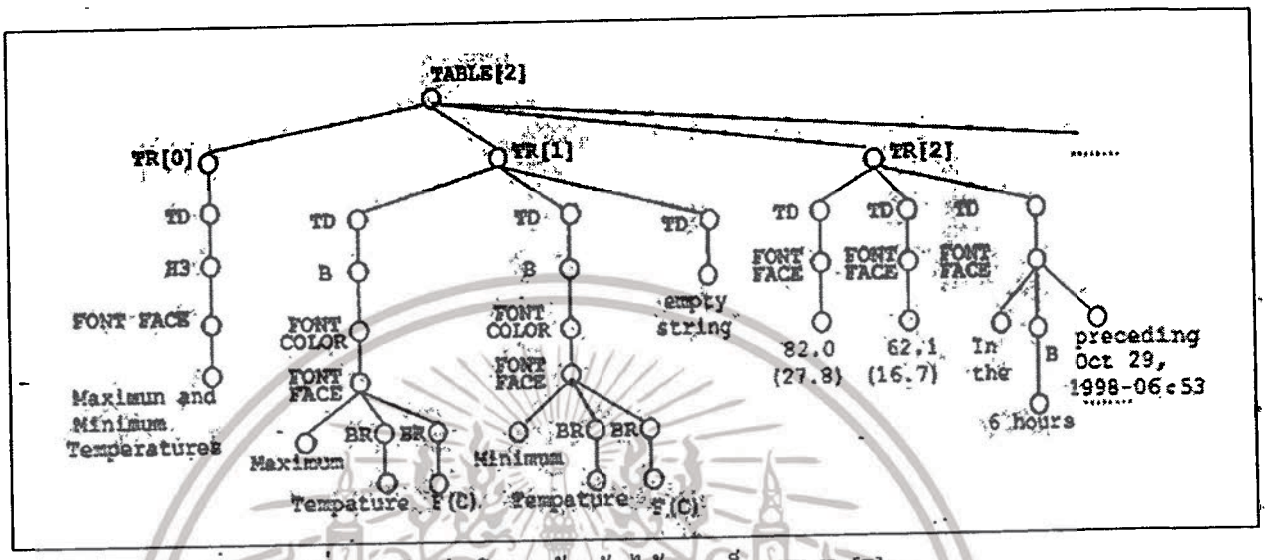
2.1. การสกัดข้อมูล (Information Extraction: IE)

การสกัดข้อมูล [6] คือ กระบวนการที่นำข้อความที่อยู่ในเอกสารต่างๆ เช่น เอกสารเอ็กซ์เอ็มแอล (Extensible Markup Language: XML) หรือเอกสารเอชทีเอ็มแอล (HyperText Markup Language: HTML) เป็นต้น นำมาประมวลเพื่อให้ได้ข้อความที่มีลักษณะชัดเจน ตรงตามกับผู้ใช้ต้องการแสดงออกมาเป็นผลลัพธ์ ซึ่งผลลัพธ์อาจถูกเก็บอยู่ในฐานข้อมูล หรือแสดงให้ผู้ใช้เห็นได้โดยตรง ในปัจจุบันมีการใช้เทคนิคสกัดข้อมูลนำไปประยุกต์ใช้กับเอกสารหลายชนิด แต่มีเอกสารหนึ่งที่ได้รับความนิยมนำมาใช้ในการสกัดข้อมูลเป็นอย่างมากคือเอกสาร HTML เนื่องจากในปัจจุบันเทคโนโลยีอินเตอร์เนตเติบโตอย่างรวดเร็ว จึงทำให้ผู้ใช้สามารถสร้างและเข้าถึงข้อมูลต่างๆ ได้สะดวก ซึ่งในแต่ละเว็บไซต์ต่างๆ ก็มีข้อมูลอยู่จำนวนมากทำให้ผู้ไม่สามารถเลือกเข้าถึงข้อมูลที่ผู้ใช้ต้องการได้ทั้งหมด ดังนั้นจึงมีผู้คิดค้นเทคนิคที่ใช้สกัดข้อมูลบนเว็บไซต์ เพื่อให้ผู้ใช้สามารถเลือกรับข้อมูลในส่วนเฉพาะสาระสำคัญ ซึ่งเทคนิคที่ใช้สกัดสารสนเทศบนเว็บไซต์แบ่งออกเป็น 3 เทคนิคดังต่อไปนี้ เทคนิคแรกคือ DOM tree path เทคนิคถัดมาคือ HTML tags and Literal words และเทคนิคสุดท้ายคือ Syntactic semantic analysis ซึ่งในแต่ละเทคนิคมีรายละเอียดดังต่อไปนี้

เทคนิค DOM tree path เป็นเทคนิคที่นำแท็ก HTML ที่อยู่ในแต่ละเว็บไซต์นำมาสร้างเป็นโครงสร้างของต้นไม้ (tree) ถัดมาผู้ใช้ต้องเป็นผู้กำหนดแท็ก HTML ที่ต้องการใช้สกัดข้อความออกมา สำหรับตัวอย่างของวิธีการสกัดข้อมูลด้วย DOM tree path ในที่นี้ขอยกตัวอย่างการสกัดสารสนเทศของรายงานสภาพอากาศบนเว็บไซต์ [7] ซึ่งรายละเอียดซอร์สโค้ดที่ของเว็บเพจแสดงดังรูปที่ 2.1 และหลังจากนั้นจะนำซอร์สโค้ดดังกล่าวมาแปลงเปลี่ยนให้อยู่ในรูปแบบโครงสร้างของต้น ดังแสดงในรูปที่ 2.2

```
<TABLE><TR><TD COLSPAN=3><H3><FONT FACE="Arial, Helvetica">Maximum and Minimum Temperatures</FONT>
</H3> </TD></TR><TR><TD ALIGN=CENTER BGCOLOR="#FFFFFF"><B><FONT COLOR="#0000A0"><FONT FACE=
"Arial,Helvetica">Maximum<BR>Temperature<BR>F(C)</FONT></FONT></B></TD><TD ALIGN=CENTER BGCOLOR=
"#FFFFFF"><B><FONT COLOR="#0000A0"><FONT FACE="Arial,Helvetica">Minimum<BR>Temperature<BR>F(C)
</FONT></FONT></B></TD><TD></TD></TR><TR><TD ALIGN=CENTER><FONT FACE="Arial,Helvetica">82.0(27.8)
</FONT></TD><TD ALIGN=CENTER><FONT FACE="Arial, Helvetica">62.1(16.7)</FONT></TD><TD><FONT FACE=
"Arial, Helvetica">In the <B>6 hours</B> preceding Oct 29, 1998 - 06:53 PM EST / 1998.10.29 2353
UTC</FONT></TD></TR><TR><TD ALIGN=CENTER><FONT FACE="Arial,Helvetica">80.1(26.7)</FONT></TD>
<TD ALIGN=CENTER><FONT FACE="Arial, Helvetica">45.0(7.2)</FONT></TD><TD><FONT FACE="Arial,
Helvetica">In the <B>24 hours</B> preceding Oct 28, 1998 - 11:53 PM EST / 1998.10.28 0453 UTC</FONT>
</TD></TR><TR><TD COLSPAN=3><HR SIZE=1 NOSHADE WIDTH="100%"></TD></TR></TABLE> .....
```

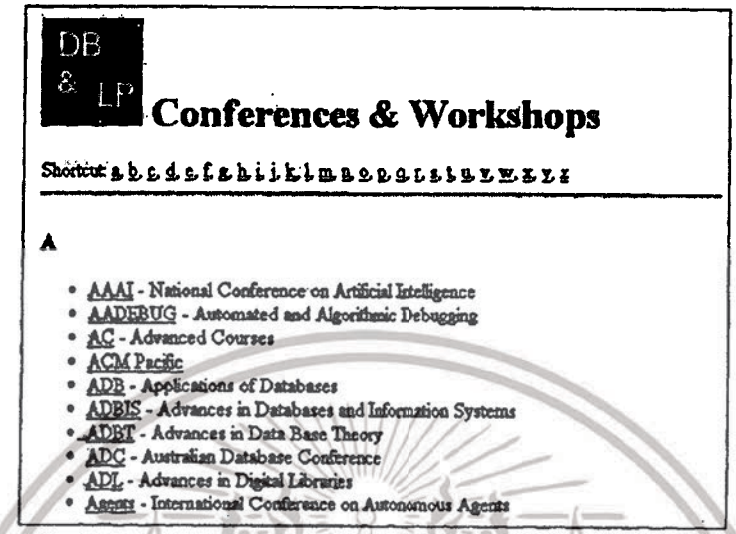
รูปที่ 2.1 ตัวอย่างซอร์สโค้ดของเว็บเพจสภาพอากาศบนเว็บไซต์ [7]



รูปที่ 2.2 ตัวอย่างโครงสร้างต้นไม้ของแท็ก HTML [7]

จากรูปที่ 2.2 แสดงถึงโครงสร้างของต้นไม้ที่สร้างมาจากซอร์สโค้ดจากรูปที่ 2.1 โดยการสร้างต้นไม้จะสร้างโดยไล่เรียงตามแท็กที่อยู่บนสุด ไปจนถึงแท็กย่อยของแต่ละแท็ก ซึ่งจากรูปจะเห็นว่าถ้าต้องการข้อมูลของอุณหภูมิที่สูงสุด สามารถท่องเข้าไปตามแท็กโดยใช้แท็ก TABLE[2], TR[2], TD[0] ซึ่งจะได้ผลลัพธ์เป็น 82.0 (27.8)

เทคนิคถัดมาคือ HTML tags and Literal words เป็นเทคนิคที่ผู้ใช้ต้องมีการกำหนดหนดแท็ก HTML ที่ต้องการสกัด พร้อมทั้งระบุว่าลักษณะของแท็กที่เกิดขึ้นว่ามีลักษณะเกิดขึ้นแบบใด เช่น แท็กที่ต้องการเก็บข้อความเกิดขึ้นได้ครั้งเดียว หรือมากกว่าหนึ่งครั้ง นอกจากนี้เทคนิค HTML tags and Literal words ยังคงต้องกำหนดประเภทของข้อมูลที่ต้องการ เพื่อที่จะได้จัดเก็บข้อมูลได้อย่างถูกต้อง สำหรับตัวอย่างของการสกัดข้อมูลด้วยเทคนิค HTML tags and Literal words ในที่นี้ขอยกตัวอย่างการสกัดข้อมูลของการประชุมวิชาการ (Conference) [8] ซึ่งหน้าเว็บเพจที่แสดงข้อมูล แสดงในรูปที่ 2.3 และซอร์สโค้ดของหน้าเว็บเพจที่ใช้แสดงผลแสดงดังรูปที่ 2.4



รูปที่ 2.3 ตัวอย่างหน้าเว็บเพจของการประชุมวิชาการ [8]

```
<html>
<head>
<title>DB&amp;LP: Conferences &amp;amp; Workshops</title>
</head>
<body bgcolor="B4CFF4D" text="#000000" link="#000000">
<h1><a href="../index.html"><img alt="DB&LP" src="../logo.gif"></a>
Conferences &amp;amp; Workshops</h1>
... omissis ...

<hr>
<h3><a name="a">A</a></h3>
<ul>
<li><a href="aaai/index.html">AAAI</a> - National Conference on Artificial
    Intelligence
<li><a href="aadebug/index.html">AADEBUG</a> - Automated and Algorithmic
    Debugging
<li><a href="ac/index.html">AC</a> - Advanced Courses
<li><a href="pacific/pacific96.html">ACM Pacific</a>
<li><a href="adb/index.html">ADB</a> - Applications of Databases
... omissis ...
<li><a href="http://www.ecn.edu.au/conweb96/">AusWeb</a> - Australian World
    Wide Web Conference - link
</ul>

<h3><a name="b">B</a></h3>
<ul>
<li><a href="berkeley/index.html">Berkeley Workshop</a> on Distributed Data
    Management and Computer Networks
<li><a href="bivit/index.html">BIVIT</a> - Basque International Workshop on
    Information Technology
<li><a href="bncod/index.html">BNCOD</a> - British National Conference on Databases
<li><a href="btw/index.html">BTW</a> - Datenbanksysteme in Buero;ru, Technik
    und Wissenschaft (German Database Conference)
</ul>
... omissis ...

<h3><a name="i">I</a>, <a name="j">J</a>, <a name="k">K</a></h3>
... omissis ...
</body>
</html>
```

รูปที่ 2.4 ตัวอย่างซอร์สโค้ดของหน้าเว็บเพจการประชุมวิชาการ [8]

สำหรับข้อมูลที่ต้องการสกัดออกจากหน้าเว็บเพจดังกล่าว 4 ข้อมูลประกอบด้วย ตัวอักษรที่เป็นอินเด็กซ์ของงานประชุมวิชาการ เช่น A B C เป็นต้น ข้อมูลถัดมาคือชื่อย่อของงานประชุมวิชาการ ข้อมูลในส่วนที่สามคือชื่อทางการของงานประชุมวิชาการ และข้อมูลในส่วนสุดท้ายคือลิงค์เชื่อมโยงไปหน้าเว็บไซต์ของแต่ละงานประชุมวิชาการ จากรูปที่ 2.4 จะสังเกตเห็นว่า ข้อมูลที่เป็นอินเด็กของงานประชุมวิชาการจะอยู่ในแท็ก <h3> ในส่วนข้อมูลของชื่อย่อของงานประชุมวิชาการอยู่ในแท็ก <a href> ที่อยู่ข้างในของแท็ก <li> นอกจากนี้แท็ก <li> ยังเก็บชื่อที่เป็นทางการของงานประชุมวิชาการอีกด้วย และสำหรับลิงค์ที่เชื่อมโยงไปหน้าเว็บไซต์หลักของงานประชุมวิชาการจะถูกจัดเก็บอยู่ในแท็ก <a href> ซึ่งจากแท็กดังกล่าวสามารถนำมาเขียนเป็นแพ็ตเทิร์น (pattern) เพื่อสกัดข้อมูลได้อย่างถูกต้องดังแสดงในรูปที่ 2.5

```
PAGE AllConferences

$AllConferences :   *<hr> ( $ConfWithInitial )+ ;

$ConfWithInitial :  <h3><a name="*">$Initial</a></h3>
                    [<ul> (
                              <li><a href="$ConfURL">$Acronym</a> - $ConfName $TP
                           )* </ul>] ;

$Initial :          *(?</a>) ;
$ConfURL :          *(?") ;
$Acronym :          *(?</a>) ;
$ConfName :         *(?<li> | </ul>) ;

$TP :               {
                      $Initial              char(1);
                      $Acronym              char(100);
                      $ConfName, $ConfURL   char(255);
                    }

END
```

รูปที่ 2.5 แพ็ตเทิร์นที่ใช้สกัดข้อมูลบนหน้าเว็บเพจประชุมวิชาการ [8]

จากรูปที่ 2.5 สามารถอธิบายรูปแบบที่สร้างขึ้นได้ดังนี้ ขั้นแรกเป็นการหากลุ่มรายชื่อของงานประชุมวิชาการ ($AllConference) โดยใช้แท็ก <hr> ซึ่งข้อมูลที่หาได้จะเก็บอยู่ในตัวแปร $ConfWithInitial ซึ่งจะสังเกตเห็นสัญลักษณ์ * เป็นสิ่งที่บ่งบอกถึงรูปแบบของแท็ก HTML อาจเกิดขึ้นได้หลายครั้ง หรือไม่เกิดขึ้นเลยก็ได้ ในส่วนของสัญลักษณ์ + เป็นสิ่งที่บ่งบอกถึงรูปแบบแท็ก อาจเกิดขึ้นได้มากกว่าหนึ่งครั้งขึ้นไป ถัดมาจะนำกลุ่มแท็กที่จัดเก็บอยู่ใน $ConfWithInitial มาหารายละเอียด ข้อมูลแท็กที่บ่งบอกถึงเป็นอินเด็กของงานประชุมวิชาการ ซึ่งข้อมูลดังกล่าวจะถูกแทนที

ด้วยตัวแปร $Initial ส่วนลิงค์เชื่อมโยงไปยังหน้าประชุมวิชาการจะเก็บในตัวแปร $ConfURL ในส่วนของตัวแปร $Acronym และ $ConfName จะเป็นที่เก็บข้อมูลของชื่อย่อของงานประชุมวิชาการ และชื่อเต็มของงานประชุมวิชาการตามลำดับ ถัดมาก็จะนำข้อมูลที่เก็บอยู่ในตัวแปรดังกล่าวนำมาสกัดเพื่อนำในส่วนของเฉพาะข้อความออกมา ซึ่งการเก็บข้อความที่ต้องการจะใช้เครื่องหมายวงเล็บเป็นสิ่งที่บ่งบอกถึงตำแหน่งที่ต้องการดึงข้อมูลออกมา หลังจากนั้นจะนำข้อมูลดังกล่าวมาจัดเก็บอยู่ในตัวแปร $TP ซึ่งข้อมูลดังกล่าวประกอบด้วยอินเด็ก ชื่อย่อ ชื่อเต็มร่วมถึงลิงค์ที่เชื่อมโยงของแต่ละงานประชุมวิชาการมาจัดเรียงเพื่อนำข้อมูลดังกล่าวแทรกลงฐานข้อมูลได้อย่างถูกต้อง ซึ่งผลลัพธ์ของการสกัดข้อมูลของงานประชุมวิชาการแสดงในรูปที่ 2.6

รูปที่ 2.6 ผลลัพธ์ของการสกัดข้อมูลของงานประชุมวิชาการ [8]

Syntactic semantic analysis เป็นเทคนิคที่ใช้กฎนำมาสกัดข้อความออกจากเอกสาร HTML ซึ่งกฎที่นำมาเขียนจะมีลักษณะคล้ายคลึงกับ regular expreesion ซึ่งการสกัดข้อความจากเอกสาร HTML ด้วยวิธีจะมีลักษณะที่หยืดหยุ่นกว่าเทคนิค DOM tree path และเทคนิค HTML tags and Literal words เนื่องจากการสกัดข้อความด้วยวิธี Syntactic semantic analysis จะไม่ได้อาศัยแท็ก HTML ในการสกัดข้อความจากเอกสาร HTML สำหรับตัวอย่างขั้นตอนของการสกัดด้วย Syntactic semantic amalysis ในที่นี้ข้อยกตัวอย่างการสกัดข้อมูลการจ้องโรงแรมที่อยู่บนเว็บไซต์ [9] ซึ่งข้อมูลที่ต้องการสกัดคือจำนวนเตียง และราคาของห้องพักแต่ละห้อง สำหรับซอร์สโค้ดที่จะนำมาสกัดแสดงดังรูปที่ 2.7

```
Capitol Hill - 1 br twnhme. fplc D/W W/D. Undrgrnd pkg
incl $675. 3 BR, upper flr of turn of ctry HOME. incl gar,
grt N. Hill loc $995. (206) 999-9999 <br>
<i> <font size=-2> (This ad last ran on 08/03/97.)
</font> </i> <hr>
```

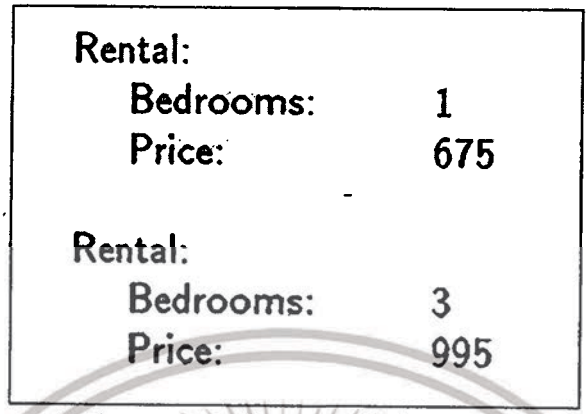
รูปที่ 2.7 ตัวอย่างซอร์สโค้ดของเว็บเพจราคาห้องพักในโรงแรม [9]

จากรูปที่ 2.7 จะสังเกตว่าข้อมูลที่จะถูกสกัดมีทั้งหมด 2 ข้อมูลคือ ห้องพักที่หนึ่งมีจำนวนเตียง 1 เตียง และราคา 675 ดอลล่า ห้องพักที่สองมีจำนวนเตียง 3 เตียง และราคา 995 ดอลล่า ซึ่งกฎที่จะนำมาสกัดข้อมูลดังกล่าวแสดงดังรูปที่ 2.8 จากรูปกฎออกเป็น 3 ส่วน ส่วนแรกคือส่วนของหมายเลขของกฎ (ID) ส่วนที่สองเป็นส่วนของรูปแบบของกฎ (Pattern) ส่วนสุดท้ายคือผลลัพธ์ของการสกัดข้อมูล (Output) ซึ่งส่วนที่สำคัญที่สุดคือส่วนของ Pattern ซึ่งจะสังเกตเหตุสัญลักษณ์ต่างๆ ซึ่งประกอบด้วยเครื่องหมาย 3 เครื่องหมายหลักคือ เครื่องหมายดอกจัน เครื่องหมายอัญประกาศเดี่ยว และเครื่องหมายวงเล็บ ซึ่งแต่ละเครื่องหายก็มีความหมายแตกต่างกันไปเช่น เครื่องหมายดอกจันแสดงถึง ให้ข้ามตัวอักษรทุกตัวจนกว่าจะพบรูปแบบข้อความที่ต่อท้ายเครื่องหมายดอกจัน เครื่องหมายอัญประกาศเดี่ยวหมายถึงต้องมีข้อความที่ตรงกับข้อความที่อยู่ในอัญประกาศเดียว และวงเล็บหมายถึงให้ข้อมูลที่ต้องการนำมาแสดงเป็นผลลัพธ์ ซึ่งจากกฎที่แสดงในรูปที่ 2.8 สามารถอธิบายได้ว่า ให้ข้ามตัวอักษรทุกตัวจนกว่าจะพบตัวเลข (Digit) ที่มีข้อความ BR ต่อท้าย แล้วเก็บตัวเลขดังกล่าวเพื่อใช้แสดงเป็นผลลัพธ์ และถัดมาให้ข้ามตัวอักษรทุกตัวจนกล่าวจะพบตัวอักษรดอลล่า ($) และให้เก็บตัวเลขที่ต่อท้ายเครื่องหมายดอลล่าเพื่อนำมาแสดงเป็นผลลัพธ์

ในส่วนของการแสดงผลลัพธ์ (Output) เป็นส่วนที่นำข้อมูลที่สกัดได้ในส่วนของ Pattern มาแสดงผล ซึ่งจะสังเกตเห็นว่าข้อมูลที่สกัดได้ในส่วนที่ของ Pattern มีสองข้อมูล คือในส่วนของ (Digit) และ (Number) ซึ่งสองข้อมูลดังกล่าวสามารถนำมาแทนที่เป็นตัวแปรจะได้ $1 และ $2 ตามลำดับ ดังนั้นข้อมูลที่อยู่ในส่วน (Digit) จะถูกแทนด้วยตัวแปร $1 ซึ่งใช้นำเสนอข้อมูลของจำนวนเตียงของห้องพัก และข้อมูลในส่วน (Number) จะถูกแทนที่ด้วยตัวแปร $2 ซึ่งใช้นำเสนอข้อมูลราคาของห้องพัก ซึ่งผลลัพธ์ของการสกัดข้อมูลดังกล่าวแสดงในรูปที่ 2.9

```
ID:: 1
Pattern:: * ( Digit ) ' BR' * '$' ( Number )
Output:: Rental {Bedrooms $1} {Price $2}
```

รูปที่ 2.8 ตัวอย่างกฎที่ใช้สกัดข้อมูลบนหน้าเว็บเพจราคาห้องพักในโรงแรม [9]



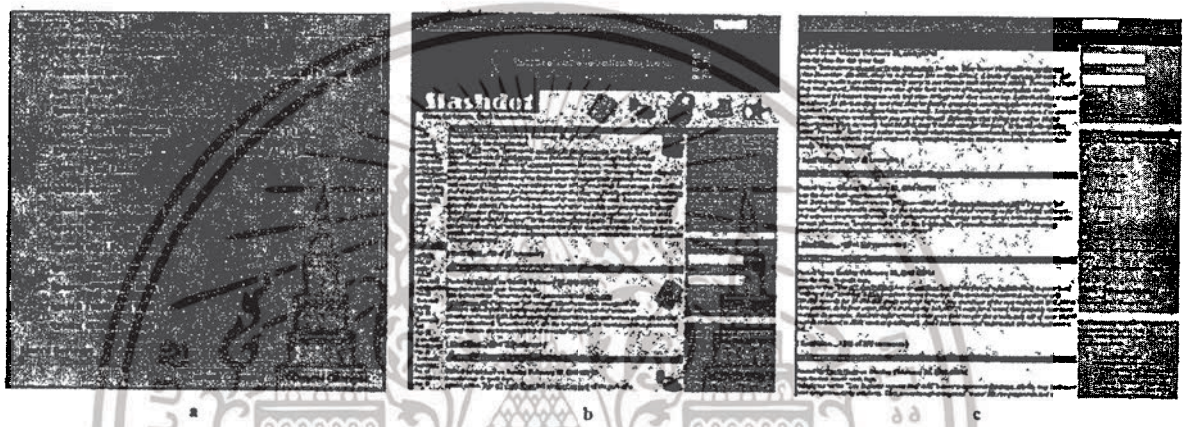
รูปที่ 2.9 ตัวอย่างผลลัพธ์ของการสกัดข้อมูลบนหน้าเว็บเพจราคาห้องพักในโรงแรม [9]

2.2. งานวิจัยที่เกี่ยวข้อง

ในปัจจุบันมีหลายงานวิจัยที่นำพยายามพัฒนากระบวนการของการสกัดข้อมูลบนเว็บไซต์อย่างแพร่หลาย โดยเริ่มจากงานวิจัยที่พยายามที่จะคัดกรองแท็ก HTML ที่เกี่ยวข้องกับเนื้อหาหลักในแต่ละเว็บไซต์ ตามที่ผู้ใช้ต้องการ [10][11] ไปจนถึงงานวิจัยที่มีทำเทคนิคอื่นๆ มาประยุกต์ใช้ร่วมกับการคัดกรองแท็ก HTML เพื่อใช้การสกัดข้อมูลบนเว็บไซต์ให้มีความสมบูรณ์มากยิ่งขึ้น [12] ซึ่งแต่ละงานวิจัยมีรายละเอียดดังต่อไปนี้ [10] ได้นำเสนอระบบที่สามารถให้ผู้ใช้เลือกแท็ก HTML ที่ต้องการแสดงผลหรือสามารถลบแท็ก HTML ที่ไม่ต้องการให้แสดงผลได้ เช่น แท็ก HTML ที่เกี่ยวข้องกับรูปภาพ หรือแท็ก HTML ที่เชื่อมโยงไปยังเว็บไซต์อื่นๆ เพื่อให้เว็บเพจสามารถแสดงผลบนเครื่องพีดีเอ (Personal Digital Assistant: PDA) หรือบนโทรศัพท์มือถือได้อย่างราบรื่น เป็นต้น นอกจากนี้งานวิจัยนี้ยังได้นำเสนอกระบวนการในการวิเคราะห์ลิงค์ที่น่าจะเป็นลิงค์ของการโฆษณา และนำเสนอกระบวนการวิเคราะห์ตารางที่เป็นเนื้อหาหลักของแต่ละเว็บเพจ ซึ่งรายละเอียดของกระบวนการของทั้งสองกระบวนการมีดังต่อไปนี้

ในส่วนกระบวนการของการวิเคราะห์ลิงค์ที่น่าจะเป็นลิงค์ที่เชื่อมโยงไปยังโฆษณา งานวิจัยนี้ได้ใช้เครื่องมือ openXML ซึ่งเป็นเครื่องมือ HTML Parser ในการสกัดลิงค์ของเว็บไซต์ที่อยู่ในแท็ก “src” และ “href” เมื่อได้ข้อมูลลิงค์เว็บไซต์ดังกล่าวมาแล้ว จะนำลิงค์เว็บไซต์ดังกล่าวไปเปรียบเทียบกับรายการของลิงค์โฆษณาที่ได้รวบรวมไว้ก่อนหน้านี้ ถ้าลิงค์ของเว็บไซต์ที่สกัดออกมาได้ไปตรงกับรายการของลิงค์โฆษณาที่ได้จัดเก็บไว้ ลิงค์โฆษณาดังกล่าวก็จะถูกระบบลบออกไปทันที สำหรับรายละเอียดของกระบวนการวิเคราะห์ตารางที่เป็นเนื้อหาหลักของแต่ละเว็บเพจ แบ่งออกเป็น 2 ขั้นตอนหลักคือ ขั้นแรกจะให้ผู้ใช้กำหนดจำนวนตัวอักษรที่อยู่ในตารางหลัก ขั้นที่สองใช้ openXML

ในการดึงข้อความที่อยู่ในตารางออกมา หลังจากนั้นจะนำข้อความที่สกัดออกมาได้นำมานับจำนวนตัวอักษร ถ้าจำนวนในตัวอักษรนั้นมีมากกว่าจำนวนตัวอักษรที่ผู้ใช้ได้กำหนดไว้ ก็จะถือว่าตารางดังกล่าวเป็นตารางที่มีเนื้อหาหลักโดยทันทีนอกจากนี้ในส่วนของการแสดงผลผู้ใช้ก็สามารถเลือกให้ระบบแสดงผลลัพธ์เฉพาะในส่วนของเนื้อหา หรือแสดงผลลัพธ์ให้กับผู้ใช้ในรูปแบบของซอร์สโค้ด HTML สำหรับตัวอย่างหน้าจอที่ใช้ติดต่อผู้ใช้แสดงในรูปที่ 2.10



รูปที่ 2.10 (a) การกำนดแท็กที่ผู้ใช้ต้องการแสดงผลลัพธ์และไม่ต้องการแสดงผลลัพธ์ (b) หน้าเว็บเพจก่อนการลบแท็กของรูปภาพ (c) หน้าเว็บเพจหลังลบแท็กของรูปภาพ [10]

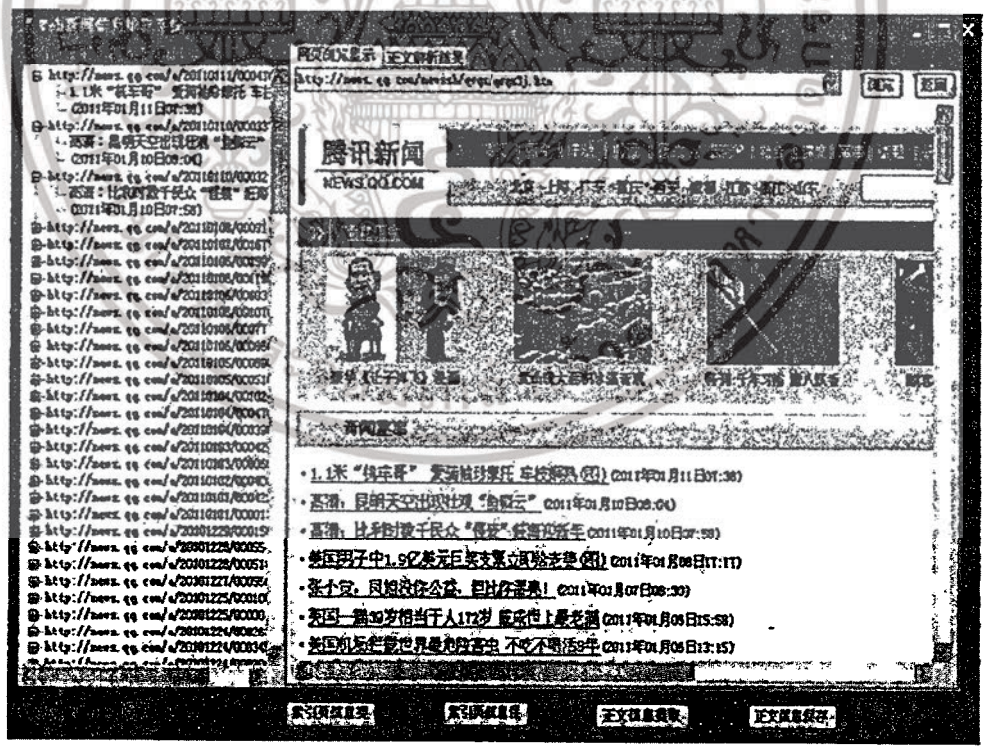
ซึ่งจากรูปที่ 2.10 (a) แสดงถึงการตั้งค่าแท็กที่ผู้ใช้ต้องการให้แสดงผลในส่วนของผลลัพธ์ และไม่ต้องการให้แสดงผลในส่วนของผลลัพธ์ของแต่ละเว็บเพจ ในส่วนของรูปที่ 2.10 (b) แสดงถึงหน้าเว็บเพจที่เป็นต้นฉบับ แต่หลังจากเลือกแท็ก HTML ตามที่ผู้ใช้ได้ตั้งค่าไว้ผลลัพธ์ดังกล่าวแสดงผลในรูปที่ 2.11 (c)

[11] ได้นำเสนอกระบวนการสกัดข้อมูลจากเว็บข่าวภาษาจีน ที่มีลักษณะกึ่งโครงสร้าง (semi-structured) ซึ่งงานวิจัยนี้ได้ใช้ *HTML Parser* และ regular expression ในการสกัดข้อมูลที่เกี่ยวข้องกับข่าวในเว็บไซต์ 4 เว็บไซต์ ซึ่งประกอบด้วย news.sina.com.cn, news.sohu.com, news.qq.com และ news.163.com ซึ่งขั้นตอนการสกัดข่าวจากหลายเว็บไซต์ประกอบด้วย 2 ขั้นตอนหลัก ดังต่อไปนี้

ขั้นตอนการสกัดข้อมูลในหน้าแรกของเว็บไซต์ (GENERAL EXTRACTION MODULE INDEX PAGE) ขั้นตอนนี้จะทำการสกัดชื่อหัวข้อข่าว ลิงค์ของข่าว และเวลาของข่าวแต่ละหัวข้อข่าวที่ปรากฏบนหน้าแรกของเว็บไซต์ ซึ่งขั้นตอนนี้แบ่งออกเป็น 2 ขั้นตอนย่อยดังนี้ ขั้นตอนแรกคือ

ขั้นตอนการกำหนด encoding ของแต่ละเว็บไซต์ข่าว (WEB PAGE CODING) เนื่องจากงานวิจัยนี้เป็นข่าวที่เกี่ยวข้องกับภาษาจีน ดังนั้นเว็บเพจส่วนใหญ่จะเข้ารหัสเนื้อหาด้วย GB2312 แต่ก็มีบางเว็บไซต์ที่เข้ารหัสข้อมูลด้วย UTF-8 หรือ บางเว็บไซต์ไม่ได้กำหนดการเข้ารหัสของเนื้อหา ดังนั้นเพื่อการสกัดข้อมูลมีประสิทธิ์ภาพ งานวิจัยนี้จึงทำการตรวจสอบว่า แต่ละเว็บไซต์มีการกำหนดการเข้ารหัสของเนื้อหาหรือไม่ โดยดูจากแท็ก <meta> ถ้าเว็บไซต์นั้นมีการกำหนดการเข้ารหัสข้อมูลของเนื้อหา ก็จะกำหนดการเข้ารหัสข้อมูลของการสกัดข้อมูลตามการเข้ารหัสของแต่ละเว็บไซต์นั้น แต่ถ้าเว็บไซต์ในไม่ได้กำหนดการเข้ารหัสของข้อมูล ก็จะกำหนดการเข้ารหัสข้อมูลของการสกัดข้อมูลด้วย GB2312

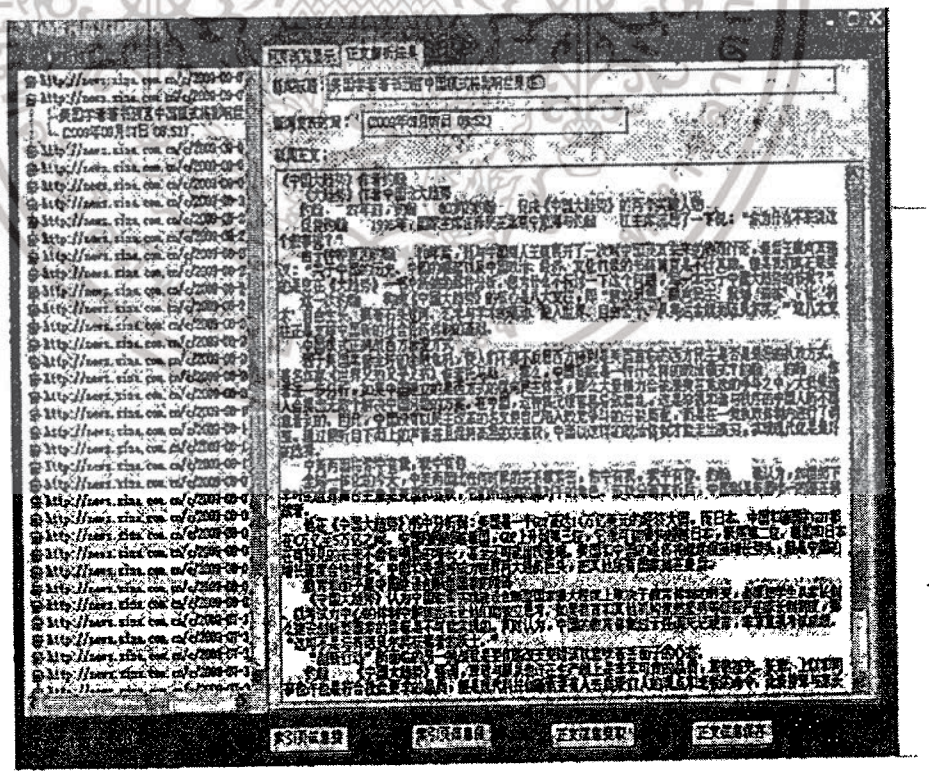
ขั้นตอนที่สองคือการสกัดลิงค์ของข่าวที่อยู่ในหน้าแรก หัวข้อข่าวที่สามารถเชื่อมโยงไปยังรายละเอียดของหัวข้อข่าวแต่ละหัวข้อนั้น ส่วนใหญ่อยู่ในแท็ก <table> หรือ <li> ดังนั้นงานวิจัยนี้จึงสนใจหัวข้อข่าวที่อยู่ในแท็กดังกล่าว ซึ่งในแต่ละหัวข้อข่าวจะมีลิงค์ที่สามารถเชื่อมโยงไปยังรายละเอียดของเนื้อหาในแต่ละหัวข้อ ซึ่งงานวิจัยนี้จึงทำการสกัดลิงค์ที่เชื่อมโยงไปยังรายละเอียดของเนื้อหาข่าว โดยใช้ HTML parser ขั้นตอนที่สุดท้ายคือการตรวจสอบความสมบูรณ์ของลิงค์ โดยใช้ regular expression ซึ่งผลลัพธ์ของขั้นตอนนี้แสดงในรูปที่ 2.11



รูปที่ 2.11 การสกัดลิงค์ที่เชื่อมโยงข่าวในเว็บไซต์ [11]

จากรูปที่ 2.11 พาเนล (panel) ทางด่านขวามือแสดงถึงหน้าแรกของเว็บข่าว ที่ต้องการสกัดลิงค์ที่สามารถเชื่อมโยงไปยังรายละเอียดของแต่ละข่าว รวมถึงยังสกัดชื่อหัวข้อข่าวและเวลาของข่าวของแต่ละหัวข้อข่าว ส่วนพาเนลทางขวามือคือ รายการของลิงค์ที่สกัดมาจากหน้าแรกของเว็บข่าว ซึ่งรายการของลิงค์ดังกล่าวถูกนำไปใช้ในการสกัดเนื้อหาข่าวในกระบวนการถัดไป

ขั้นตอนการสกัดข้อมูลในหน้าของเนื้อหา (GENERAL EXTRACTION MODULE CONTENT PAGE) เมื่อได้ลิงค์ที่สามารถเชื่อมโยงไปยังรายละเอียดของจากขั้นตอนที่ 1 เรียบร้อยแล้ว ขั้นตอนนี้จะทำการดึงเนื้อหาข่าวที่อยู่ในแต่ละเพจออกมา โดยใช้ HTML parser ซึ่งมี 5 ขั้นตอนหลักดังต่อไปนี้ ขั้นตอนแรกคือ ดึงเนื้อซอสโค้ด (source code) ทั้งหมดที่อยู่ในแต่ละเพจออกมา ขั้นตอนที่สองค้นหาแท็ก div ที่อยู่ในแต่ละเพจ เหตุผลที่ค้นหาเฉพาะแท็ก div เพราะว่าเนื้อหาข่าวส่วนใหญ่จะอยู่ในแท็กดังกล่าว ถัดมาจะทำการนับคำที่อยู่ในแต่ละแท็ก div ถ้าแท็ก div ไหนมีคำปรากฏอยู่ในแท็กมากที่สุดก็จะถูกกำหนดเป็นเนื้อหาหลัก และขั้นตอนสุดท้ายคือการสกัดข้อมูลที่อยู่แท็กโดยใช้ HTML parser ซึ่งผลลัพธ์แสดงดังรูปที่ 2.12



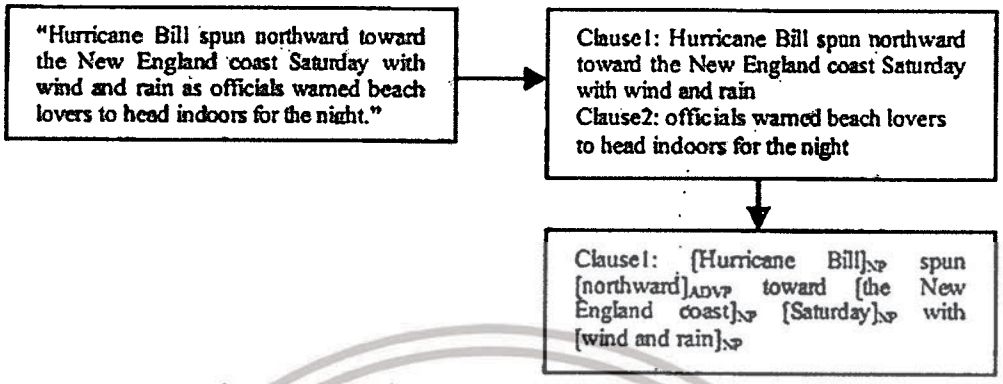
รูปที่ 2.12 ผลลัพธ์ของการสกัดเนื้อหาข่าวในแต่ละเว็บเพจ [11]

จากรูปที่ 2.12 พาเนลทางขวามือคือ ผลลัพธ์ของการสกัดเนื้อหาข่าวของแต่ละหัวข้อข่าว ตามรายการของลิงค์ที่อยู่ในพาเนลทางซ้ายมือ แต่ถึงอย่างไรก็ตามกระบวนการของงานวิจัยนี้ ยังไม่สามารถสกัดรายละเอียดของเนื้อหาข่าวได้ ซึ่งเป็นสิ่งที่จำเป็นต่อข้อมูลพืชสมุนไพรไทย ซึ่งในเนื้อหาหลักของแต่ละเว็บไซต์มีรายละเอียดของพืชสมุนไพรไทยต่างๆ มากมาย เช่น ชื่อของพืชสมุนไพรไทยต่างๆ หรือส่วนที่ใช้รักษาอาการของแต่ละพืชสมุนไพรไทย

จากสองงานวิจัย [10][11] ที่กล่าวถึงในข้างต้นทำให้ผู้วิจัยเกิดแนวคิดที่จะใช้แท็ก HTML และ regular expression เข้ามาช่วยในการสกัดข้อมูลพืชสมุนไพรไทย แต่ในงานวิจัยของผู้จัดทำมีบางเว็บไซต์ที่ไม่สามารถข้อมูลพืชสมุนไพรไทยโดยใช้แท็ก HTML เนื่องจากบางแท็ก HTML สามารถบ่งบอกถึงข้อมูลพืชสมุนไพรไทยได้หลายอย่าง แต่ถึงอย่างไรก็ตามได้มีงานวิจัยที่สกัดข้อมูลจากเว็บไซต์โดยไม่ใช้แท็ก HTML นั้นคือ [12] งานวิจัยนี้ได้นำเสนอกระบวนการสกัดข้อมูลการพยากรณ์อากาศบนเว็บเพจ ซึ่งงานวิจัยนี้ได้ใช้ออนโทโลยี รวมกับ semantic tagger ในการนำมาช่วยการสกัดข้อมูลให้สมบูรณ์มากขึ้น กระบวนการของการสกัดข้อมูลพยากรณ์บนเว็บไซต์ของงานนี้แบ่งออกเป็น 3 ส่วนหลัก ขั้นตอนแรกคือการดึงข้อความและเลือกประโยคที่เกี่ยวข้องกับข้อมูลของการพยากรณ์อากาศ ขั้นที่สองคือการสกัดอนุประโยคและวลีที่เกี่ยวข้องกับข้อมูลของการพยากรณ์อากาศบนเว็บเพจ และขั้นตอนสุดท้ายคือการสกัดเหตุการณ์ของข้อมูลการพยากรณ์อากาศบนเว็บเพจ

สำหรับกระบวนการของขั้นตอนแรก งานวิจัยนี้ได้ใช้เครื่องมือ AlchemyAPI ในการดึงเฉพาะข้อความที่อยู่ในแต่ละเว็ลเพจ หลังจากนั้นก็นำข้อความดังกล่าวแบ่งออกมาเป็นประโยค เพื่อหาประโยคที่มีความเกี่ยวข้องกับข้อมูลของการพยากรณ์อากาศ ซึ่งการเลือกประโยคที่มีความเกี่ยวข้องนั้น งานวิจัยนี้ได้ใช้อินสแตน (instance) ที่จัดเก็บอยู่ในออนโทโลยีมาช่วยเลือกหาประโยคที่เกี่ยวข้อง โดยนำประโยคแต่ละประโยคมาตรวจสอบดูว่าประโยคดังกล่าวมีอินสแตนที่อยู่ในออนโทโลยีหรือไม่ ถ้ามีก็แสดงว่าประโยคดังกล่าวนั้นเป็นประโยคที่มีความเกี่ยวข้องกับข้อมูลพยากรณ์อากาศ

หลังจากได้ประโยคที่มีความเกี่ยวข้องกับข้อมูลของการพยากรณ์อากศเรียบร้อยแล้ว ถัดมาจะนำประโยคดังกล่าวนำมาแจงแจกเพื่อหาอนุประโยคและวลีของข้อมูลที่มีความเกี่ยวข้อง สำหรับการแจงแจกประโยคนั้นเพื่อหาอนุประโยคในแต่ละประโยค งานวิจัยนี้ได้ใช้ Stanford Parser ซึ่งเป็นเครื่องมือที่ช่วยวิเคราะห์โครงสร้างของประโยคที่ต้องการ สำหรับตัวอย่างขั้นตอนของการหาอนุประโยคและวลีแสดงดังรูปที่ 2.13

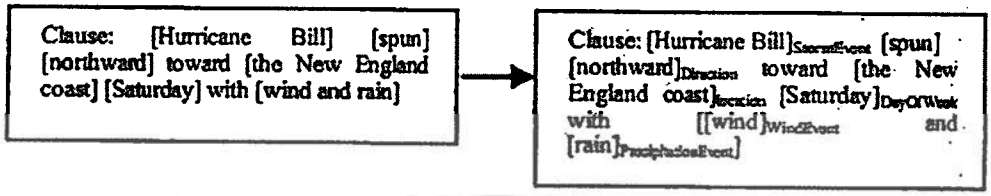


รูปที่ 2.13 ตัวอย่างขั้นตอนของการหาอนุประโยคและวลี [12]

จากรูปที่ 2.13 ตัวอย่างประโยคที่นำมาใช้เพื่อหาอนุประโยคและวลีนั้นคือ "Hurricane Bill spun northward toward the New England coast Saturday with wind and rain as officials warned beach lovers to head indoors for the night" ซึ่งเมื่อผ่านกระบวนการวิเคราะห์หาอนุประโยคแล้วสามารถแบ่งออกได้เป็น 2 ประโยคคือ ""Hurricane Bill spun northward toward the New England coast Saturday with wind and rain" และปรหาอนุประโยคที่สองคือ "officials warned beach lovers to head indoors for the night" หลังจากนั้นนำหาอนุประโยคดังกล่าวมาแจงแจงวลี ซึ่งจากตัวอย่างได้นำอนุประโยคแรกมาแจงเป็นวลี ซึ่งสามารถแบ่งวลีได้ 3 ประเภทคือ NP (Noun Phrase) ซึ่งประกอบด้วย Hurricane, the New England coast, Saturday และวลีประเภทสุดท้ายคือ ADVP (Adverb Phrase) ซึ่งประกอบด้วย northward

และขั้นตอนสุดท้ายนำวลีจากขั้นตอนที่สองมากำกับแท็กเชิงความหมาย เพื่อใช้บ่งบอกถึงเหตุการณ์ของการพยากรณ์อากาศ เช่น สถานที่ของการพยากรณ์อากาศ หรือวันที่ของการพยากรณ์อากาศ เป็นต้น ซึ่งการกำกับแท็กเพื่อบ่งบอกถึงเหตุการณ์มี 3 วิธีการคือ การกำกับแท็กโดยใช้คอนเซ็ปต์ของออนโทโลยี การกำกับแท็กโดยใช้ regular expression และการกำกับแท็กโดยใช้อินสแตนที่อยู่ในออนโทโลยี สำหรับวิธีการของการกำกับแท็กโดยใช้คอนเซ็ปของออนโทโลยี จะนำวลีที่ได้จากกระบวนการที่สองนำมาค้นหาคอนเซ็ปของออนโทโลยี ถ้าวลีตรงกับคอนเซ็ปที่อยู่ในออนโทโลยี วลีดังกล่าวก็จะถูกแท็กเป็นคอนเซ็ปที่อยู่ออนโทโลยี สำหรับวิธีการกำกับแท็กโดยใช้ regular expression จะนำวลีจากขั้นตอนที่สองมาเช็ครูปแบบตาม regular expression ที่ได้กำหนดไว้ ถ้าวลีดังกล่าวมีความสอดคล้องกับรูปแบบของ regular expression ที่ได้กำหนดไว้ วลีดังกล่าวจะถูกกำหนดเป็นข้อมูลของ data property และวิธีการสุดท้ายคือการกำกับแท็กโดยใช้อินสแตนที่อยู่ในออนโทโลยี วิธีการนี้จะคล้ายกับวิธีการของการกำกับแท็กโดยใช้คอนเซ็ปต์ของออนโทโลยี ซึ่งจะนำวลีที่ได้จากขั้นตอนที่สองมาหา

ค้นหาอินสแตนที่อยู่ในออนโทโลยี ถ้าวลีดังกล่าวตรงกับอินสแตนที่อยู่ในออนโทโลยี วลีดังกล่าวก็จะถูกแท็กเป็นอินสแตนที่อยู่ออนโทโลยี สำหรับตัวอย่างการกำกับแท็กเชิงความหมายแสดงในรูปที่ 2.14



รูปที่ 2.14 ตัวอย่างการกำกับแท็กเชิงความหมายโดยใช้ออนโทโลยี [12]

จากรูปที่ 2.14 เป็นผลลัพธ์ของการนำวลีที่ได้จากขั้นตอนที่สองนำมากำกับแท็กเชิงความหมาย ซึ่งจะพบว่า Herricane Bill จะถูกกับแท็กเชิงความหมายเป็นเหตุการณ์ของ stromEvent ส่วน northward จะถูกกำกับแท็กเชิงความหมายเป็นเหตุการณ์ของ Direction ในส่วนของ the New England coast จะถูกกำกับแท็กเชิงความหมายเป็นเหตุการณ์ของ location และในส่วนของ Saturday, wind และ rain จะถูกกำกับแท็กเชิงความหมายเป็นเหตุการณ์ของ DayOfweek, WindEvent และ PrecipitationEvent ตามลำดับ

แต่ถึงอย่างไรก็ตามงานวิจัยของผู้จัดทำไม่สามารถใช้ออนโทโลยีในการกำกับแท็กเชิงความหมายเนื่องจากข้อมูลของพืชสมุนไพรไทยมีความหลากหลาย เช่น ข้อมูลของอาการของพืชสมุนไพรไทยที่สามารถรักษาได้นั้น สามารถมีชื่ออาการได้เป็นจำนวนมากเกินกว่าที่จะคาดการณ์ได้

บทที่ 3
วิธีดำเนินการวิจัย

ในบทนี้จะกล่าวถึงกระบวนการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ สำหรับเว็บไซต์ที่งานวิจัยฉบับนี้ใช้ในการทำวิจัยมีทั้งหมด 5 เว็บไซต์ ซึ่งประกอบด้วย

- ฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี (www.dru.ac.th)
- ฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี (http://www.rspg.or.th/plants_data/herbs/herbs_200.htm)
- ฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ (http://www.pharmacy.mahidol.ac.th/siri/index.php?page=home)
- ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย (http://www.medplant.mahidol.ac.th/tpex/ toxic_mnu3.asp)
- ฐานข้อมูลพืชสมุนไพรไทยพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม (http://www.samunpri.com/)

3.1 การสำรวจโครงสร้างแท็ก HTML ของเว็บไซต์ที่เกี่ยวข้อง

สำหรับข้อมูลพืชสมุนไพรไทยที่งานวิจัยฉบับนี้จะไปสกัดออกมาในแต่ละเว็บไซต์ประกอบด้วย ชื่อทางการ ชื่อสามัญ ชื่อวิทยาศาสตร์ ชื่อวงศ์ ชื่อท้องถิ่น ความเป็นพิษต่อระบบร่างกาย ส่วนที่ใช้รักษา และชื่ออาการ ซึ่งในแต่ละเว็บไซต์จะใช้แท็ก HTML ที่ใช้กำหนดข้อมูลพืชสมุนไพรไทยที่แตกต่างกันไป จากการสำรวจโครงสร้าง HTML จาก 5 เว็บไซต์ที่กล่าวถึงในข้างต้น สามารถแบ่งโครงสร้าง html ได้เป็น 2 ประเภท คือ โครงสร้างแท็ก HTML ที่แน่นอน และโครงสร้างแท็ก HTML ที่ไม่มีความแน่นอน ดังแสดงในรูปที่ 3.1 และรูปที่ 3.2 ตามลำดับ

```
<tr>
<td >……. ชื่อภาษาอังกฤษ ……..</td>
<td>……. Ylang Ylang …….</td>
</tr>
<tr>
<td >……. ชื่อวิทยาศาสตร์ ……..</td>
<td>……. Cananga odorata (Lamk.) Hook. f. et. Th.
ANNONACEAE …….</td>
</tr>
```

รูปที่ 3.1 โครงสร้างแท็ก HTML ที่แน่นอน

```
<li><strong>ต้น</strong>: เป็นพรรณไม้ล้มลุก มีลำต้นอยู่ใต้ดินซึ่งเรียกว่าเหง้า ลำต้นจะมีความสูง
ประมาณ 50-100 ซม. ลักษณะเหง้าที่อยู่ใต้ดินจะกลมและแบน ลำต้นแท้จะมีลักษณะเป็นร้อ ๆ เนื้อในจะเป็นสี
ขาวหรือเหลืองอ่อน ส่วนท้ายของร้อนั้นจะเป็นยอดหรือต้นเทียมใหญ่เท่าแท่งดินสอดำ และกาบหรือโคนใบหุ้ม
</li>
……
……
……
<li><strong> ต้น</strong> รับผายลม บรรเทาอาการจุกเสียดแน่นเฟ้อ บำรุงไฟธาตุรักษาผิว คอเปื่อย ช่วย
ย่อยอาหาร ฆ่าพยาธิ รักษาโรคตา บิด ลมป่วง ท้องร่วงอย่างแรง อาเจียน</li>
```

รูปที่ 3.2 โครงสร้างแท็ก HTML ที่มีไม่แน่นอน

จากรูปที่ 3.1 จะสังเกตเห็นว่าแต่ละแท็ก <tr> จะประกอบด้วยแท็ก <td> 2 แท็ก ซึ่งแท็กแรกจะแสดงถึงชื่อหัวข้อ เช่น ชื่อภาษาอังกฤษ หรือ ชื่อวิทยาศาสตร์ เป็นต้น ส่วนแท็ก <tr> สุดท้ายแสดงถึงเนื้อหาของแต่ละหัวข้อ ซึ่งโครงสร้าง HTML ที่ไม่มีความซับซ้อน งานวิจัยนี้ได้ใช้ JSOUP API [13] เป็น HTML Parser ในการสกัดข้อมูลแต่ละหน้า และจากรูปที่ 3.2 จะสังเกตเห็นว่าแท็ก 1 ชนิด สามารถที่จะประเภทข้อมูลได้หลายชนิด เช่น ข้อมูลในแท็ก <li> อาจจะเป็นได้ทั้งข้อมูลของลักษณะทางพฤกษศาสตร์ และข้อมูลที่เกี่ยวข้องกับสรรพคุณที่ใช้รักษาอาการ เป็นต้น สำหรับโครงสร้าง HTML ที่มีความซับซ้อน จะใช้ไฟล์เทมเพลตในการสกัดข้อมูลพืชสมุนไพรไทย ซึ่งในไฟล์เทมเพลตจะประกอบด้วยสัญลักษณ์และคำบ่งชี้ของแต่ละหัวข้อ (indicator word of topic) ซึ่งแสดงในตารางที่ 3.1

ตารางที่ 3.1 แท็กที่บ่งชี้ชื่อของหัวข้อพืชสมุนไพรไทย

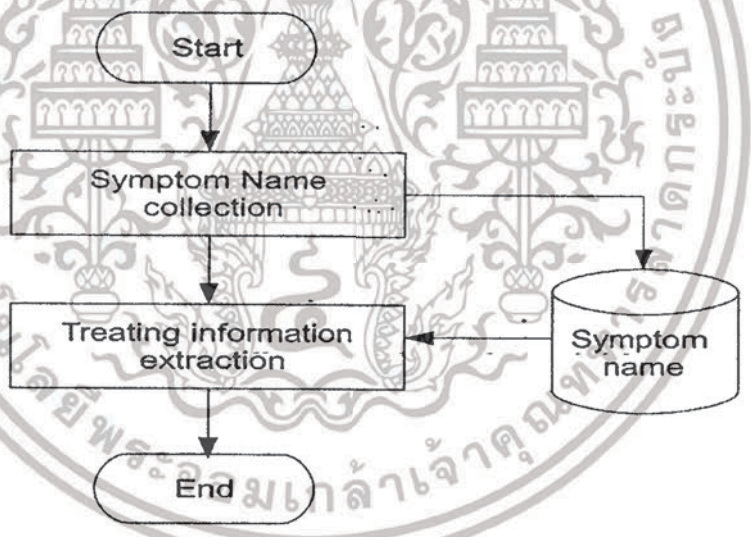
[illegible]	[illegible]	[illegible]	[illegible]
%SNTittle:	เป็นสัญลักษณ์ที่ใช้หาหัวข้อของชื่อวิทยาศาสตร์	ชื่อวิทยาศาสตร์	ชื่อวิทยาศาสตร์ : Hibiscus sabdariffa Linn.
%FNTittle:	เป็นสัญลักษณ์ที่ใช้หาหัวข้อของชื่อวงศ์	ชื่อวงศ์	วงศ์ : Malvaceae
%CNTittle	เป็นสัญลักษณ์ที่ใช้หาหัวข้อของชื่อทั่วไปที่เป็นภาษาอังกฤษ	ชื่อสามัญ ชื่ออังกฤษ	ชื่อสามัญ : Jamaican Sorrel, Rosella
%GNTittle:	เป็นสัญลักษณ์ที่ใช้หาหัวข้อของชื่อท้องถิ่น	ชื่ออื่นๆ	ชื่ออื่น ๆ : กระเจี๊ยบเปรี้ยว(ภาคกลาง), ส้มเก็งเค็ง(ภาคเหนือ), ส้มปู(เงี้ยว-แม่ฮ่องสอน), ส้มตะเลงเครง(ตาก) , ผักเก็งเค็ง,ส้มพอเหมาะ
%PATTittle:	เป็นสัญลักษณ์ที่ใช้หาหัวข้อของสรรพคุณ	สรรพคุณ	สรรพคุณ : ยอดและใบช่วยย่อยอาหาร ละลายเสมหะ ขับปัสสาวะ หล่อลื่นลำไส้

จากตารางที่ 3.1 สัญลักษณ์ที่ใช้สำหรับกำหนด หรือหาหัวข้อของหัวข้อสำหรับโครงสร้าง HTML ที่มีความไม่แน่นอน สัญลักษณ์ที่ใช้ในงานวิจัยนี้ประกอบด้วย 5 สัญลักษณ์ ซึ่งแต่ละสัญลักษณ์จะประกอบคำบ่งชี้ของแต่ละหัวข้อ เช่น สัญลักษณ์ %SNTittle: เป็นสัญลักษณ์ที่ใช้หาชื่อวิทยาศาสตร์ เป็นต้น สำหรับเว็บไซต์ที่มีแท็ก HTML แน่นอนมีทั้งหมด 4 เว็บไซต์ ซึ่งประกอบด้วย ฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี เว็บไซต์ถัดมาคือฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี เว็บไซต์ที่สามคือฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกข

ชาติ และเว็บไซต์สุดท้ายคือ ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย ในส่วนของเว็บไซต์ที่มีแท็ก HTML ที่ไม่แน่นอน คือ ฐานข้อมูลพืชสมุนไพรไทยพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม

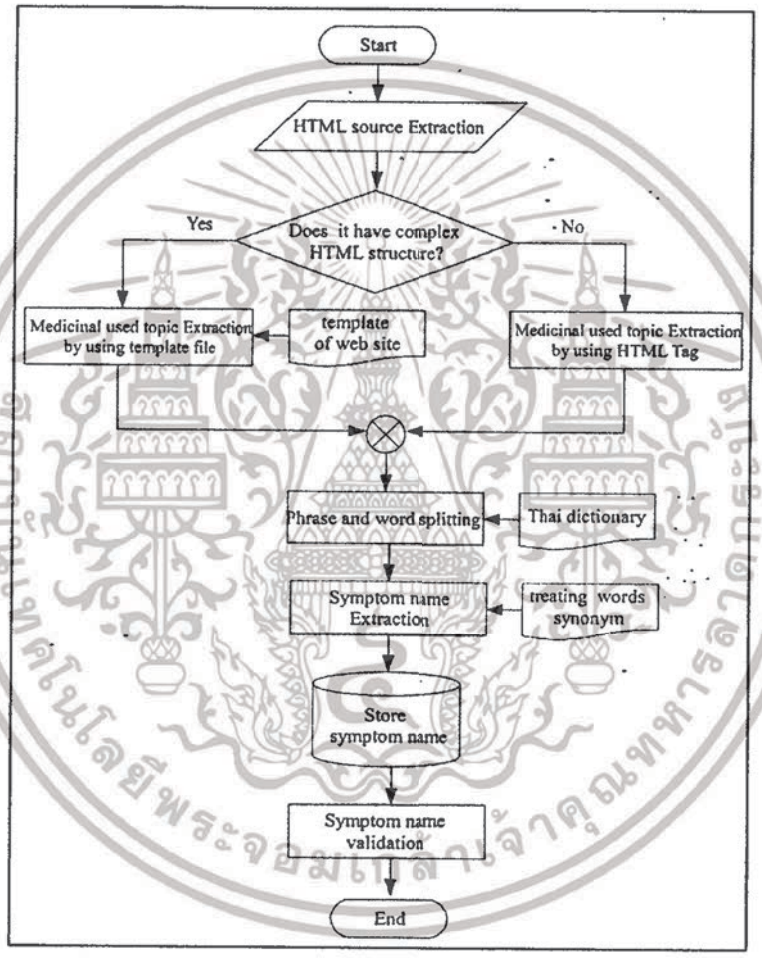
3.2 กระบวนการของการสกัดพืชสมุนไพรไทยจากหลายเว็บไซต์

สำหรับกระบวนการของการสกัดพืชสมุนไพรไทย แสดงดังรูปที่ 3.3 ขั้นตอนการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ ซึ่งขั้นตอนของการสกัดพืชสมุนไพรไทยแบ่งออกเป็น 2 ส่วนหลัก คือ การรวบรวมรายชื่อของอาการ (Symptom name collection) และการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย (Treating information extraction) ซึ่งประกอบด้วย ชื่อทางการ ชื่อสามัญ ชื่อวิทยาศาสตร์ ชื่อวงศ์ ชื่อท้องถิ่น ความเป็นพิษต่อระบบร่างกายส่วนที่ใช้รักษา และชื่ออาการ



รูปที่ 3.3 ขั้นตอนการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์

3.2.1. การรวบรวมชื่อของอาการ (Symptom name collection) เป็นกระบวนการสกัด และรวบรวมชื่ออาการจากหลายเว็บไซต์ เพื่อนำผลลัพธ์ที่เป็นชื่ออาการไปใช้ตรวจสอบว่าพืชสมุนไพรไทยในแต่ละชนิดสามารถรักษาอาการใดได้บ้าง ในขั้นตอนของการสกัดข้อมูลที่เกี่ยวข้อง ซึ่งขั้นตอนของการรวบรวมชื่ออาการแสดงในรูปที่ 3.4



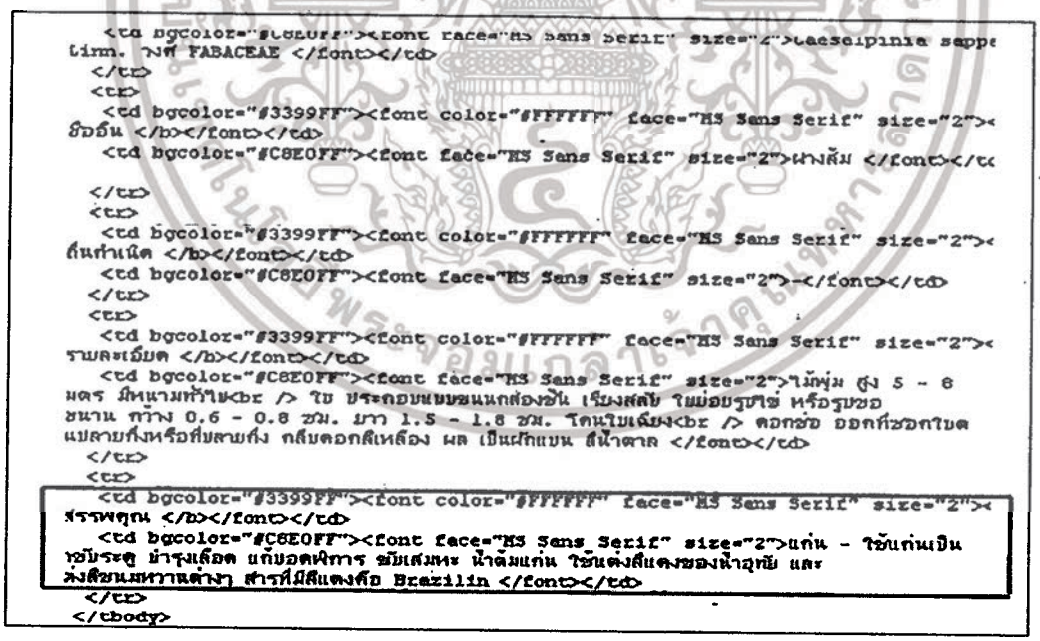
รูปที่ 3.4 ขั้นตอนของการรวบรวมชื่ออาการ

จากรูปที่ 3.4 ขั้นตอนของการรวบรวมชื่ออาการ สำหรับกระบวนการรวบรวมรายชื่อของอาการแบ่งออกเป็น 5 กระบวนการหลัก ดังต่อไปนี้ กระบวนการแรกคือการดึงซอร์สโค้ดแต่ละเว็บเพจ (HTML Source Extraction) กระบวนการถัดมาคือการกำหนดหัวข้อของสรรพคุณ (Medicinal-used Topic Extraction) กระบวนการที่สามคือการแบ่งแยกประโยคและตัดคำ (Phrase and word splitting)

กระบวนการที่สี่คือ การสกัดชื่ออาการ (Symptom Name Extraction) และกระบวนการสุดท้ายคือการตรวจสอบรายชื่ออาการ (Symptom Name Validation) ซึ่งแต่ละกระบวนการมีรายละเอียดดังต่อไปนี้

1. การดึงซอร์สโค้ดแต่ละเว็บเพจ (HTML Source Extraction) เป็นขั้นตอนของการดึงซอรสโค้ดของแต่ละเว็บเพจ เพื่อนำไปใช้ในการหาหัวข้อของสรรพคุณ ซึ่งกระบวนการนี้ได้ใช้ JSOUP API เป็นเครื่องมือที่ใช้ดึงซอร์สโค้ดมาจากแต่ละเว็บเพจ

2. การกำหนดหัวข้อของสรรพคุณ (Medicinal-used Topic Extraction) เป็นขั้นตอนที่หาเนื้อหาที่เกี่ยวข้องกับสรรพคุณที่อยู่ในแต่ละเว็บเพจ เพื่อนำเนื้อหาที่อยู่ในหัวของสรรพคุณในแต่ละเว็บเพจไปใช้ในการหาชื่ออาการในขั้นตอนถัดไป สำหรับขั้นตอนการกำหนดหัวข้อของสรรพคุณแบ่งออกเป็น 2 วิธีคือ วิธีการกำหนดหัวข้อสรรพคุณสำหรับโครงสร้าง HTML ที่ไม่แน่นอน และการกำหนดหัวข้อสรรพคุณสำหรับโครงสร้าง HTML ที่แน่นอน สำหรับการกำหนดหัวข้อสรรพคุณสำหรับโครงสร้าง HTML ที่แน่นอน จะใช้ JSOUP API ในการค้นหาแท็ก HTML ที่บ่งบอกถึงหัวข้อและเนื้อหาที่เกี่ยวข้องกับสรรพคุณ ดังแสดงในรูปที่ 3.5 และผลลัพธ์ของการสกัดโครงสร้าง HTML ที่แน่นอนแสดงในรูปที่ 3.6

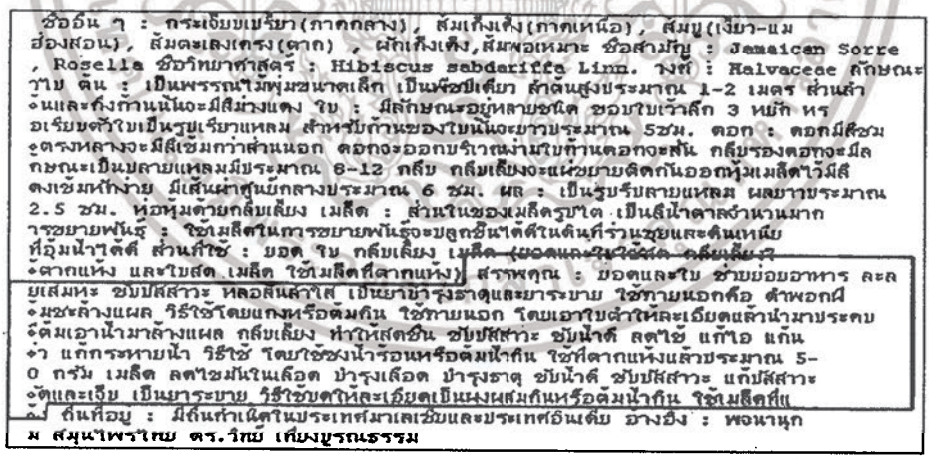


รูปที่ 3.5 ตัวอย่างข้อมูลของหัวข้อสรรพคุณ ที่อยู่ในโครงสร้าง HTML ที่แน่นอน

สรรพคุณ แก่น - ใช้แก่นเป็นยาชับระดู บำรุงเลือด แก้บอดพิการ ชับเสมหะ น้ำต้มแก่น ใช้แต่งสีแดงของน้ำอุหัย และแต่งสีชนมหาหาแต่งๆ สารที่มีสีแดงคือ Brazilin

รูปที่ 3.6 ตัวอย่างข้อมูลที่ถูกสกัดจากโครงสร้าง HTML ของหัวข้อสรรพคุณ

สำหรับการกำหนดหัวข้อสรรพคุณสำหรับโครงสร้าง HTML ที่ซับซ้อน จะใช้เทมเพลตของแต่ละเว็บไซต์ ในการกำหนดหัวข้อและเนื้อหาที่เกี่ยวข้องกับสรรพคุณซึ่งขั้นตอนของการกำหนดหัวข้อสรรพคุณสำหรับโครงสร้าง HTML ที่ซับซ้อน แบ่งออกเป็น 3 กระบวนการย่อยคือ ขั้นแรกจะทำการลบแท็ก HTML ที่อยู่ในแต่ละเว็บเพจทั้งหมด ซึ่งแสดงในรูปที่ 3.7 จากรูปดังกล่าวข้อมูลที่ล้อมรอบด้วยกรอบสีดำแสดงถึงข้อมูลที่เกี่ยวข้องกับสรรพคุณ ขั้นที่สองจะทำการตัดคำโดย (word segmentation) โดยใช้เครื่องมือ LEXTO[14] และขั้นตอนสุดท้ายจะนำคำที่ตัดมาทำการเปรียบเทียบกับคำบ่งชี้ของหัวข้อ (Indicator word of topic) ตามที่กล่าวในข้างต้น ถ้ามีคำที่เหมือนกับคำบ่งชี้หัวข้อของสรรพคุณก็จะถูกแท็กด้วยสัญลักษณ์ %PATTittle: แล้วทำการเก็บคำถัดไปเรื่อยๆ จนพบคำบ่งชี้ที่ไม่เกี่ยวข้องกับหัวข้อของสรรพคุณ ซึ่งผลลัพธ์แสดงของกระบวนการนี้แสดงดังรูปที่ 3.8

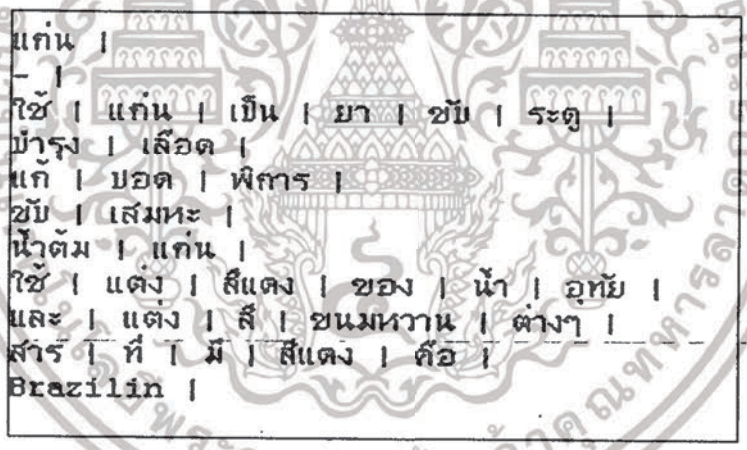

ชื่ออื่น ๆ : กระเจี๊ยบเปรี้ยว(ภาคกลาง), ส้มเก็งเค็ง(ภาคเหนือ), ส้มปู(เงี้ยว-แม
ฮ่องสอน), ส้มตะเลงเครง(ตาก) , ผักเก็งเค็ง,ส้มพอเหมาะ ชื่อสามัญ : Jamaican Sorre
, Roselle ชื่อวิทยาศาสตร์ : Hibiscus sabdariffa Linn. วงศ์ : Malvaceae ลักษณะ
วิสัย ต้น : เป็นพรรณไม้พุ่มขนาดเล็ก เป็นพืชปีเดียว ลำต้นสูงประมาณ 1-2 เมตร ลำต้น
และกิ่งก้านมีสีม่วงแดง ใบ : มีลักษณะอยู่หลายชนิด ชอบใบเว้าลึก 3 หยัก หร
อเรียงตัวใบเป็นรูปเรียวแหลม สำหรับก้านของใบนั้นจะยาวประมาณ 5ซม. ดอก : ดอกมีสีชม
ตรงหลางจะมีสีเข้มกว่าส่วนนอก ดอกจะออกบริเวณง่ามใบก้านดอกจะสั้น กลีบรองดอกจะมีล
กษณะเป็นปลายแหลมมีประมาณ 8-12 กลีบ กลีบเลี้ยงจะแผ่ขยายติดกันออกหุ้มเมล็ดไว้มีสี
ดงเข้มหักง่าย มีเส้นผ่าศูนย์กลางประมาณ 6 ซม. ผล : เป็นรูปรีปลายแหลม ผลยาวประมาณ
2.5 ซม. หุ้มหุ้มด้วยกลีบเลี้ยง เมล็ด : ส่วนในของเมล็ดรูปไต เป็นสีน้ำตาลจำนวนมาก
การขยายพันธุ์ : ใช้เมล็ดในการขยายพันธุ์จะปลูกขึ้นได้ดีในดินที่ร่วนซุยและดินเหนีย
ที่อุ้มน้ำได้ดี ส่วนที่ใช้ : ยอด ใบ กลีบเลี้ยง เมล็ด (ยอดและใบใช้สด กลีบเลี้ยงใ
ตากแห้ง และใบสด เมล็ด ใช้เมล็ดที่ตากแห้ง) สรรพคุณ : ยอดและใบ ช่วยย่อยอาหาร ละล
ยเสมหะ ชับปัสสาวะ หลอดลมอักเสบ เป็นยาบำรุงธาตุและยาระบาย ใช้ภายนอกคือ ตำพอกฝี
มดชะล้างแผล วิธีใช้โดยแกงหรือต้มกิน ใช้ภายนอก โดยเอาใบตำให้ละเอียดแล้วนำมาประคบ
ต้มเอาน้ำมาล้างแผล กลีบเลี้ยง ทำให้สุกขึ้น ชับปัสสาวะ ชับน้ำดี ลดไข้ แก้ไอ แก้น
ว่า แก้กระหายน้ำ วิธีใช้ โดยใช้ชงน้ำร้อนหรือต้มน้ำกิน ใช้ที่ตากแห้งแล้วประมาณ 5-
0 กรัม เมล็ด ลดไขมันในเลือด บำรุงเลือด บำรุงธาตุ ชับน้ำดี ชับปัสสาวะ แก้ปัสสาวะ
ตัและเจ็บ เป็นยาระบาย วิธีใช้ชงให้ละเอียดเป็นผงผสมกินหรือต้มน้ำกิน ใช้มเมล็ดที่แ
ง้ ถิ่นที่อยู่ : มีถิ่นกำเนิดในประเทศมาเลเซียและประเทศอินเดีย อ้างอิง : พจนานุก
ม สมุนไพรไทย ดร.วิทย์ เทียงบูรณธรรม

รูปที่ 3.7 ตัวอย่างหัวข้อสรรพคุณที่อยู่ในโครงสร้าง HTML ที่ไม่แน่นอน

สรรพคุณ :%PATTitle: ยอดและใบ ช่วยย่อยอาหาร ละลายเสมหะ ชับปัสสาวะ หล่อลื่นลำไส้ เป็นยาบำรุงธาตุและยาระบาย ใช้ภายนอกคือ ตำพอกฝี ต้มชะล้างแผล วิธีใช้โดยแกงหรือต้มกิน ใช้ภายนอก โดยเอาใบตำให้ละเอียดแล้วนำมาประคบฝีต้มเอาน้ำมาล้างแผล กลั้วเลี้ยง ทำให้สดชื่น ชับปัสสาวะ ชับน้ำดี ลดไข้ แก้ไอ แก้หืด แก้กระหายน้ำ วิธีใช้ โดยใช้ชงน้ำร้อนหรือต้มน้ำกิน ใช้ที่ตากแห้งแล้วประมาณ 5-10 กรัม เมล็ด ลดไขมันในเลือด บำรุงเลือด บำรุงธาตุ ชับน้ำดี ชับปัสสาวะ แก้ปัสสาวะขัดและเจ็บ เป็นยาระบาย วิธีใช้บดให้ละเอียดเป็นผงผสมกินหรือต้มน้ำกิน ใช้เมล็ดที่แห้ง

รูปที่ 3.8 ตัวอย่างข้อมูลหลังจากการสัดข้อมูลที่อยู่ในโครงสร้าง HTML ที่ไม่แน่นอน

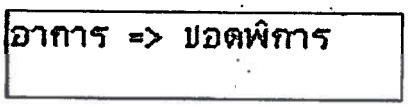
3. การแบ่งแยกประโยคและตัดคำ (Phrase and word splitting) ขั้นตอนนี้จะทำการแบ่งประโยคเนื้อหาของสรรพคุณที่ได้จากกระบวนการของการกำหนดหัวข้อของสรรพคุณ โดยใช้ช่องว่าง (white space) และนำประโยคดังกล่าวนำมาสู่กระบวนการตัดคำ เพื่อนำคำดังกล่าวไปหาชื่ออาการในขั้นตอนต่อไป ซึ่งผลลัพธ์แสดงดังรูปที่ 3.9



รูปที่ 3.9 ผลลัพธ์ของการแยกประโยคและตัดคำของเนื้อหาสรรพคุณ

4. การสกัดชื่ออาการ (Symptom Name Extraction) ขั้นตอนนี้เป็นขั้นตอนที่หาชื่ออาการโดยใช้คำบ่งชี้ของการรักษา (indicator word of treatment) ซึ่งคำที่บ่งบอกอาการที่งานวิจัยนี้ใช้มีทั้งหมด 13 คำบ่งบอกอาการ ซึ่งประกอบด้วย แก้ รักษา รักษาโรค รักษาอาการ บำบัด บำบัดอาการ บำบัดโรค บรรเทา บรรเทาโรค บรรเทาอาการ เป็นยา เป็นยาแก้ และเป็นยาแก้โรค เป็นยาแก้อาการ เนื่องจากสมมติฐานของงานวิจัยฉบับนี้คือ ชื่ออาการจะตามด้วยคำบ่งชี้ของการรักษา เช่น แก้ปอดพิการ รักษา

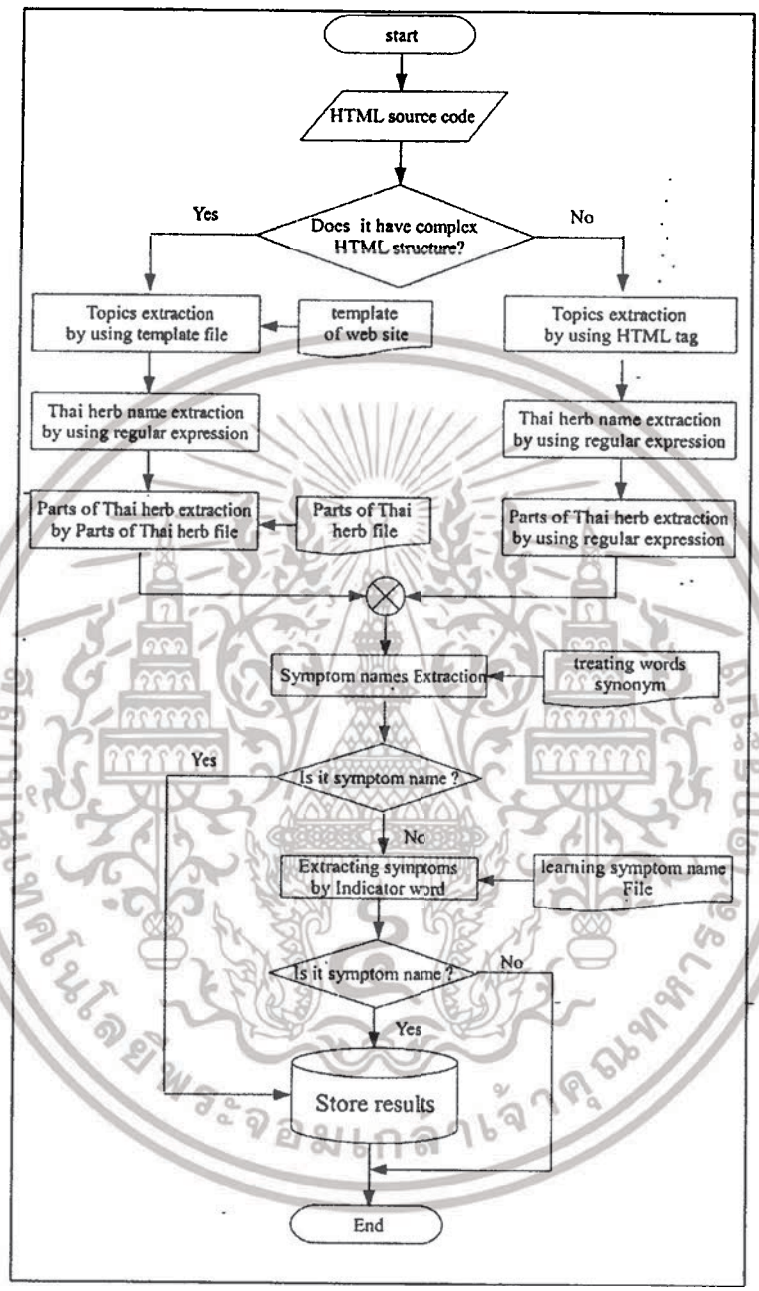
อาการปวดท้อง เป็นต้น ดังนั้นถ้ามีคำที่ต่อจากคำบ่งชี้ของการรักษาจะถูกเก็บไว้ในฐานความรู้ เพื่อนำไปใช้งานต่อไป สำหรับผลลัพธ์ของขั้นตอนนี้แสดงดังรูปที่ 3.10



รูปที่ 3.10 แสดงผลลัพธ์ของการสกัดข้อมูลโดยใช้คำบ่งชี้

5. การตรวจสอบรายชื่ออาการ (Symptom Name Validation) ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการรวบรวมรายชื่ออาการ ซึ่งขั้นตอนจะทำการตรวจสอบความถูกต้องชื่ออาการที่จัดเก็บไว้ในฐานความรู้ โดยใช้มนุษย์เป็นผู้ตรวจตรวจสอบ ถ้าชื่ออาการในฐานความรู้นั้นไม่ถูกต้องก็จะถูกลบออกโดยทันที ซึ่งผลลัพธ์ของกระบวนการนี้จะถูกเรียกว่าไฟล์การเรียนรู้ชื่ออาการ (Learned Symptom Name)

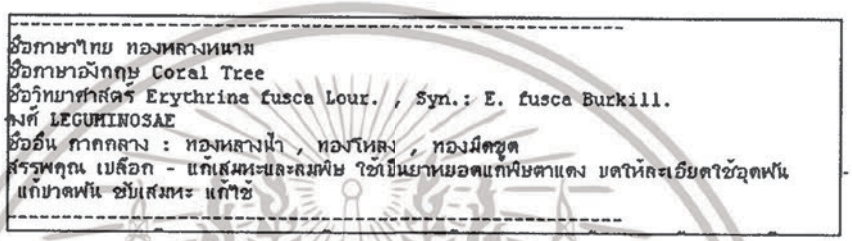
3.2.2. การสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย (Treating Information Extraction) ขั้นตอนนี้เป็นขั้นตอนของการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ข้อมูลพืชสมุนไพรไทยที่วิจัยฉบับนี้ได้สกัดออกมาประกอบด้วย ชื่อทางการ ชื่อวิทยาศาสตร์ ชื่อท้องถิ่น ชื่อวงศ์ ส่วนที่ใช้รักษาอาการ และความเป็นพิษของพืชสมุนไพรไทย สำหรับกระบวนของการของการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทยแสดงในรูปที่ 3.11 แบ่งออกเป็น 3 กระบวนการหลัก ดังต่อไปนี้ กระบวนการแรกคือการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้อง กระบวนการถัดมาคือ การสกัดข้อมูลพืชสมุนไพรไทยโดยใช้ regular expression และ กระบวนการสุดท้ายคือการสกัดส่วนที่ใช้และชื่ออาการ ซึ่งแต่ละกระบวนการมีรายละเอียดดังต่อไปนี้



รูปที่ 3.11 กระบวนการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย

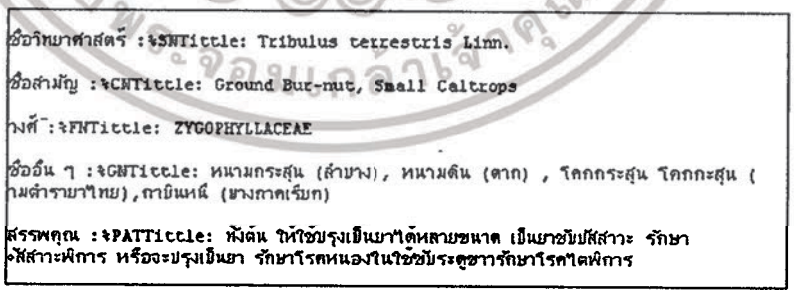
1. การกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้อง ขั้นตอนนี้จะทำการหาหัวข้อและเนื้อหาที่สนใจที่อยู่ในแต่ละเว็บเพจ ประกอบด้วยชื่อทางการ ชื่อวิทยาศาสตร์ ชื่อท้องถิ่น ชื่อวงศ์ ส่วนที่ใช้รักษาอาการ และความเป็นพิษของพืชสมุนไพรไทย ขั้นตอนของการกำหนดข้อมูลทีเกี่ยวข้องแบ่ง

ออกเป็น 2 วิธีคือ วิธีการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้องสำหรับโครงสร้าง HTML ที่ไม่ซับซ้อน และการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้องสำหรับโครงสร้าง HTML ที่ซับซ้อน สำหรับวิธีการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้องสำหรับโครงสร้าง HTML ที่ไม่ซับซ้อน จะใช้ JSOUP API ในการค้นหาแท็ก HTML ที่เกี่ยวข้องกับกับข้อมูลพืชสมุนไพรไทย ซึ่งผลลัพธ์ของการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้องสำหรับโครงสร้าง HTML ที่แน่นอนในรูปที่ 3.12



รูปที่ 3.12 ผลลัพธ์ของการกำหนดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ในเว็บเพจที่มีโครงสร้าง HTML ที่แน่นอน

สำหรับวิธีการกำหนดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้องสำหรับโครงสร้าง HTML ที่ไม่แน่นอนจะใช้สัญลักษณ์จะใช้ไฟล์เทมเพลตของแต่ละเว็บไซต์ ซึ่งกระบวนการนี้จะคล้ายกับขั้นตอนของการกำหนดหัวข้อของสรรพคุณ แต่ในกระบวนการนี้จะทำการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ซึ่งผลลัพธ์แสดงดังรูปที่ 3.13

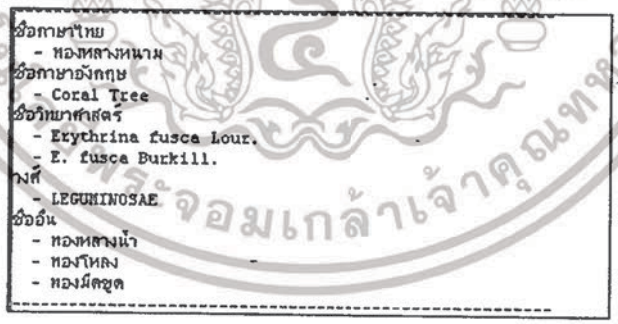


รูปที่ 3.13 ผลลัพธ์ของการกำหนดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ในเว็บเพจที่มีโครงสร้าง HTML ที่ไม่แน่นอน

2. การสกัดข้อมูลพืชสมุนไพรไทยโดยใช้ regular expression ขั้นตอนนี้เป็นการสกัดชื่อของพืชสมุนไพรไทยต่างๆ และความเป็นพิษต่อระบบร่างกายของพืชสมุนไพรไทยโดยใช้ regular expression สำหรับตัวอย่าง regular expression ที่ใช้แสดงในตารางที่ 3.2 สำหรับ regular expression ทั้งหมดที่งานวิจัยฉบับนี้ได้นำมาใช้แสดงในภาคผนวก ก และผลลัพธ์ของการสกัดชื่อพืชสมุนไพรไทยแสดงในรูปที่ 3.14

ตารางที่ 3.2 ตัวอย่าง Regular expression ที่ใช้สกัดข้อมูลพืชสมุนไพรไทย

ตัวอย่างข้อความ	Regular expression	ชื่อหัวข้อ
มะละกอ	([ก-ฮ]+)	ชื่อทางการ
Coconut	([A-z]+)	ชื่อสามัญ
Artocarpus lakoocha Roxb.	(([.|a-z|A-Z|(|)]+)\\s*)+.	ชื่อวิทยาศาสตร์
กระเจี๊ยบแดง, ส้มพอเหมาะ, ส้มเก็ง (เหนือ), ส้มปู (เงี้ยว), ส้มพอดี (อีสาน)	(([ก-ฮ]+)\\s*)[()ก-ฮ]*\\s*,*\\s*+	ชื่อท้องถิ่น
ระคายเคืองต่อผิวหนัง	([ก-ฮ]+)	ความเป็นพิษต่อระบบร่างกาย



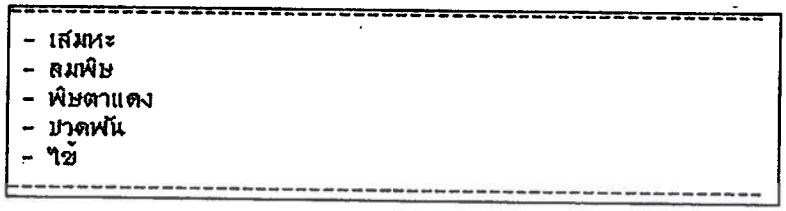
รูปที่ 3.14 ผลลัพธ์ของการสกัดชื่อพืชสมุนไพรไทย

3. การสกัดส่วนที่ใช้และชื่ออาการ ขั้นตอนนี้เป็นขั้นตอนสุดท้ายของการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทย ซึ่งขั้นตอนนี้จะทำการสกัดส่วนที่ใช้ และชื่ออาการที่พืชสมุนไพรไทยแต่ละชนิดสามารถรักษาได้ โดยขั้นตอนของการหาส่วนประกอบของพืชสมุนไพรไทยที่ใช้รักษาอาการ มี 2 วิธี ดังต่อไปนี้ วิธีแรกสำหรับเว็บเพจที่มีโครงสร้าง HTML ไม่ซับซ้อนจะใช้ regular

expression ในการหาส่วนประกอบของพืชสมุนไพรไทย ตัวอย่างเช่น "*([ก-ธ]+)\ใช*–\ใช*(([ก-ธ|a-z|A-Z|0]+\ใช*)*)*" เป็น regular expression ที่สามารถใช้ได้กับข้อความ "*เปลือก – แก้เสมหะและลมพิษ ใช้เป็นยาหยอดพิษแก้ตาแดง บดให้ละเอียดใช้อุดฟัน แก้ปวดฟัน ขับเสมหะ แก้ไข้*" ส่วนเว็บเพจที่มีโครงสร้าง HTML ที่ซับซ้อนจะใช้คำบ่งชี้ของส่วนประกอบของพืชสมุนไพรไทย เช่น ใบ ราก ลำต้น สำหรับขั้นตอนการสกัดส่วนที่ใช้รักษาอาการของโครงสร้าง HTML ที่ซับซ้อน แบ่งออกเป็น 3 กระบวนการย่อย ดังต่อไปนี้ กระบวนการแรกจะทำการแยกประโยคด้วยช่องว่าง และนำประโยคเหล่านั้นเข้าสู่กระบวนการตัดคำ เพื่อให้ง่ายต่อการหาส่วนที่ใช้และชื่ออาการในขั้นตอนต่อไป กระบวนการถัดมาคือ การสกัดหาส่วนประกอบของพืชสมุนไพรไทยที่ใช้รักษาอาการของผู้ใช้ ถ้ามีคำใดที่ตรงกับคำบ่งชี้ของส่วนประกอบของพืชสมุนไพรไทยก็จะทำการเก็บคำนั้นเป็นคำที่บ่งบอกถึงส่วนที่ใช้ ซึ่งผลลัพธ์ของการสกัดส่วนที่ใช้แสดงในรูปที่ 3.15 โดยผลลัพธ์ของขั้นตอนนี้แบ่งออกเป็น 2 ส่วน ส่วนแรกเป็นส่วนที่แสดงถึงส่วนประกอบของพืชสมุนไพรไทยที่ใช้รักษาอาการ และส่วนสุดท้ายแสดงถึงเนื้อหาของอาการ ซึ่งผลลัพธ์ในส่วนของเนื้อหาของอาการจะถูกนำไปประมวลผลในกระบวนการถัดไป

รูปที่ 3.15 ผลลัพธ์ของการสกัดส่วนที่ใช้และชื่ออาการ

กระบวนการสุดท้ายเป็นการหาชื่ออาการ กระบวนการนี้เป็นการหาชื่ออาการจากเนื้อหาของอาการ ซึ่งกระบวนการนี้จะคล้ายคลึงกับกระบวนการรวบรวมชื่ออาการ กล่าวคือถ้ามีคำใดที่ต่อท้ายคำบ่งชี้ของอาการ เช่น แก้ รักษา บำบัด เป็นต้น ก็จะทำการเก็บเป็นชื่ออาการ และนอกจากนี้ ถ้าคำใดที่ไม่มีคำบ่งชี้ของอาการนำหน้า ก็จะนำคำนั้นมาทำการเปรียบเทียบกับฐานความรู้ของอาการที่ได้จัดเก็บอยู่ในไฟล์เรียนรู้ชื่ออาการ (Learned Symptom Name) ถ้าคำใดตรง คำเหล่านั้นจะถูกเก็บเป็นชื่ออาการ ซึ่งผลลัพธ์แสดงดังรูปที่ 3.16 และรูปที่ 3.17



รูปที่ 3.16 ผลลัพธ์ของการสกัดชื่ออาการโดยใช้คำบ่งชี้

รูปที่ 3.17 ผลลัพธ์ของการสกัดชื่ออาการ โดยใช้ไฟล์ "Learned symptom name"

บทที่ 4
ผลการวิจัย

ในการวัดประสิทธิภาพของการสกัดข้อมูลพืชสมุนไพรไทยมีจุดประสงค์เพื่อตรวจสอบความแม่นยำ และความครบถ้วนของการสกัดข้อมูลชื่อทางการ ชื่อสามัญ ชื่อท้องถิ่น ชื่อวงศ์ของพืชสมุนไพรไทย ส่วนที่ใช้รักษาอาการ และอาการที่พืชสมุนไพรไทยแต่ละชนิดรักษาได้ รวมถึงความเป็นพิษต่อระบบร่างกายของมนุษย์ ในการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ ในวิทยานิพนธ์นี้ได้ใช้ข้อมูลพืชสมุนไพรไทยจาก 5 เว็บไซต์ ดังต่อไปนี้

1. ฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี (www.dru.ac.th)
2. ฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี (http://www.rspg.or.th/plants_data/herbs/herbs_ 200.htm)
3. ฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ (http://www.pharmacy.mahidol.ac.th/siri/index.php?page=home)
4. ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย (http://www.medplant.mahidol.ac.th/tpex/ toxic_mnu3.asp)
5. ฐานข้อมูลพืชสมุนไพรไทยพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม (http://www.samunpri.com/)

ในการวัดประสิทธิภาพของการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ ในวิทยานิพนธ์นี้ได้จะวัดความถูกต้อง (Precision) วัดค่าครบถ้วนในการสืบค้น (Recall) และค่าเฉลี่ยของความถูกต้องและความครบถ้วนในการสืบค้น (F-measure) ดังแสดงในสมการ ที่ 4.1 4.2 และ 4.3 ตามลำดับ

$$Precision = \frac{RelevantSelectionsBySystem}{AllSelectionBySystem} \quad (4.1)$$

$$Recall = \frac{RelevantSelectionsBySystem}{AllRelevantInWebpage} \quad (4.2)$$

$$F - measure = \frac{2 \times precision \times recall}{precision+recall} \quad (4.3)$$

โดยที่

- *RelevantSelectionsBySystem* คือ จำนวนข้อมูลพืชสมุนไพรไทยที่ระบบสกัดได้ถูกต้อง
- *AllSelectionBySystem* คือ จำนวนข้อมูลพืชสมนุนไพรไทยที่ระบบสกัด
- *AllRelevantInWebpage* คือ จำนวนข้อมูลพืชสมุนไพรไทยที่มีอยู่ในเว็บเพจ

ในการวัดประสิทธิภาพโดยใช้ precision เป็นการวัดความแม่นยำของการสกัดข้อมูลพืชสมุนไพรไทยของระบบ สำหรับการวัดประสิทธิภาพโดยใช้ recall เป็นการวัดความครบถ้วนของการสกัดข้อมูลพืชสมุนไพรไทยของระบบ และการวัดประสิทธิภาพ F-measure เพื่อใช้เป็นการเปรียบเทียบประสิทธิภาพของการสกัดข้อมูลชื่ออาการที่ระบบสามารถสกัดได้

โดยกระบวนการวัดประสิทธิภาพของการสกัดข้อมูลพืชสมุนไพรไทย วิทยานิพนธ์ฉบับนี้ได้สุ่มเลือกเว็บเพจที่อยู่ในแต่ละเว็บไซต์อย่างละ 20 เว็บเพจ ซึ่งการวัดประสิทธิภาพได้แบ่งออกเป็น 2 ส่วนหลัก

4.1 การวัดประสิทธิภาพการเปรียบเทียบของการสกัดชื่ออาการโดยใช้ไฟล์เรียนรู้ชื่ออาการและไม่ใช้ไฟล์เรียนรู้ชื่ออาการ

ส่วนแรกเป็นการเปรียบเทียบของการสกัดชื่ออาการโดยใช้ไฟล์การเรียนรู้ชื่ออาการ (Learned Symptom Name) และไม่ใช้ไฟล์การเรียนรู้ชื่ออาการ ในการวัดผลของการสกัดชื่ออาการ เนื่องจากข้อมูลในเว็บไซต์ของฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย [5] ไม่ได้มีข้อมูลของชื่ออาการ ในวิทยานิพนธ์ฉบับนี้ จึงไม่ได้ใช้ข้อมูลจากเว็บไซต์ดังกล่าวในการสกัดข้อมูลชื่ออาการ สำหรับขั้นตอนการสร้างไฟล์เรียนรู้ชื่ออาการ ในวิทยานิพนธ์ฉบับนี้ได้สุ่มเลือกเว็บเพจที่อยู่ในแต่ละเว็บไซต์อย่างละ 80 เว็บเพจ และทำการสกัดข้อมูลชื่ออาการโดยใช้เฉพาะคำบ่งชี้ของอาการ เช่น รักษา หรือ บรรเทา เป็นต้น ดังที่กล่าวถึงในบทที่ 3 ในส่วนของขั้นตอนการรวบรวมชื่ออาการ สำหรับผลลัพธ์ของการวัดประสิทธิภาพโดยใช้ไฟล์เรียนรู้ชื่ออการ และไม่ใช้ไฟล์เรียนรู้ชื่ออาการ แสดงในตารางที่ 4. 1 และตารางที่ 4. 2 ตามลำดับ

ตารางที่ 4. 1 การสกัดข้อมูลชื่ออาการโดยไม่ใช้ไฟล์เรียนรู้ชื่ออาการ

เว็บไซต์	จำนวนข้อมูล		จำนวนข้อมูลทั้งหมด	F-measure
	ข้อมูลที่ถูก	ข้อมูลที่ผิด		
[1]	87	1	128	80%
[2]	170	15	312	68%
[3]	161	18	268	72%
[4]	44	1	81	70%
ผลรวม	**462**	**35**	**789**	**72%**

ตารางที่ 4. 2 การสกัดข้อมูลอาการโดยใช้ไฟล์เรียนรู้ชื่ออาการ.

เว็บไซต์	จำนวนข้อมูล		จำนวนข้อมูลทั้งหมด	F-measure
	ข้อมูลที่ถูก	ข้อมูลที่ผิด		
[1]	99	8	128	86%
[2]	271	37	312	87%
[3]	223	30	268	86%
[4]	74	3	81	94%
ผลรวม	**667**	**78**	**789**	**86%**

จากผลการทดลองของการสกัดชื่ออาการ โดยไม่ใช้ไฟล์การเรียนรู้ชื่ออาการ ค่า F-measure อยู่ที่ 72% แต่หลังจากใช้ไฟล์การเรียนรู้ชื่ออาการ ค่า F-measure เพิ่มขึ้นมาอยู่ที่ 86% ซึ่งสามารถสรุปได้ว่า ถ้าใช้ไฟล์การเรียนรู้ในการสกัดชื่ออาการสามารถช่วยเพิ่มประสิทธิภาพในการสกัดข้อมูลชื่ออาการให้ได้ดียิ่งขึ้น

4.2 การวัดประสิทธิภาการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทยที่อยู่ในแต่ละเว็บไซต์

สำหรับการวัดผลการทดลองในส่วนที่สอง เป็นการวัดประสิทธิภาพของการสกัดข้อมูลที่เกี่ยวข้องกับพืชสมุนไพรไทยที่อยู่ในแต่ละเว็บไซต์ ซึ่งในงานวิทยานิพนธ์ฉบับนี้ จะสกัดข้อมูลพืชสมุนไพรไทยประกอบด้วย ชื่อทางการ, ชื่อวิทยาศาสตร์, ชื่อสามัญ, ชื่อท้องถิ่น, ชื่อวงศ์, ความเป็นพิษ, ส่วนที่ใช้รักษาอาการ และชื่ออาการ สำหรับผลลัพธ์ของการสกัดข้อมูลพืชสมุนไพรไทย แสดงในตารางที่ 4.3 และตารางที่ 4.4 จากสองตารางดังกล่าวสามารถอธิบายได้ว่า ผลลัพธ์ของการสกัดข้อมูลที่เกี่ยวข้องของพืชสมุนไพรไทย ค่า precision และ recall ของการสกัดชื่อทางการ และความเป็นพิษของพืชสมุนไพรไทยอยู่ที่ 100% ส่วนค่า precision และ recall ของชื่อวิทยาศาสตร์ ชื่อสามัญ ชื่ออื่นๆ และ ชื่อวงศ์ อยู่ที่ประมาณเกือบ 100% และค่า precision และ recall ของการสกัดส่วนที่ใช้อยู่ที่ 95% และ 92% ตามลำดับ ส่วนค่า precision และ recall ของการสกัดชื่ออาการอยู่ที่ 90% และ 85% ตามลำดับ

ตารางที่ 4. 3 จำนวนข้อมูลดิบของข้อมูลพืชสมุนไพรไทยที่ระบบสามารถสกัดได้

ชื่อหัวข้อ/ ชื่อเว็บไซต์		[1]	[2]	[3]	[4]	[5]	ผลรวม
ชื่อทางการ	จำนวนข้อมูลที่ถูก	20	20	20	20	20	100
	จำนวนข้อมูลที่ผิด	0	0	0	0	0	0
	จำนวนข้อมูลทั้งหมด	20	20	20	20	20	100
ชื่อวิทยาศาสตร์	ข้อมูลที่ถูก	18	20	25	20	20	103
	ข้อมูลที่ผิด	0	0	0	0	0	0
	จำนวนข้อมูลทั้งหมด	20	20	25	20	20	105
ชื่อสามัญ	ข้อมูลที่ถูก	20	23	16	-	42	101
	ข้อมูลที่ผิด	0	0	0	-	0	0
	จำนวนข้อมูลทั้งหมด	20	23	17	-	42	102
ชื่อท้องถิ่น	ข้อมูลที่ถูก	78	73	111	71	155	487
	ข้อมูลที่ผิด	0	0	5	0	0	5
	จำนวนข้อมูลทั้งหมด	81	80	113	71	155	500
ชื่อวงศ์	ข้อมูลที่ถูก	18	23	20	19	20	100
	ข้อมูลที่ผิด	0	0	0	1	0	1
	จำนวนข้อมูลทั้งหมด	18	23	20	20	20	101
ความเป็นพิษ	ข้อมูลที่ถูก					20	20
	ข้อมูลที่ผิด					0	0
	จำนวนข้อมูลทั้งหมด					20	20
ส่วนที่ใช้รักษาอาการ	ข้อมูลที่ถูก	45	67	53	43	-	208
	ข้อมูลที่ผิด	0	4	2	6	-	12
	จำนวนข้อมูลทั้งหมด	48	72	61	44	-	225
ชื่ออาการ	ข้อมูลที่ถูก	99	271	223	74		667
	ข้อมูลที่ผิด	8	37	30	3		78
	จำนวนข้อมูลทั้งหมด	128	312	268	81		789

ตารางที่ 4. 4 ค่า Precision และ Recall ของการสกัดข้อมูลพืชสมุนไพรไทย

ชื่อหัวข้อ/ข้อมูลที่ต้อง		[1]	[2]	[3]	[4]	[5]	ค่าเฉลี่ย
ชื่อทางการ	Precision	100%	100%	100%	100%	100%	100%
	Recall	100%	100%	100%	100%	100%	100%
ชื่อวิทยาศาสตร์	Precision	100%	100%	100%	100%	100%	100%
	Recall	90%	100%	100%	100%	100%	98%
ชื่อสามัญ	Precision	100%	100%	100%	-	100%	100%
	Recall	100%	100%	94%	-	100%	99%
ชื่อท้องถิ่น	Precision	100%	100%	96%	100%	100%	99%
	Recall	96%	100%	98%	100%	100%	99%
ชื่อวงศ์	Precision	100%	100%	100%	95%	100%	99%
	Recall	100%	100%	100%	95%	100%	99%
ความเป็นพิษ	Precision					100%	100%
	Recall					100%	100%
ส่วนที่ใช้รักษาอาการ	Precision	100%	94%	96%	88%	-	95%
	Recall	94%	93%	87%	96%	-	92%
ชื่ออาการ	Precision	93%	88%	86%	98%		91%
	Recall	80%	79%	83%	91%		83%

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1. สรุปผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อที่สกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ เนื่องจากข้อมูลในบางเว็บไซต์ยังมีไม่ครบถ้วน เช่น [1][2][3][4] ได้นำเสนอข้อมูลพืชสมุนไพรไทยที่สามารถรักษาอาการของผู้ใช้ได้ แต่ไม่ได้บอกความเป็นพิษของพืชสมุนไพรไทยแต่ละชนิด แต่ในทางกลับกัน [5] นำเสนอข้อมูลพืชสมุนไพรไทยที่เป็นพิษกับระบบร่างกายของพืชสมุนไพรไทย แต่ไม่ได้บอกถึงสรรพคุณของพืชสมุนไพรไทยแต่ละชนิด เป็นต้น นอกจากนี้บางเว็บเพจอาจนำเสนอสรรพคุณของพืชสมุนไพรไทยแต่ละชนิดไม่ครบถ้วน เช่น บางเว็บเพจนำเสนอสรรพคุณของขิงสามารถรักษาอาการปวดท้องได้ แต่บางเว็บเพจอาจไม่ได้นำเสนอสรรพคุณของขิงสามารถรักษาอาการดังกล่าวได้ เป็นต้น จึงอาจทำให้ผู้ใช้ที่สนใจข้อมูลพืชสมุนไพรไทยนั้นอาจต้องเข้าไปหลายเว็บไซต์เพื่อได้รับข้อมูลที่มีความครบถ้วน สำหรับเว็บไซต์ที่งานวิจัยนี้ได้ไปสกัดข้อมูลมีทั้งหมด 5 เว็บไซต์ ดังต่อไปนี้

- ฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี (www.dru.ac.th)
- ฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี (http://www.rspg.or.th/plants_data/herbs/herbs_ 200.htm)
- ฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ (http://www.pharmacy.mahidol.ac.th/siri/index.php?page=home)
- ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย (http://www.medplant.mahidol.ac.th/tpex/ toxic_mnu3.asp)
- ฐานข้อมูลพืชสมุนไพรไทยพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม (http://www.samunpri.com/)

สำหรับข้อมูลพืชสมุนไพรไทยที่งานวิจัยฉบับนี้จะไปสกัดออกมาในแต่ละเว็บไซต์ประกอบด้วย ชื่อทางการ ชื่อสามัญ ชื่อวิทยาศาสตร์ ชื่อวงศ์ ชื่อท้องถิ่น ความเป็นพิษต่อระบบร่างกาย ส่วนที่ใช้รักษา และชื่ออาการ ซึ่งแต่ละเว็บไซต์ที่กล่าวมาในข้างต้น ผู้วิจัยได้ทำการสำรวจแท็ก HTML ของแต่ละเว็บไซต์ก่อนที่จะทำการสกัดข้อมูลพืชสมุนไพรไทย ทำให้ผู้วิจัยสามารถจัดกลุ่มแท็ก HTML ที่แสดงข้อมูลพืชสมุนไพรไทยออกเป็น 2 กลุ่ม คือ กลุ่มแรกคือแท็ก HTML ที่มีรูปแบบแท็กแน่นอน

ตรงตามข้อมูลที่ต้องการจัดเก็บ และ กลุ่มที่สองคือแท็ก HTML ที่มีรูปแบบไม่แน่นอน ซึ่งโครงสร้าง HTML ที่แน่นอนนี้ งานวิจัยนี้ได้ใช้ JSOUP API เป็น HTML Parser ในการสกัดข้อมูลแต่ละหน้า และสำหรับโครงสร้าง HTML ที่มีไม่แน่นอน จะใช้เทมเพลตที่ประกอบด้วยสัญลักษณ์และคำบ่งชี้ของแต่ละหัวข้อ (indicator word of topic) ในการสกัดข้อมูลแต่ละหน้า

ซึ่งกระบวนการสำหรับการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ ประกอบด้วย 2 ขั้นตอนหลัก คือ ขั้นตอนของการรวบรวมชื่ออาการ และขั้นตอนการสกัดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้อง ในขั้นตอนของการรวบรวมชื่ออาการ เป็นกระบวนการที่สกัดและรวบรวมชื่ออาการ โดยในงานวิจัยนี้ได้สกัดชื่ออาการโดยใช้คำบ่งบอกถึงชื่ออาการ เช่น รักษาอาการ หรือ บรรเทาอาการ เป็นต้น เพราะข้อมูลชื่ออาการส่วนใหญ่จะนำหน้าด้วยคำบ่งบอกที่กล่าวถึงในข้างต้น เช่น รักษาอาการไอ หรือ อาการปวดท้อง เป็นต้น ซึ่งผลลัพธ์ของการหาชื่ออาการจะถูกจัดเก็บไว้ เพื่อถูกนำไปใช้ในขั้นตอนที่สองเพื่อตรวจสอบว่า พืชสมุนไพรไทยแต่ละชนิดสามารถรักษาอาการใดได้บ้าง นอกจากนี้ในขั้นตอนที่สองยังสกัดข้อมูลชื่อของพืชสมุนไพรไทย เช่น ชื่อที่เป็นทางการ ชื่อวิทยาศาสตร์ ชื่อวงศ์ และ ชื่อท้องถิ่นของพืชสมุนไพรไทย แต่ละชนิด โดยใช้ regular expression ซึ่งผลลัพธ์ของการสกัดข้อมูลพืชสมุนไพรไทยจะถูกจัดเก็บอยู่ในรูปแบบ XML ไฟล์

สำหรับขั้นตอนของการวัดประสิทธิภาพของการสกัดข้อมูลพืชสมุนไพรไทยจากหลายเว็บไซต์ แบ่งออกเป็น 2 ส่วนหลัก ส่วนแรกเป็นวัดประสิทธิภาพของการสกัดชื่ออาการโดยใช้ไฟล์เรียนรู้ชื่ออาการ และไม่ใช้ไฟล์เรียนรู้ชื่ออาการ ส่วนที่สองเป็นการวัดประสิทธิภาพของการสกัดข้อมูลพืชสมุนไพรไทยที่เกี่ยวข้อง ซึ่งผลการทดลองในส่วนของการสกัดชื่ออาการโดยไม่ใช้ไฟล์การเรียนรู้ชื่ออาการ ค่า F-measure อยู่ที่ 72% แต่หลังจากใช้ไฟล์การเรียนรู้ชื่ออาการ ค่า F-measure เพิ่มขึ้นมาอยู่ที่ 86% ซึ่งสามารถสรุปได้ว่าถ้าใช้ไฟล์การเรียนรู้ในการสกัดชื่ออาการสามารถช่วยเพิ่มประสิทธิภาพในการสกัดข้อมูลชื่ออาการให้ได้ดียิ่งขึ้น

สำหรับผลลัพธ์ของการวัดประสิทธิภาพของการสกัดข้อมูลที่เกี่ยวข้องของพืชสมุนไพรไทย ค่า precision และ recall ของการสกัดชื่อทางการ และความเป็นพิษของพืชสมุนไพรไทยอยู่ที่ 100% ส่วนค่า precision และ recall ของชื่อวิทยาศาสตร์ ชื่อสามัญ ชื่ออื่นๆ และ ชื่อวงศ์ อยู่ที่ประมาณเกือบ 100% และค่า precision และ recall ของการสกัดส่วนที่ใช้อยู่ที่ 95% และ 92% ตามลำดับ ส่วนค่า precision และ recall ของการสกัดชื่ออาการอยู่ที่ 90% และ 85% ตามลำดับ

5.2. ข้อเสนอแนะ

เนื่องจากการสกัดข้อมูลพืชสมุนไพรไทยของงานวิจัยนี้จำเป็นที่ต้องรู้แท็ก HTML หรือต้องรู้ชื่อหัวข้อที่มีข้อมูลพืชสมุนไพรไทยที่ต้องการไว้ล่วงหน้า ซึ่งอาจทำให้เกิดปัญหาในการสกัดข้อมูลเมื่อมีบางเว็บไซต์ในมีการเปลี่ยนแปลงแท็ก HTML ที่หรือเปลี่ยนชื่อหัวข้อของพืชสมุนไพรไทย ดังนั้นถ้าทำให้คอมพิวเตอร์สามารถวิเคราะห์แท็ก HTML ทีจัดเก็บข้อมูลพืชสมุนไพรไทยได้เองก็จะทำให้ระบบมีความหยืดหยุ่นมากขึ้น

บรรณานุกรม

[1]. มหาวิทยาลัยราชภัฏธนบุรี, "ฐานความรู้พืชสมุนไพรไทย," 2002. [ออนไลน์]. เข้าถึงได้จาก: http://dit.dru.ac.th/herb/Main.htm.

[2]. สำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริสมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี, "ฐานข้อมูลพืชสมุนไพรไทย," 2001. [ออนไลน์]. เข้าถึงได้จาก: http://www.rspg.or.th/plants_data/herbs/herbs_200.htm.

[3]. สมุนไพรไทยดอทคอม, "ฐานข้อมูลพืชสมุนไพรไทย." [ออนไลน์]. เข้าถึงได้จาก: http://www.samunpri.com.

[4]. มหาวิทยาลัยมหิดล, "อุทยานธรรมชาติวิทยาสิรีรุกขชาติ," 2010. [ออนไลน์]. เข้าถึงได้จาก: http://www.pharmacy.mahidol.ac.th/siri/index.php?page=home.

[5]. สำนักงานข้อมูลพืชสมุนไพร, "ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษ." [ออนไลน์]. เข้าถึงได้จาก: http://www.medplant.mahidol.ac.th/tpex/toxic_mnu3.asp.

[6]. L. U. O. Xiao, D. Wissmann, C. T. Se, S. Ag, M. Brown, and C. Sciences, "Information Extraction from the Web : System and Techniques," *Applied Intelligence*, vol. 21, no. 2, pp. 195–224, 2004.

[7]. L. Liu, C. Pu, and W. Han, "XWRAP: an XML-enabled wrapper construction system for Web information sources," in *Proceedings of 16th International Conference on Data Engineering (Cat. No.00CB37073)*, 2000, pp. 611–621.

[8]. G. Meccas, "GRAMMARS HAVE EXCEPTIONS," *Information Systems*, vol. 23, no. 8, pp. 539–565, 1998.

[9]. S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text," *Machine Learning - Special issue on natural language*, vol. 272, no. 1–3, pp. 233–272, 1999.

[10]. S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm, "DOM-based content extraction of HTML documents," in *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, 2003, p. 207.

[11]. H. Xia and Y. Zhang, "Design and implementation of a web news extraction system," in *Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 2011, pp. 1793–1797.

[12]. S. Kuptabut, "Ontology Directed Semantic Annotation Process," in *Proceedings of 3rd International Conference on Information Sciences and Interaction Sciences*, 2010, pp. 251–255.

[13]. J. Hedley, "Jsoup," 2009. [ออนไลน์]. เข้าถึงได้จาก: http://jsoup.org/.

[14]. ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ, "LexTo : เล็กซ์โต - โปรแกรมตัดคำสำหรับข้อความภาษาไทย." [ออนไลน์]. เข้าถึงได้จาก: http://www.sansarn.com/lexto/.

ภาคผนวก

ภาคผนวก ก
Regular expression ที่ใช้สกัดข้อมูลพืชสมุนไพรไทย

1.1. Regular expression ที่ใช้สกัดข้อมูลของฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี (www.dru.ac.th)

ตารางที่ ก.1 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของมหาวิทยาลัยราชภัฏธนบุรี

ชื่อภาษาไทยทั่วไป	Regular expression	ประเภท
มะลิลา, มะลิซ้อน	(([ก-ธ]+)\\s*),*\\s*+	ชื่อทางการ
มะละกอ	([ก-ธ]+)	ชื่อทางการ
Broad Leaf Mahogany, False Mahogany, Honduras Mahogany	([A-z]+\\s*)+,*	ชื่อสามัญ
Coconut	([A-z])+	ชื่อสามัญ
Artocarpus lakoocha Roxb.	(([.]a-z|A-Z|(|)]+)\\s*)+.	ชื่อวิทยาศาสตร์
Aganosma marginnata G. Don., Syn.: A. marginata Vanpruk., marginata Craib., marginata Burkill	(([.]a-z|A-Z]+)\\s*)+.,	ชื่อวิทยาศาสตร์
ฟ้าทะลายโจร หญ้ากับงู น้ำลายพังพอน	(([ก-ธ]+)\\s*)\\s*+	ชื่อท้องถิ่น
ราชบุรี : มะเดื่อดิน มะเดื่อเถา, ภาคเหนือ : เดื่อดิน, ใต้-พายัพ : เดื่อเครือ เดื่อเถา เดื่อไม้ โมกเครือ, กระบี่ : เดือยดิบ, อีสาน-โคราช : ไส้ตัน, สุราษฎร์ธานี : เดือยบิด	([ก-ธ|-]+\\s*:\\s*)*(([ก-ธ]+)\\s*),*\\s*+	ชื่อท้องถิ่น
ภาคกลาง-มังลัก, แมงลัก ภาคเหนือ-กอมก้อ	([ก-ธ]+-)*(([ก-ธ]+)\\s*),*\\s*+	ชื่อท้องถิ่น
กระเจี๊ยบแดง, ส้มพอเหมาะ, ส้มเก็ง (เหนือ), ส้มปู (เงี้ยว), ส้มพอดี (อีสาน)	(([ก-ธ]+)\\s*)[()ก-ธ]*\\s*,*\\s*+"	ชื่อท้องถิ่น

1.2. Regular expression ที่ใช้สกัดข้อมูลของฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี (http://www.rspg.or.th/plants_data/ herbs/herbs_ 200.htm)

ตารางที่ ก.2 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ฐานข้อมูลพืชสมุนไพรไทยของสำนักงานโครงการอนุรักษ์พันธุกรรมพืชอันเนื่องมาจากพระราชดำริ สมเด็จพระเทพรัตนราชสุดาฯ สยามบรมราชกุมารี

ตัวอย่างข้อมูล	Regular expression	ชื่อที่ได้
โหระพา	([ก-ฮ]+)	ชื่อทางการ
Sweet Basil	([A-z]+)	ชื่อสามัญ
Blue Pea, Butterfly Pea	([A-z]+\\s*)+,*	ชื่อสามัญ
Syzygium aromaticum (L.) Merr.& L.M.Perry	([^ก-ฮ|:]+\\s*)+.	ชื่อวิทยาศาสตร์
CAPPARACEAE	([A-z]+)	ชื่อวงศ์
แดงชัน (เชียงใหม่); อัญชัน (ภาคกลาง); เอื้องชัน (ภาคเหนือ)	(([ก-ฮ]+)\\s*)[()ก-ฮ]*\\s*;*\\s*+	ชื่อท้องถิ่น
คำเงาะ คำแงะ คำไทย คำแฝด คำยง ชาตี จำปู้ ส้มปู้ (เขมร)	(([^\\(\\)]+)\\s*)[()ก-ฮ|-]*	ชื่อท้องถิ่น
ขะนู (ชอง-จันทบุรี) ขะเนอ (เขมร) ชีคีย, ปะหน่อย (กะเหรี่ยง-แม่ฮ่องสอน) นะยวยซะ (กะเหรี่ยง-กาญจนบุรี)	(([ก-ฮ]+)\\s*)[()ก-ฮ|-]*\\s*,*\\s*+	ชื่อท้องถิ่น

1.3. Regular expression ที่ใช้สกัดข้อมูลของฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ (http://www.pharmacy.mahidol.ac.th/siri/ index.php?page=home)

ตารางที่ ก.3 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ของฐานข้อมูลพืชสมุนไพรไทยของอุทยานธรรมชาติวิทยาสิรีรุกขชาติ

[illegible]	[illegible]	[illegible]
ขิง	([ก-ฮ]+)	ชื่อทางการ
Hibiscus esculentus L. (ชื่อพ้อง Abelmoschus esculentus (L.))	\\(*([^ก-ฮ]+\\s+)\\)*	ชื่อวิทยาศาสตร์
Eclipta prostrata (L.) L. Roem.	(([.|a-z|A-Z|()]+)\\s*)+.	ชื่อวิทยาศาสตร์
HEMEROCALLIDACEAE (PHORMIACEAE)	\\(*([^()]+)\\)*	ชื่อวงศ์
THYMELAEACEAE	([A-z])+	ชื่อวงศ์
LABIATAE [LAMIACEAE]	\\[*([^\\[\\]]+)\\]	ชื่อวงศ์
กะเม็งตัวเมีย, คัดเม็ง, บังกีเช้า, หญ้าสับ, ย่อมเกี่ยว	(([ก-ฮ]+)\\s*),*\\s*+	ชื่อท้องถิ่น
กระเบาน้ำ กระเบาเป้าแข็ง กระเบากาหลง เบา	(([ก-ฮ]+)\\s*)\\s*+	ชื่อท้องถิ่น

1.4. ฐานข้อมูลพืชสมุนไพรไทยพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม (http://www.samunpri.com/)

ตารางที่ ก.3 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ของฐานข้อมูลพืชสมุนไพรไทยของสมุนไพรไทยดอทคอม

[illegible]	Regular expression	ประเภท
กระเจี๊ยบแดง	([ก-ธ]+)	ชื่อทางการ
Jamaican Sorrel, Rosella	([A-z]+\\s*)+,*	ชื่อสามัญ
Shoe Flower	([A-z]+)	ชื่อสามัญ
Hibiscus sabdariffa Linn.	(([.|a-z|A-Z|(|)]+)\\s*)+.	ชื่อวิทยาศาสตร์
Senna tora (L.) Roxb. ชื่อพ้อง : Cassia tora L.	([^ก-ธ|:]+\\s*)+.	ชื่อวิทยาศาสตร์
Malvaceae	([A-z]+)	ชื่อวงศ์
กระเจี๊ยบเปรี้ยว(ภาคกลาง), ส้มเก็งเค็ง (ภาคเหนือ), ส้มปู(เงี้ยว-แม่ฮ่องสอน)	(([ก-ธ]+)\\s*)[()ก-ธ|-]*\\s*,*\\s*+	ชื่อท้องถิ่น
สลิด ผักสลิดคาเลา สลิดป่า ผักสลิด กะจอน ขะจอน ผักชิก	(([ก-ธ]+)\\s*)	ชื่อท้องถิ่น
มะหนุน(ภาคเหนือ -ภาคใต้) ขะนู(ชอง-จันทบุรี) นากอ(มลายู-ปัตตานี)	(([ก-ธ]+)\\s*)[()|ก-ธ|-]*\\s*	ชื่อท้องถิ่น
ผ้าลายห่อท้อง มันแดง แหนเนือ (ภาคเหนือ) ขมิ้นเครือ ขมิ้นฤาษี (ไทยภาคกลาง) เดิมวอ โกรด (เขมร)	([ก-ธ]+)\\s+(\\((.*?)\\))*	ชื่อท้องถิ่น
สลิด ผักสลิดคาเลา สลิดป่า ผักสลิด กะจอน ขะจอน ผักชิก	(([ก-ธ]+)\\s*)	ชื่อท้องถิ่น
มะหนุน(ภาคเหนือ -ภาคใต้) ขะนู(ชอง-จันทบุรี) นากอ(มลายู-ปัตตานี)	(([ก-ธ]+)\\s*)[()|ก-ธ|-]*\\s*	ชื่อท้องถิ่น

1.5. ฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย (http://www.medplant.mahidol.ac.th/tpex/ toxic_mnu3.asp)

ตารางที่ ก.4 Regular expression ที่ใช้สกัดข้อมูลในเว็บไซต์ของฐานข้อมูลพืชสมุนไพรไทยที่เป็นพิษของสำนักงานข้อมูลพืชสมุนไพรไทย

[illegible]	[illegible]	[illegible]
กระดาด*_	([ก-ฮ]+)	ชื่อทางการ
GREAT-LEAVED CALADIUM-	([A-z]+)	ชื่อสามัญ
Jamaican Sorrel, Rosella	(([.|a-z|A-Z|(|)]+)\\s*)+.	ชื่อวิทยาศาสตร์
ARACEAE	([A-z]+)	ชื่อวงศ์
กระดาด*, กระดาดแดง, คือ, โทปัะ, บอนกาวี, เผือกกะลา, เผือกโทป่าด, มันโทป่าด	(([ก-ฮ]+)\\s*)\\s*,*\\s*+	ชื่อท้องถิ่น
ระคายเคืองต่อระบบทางเดินอาหาร	([ก-ฮ]+)	ความเป็นพิษต่อระบบร่างกาย

ประวัตินักวิจัย

ชื่อ - นามสกุล (ภาษาไทย) รศ.ดร พรฤดี เนติโสภากุล

ชื่อ - นามสกุล (ภาษาอังกฤษ) Assoc. Prof. Ponrudee Netisopakul, Ph.D.

สถานที่ติดต่อ คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

หมายเลขโทรศัพท์ ไปรษณีย์อิเล็กทรอนิกส์ ponrudee@it.kmitl.ac.th

ประวัติการศึกษา

- Ph.D. (Computer Information and Science), Case Western Reserve University, Cleveland, OH, USA.
- M.S. (Computer Information Science), University of Delaware, Newark, DE, USA.
- M.S. (Computer Science), University of Southern California, Los Angeles, CA, USA.
- B.S. with Honor (Statistics), Chulalongkorn University, Bangkok, THAILAND.

ประสบการณ์งานวิจัยที่เกี่ยวข้อง

Kuptabuth, S., Netisopakul, P., "On Factors Affect Document Clustering: Comparison of Summary versus Full Documents", 6th International Joint Conference on Computer Science and Software Engineering, May 13-15, 2009, Phuket, Thailand.

Lertlilitrungroj, W., Netisopakul, P., "Simulation Modeling for Tollway Collection Decision Support System", 6th International Joint Conference on Computer Science and Software Engineering, May 13-15, 2009, Phuket, Thailand.

Netisopakul, P., Saapajit, W., "Pre-diagnosis Doctor Simulation using Case-Based Techniques", 2009 World Congress on Computer Science and Information Engineering, March 31-April 2, 2009, Los Angeles, USA.

Netisopakul, P., Lertvikul, S., "Development of Vendor Managed Inventory using Web Service", 2008 International Conference Global Research in Business and Economics, Sept 17-19, 2008, Orlando, USA.

Netisopakul, P., "Software Engineering: Theory, Principles, and Practices (in Thai)", Translated from the original text book by Pressman, Top Publishing Co. Ltd., 2006

Netisopakul, P., Siriumpunkul, N., "Educational Service Web Database Prototype", Third International Conference on Intelligent Computing, August 21-24, 2007, China.

Netisopakul, P., "Web Metrics Support System (WMSS): Case Study at Faculty of Architecture, Chiang Mai University", Hawaii International Conference on Business, May 25-28, 2006, Honolulu, Hawaii.

Leenawong, C., Netisopakul, P., "Multiobjective Optimization Models for Production Planning at the Dairy Plant of Thailand's Royal Chitralada Projects", International Congress on Logistics and SCM System, April 30-May 6 2006, Kaohsiung, Taiwan.

Netisopakul, P., Leenawong, C., "Application of Nearest Neighbor Algorithm in E-Tourism Advisory System", International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), July 4-7, 2005, Korea.

Leenawong, C., Wattanasiripong, N., Netisopakul, P., "Interaction-based Algorithm for Replacing Components in the Multiple Complex System Model", International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), July 4-7, 2005, Korea.

Leenawong, C. and Netisopakul, P. "Modeling and Optimization Approaches for Team Building Problem" Asia Pacific Industrial Engineering and Management Systems (APIEMS) 12-15 December 2004, USA. Pp. 34.14.1 - 34.14.12.

Netisopakul P. and Leelapat W. "User Adaptive Web Search Engine Architecture" Proceeding of International Conference on Computing, Communications and Control Technologies, August 2004, USA. Vol. VII pp. 92-97.

Punjataewakupt S. and Netisopakul P. "Knowledge Modules in Educational Software" Proceeding of International Conference on Computing, Communication and Control Technologies, August 2004, USA. Vol. VII pp. 103-106.

Netisopakul P. "Visualizing Dynamic Objects in Object-Oriented Program" Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, July 2003. USA. Vol. XIII pp. 321-325.