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