Thai Herb Information Extraction from Multiple Websites

Phakphoom Chainapaporn and Ponrudee Netisopakul
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
chphakphoom@gmail.com, ajkmake@gmail.com

***Abstract*—Thai herbs have increasingly gained public attention. Recently, there are a number of Thai herb websites. Each website has similar information but quite different details. For example, some webpages do not provide information indicating which part of Thai herb can treat the specified symptom. In order to collect more complete Thai herb information, we have developed information extraction process to extract Thai herb information from multiple websites. The process employed a HTML parser and file templates to recognize useful information in various webpage formats. Preliminary experiments gave satisfactory precision and recall over 85 percent.**

Keywords Web information extraction; HTML parser; Thai herbs.

I. INTRODUCTION

Most Thai people are familiar with Thai herbs, which have been used for curing many simple symptoms, such as indigestion, stomachache, fever and so on. Many organizations [1-4] have published useful Thai herb information on their websites. Although these information are similar and somewhat consistent, none have completed information regarding names of Thai herbs, symptoms, and parts of used. For example, one webpage only gives scientific name but does not give common name; while another webpage only indicate which Thai herb to be used for curing which symptom, but does not provide which part of the Thai herb to be used.

This paper proposes Thai herb information extraction process to recognize useful Thai herb information and extract them from multiple websites. The process employed an open source HTML parser called JSOUP [5] and several template files. The overall process has two main phase; symptom name collection phase and treatment information extraction phase. Output of the first phase is a file containing symptom names, which collecting from multiple websites using synonyms of a word 'treat'. The second phase extracted Thai herb treatment information, including Thai herb names and medicinal used. Names include science name, common name, and family name. Medicinal used includes part of used and symptom name.

The challenge of Thai herb information extraction is that each websites have different HTML structures. Some are simple and some are complicated. When one Thai herb can treat many symptoms, the symptom names are listed continuously, with or without delimiter. However, each symptom can be treated using different Thai herb parts. The process needs to recognize which content is the Thai herb part and which content is the symptom.

Not only that Thai herb information from multiple sources has no fixed pattern content, but there is also inherently nested content and listed content structure within medicinal-used topic of each page. That is, medicinal-used topic can contain many part-of-used topics, and each part-of-used topic can contain many symptoms that it can cure.

This paper divides into 3 parts; the next section presents related work on web information extraction. Section 3 describes the proposed process for extracting Thai herb information from multiple websites. The last section describes the experiment process and evaluates precision and recall of the proposed process.

II. RELATED WORK

Web information extraction research has many approaches, ranking from a simple HTML tag extraction approach [6] to a sophisticate semantic annotation approach [7]. In between there are web content extraction approach [8] and webpage topic identification approach [9].

The work by [6] proposes the HTML tags extraction approach using DOM trees analysis. This approach can extract 'useful and relevant' content from webpage, which originally contains other irrelevant images and links. Basically, it worked by filtering out 'unwanted' tags, such as images, links, scripts, styles. In addition, while parsing the DOM tree, it searched for additional information and design algorithm to specifically remove advertisement links, empty tables and so on.

A more complex content extraction approach proposed by [8] is a web news extraction system. It extracted news content from a news portal page. It combined regular expression, HTML parser, search engine, and page structure analysis to classify news index page and news content page. Then, the contents in the content pages are extracted. It also extracted contents from multiple page news, and related page news.

Although our approach also use regular expression and HTML parser technologies to process webpages, there are significantly differences. While [8] used regular expression to validate URL and domain name, we use regular expression to extract Thai herb names, symptom names and parts of used. In

978-1-4673-2167-9/12/$31.00 ©2012 IEEE

addition, while [8] extracted the whole content from HTML body tag, our work must further separate topics and subtopics within the same tag.

In work by [9], regular expressions are given as a specification to a LEX tool to identify different types of headings and words on geography information webpage. Then, a more hierarchical specification is given to a YACC tool to generate a parser. The parser can extract specific geography information, such as location, map, area, and so on, from web sources. However, in order to generate a good parser, both LEX and YACC must be given an appropriate specification. An extra pre-learning step called structuring step samples web pages from a web source and analyzes the page to determine format of interesting geography information topics, which will be used to construct specification for LEX and YACC.

Our work did not employ LEX and YACC tool because we felt that it mainly used for languages with fixed grammar, such as programming languages, rather than human languages. Moreover, after we sample Thai herb webpages to investigate, we found that we cannot construct a grammar for LEX and YACC. Mainly because Thai language has no fixed pattern to differentiate words indicating symptoms from words indicating parts of used. This problem is more in the area of semantic annotation.

The work by [7] proposes a semantic annotation process using ontology in a weather forecast domain. The ontology is knowledge base for the weather domain. There were two steps, the first step extracted contents from webpage and selected relevant sentences. The second step compared noun phrase in the sentence to the concept and properties in the weather ontology.

Our work employed list of symptoms in the second phase. This list can also be considered as a knowledge base, although in a very simple form. This list is constructed automatically in the first phase, when our proposed process 'learns' from multiple web sources data.

.

III. THE PROBLEM

During the preliminary investigation of Thai Herb websites, we found the following problems for the task of extracting Thai herb treatment information.

The first problem is that the available HTML parser tool, such as JSOUP, can only extract the Thai herb treatment information from some websites, but not all websites. The websites that it can process must have explicit structure for different types of information. However, some websites do not have this explicit structure, as illustrate in figure 1 and figure 2.

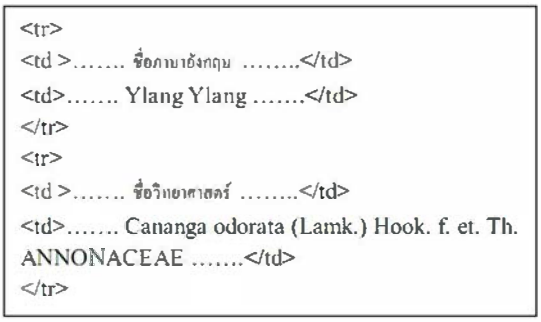

Figure 1. A simple HTML structures

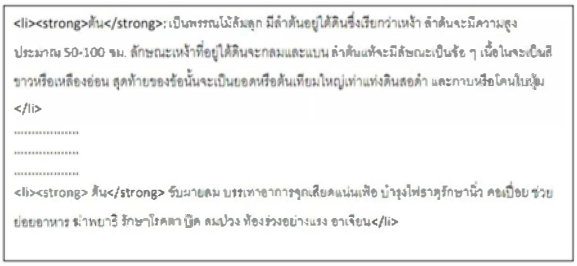

Figure 2. A complicated HTML structures

Figure 1 shows a part of HTML source designed to give information in the following format:

<tr>

<td> topic name </td>

<td> content of the above topic </td>

</tr>

When the topic name can be common name, science name, family name, and medicinal used.

Hence, a HTML parser tool can extract this type of website rather reliably.

In contrast, Figure 2 shows a HTML source with inconsistent structure. For example, it embeds different type of information in the same tag type, i.e., <li>. Hence, the HTML parser cannot reliably extract needed information.

The second problem is how to extract parts of used and symptom names, which embedded inside the medicinal used, as shown in figure 3.

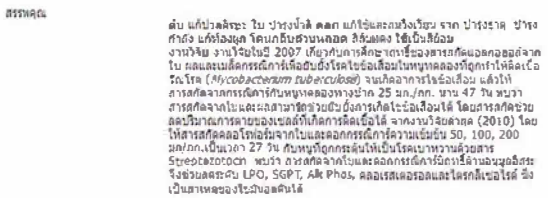

Figure 3. The topic and content in webpage

The medicinal used topic in figure 3 states the part of used, then follows by a list of symptom names together with other information, which we need not to collect. A simple regular expression may be able to separate parts-of-used from the rest of information (from some websites). However, it is neither able to separate each symptom names from another, nor separate symptoms from other unwanted information.

IV. THE PROCESS

Our Thai herb extraction process extracted and collected Thai herb treatment information from Thai herb websites [1-4]. The process is illustrated in figure 4.

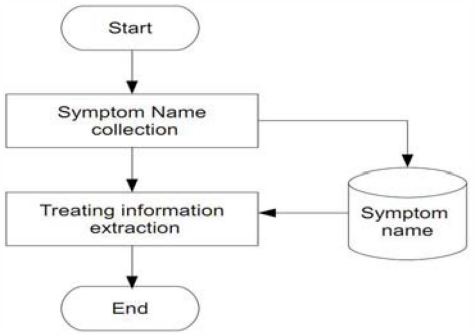

Figure 4. The overall Thai herb extraction process

A. Symptom name collection

Symptom name collection collected symptom names from five websites by randomly selected 80 webpages from each website, that is, there are totally 400 webpages. We can think of this process as a learning process to learn which words are symptoms. Figure 5 illustrates the working details, which can be divided into 5 steps: HTML source extraction, medicinal-used topic identification, phrase and word splitting, symptom name extraction, and symptom name validation.

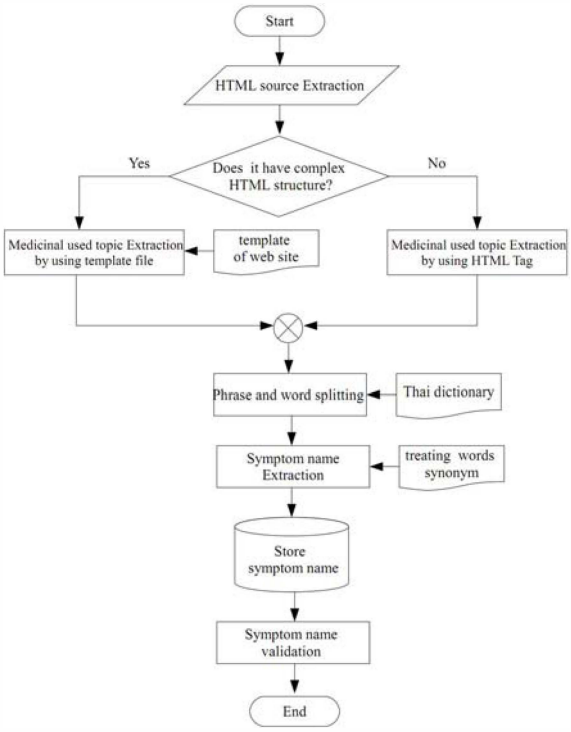

Figure 5. Symptom name collection procedure.

1) HTML Source Extraction: This step used JSOUP API [5] to extract HTML source from each web page

2) Medicinal-used Topic Identification: The HTML source code form previous step is examined for its complexity. If it has a simple structure, JSOUP API search for HTML tags containing medicinal-used. The examples of tags that may contain medicinal-used information are <td>, <p>, <ul>. The example of HTML tags and the result of medicinal-used identification are shown in Figure 6 and 7, respectively.

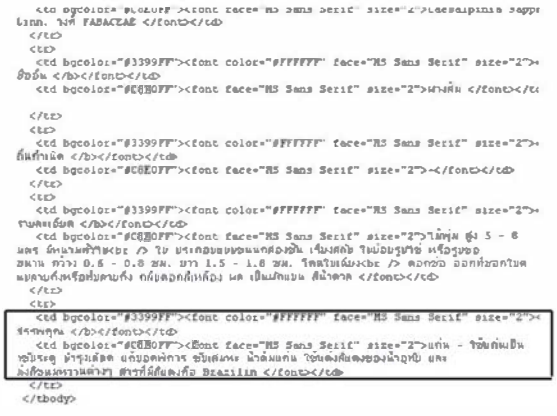

Figure 6. A sample of a simple HTML tag containing medicinal-used information

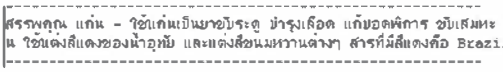

Figure 7. The result of Medicinal-used Identification using simple HTML tag analysis

For complex HTML source, a template file containing indicator word list for medicinal-used section must be used to identify medicinal-used subtopics in the HTML source. There are three sub-steps: HTML tags removal step, word splitting step, and relevant topic extraction step.

The first step removes HTML tag from HTML source. These tags can be ignored because they contain no useful information. The result left only free text as shown in Figure 8. The second step splits Thai free continuous text into words using Lexito [10], the open source Thai word segmentation tool. This step is necessary for Thai language.

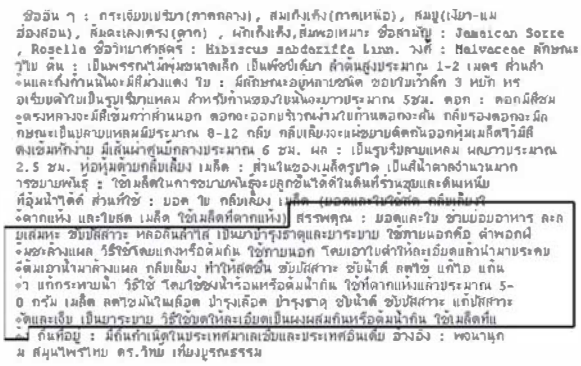

Figure 8. The result left after HTML tags removal for a complex HTML source

The third step compares each word with indicator word in the above mentioned template file. If the word matches with any indicator word, then the words that followed are supposed to be the detail of that indicator word subtopic. The process then inserts corresponding symbols, such as "%PATTittle: "to indicate the beginning of the subtopic content. The end of the current subtopic content is found when the next keyword is found or if the process reaches the end of the content source. The output result of relevance topic extraction step is shown in Figure 9.

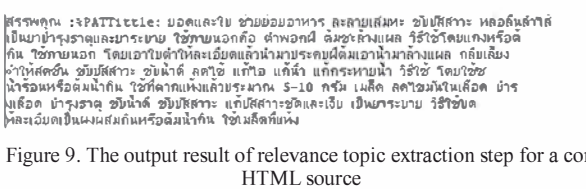

Figure 9. The output result of relevance topic extraction step for a complex HTML source

3) Phrase and word splitting: The purpose of this step is to split a long content phrase into shorter phrases and split a shorter phrase into separate words. The shorter phrases can be obtained easily using whitespace delimiter, while the word splitting step uses the Lexito tool mentioned above. Figure 10 shows the result of this step. This step is necessary for Thai language

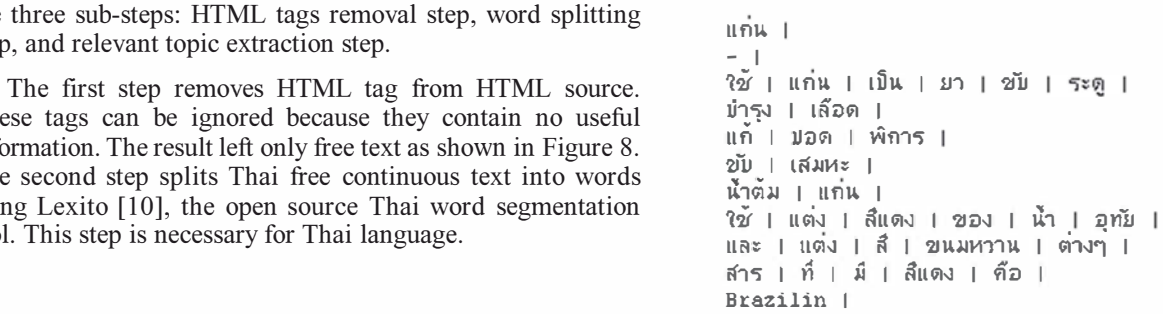

Figure 10. The result of phrase and word splitting from previous Medicinal-used content

4) Symptom Name Extraction: This process extracts symptom names using treating word indicator, such as cure (รักษา), remedies (แก้โรค), and assuages (บำบัด). Our assumption is that the words that follow the word 'cure' in that phrase are the symptom name. Each symptom name is written into a symptom name list file. Hence, the output file contains 'learn' words to be used in the next process: treatment information extraction process. The sample result of this step is shown in Figure 11.

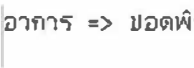

Figure 11. The result of symptom name collection

5) Symptom Name Validation: The 'learn' word in the previous step must be validated manually. Incorrect result is deleted from the file.

B. Thai herb treatment information extraction

After 'learn' the symptom name, now the process is ready to extract Thai herb treatment information from webpages. The overall process is shown in Figure 12.

The process starts from extracting HTML source code from the webpage, identifying relevant treatment information, extracting Thai herb names, extracting part of used, and extracting symptom name. Some steps are similar to steps in the previous process; hence, we will explain only some key steps here.

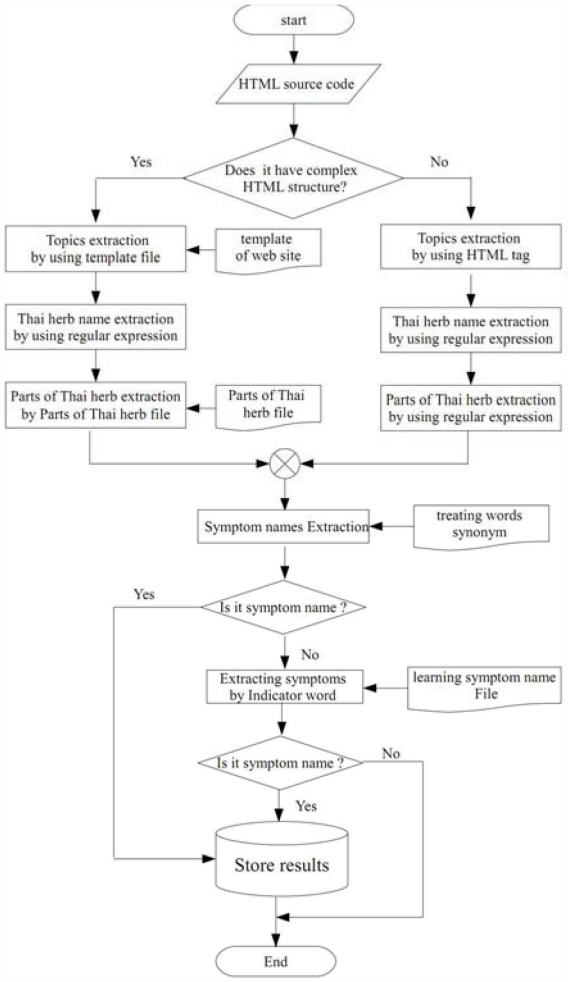

Figure 12. The Thai herb information extraction process

1) Relevant Treatment Information Identification: This process extracts relevant topic and its content from each webpage, including science name, common name, family name and medicinal-used. Similar to the medicinal-used topic identification step in the previous process, there are 2 approaches, one for a simple HTML source and the other is for a complex HTML source. The only difference is that this process extracts not only medicinal-used content but also extracts other relevant treatment content mentioned before. Hence, the file template contains more indicator words for each treatment information, such as names and medicinal-used. Table 1 shows the indicator words and the symbols to insert when the relevance information is found. The output result of this step from a complex HTML source and a simple HTML source are shown in Figure 13 and Figure 14, respectively.

TABLE I. TAG NAMES ARE USED TOPIC IDENTIFICATION .

Symbol to be inserted	Description	Indicator word
%SNTittle:	science name indicator	science name: (ชื่อวิทยาศาสตร์:)
%FNTittle:	family name indicator	Family name: (ชื่อวงศ์:) ,Family (วงศ์(
%CNTittle	common name indicator	common name :)ชื่อสามัญ:), English name (ชื่ออังกฤษ(
%GNTittle:	other name indicator	Other name: (ชื่ออื่น:)
%PATTittle:	medicinal used section indicator	Medicinal used: (ชื่อสรรพคุณ :)

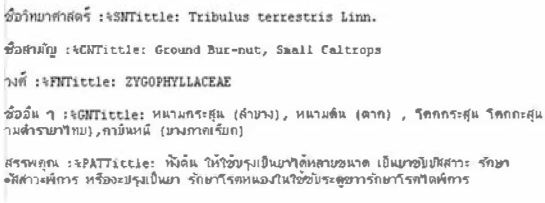

Figure 13. The result of relevant treatment information identification step from a complex HTML source

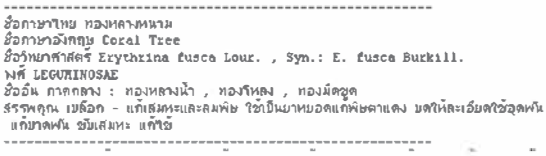

Figure 14. The result of relevant treatment information identification step from a simple HTML source

2) Thai herb name extraction using regular expression: This process extracts the various names of a Thai herb such as science name, common name, family name and other names using regular expression. For example, a regular expression such as *"([ก-ฮ]+)/s*,*"* can be used to extract the following content: "*ทองหลางน้ำ, ทองโหลง, ทองบีดขูด*", which are common names of a Thai herb. The result of this process is shown in Figure 15.

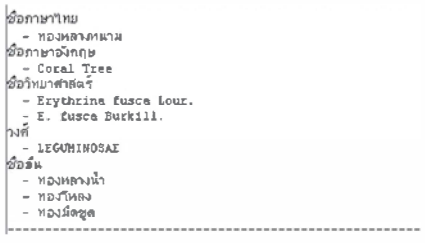

Figure 15. The result of various Thai herb name Extraction.

3) Part-of-used and Symptom Name Extraction: This step is used to extract Thai herb's part of used together with the symptom names it can treat. The technique used here is similar to the technique explained in the step 3 and 4 in the previous process. In short, the process needs to split a long phrase into separate words, and then search for indicator words indicating part-of-used.

For a simple HTML source, we can also use regular expression to extract part-of-used. The example of regular expression is *"([ก-ฮ(+[\\s*–\\s*(([ก-ฮ/a-z/A-Z/()]+\\s*)*)"*. This regular expression extracts these content *"เปลือก – แก้เสมหะ และลมพิษ ใช้เป็นยาหยอดพิษแก้ตาแดง บดให้ละเอียดใช้อุดฟัน แก้ปวดฟัน ขับเสมหะ แก้ไข้"*.

Figure 16. The result of finding parts of Thai herb by using regular expression

Note that the result of this regular expression (shown in Figure 16) can be divided into 2 parts, the first part "ส่วนที่ใช้ => เปลือก" indicates part- of-used; the second part is a list of the symptom names, which will be processed further by splitting into words and extract symptom names. Now the extraction process not only use 'treating' word synonyms, but also use the 'learn' symptom names from the first process. The result of extracting symptom names using 'treating' word synonyms is shown Figure 17 and the result of extracting symptom names using previously 'learn' symptom name file is shown in Figure 19.

Figure 17. The result of extracting symptom names using 'treating' word synonyms

Figure 19. The result of extracting symptom names using 'learn' symptom name file

V. EXPERIMENTS AND EVALUATION

In our experiments, we used Thai herb information from 4 websites, each with 20 webpages, selected randomly. To evaluate a performance, precision and recall are calculated as follows. A precision is defined as the ratio of relevant treatment information extracted by our system and the total number of extracted information by our system.

$$\text{precision} = \frac{\text{relevant selections by our system}}{\text{all selections by our system}} \quad (1)$$

A recall is defined as the ratio of relevant treatment information extracted by our system and total the number of corresponding relevant treatment information on webpages.

$$\text{recall} = \frac{\text{relevant selections by our system}}{\text{all relevant items from webpage}} \quad (2)$$

To better understanding the comparative performance, we also used f-measure to combine precision and recall measures. The f-measure is defined as follows:

$$\text{f} - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

The experiment divided into 2 parts. The purpose of the first part is to compare the effectiveness of symptom name extraction with and without symptom name collection phase. The result is shown in table 2 and table 3.

TABLE II. THE F-MEASURE FOR SYMPTOM NAME EXTRACTION WITHOUT TREATING WORDS SYNONYM .

website	collection		Total data	F-measure
	correct	incorrect		
[1]	87	1	128	80%
[2]	170	15	312	68%
[3]	161	18	268	72%
[4]	44	1	81	70%
Summation	462	35	789	72%

TABLE III. THE F-MEASURE FOR SYMPTOM NAME EXTRACTION WITH TREATING WORDS SYNONYM.

website	collection		Total data	F-measure
	correct	incorrect		
[1]	99	8	128	86%
[2]	271	37	312	87%
[3]	223	30	268	86%
[4]	74	3	81	94%
Summation	667	78	789	86%

We found that without the symptom name collection phase, the extraction process gave the F-measure of 72%. After the process 'learned' about the symptom names, the F-measure was 86%, gave a 14% improvements.

The second part of the experiment used both phase of processes. Table 4 shows raw numbers of extraction result for Thai herb information extraction system from the experiment and table 5 shows the precision and recall measure of our proposed Thai herb information extraction system on each website.

Precision of names are 100% or almost 100%. This is because most names that have been extracted are correct. However, the recalls are from 98 to 99%, which are pretty high. This is because the name patterns are not complicated; hence they are recognized by our process.

Some examples of science names, common name and family name that cannot be extracted are “Naringi crenulata (Roxb.) Nicolson (Hesperethusa crenulata (Roxb.) Roem.)” or “Hibiscus esculentus L. (ชื่อพ้อง Abelmoschus esculentus (L.) Moench)”. This is because the names in parentheses can mean either short name or another name. For this reason, the process only extracts the name outside parentheses.

Some pattern of other name that can be extracted incorrectly are “ภาคกลาง สะระแหน่สวน -, สะระแหน่ญวน” or “เม็ด เหม็ด เสม็ดขาว”.

Precisions and recalls of parts-of-used are over 90%. Examples of parts-of-used that are extracted incorrectly are “ตำรายาพื้นบ้านใช้รักษาอาการไอ” or “ราก ใช้รักษาอาการปวดท้อง”. This is because word in tag can be parts of Thai herb or other information.

TABLE IV. RAW NUMBERS OF EXTRACT RESULT FOR THAI HERB INFORMATION EXTRACTION SYSTEM FROM EACH WEBSITE.

Topic names / website			[1]	[2]	[3]	[4]	summation
Science name	Collection	correct	18	20	25	20	83
		incorrect	0	0	0	0	0
	Total data		20	20	25	20	85
Common name	Collection	correct	20	23	16	-	59
		incorrect	0	0	0	-	0
	Total data		20	23	17	-	60
Other name	Collection	correct	78	73	111	71	332
		incorrect	0	0	5	0	5
	Total data		81	80	113	71	345
Family name	Collection	correct	18	23	20	19	80
		incorrect	0	0	0	1	1
	Total data		18	23	20	20	81
Part of Thai herb	Collection	correct	45	67	53	43	208
		incorrect	0	4	2	6	12
	Total data		48	72	61	44	225
Symptom name	Collection	correct	99	271	223	74	667
		incorrect	8	37	30	3	78
	Total data		128	312	268	81	789

TABLE V. THE PRECISION AND RECALL OF THAI HERB INFORMATION EXTRACTION FOR EACH WEBSITE.

Topic names / website		[1]	[2]	[3]	[4]	summation
Science name	precision	100%	100%	100%	100%	100%
	recall	90%	100%	100%	100%	98%
Common name	precision	100%	100%	100%	-	100%
	recall	100%	100%	94%	-	98%
Other name	precision	100%	100%	96%	100%	99%
	recall	96%	91%	98%	100%	96%
Family name	precision	100%	100%	100%	95%	99%
	recall	100%	100%	100%	95%	99%
Part of Thai herb	precision	100%	94%	96%	88%	95%
	recall	94%	93%	87%	96%	92%
Symptom name	precision	93%	88%	89%	98%	90%
	recall	80%	87%	83%	91%	85%

VI. Conclusion and Future work

This paper proposes a Thai herb extraction process that can extract Thai herb treatment information from multiple websites using HTML parser and template file. The proposed process is able to correctly extract Thai herb information about 95% in precision and 90% in recall.

The future work will consider an algorithm to combine similar and different symptoms for Thai herb recommendation system [11][12]. In the case of inconsistent information from different source, an algorithm for conflict resolution must be developed.

References

[1] T.Chakthong (2010) Thai herb knowledge base. Retrieved April 01, 2001, from Dhonburi Rajabhat University http://dit.dru.ac.th/herb/Main.htm

[2] Herbal Medicine Lists and uses. Retrieved April 01, 2001, from Plant Genetic Conservation Project Office http://www.rspg.or.th/plants_data/herbs/herbs_200.htm

[3] Thai herb database. Retrieved April 01, 2001, from http://www.samunpri.com

[4] Sirirukkachad Nature Park. Retrieved April 01, 2001, from Mahidol University http://www.pharmacy.mahidol.ac.th/siri/index.php?page=home

[5] Jonathan Hedley (2009). JSOUP. Retrieved April 01, 2001, from http://jsoup.org/

[6] S. Gupta, G.E.Kaiser, D.Neistadt, and P. Grimm, “Dom-based content extraction of html documents”, The 12th international conference on World Wide Web, New York, 2003, pp. 207-214.

[7] S. Kuptabut, and P. Netisopakul, "Ontology directed semantic annotation process", The 3rd International Conference on Information Sciences and Interaction Sciences, Chengdu, 2010, pp. 251 - 255.

[8] H. Xia, and Y. Zhang, "Design and Implementation of a Web News Extraction System", The 8th international conference on Fuzzy Systems and Knowledge Discovery, Shanghai, 2011, pp. 1793 - 1797.

[9] N. Ashish, and C.A. Knoblock, "Semi-automatic wrapper generation for Internet information sources", The 2nd international conference on Cooperative Information Systems, Southern California, 1997, pp. 160 – 169.

[10] Lexito. Retrieved April 01, 2001, from NECTEC http://www.sansarn.com/lexto/

[11] P. Chainapaporn, and P. Netisopakul, “Multi-Agent Architecture for Thai Herb Recommendation”,The 9th Joint International Symposium on Natural Language Processing and Agricultural Ontology Service, Bangkok, 2011, pp. 1-6.

[12] P. Chainapaporn, and P. Netisopakul, “Expert system for personal Thai herb Recommendation”, The 4th National Conference on Information Technology, Cha-am, 2012, pp. 286-291.