Research Journal of Applied Sciences, Engineering and Technology 9(7): 500-506, 2015
DOI:10.19026/rjaset.9.1432
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 03, 2014 Accepted: September 25, 2014 Published: March 05, 2015

Research Article

The Development of Medicinal Plants Database for Use in Primary Health Care from Various Herbal Websites

Taweechai Anguranak and Nisanart Tachpetpaiboon
Suan Sunandha Rajabhat University, Dusit Province, Thailand

Abstract: The purpose of this research was to develop a Medicinal Plants Database for primary health care from various herbal websites. The system received needed data from 3 Thai herbal websites that could be reliably supplied to the database. The result from extracting data showed the amount of precision, recall and F-measure to be more than 95% and provided information about plant data which was extracted from those websites into the relational database. The users could search for data which was related to herbs from the application. The assessment of efficiency was conducted by using a developed system from the data of questionnaires by asking 25 users. The total results show that the average of common users equals 4.17 and the standard deviation equals 0.60. This developed system showed a very good quality and in the result available to use.

Keywords: Extracting data from website, medicinal plants in primary health care, Thai herb

INTRODUCTION

At present, people worldwide are interested in Thai herbs and generally use them to cure diseases. The Ministry of Public Health has selected 67 kinds of simple herbs which are used in primary health care, which are safe Thai herbs and easily found locally. The purpose was to encourage and give suggestions for the use of herbs for basic health care, or curing the common diseases. Moreover, Thai herbs can be used efficiently in different fields such as herbal consumption and food supplements.

Searching for herbs on the internet through web browsers from the only source or only one website may not get the complete results and appropriate medicinal plants data for use in primary health care which are needed. This includes websites without a section on medicinal or common use. To make the medicinal plants data in primary health care complete, there is a need to compose data by extracting Thai herbal data from 3 website especially the medicinal plants data for use in primary health care. Its purpose is for use as a data source of herbal primary health care and for maintaining and passing on Thai knowledge.

This research of development of medicinal plants database in primary health care from various herbal websites was done to create an efficient database, by usefully applying variables such as the search of medicinal plants names and scientific names.

There are numerous approaches to extract main content. Typically, extracting main content from HTML Document (Louvan, 2009) in which the input is HTML documents and output is text strings that are extracted. From the HTML document, the irrelevant parts are removed out of the web document. There are two processes, namely classification tasks and heuristics rules. Classification tasks can split the HTML Document Structure into a small segment and use the Document Object Model (DOM) for separately checking a good segment which is appropriate with the context as a uniform semantic unit.

After this process, there will be irrelevant stained data such as embedded advertisements, links to related articles, headers, etc. These irrelevant data will be removed by using heuristic rules, for the purpose of getting the text string of main content. There are 5 steps in the process of extracting data of Thai herbs (Chainapaporn and Netisopakul, 2012) for the purpose of learning processes in the study, the words about disease. These steps were composed by the name of the indication group for 5 websites by randomly selected 80 webpages, from each website; there were a total of 400 webpages:

- Using JSOP API for extracting HTML Source from each webpage.
- If it has a simple structure, JSOP API searches for HTML tags containing medicinal use. But if it has complex HTML, template file searches for medicinal-use containing word lists. By the work of removing HTML tags, the result is text strings and using a Thai word segmentation tool which called “Lexto” for splitting into words. It takes those words to compare with the word list in template.

Corresponding Author: Taweechai Anguranak, Suan Sunandha Rajabhat University, Dusit Province, Thailand

Fig. 1: HTML structure of webpage

If the word matches with indication list, it will insert the subtopic content symbol.

When the next keyword is found, it is the end of current subtopic content, or the end of content source.

- The phrase and word splitting for the Thai language is a step of splitting a long content phrase into shorter phrases and splitting a shorter phrase into separate words.
- The symptom name extraction is a step of extracting the symptom name by using the word that expresses curing, such as treatment, remedy and recording the symptom name into list file that is the result of learned words.
- The symptom name validation can be manually validated as the correct word of the symptom name in list file from the fourth step.

MATERIALS AND METHODS

This research uses the step of extracting data from websites by using PHP Simple HTML DOM Parser (2013) and the step of searching the heading position for extracting data which is needed to store to relational database. The data which is needed to extract from three websites for creating the medicinal plants database in primary health care that are featured in these related websites.

Websites (Herbal Medicine List and Uses, 2012) have common information. they are in order of names of heading, scientific name, synonym name, common name, family name and other name, but there will not be completed heading in some webpages.

Websites (Siri Rackhachati Nature Park, 2012) have botanical characteristics. they are in order of names of heading, habit, leaf, flower, fruit and medicinal use.

Websites (Herbal Maejo University, 2012) have ecological information. they are in order names of heading; cultivated varieties, plant propagation, planting season, planting, harvesting period, harvest and produce.

The process of extracting information of Thai herbs: Most webpages from each website which are used to extract data have the HTML structure that shows the same. Figure 1 is the webpage structure of website (Herbal Medicine List and Uses, 2012). The word in the rectangle is the heading for the searching position and it is in order of heading for the searching position which some webpage may not be available for some heading.

It's the design of flow chart by extracting data from Thai herb webpage.

According to Fig. 2, there are steps of extracting for Thai herb websites:

- Input the URL of webpage from Thai herb website. For example, URL in Fig. 1.
- Removing HTML tags from the webpage by using PHP Simple HTML DOM Parser (2013), the result is the plain text as in Fig. 3 and there is the data which is needed to extract from the wanted heading (the word in the rectangle is the heading).
- Using an array name $title to collect the heading of word lists that are used for searching the positions in plain text such as website (Herbal Medicine List and Uses, 2012). It is used for searching the word list by order: scientific name (ชื่อวิทยาศาสตร์), synonym name (ชื่อพ้อง), common name (ชื่อสามัญ), family name (ชื่อวงศ์), other name (ชื่ออื่นๆ).

Using function strops() (Philip Olson, 1997-2014) for finding position of word list in $title array which is in order heading in plain text and store in array name $positionTitle. For example, strpos (Data, Find, Start).

Data: The plain text of herbal information from Fig. 3.

Find: The word list in $title which is the heading, used for finding the position.

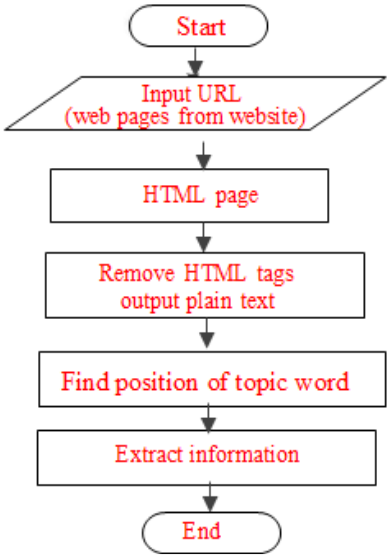

Fig. 2: The flow chart of extracting data from webpage

- Using array name $contentHerb store content of data and use function content() for extracting the content of data. For Example, code of program in PHP:

```
Function content ($title, $positionTitle, $data) {
    $count = count ($title);
    for ($i = 0; $i < $count - 1; $i++) {
    if ($positionTitle [$i] >0) {
        $contentHerb [$i] =                          =>
substr ($data, $positionTitle [$i] + $title [$i],    =>
        $positionTitle [$i + 1] -                    =>
        ($positionTitle [$i] + $title [$i]));}
        else {
        $contentHerb [$i] = ”“;}    }
    $contentHerb [$count - 1] =                      =>
substr ($data, $positionTitle [$count - 1] +         =>
$title [$count - 1],
    $positionTitle [$count - 1]-                     =>
($positionTitle [$count] + $title [$count]));
return $contentHerb;}
```

For example, Using function substr() (Philip Olson, 1997-2014) for extracting data from the topic of the scientific name and the next topic is the common name. substr (String, Start, Length):

String: The plain text of herbal data.
Start: The position of the first heading name + the length of the first. heading name.
Length: The position of the following heading name.

From Fig. 4, the first heading is the scientific name and the next heading is the common name. So, the result of extracted information is “Hibicus sabdariffa L.” Which uses function.

สรรพคุณสมุนไพร 200 ชนิด กลับหน้าหลัก : บทนำ ความรู้ทั่วไป สรรพคุณสมุนไพรแบ่งตามกลุ่มอาการ กลุ่มยาขับปัสสาวะ กระเจี๊ยบแดง ทองกวาว ตะไคร้ ทานตะวัน สามสิบ สับปะรด สมอพิเภก หญ้าหนวดแมว อ้อยแดง กลุ่มยาขับปัสสาวะ กระเจี๊ยบแดง ชื่อวิทยาศาสตร์ : Hibiscus sabdariffa L. ชื่อสามัญ : Jamaican Sorel, Roselle วงศ์ : Malvaceae ชื่ออื่น : กระเจี๊ยบ กระเจี๊ยบเปรี้ย ผักเก็งเค็ง ส้มเก็งเค็ง ส้มตะเลงเครง ลักษณะทางพฤกษศาสตร์ : ไม้ล้มลุก อายุปีเดียว สูง 1-2 เมตร เปลือกต้นเรียบ ลำต้นและกิ่งสีม่วงแดง ใบ เป็นใบเดี่ยว ออกเรียงสลับ ใบหยักเว้าลึก 3-5 แฉก แต่ละแฉกกว้าง 0.5-3 ซม. ยาว 3-8 ซม. โคนใบมน ปลายใบแหลม ก้านใบยาว 4-15 ซม. ดอก ออกเดี่ยวตามซอกใบ มีริ้วประดับสีแดง กลีบเลี้ยงโคนเชื่อมติดกัน ปลายแยก 5 แฉก สีแดงเข้ม อวบน้ำ กลีบดอก 5 กลีบ สีเหลือง ตรงกลางดอกสีม่วงแดง เกสรเพศผู้จำนวนมาก ผล รูปไข่ สีแดงเข้ม มีกลีบเลี้ยง ติดทนขนาดใหญ่รองรับอยู่จนผลแก่ ผลแห้งแตกได้ เมล็ดสีน้ำตาลจำนวนมาก ส่วนที่ใช้ : กลีบเลี้ยงของดอก หรือกลีบที่เหลืออยู่ที่ผล ใบ ดอก ผล เมล็ด สรรพคุณ : กลีบ

Fig. 3: The plain text after removing HTML tags

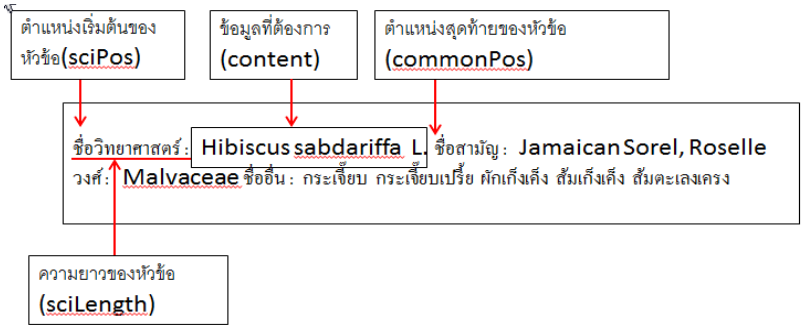

Fig. 4: Extracting content data of the heading of the scientific name

ชื่อวิทยาศาสตร์ : Hibiscus sabdariffa L.
ชื่อสามัญ : Jamaican Sorel, Roselle
วงศ์ : Malvaceae
ชื่ออื่น : กระเจี๊ยบ กระเจี๊ยบเปรี้ย ผักเก็งเค็ง ส้มเก็งเค็ง ส้มตะเลงเครง

Fig. 5: The result by extracting data from webpage (Herbal Medicine List and Uses, 2012)

substr ($data, $sciPos + $sciLength, => $commonPos - ($sciPos + $sciLength));

From Fig. 1, the result of all data which is extracted from webpage as Fig. 5.

Experiments and evaluation: From the experiment, data was extracted from Thai herb websites using 3 websites by randomly selecting 25 webpages from each website and displaying them by using Precision, Recall and F-measure (Precision and Recall, 2013).

Precision is the ratio of related medicinal plants data which is extracted by the system and all of the Thai herb data is extracted by the system.

Recall is the ratio of related Thai herb data which is extracted by system and all of the related Thai herb data on webpage.

The F-measure is defined as a harmonic mean of precision:

$$F - Measure = \frac{2 \, x \, \mathrm{Pr}\, ecision \; x \, \mathrm{Re}\, call}{\mathrm{Pr}\, ecision + \mathrm{Re}\, call}$$

The experiment of F-measure as the Table 1.

Website (Herbal Medicine List and Uses, 2012; Siri Rackhachati Nature Park, 2012; Herbal Maejo University, 2012) has the similar structure of each web page which makes the data extraction correctly precise. So, the total average of system efficiency as F-measure is 99% which is a very high average. It is to be considered as good quality as a Thai herb data extraction system.

Architecture system: The design of architecture system can be split in various parts; there are

Table 1: F-measure of extracting Thai herb data

	Collection					
Website	Correct	Incorrect	Total data	Precision	Recall	F-measure
Herbal Medicine List and Uses (2012)	105	1	106	0.990566	0.990566	0.990566
Siri Rackhachati Nature Park (2012)	75	-	75	1	1	1
Herbal Maejo University (2012)	165	-	165	1	1	1
Summation	345	1	346	0.997110	0.997110	0.997110

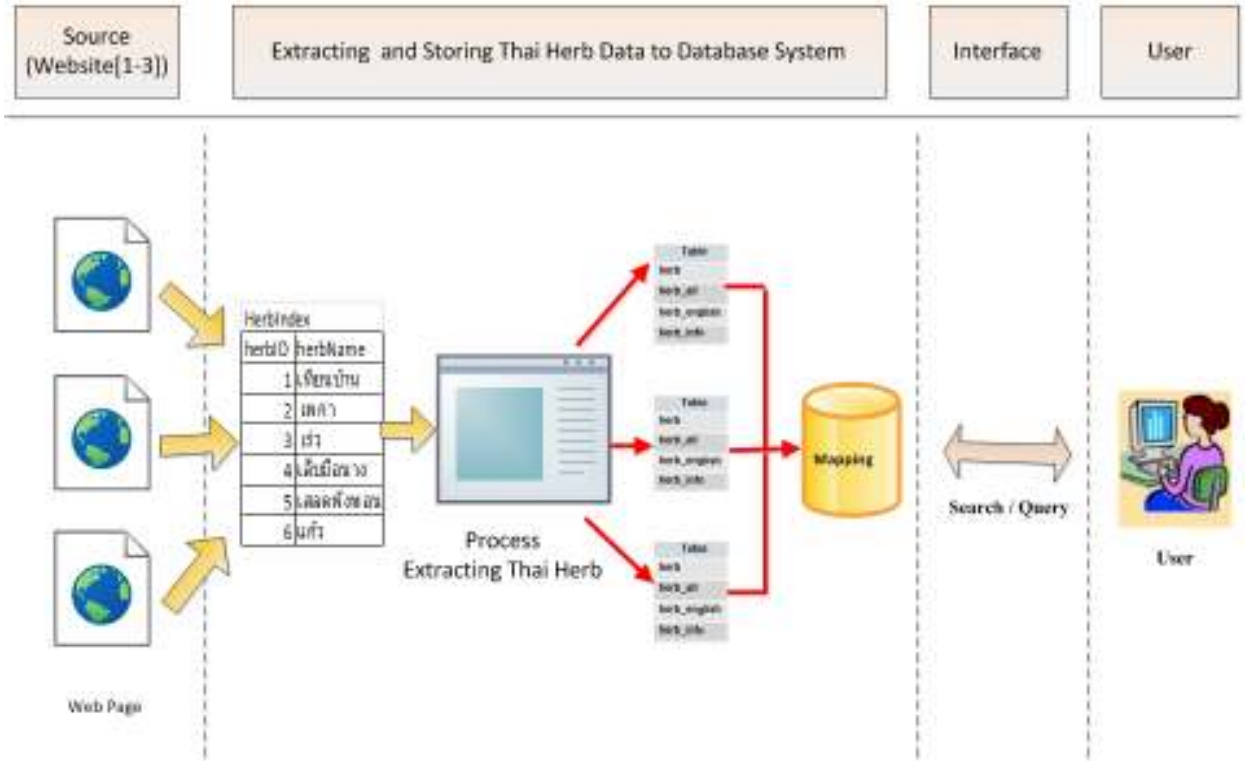

Fig. 6: Architecture system

Fig. 7: The application for extracting and saving data into database

different roles from each part as shown in Fig. 6 by these following details.

Source: Select webpages from websites (Herbal Medicine List and Uses, 2012; Siri Rackhachati Nature Park, 2012; Herbal Maejo University, 2012) which you want to extract the data of medicinal plants names in the medicinal plants list for primary health care.Process Extracting Thai Herb: It extracts Thai herb data and stores it in the relational database.

Interface: Searches and shows the medicinal plants data for primary healthcare; it is the part that searches data through the web application for users. The system will obtain a command which shows the conditional data and wanted reason through web browser.

User: The users who want to search for medicinal plants data for use in primary health care.

Steps of work in the system:

First: Input is URL (webpage) from sources of websites (Herbal Medicine List and Uses, 2012; Siri Rackhachati Nature Park, 2012; Herbal Maejo University, 2012).

Second: Check the medicinal plants names of web pages to match with index Herb table that stores medicinal plants data for use in primary health care.

Third: Extract medicinal plants data for use in primary health care as specified and stored into relational database by application on website as in Fig. 7.

Fourth: The users can search for herbal data from the system. If the medicinal plants name matches the medicinal plants names in index Herb, the third step is available to do.
Else go to the first step.

Web application for extracting data from wanted webpages from websites (Herbal Medicine List and Uses, 2012; Siri Rackhachati Nature Park, 2012; Herbal Maejo University, 2012) that is specified or retrieved information from data files which are already stored, has a family name as html (html file) from a web browser. There are medicinal plants data which have different details as the user desires from each website such as botanical basic data, characteristic of the medicinal plants and ecology. Figure 7 is the extraction of wanted data from webpage of the website (Herbal Medicine List and Uses, 2012). The extraction of information will obviously specify the wanted medicinal plants heading and have to match with every letter of medicinal plants heading on the data of the webpage. If the retrieved data is incorrect, it can check and adjust it correctly from the screen of the web application and save data into the database. Moreover, it can adjust or add the data and show the medicinal plants data which is stored in the database. The code of program which used for extracting data is written in PHP language.

System development: The development of medicinal plants database for primary health care is from various herbal websites. The system has been developed in application form by using PHP language, Javascript,

Table 2: The criterion of evaluation marking

Scale of evaluation	Signification
5	Strongly agree
4	Agree
3	Undecided
2	Disagree
1	Strongly disagree

Table 3: The interpretational criterion of data by considering the average

Scale of evaluation	Signification
4.21-5.00	Excellent
3.41-4.20	Very good
2.61-3.40	Average
1.81-2.60	Fair
1.00-1.80	Poor

User interface design and MySQL database. It uses the phpMyAdmin program to managing the MySQL database.

System testing: After the processing in the performance of a variety of tests and adjusting the system perfectly, we have examined a group of 25 users by the user Satisfaction Questionnaire as a tool for compiling data.

System evaluation: System Evaluation is an acceptance test by users for evaluating the efficiency of a developed system, by splitting 4 parts of system evaluation as follows:

- Functional requirement test
- Functional test
- Performance test
- Usability test

Evaluation is specified by the standard of quantitative marking, the five ranks sort of rating scale assessment following the step of Likert (1932). There are details to specify the score rank and the scale of weight in system contentment as shown in Table 2 and 3.

Fig. 8: The search of Thai searching, the word "กระเพรา"

Fig. 9: The search shows details of medicinal plants "กระเพรา"

Table 4: The result of evaluation of satisfaction with the system

	The common users	
The evaluation of the system	$\bar{x}$	S.D.
Functional requirement test	4.15	0.61
Functional test	4.16	0.56
Performance test	4.10	0.68
Usability test	4.26	0.54
Total average	4.17	0.60

S.D.: Standard deviation

RESEARCH RESULTS

As result, the development of medicinal plants database for primary health care came from various herbal websites. When the users use this application, they can register to be a member and search for medicinal plants data which used in primary health care as lists of medicinal plants by alphabetical order, scientific name, synonym name and local name. When the admin user or recorder logs in to the application, they are able to adjust the medicinal plants data which is used in primary health care.

For example: The search of Thai searching, the word "กระเพรา" or "Basil", the result are shown in Fig. 8. The search shows details of medicinal plants "กระเพรา" or "Basil" as shown in Fig. 9.

CONCLUSION

The purpose is to evaluate the developed efficiency of development of the medicinal plants database which is used in primary health care. The average equals to 4.17 and the standard deviation equals to 0.60 from the group of 25 users. So, we can conclude that this developed system has a very good quality of satisfaction as shown in Table 4.

ACKNOWLEDGMENT

This study is supported as part of a project funded by The Institute of Research and Development Suan Sunandha Rajabhat University (www.ssru.ac.th).

REFERENCES

Chainapaporn, P. and P. Netisopakul, 2012. Thai herb information extraction from multiple websites. Proceeding of 4th International Conference on Knowledge and Smart Technology (KST, 2012), pp: 16-23.

Herbal Maejo University, 2012. Retrieved from: http://www.mmp.mju.ac.th/ (Accessed on: April 15, 2012).

Herbal Medicine List and Uses, 2012. Retrieved from: http://www.rspg.or.th/plants_data/herbs/herbs_200.htm (Accessed on: August 11, 2012).

Likert, R., 1932. A technique for the measurement of attitudes. Arch. Psychol., 140: 1-55.

Louvan, S., 2009. Extracting the Main Content from HTML Documents. Retrieved from: http://wwwis.win.tue.nl/bnaic2009/papers/bnaic2009_paper_113.pdf.

Philip Olson, 1997-2014. The PHP Documentation Group. Retrieved from: http://www.php.net/manual/en/ (Accessed on: May 07, 2013).

PHP Simple HTML DOM Parser, 2013. Retrieved from: http://simplehtmldom.sourceforge.net/manual.htm (Accessed on: March 20, 2013).

Precision and Recall, 2013. Retrieved from: http://en.wikipedia.org/wiki/ Precision_and_recall (Accessed on: April 22, 2013).

Siri Rackhachati Nature Park, 2012. Retrieved from: http://pharmacy.mahidol.ac.th/siri/ (Accessed on: February 8, 2012).