Applying Collective Intelligence for Search Improvement on Thai Herbal Information

Verayuth Lertnattee[1], Sinthop Chomya[1], Thanaruk Theeramunkong[2] and Virach Sornlertlamvanich[3]
[1]Faculty of Pharmacy, Silpakorn University
Sanamchandra Palace, Muang, Nakorn Pathom, 73000
[1]E-mail: verayuth@email.pharm.su.ac.th, sinthop@email.pharm.su.ac.th
[2]Sirindhorn International Institute of Technology, Bangkadi, Maung, Pathumthani 12000 Thailand
[2]E-mail: thanaruk@siit.tu.ac.th
[3]Thai Computational Linguistics Laboratory, Maung, Pathumthani 12000, Thailand
[3]E-mail: virach@tcllab.org

***Abstract*— Knowledge about herbal medicine can be contributed from experts in several cultures. With the conventional techniques, it is hard to find the way which the experts can build a self-sustainable community for exchanging their information. In this paper, the Knowledge Unifying Initiator for Herbal Information (KUIHerb) is used as a platform for building a web community for collecting the intercultural herbal knowledge with the concept of a collective intelligence. With this system, herb identification, herbal vocabulary and medicinal usages can be collected from this system. KUIHerb provides herbal vocabulary which is dynamically and confidentially applied for searching improvement on the Thai herbal search engine. Three strategies are utilized: (1) providing a set of technical terms in Thai with can be added into the dictionary. These terms are utilized by Thai word segmentation for improving the indexing process (2) A set of synonyms of these technical terms in both Thai and English is built for helping users from a lot of keywords of the same term and (3) a set of keywords from herbal usages can be combined with the name keyword. From the results, information collected from KUIHerb is useful for searching.**

Keywords- collective intelligence; herbal information; search engine; web application

I. INTRODUCTION

Herbal information is a special type of information dealing with medicinal herbs. Internet is one of the excellent source of herbal information since it provides the newest information for researchers or patients. With the fast growth of information on the Internet, there has been extreme needed to find and organize relevant Internet information. For this purpose, Internet search engines are constructed for finding information we need. General-purpose search engines such as Google or Altavista are frequently used as a tool for finding herbal information on the Internet. However, they can cause the user major problems. The results from these search engines consist of diverse topics [1]. Consider the situation where we want to search information about indications of an herb such as a lemon. Using Google with the simplest input "lemon", we will find a large number of pages but only few pages about its medicinal uses. Moreover, herb names have several synonyms. Only the well-trained users can use keywords efficiently by using Boolean searching with appropriated keywords. If users input fewer keywords, only some links to documents will return to users. Most of Thai traditional herbal information on Internet is written in Thai. However, when we use keywords in Thai, fewer documents are found because some terms do not recognized by the general-purpose system. From these examples, it is hard for a non herbalist or a non expertise on herbal medicine, uses a set of appropriated keywords for exploring herbal information efficiently. Herb names and their medicinal usages may be distinct according to their cultural background. Some are named different and hardly found the relation between each other. Some are complimentary knowledge of their usages. These herb names and terminology are useful for searching herbal information on the Internet.

This paper addresses the problems of a design for a collecting herbal information system. It should provide some mechanisms for selecting a set of dynamic and highly confident terms that we can apply for improving searching on the search engine. For the propose, the Knowledge Unifying Initiator for Herbal Information (KUIHerb), a system for collective intelligence on herbal medicine, is used as a platform for building a web community for collecting the intercultural knowledge. KUIHerb provides not only features for sharing and developing herbal information but also a medicinal herb terminology. In Web 2.0 system, a specific-domain search engine is usually implemented as a part of the main system. We also describe the method for applying knowledge from KUIHerb to the search engine. Three strategies are used on a prototype search engine for Thai herbal information: (1) improving the efficiency of the Thai word segmentation which is used by Thai herbal search engine (2) a set of synonyms of these technical terms in both Thai and English is built for helping users from several keywords of the same term and (3) a set of keywords from herbal usages can be combined with the name keyword.

In the rest of this paper, the concept of Web 2.0 and 3.0 system is described in Section II. Section III gives a detail of herbal information. Section IV presents the design of a model for collecting herbal information. Section V explains the search engine for Thai herbal information. Section VI gives a detial of the implementation. The experimental results is described in Section VII. A conclusion and future works is made in Section VIII.

978-0-7695-3836-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CIT.2009.52

II. COLLECTIVE INTELLIGENCE WITH WEB 2.0 AND 3.0

In Web 2.0 era, the Internet users easily share opinions and resources. Consequently, users can collectively contribute to the Web community and generate massive content behind their virtual collaboration [2]. For a system with collective intelligence, implementing scalability can indeed be challenging, but sensibility comes at variable sophistication levels. Several approaches are dealing with the sensibility e.g., user feedback, recommender systems, search engine, and mashups. As suggested by Gruber T., the true collective intelligence can be considered if the data collected from all those participants is aggregated and recombined to create new knowledge and new ways of learning that individual humans cannot do by themselves [3]. However, it provided only a little bit on control of information in Web 2.0.

Nowadays, we are going to the new generation of Web technology i.e., Web 3.0 or the future Web. Although it has already received quite a number of definitions, some useful features of Web 3.0 are described as follow. It can be considered as "The data Web" instead of "The document Web" in Web 2.0. The control of sharing information is better. The decision for the opinions which are provided in Web 3.0, is more accurate. The intelligence Web is a new important feature in Web 3.0 while in Web 2.0, it is only the social Web [4]. Unlike Web 2.0 which participants are usually general Internet users, wisdom of the expert is essential for constructing more valuable knowledge. From these features of Web 3.0, it should be a better collective intelligence system for building new knowledge by way of Information Technology (IT), especially medical knowledge, and herbal knowledge should be no exception.

III. HERBAL INFORMATION

In Thailand, many of traditional medical treatments have been derived the origins in India. The derivation has been diversified through out many cultures since then [5]. For instance, the same species of an herb may be known by different names in different areas. On the other hand, a certain herbal name may mean one thing in one area but something completely different in another. The relationship between herbs and their names is Many-To-Many i.e., a plant may have several names while a name may be several plants. For example, *Dracaena loureiri* Gagnep. We use its hard wood for fever and call Chan dang. Some time we call this plant in other names up to the area of country, e.g., Chan pha (northern part), and Lakka chan (central part) [6]. Lack of information about native herbs has made them more difficult for applying. Herbal specialists usually seek herbal information in a standard monograph. The herbal monograph deals with information to determine the proper identity of a plant genus or genus and species, including part used, indication, method for preparation and so on. However, these sources of information are limited. In the case of the herb does not appear in the pharmacopoeia, it is hard to seek accurately information about the herb. This causes general users and herbal specialists find information of herbs and their products on the Internet.

IV. DESIGN A MODEL FOR COLLECTIING HERBAL INFORMATION

To design a model for collecting herbal information from multicultural community, three components are taken into account. The detail for each component is described as follow.

A. Construction of Image Library

The images of an herb are excellent sources for sharing knowledge about herb identity. From the images, the users can discuss which species (including variety) it should be. The scientific name of an herb and its images are used for common understanding. Furthermore, the users can discuss about which herb should be the real herb that appears in traditional herbal formulas.

B. Sharing and Voting for Herbal Information

Two main parts are applied majority voting to select a set of high ranged opinions i.e., local names and usages of herbs. As an opposite direction, vocabulary collection is a list of terms for the same object which is express with vocabulary management from the content. In the herbal world, the content is usually the scientific name and its pictures which can be use for identifying. As a result, the medicinal herb vocabulary especially local names of an herb, will be collected. These terms can be applied in herbal information retrieval and data mining system, which are crucial in the area of ethnopharmacology and modern pharmacology.

Several topics of medicinal usages should be collected such as part used, indication and methods for preparation. Information about each topic should be discussed from contributors in community. A contributor can post an opinion on the selected topic. Any opinions committed to voting. Opinions can be different but majority votes will cast the belief of the communities. These features naturally realize the online collaborative works to create the knowledge communities.

C. Reliable Improvement

Several mechanisms are used for improving the reliability of the system. Firstly, the users who would like to share information or their opinions need to be members. The members can contribute and also modify their information given to the system. Secondly, some main topics such as symptoms which the herb deals with, will be defined by specialists for herbal medicine. Finally, some reliable or standard references will be added for further finding information.

V. A SEARCH ENGINE FOR THAI HERBAL INFORMATION

To investigate the usefulness of KUIHerb on search improvement, a Thai herbal search engine is constructed. A search on the Internet using as a general-purpose search engines often gives results which are far from satisfactory. Domain specific searching is the way to achieve more efficient in retrieval [7]. Our implementation is based on the crawlers for specific websites which are selected by herbal specialists in our faculty. We focus on two languages i.e., Thai and English. In

term of knowledge, we need to know about herb names, part used, indication, toxicity and so on. For this purpose, a list of URLs is set. The crawlers build indices only the Web sites or some path of the Web sites appear on the list. The rest of this section, three aspects for search improvement on Thai herbal information are described.

A. Adding Thai Herbal Terms for Thai Word Segmentation

In this section, we present the Thai word segmentation [8] that we use for pre-processing step before indexing. Word segmentation is a serious problem in some Asian languages that have no explicit word boundary delimiter, e.g., Chinese, Japanese, Korean and Thai. Let $C = c_1c_2...c_m$ is an input character string, and $w_i = w_1w_2...w_n$ is a possible word segmentation. The formal definition of this problem is defined as:

$$\arg\max_{w_i} P(w_i \mid C) = \arg\max_{w_i} \frac{P(w_i)P(C \mid w_i)}{P(C)}$$

Since $P(C \mid w_i)$ is equal to 1 and $P(C)$ is constant for every argument. The equation above can simplified to

$$\arg\max_{w_i} P(w_i \mid C) = \arg\max_{w_i} P(w_i)$$

Dictionary-based approach is suitable for indexing applications. Three well-known algorithms: longest matching, maximal matching and probabilistic word segmentation are usually applied [9]. In this paper, we use the longest matching algorithm for implementing the Thai search engine since it is simple and fast algorithm. With KUIHerb, a set of high score local names should be added into the dictionary which is used by Thai word segmentation in search engine. This also makes the search engine indices more meaningful terms into the database.

B. Collecting Herbal Terminology for Query Improvement

It is a non trivial work for collecting the names and medicinal usages of an herb in different culture. The herbal terminology is applied for search improvement in two approaches i.e., a list of synonyms from herb names and a set of keywords from indication. At least two advantages for using knowledge from collective intelligence. Firstly, it is dynamic. Therefore, it can be changed at the point of time. Finally, we can select only a subset of terms which are highly confident. To improve retrieval performance, a list of synonyms from herb names can be applied with two operators i.e., the OR as well as the AND operators. The detail for each operator is described as follow.

For the OR operator, a list of synonyms of an herb is used for expanding the search term. This is one technique of query expansion which is a concept to enhanced search terms from users. Various herb names for an herb should be expanded for finding more documents on the Internet. For example, an herb can be referred by various names e.g., English names, Thai names and a scientific name. A common name and local names for each language can be used. Without a list of synonyms, it is hard to input all of these names for an herb. Let *herbA* has two names i.e. *herb1* and *herb2*, the number of documents that contains *herbA* ($|D_{herbA}|$) can be defined as $|D_{herb1}|+|D_{herb2}|-|D_{herb1+herb2}|$, where $|D_{herb1}|$, $|D_{herb1}|$ and $|D_{herb1+herb2}|$ are the numbers of documents that contain term *herb1*, *herb2* and both *herb1* and *herb2*, respectively. Normally, $|D_{herbA}|$ is larger than either $|D_{herb1}|$ or $|D_{herb2}|$. This means that using only one term, it is hard to cover all information we need about the herb. Although a list of synonyms help us to find more documents about an herb, several pages may not be relevant. We will discuss about this topic in the next section about the AND operator.

From our observation about Thai herbal information on the Internet, informative documents about medicinal uses of an herb are usually found on the herbal monograph. In the herbal monograph, several names (synonyms) of an herb are usually provided. To find the pattern of terms which occurred together, association rules are applied. Several research works develop mining techniques on association rules such as [10].

In case of herbal monographs, a set of terms is available on the data set. Each term has a boolean variable (0 or 1) representing the absence (0) or presence (1) of that term. Each monograph can be represented by a boolean vector of values assigned to these terms. The patterns reflect terms that are frequently associated or occurred together. These patterns are in the form of association rules. For example, a monograph which presents the Thai common name (TCN) also tend to provide Thai local name(s) (TLN). This can be represented in association rule below:

$$TCN \Rightarrow TLN_{[\text{support}=20\%,\text{confidence}=50\%]}$$

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 20% means that 20% of all the transactions in the database which is used for analyzing, indicate that Thai common name and Thai local name(s) are occurred together. A confidence of 50% represents that 50% of the documents which find a Thai common name also find Thai local name(s). Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts.

C. A set of Keywords Suggestion

With the popularity of herbal medicine is increasing, patients or ordinary people would like to seek information their need about herbal products. Some people search herbal information on the Internet which is the fastest way to find herbal information with different topics. This situation leads to an increasing number of inquiries from the people to search engines on the Internet. However, it is not easy for users with no traditional medicine training, to find a set of highly relevant information within a short period of time. The collective intelligence can help the people by providing a set of high possible opinions about the topics of interest, especially part used, indication and method for preparation. The terminology can help the people to set appropriated keywords to the search engine.

VI. THE IMPLEMENTATION

The KUIHerb has been implemented using all open source software components. The scripting language is PHP. The data are collected in a database which is constructed with MySQL. Four components are implemented in the KUIHerb i.e., sharing and collecting information, providing information, searching information and Web site statistics. In this paper, a general view for accessing information as well as a main part for sharing and collecting information will be presented.

To express the usefulness of KUIHerb for search improvement, an open source search engine software namely, mnoGoSearch[1] 3.3.7, is used for building a Thai search engine on Linux TLE[2] version 7.0 (Thai Language Extension). The database management system is PostgreSQL[3] version 8.1 on another node. We set a list of links to major herbal information websites to guide the indexer. Thirty Web sites are written in Thai and the others in English. Over a night of automatic indexing, 20780 documents have been stored in database.

VII. EXPERIMENTAL RESULTS

Main components in KUIHerb and their applications for search improvement are described as follow.

A. Sharing Herb Local Names

For the first version of KUIHerb, six topics are taken into account i.e., general characteristics, pictures, local name, medicinal usages (i.e., parts used with their indications and methods for preparation), toxicity, and additional information. Among these topics, a poll-based system is implemented on local names and medicinal usages. A contributor may choose to work individually by posting his/her opinion about those topics. Any opinions or suggestions are committed to voting. Opinions can be different but majority votes will cast the belief of the communities. These features naturally realize the online collaborative works to create the knowledge communities. In this version, all members are given equal weight. The opinion with higher score will be moved up to upper part of the window. The Figure 1 represents a list of local names for an herb and the scores. Two experiments for applying the local names for search improvement.

With six month of contribution from members, 4083 unique local names were provided. When we compared the list of these local names to the Thai dictionary included with the mnoGoSearch (32895 unique terms), only 4 local names were found on the dictionary. Therefore, 4079 new terms were added into the Thai dictionary for improving the process of indexing.

In order to represent the effectiveness of our system, we randomly select five herbs. Each herb has at least one scientific name, one English name and three Thai names i.e., one Thai common name and two Thai local names (with the highest scores). Table I and III list the five herbs. Five columns are used i.e., scientific name (Sci. Name), English name (Eng. Name), The most common Thai name (Thai comm.), the first synonym (Thai Syn. 1) and the second synonym (Thai Syn. 2).

To evaluate searching with the OR operator, the various types of herb names are used as keyword(s) to our search engine in both options, without and with their synonyms. The search results are shown in Table I. Each number represents the number of hits when using the keyword. In case of without synonyms, the numbers of search results for using each term as a keyword are shown in column 2-6. When synonyms are used (Use Syn.), all keywords are applied with OR operators. The query can be represented by

Query = Thai comm. OR Thai Syn. 1 OR Thai Syn. 2 OR Eng. Name OR Sci. Name

From the result, some conclusions can be made: (1) one keyword without its synonyms may not cover all information we need (Climbing Lily, Ginger and Turmeric) (2) the common Thai name is frequently used in Thai Web pages and (3) a few Web documents are constructed with both English and Thai keywords, the language of the result pages depends on the language of the keywords (in case of without synonyms).

From the references which are suggested in KUIHerb, a set of informative documents on the Internet is collected. The total number is 1048 documents with 10817 features after removing HTML tags. The information gain is applied into this dataset for selecting 1000 main features. Apriori is used on the dataset of 1048 transactions of the main features with a support of 20% and a confidence of 80%. The number of rules is 14713779. Some rules about the names is shown in Table II (in Thai but they were translated to English). From these rules, they confirm that more than one name used in an herbal monograph. From this heuristic, the AND operator is applied to names of the herb along with OR operators. From the result in Table I, a Thai common name should be found in the document. Therefore, the query can be represented by

Query = Thai comm. AND (Thai Syn. 1 OR Thai Syn. 2 OR Eng. Name OR Sci. Name)

From the result, a set of terms about names i.e., a local name and a scientific name may occur together on the monograph which provides information about an indication, especially a scientific name. When a user inputs an herb name and the AND operator is applied, the result is shown in Table III.

TABLE II. THE EXAMPLES OF RULES FOR HERBAL NAMES ON HERBAL MONOGRAPHS

Rules	Support (%)	Confidence (%)
Scientific->Name	72.50	100.00
Thai, Scientific -> Name	24.10	100.00
Local, Scientific -> Name	26.10	100.00
Local, Scientific, Name -> Indication	26.10	98.20
Local, Scientific, Name -> Drug	26.10	93.10
Thai, Scientific, Name -> Indication	24.10	93.70
Thai, Scientific, Name -> Drug	24.10	92.10

[1]http: //mnogosearch.org
[2]http: //www.opentle.org
[3]http: //www.postgresql.org

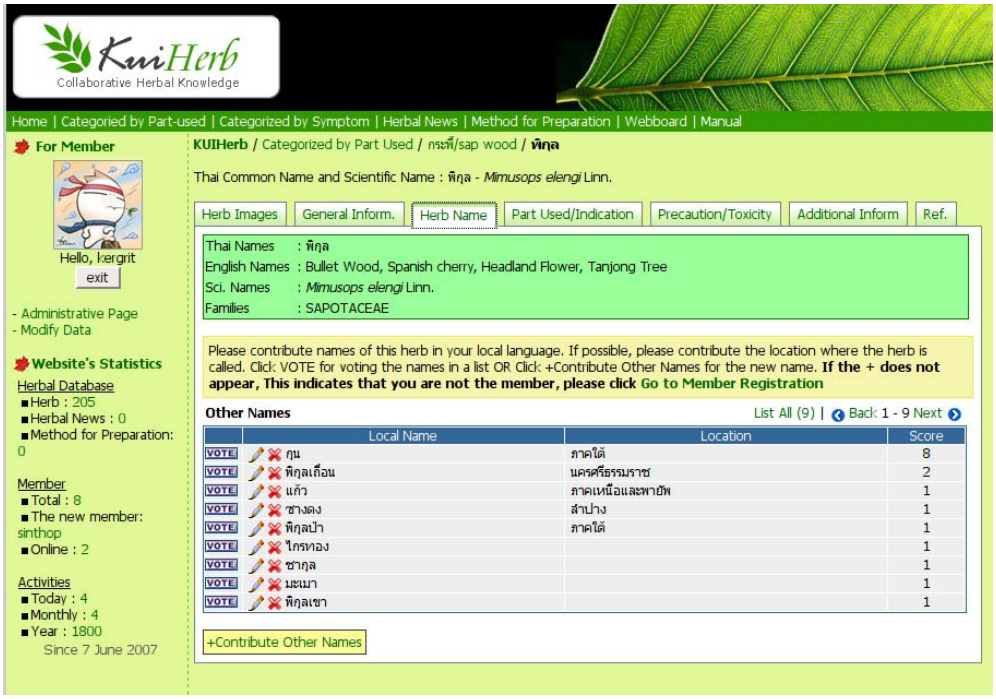

Figure 1. Majority voting system for the Thai local name

TABLE I. THE NUMBERS OF SEARCH RESULTS USING VARIOUS KEYWORDS

Herb	Sci. Name	Eng. Name	Thai Comm.	Thai Syn. 1	Thai Syn. 2	Use Syn.
Climbing Lily	3	3	4	0	6	8
Para cress	0	1	51	1	1	52
Asiatic pennywort	5	4	200	4	12	201
Ginger	8	26	237	2	2	256
Turmeric	12	20	135	212	1	293

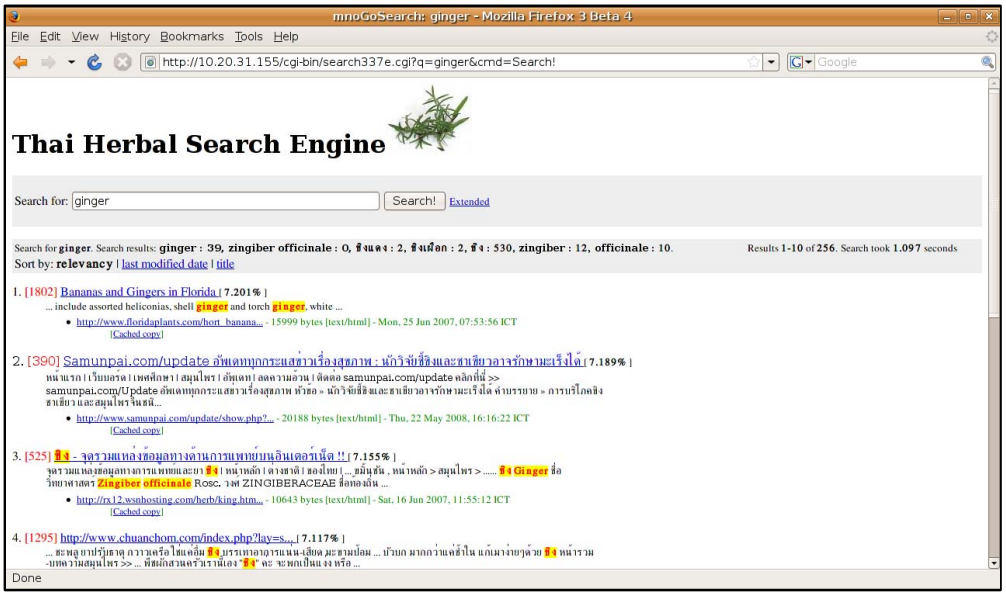

Figure 2. A result from searching with the keyword “ginger”

TABLE III. THE NUMBERS OF SEARCH RESULTS USING THE AND OPERATOR

Herb	Common Thai Name			+ AND Operator		
	Total	*Rel.*	*%Rel.*	*Total*	*Rel.*	*%Rel.*
Climbing Lily	4	2	50	3	2	67
Para cress	51	1	2	1	1	100
Asiatic pennywort	200	121	60	4	4	100
Ginger	237	140	59	7	5	71
Turmeric	135	39	29	73	16	22

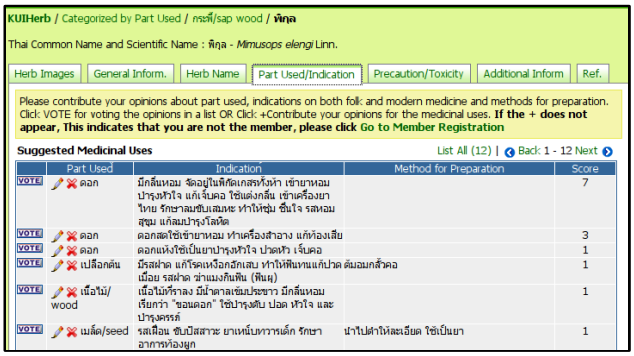

Figure 3. Opinions on herbal usages suggest several keywords.

From the result, when the set of keywords of AND operator is used, more relevant documents are retrieved from the system. From these two experiments, it can be concluded that a set of keywords about names of an herb which contributed from KUIHerb is useful for searching on both OR and AND operators.

B. Sharing Herbal Usages (Keywords Suggestion)

This topic may be the most attractive for herbal information. A registered member can provide his/her opinions about part used of the plant, its indications and how its uses (Figure 3). A list of predefined parts which may be used for treatments is provided. A member may select the part and suggest it indications. The method for preparation can be given. In case of a part with several indications and several methods for preparation, the opinion should be separated to an indication and a method of preparation for each part used. A majority voting is also applied for this topic. The content of an indication and a method for preparation, is described with free text. It will be cleaned into the standard keywords by a group of herbal specialists. When data is cleaned, the keywords can be help ordinary people to use appropriated keywords for searching information about this herb on the herbal search engine and general search engines.

VIII. CONCLUSION AND FUTURE WORKS

In this work, the KUIHerb is used as a platform for building a web community for collecting the intercultural herbal knowledge based on Web 2.0 and some features of Web 3.0 concepts. KUIHerb provides not only a feature for developing an herbal vocabulary but also a capability in expressing the information about part of herb in usages, indications and preparation. In case of multiple opinions are provided, the popular vote will select the most preferable term used in the community. With this system, herbal vocabulary is dynamically and confidentially applied for searching improvement. Three strategies are utilized. Firstly, the system provides a set of technical terms in Thai with can be added into the dictionary. These terms are utilized by Thai word segmentation for improving the indexing process of the Thai herbal search engine. Secondly, a set of synonyms of these technical terms in both Thai and English is built for helping users from a lot of keywords of the same term. Two experiments was done for applying the OR and AND operators. Finally, a set of keywords from herbal usages can be combined with the name keywords. From the results, information collected from KUIHerb is useful for searching.

In this version of KUIHerb, majority voting with equal weight from the members were used for selecting the best opinion. However, the member who has made more valuable contribution to the system should be given more weight. Furthermore, applying data mining to the collected data will be useful. These issues are left for our future works.

ACKNOWLEDGMENT

This work has been supported by Thailand Research Fund and Commission on Higher Education (CHE) under project number MRG5080125 as well as the National Electronics and Computer Technology Center (NECTEC) via research grant NT-B-22-MA-17-50-14.

REFERENCES

[1] S. Oyama, T. Kokubo, and T. Ishida, "Domain-specific web search with keyword spices," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 1, pp. 17—27, 2004.

[2] K.-J. Lin, "Building web 2.0," IEEE Computer, vol. 40, no. 5, pp. 101—102, 2007.

[3] T. Gruber, "Collective knowledge systems: Where the solcial web meets the semantic web," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 6, no. 1, pp. 4—13, 2007.

[4] D. Glustini, "Web 3.0 and medicine," British Medical Journal, vol. 335, no. -, pp. 1273—1274, 2007.

[5] H. D. Lovell-Smith, "In defence of ayurvedic medicine," The New Zealand Medical Journal, vol. 119, no. 1234, pp. 1—3, 2006.

[6] T. Smitinand, Thai Plant Names. Thailand: The forest Herbarium Royal Forest Department, 2001.

[7] M. Kim and P. Compton, "Evolutionary document management and retrieval for specialized domains on the web," International Journal of Human-Computer Studies, vol. 60, no. 2, pp. 201—241, 2004.

[8] W. Aroonmanakun, "Collocation and thai word segmentation," in Proceedings of SNLP-02, 7th International Symposiums on Natural Language Processing, Hauhin, Thailand, 2002, pp. 68—75.

[9] S. Meknavin, P. Charoenpornsawat, and B. Kijsirikul, "Feature-based thai word segmentation," in Proceedings of NLPRS-97. Proceedings of the Natural Language Processing Pacific Rim Symposium, Phuket, TH, 1997, pp. 41—46. [Online]. Available: citeseer.ist.psu.edu/ meknavin97featurebased.html

[10] A. Inokuchi, T. Washio, and H. Motoda, "An apriori-based algorithm for mining frequent substructures from graph data," in Proceedings of the PKDD-00, the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, FR, 2000, pp. 13—23. [Online]. Available: citeseer.ist.psu.edu/inokuchi00aprioribased.html