# A Web-Based Medical Diagnostic System using Data Mining Technique

Jirapond Muangprathub, Yutthaya Jareonsuk, Aekarat Sealiw
*Applied Mathematics and Informatics Laboratory, Faculty of Science and Industrial Technology, Prince of Songkla University, Suratthani Campus, Surattani, Thailand.*
*jirapond.m@psu.ac.th*

***Abstract*—The purpose of this paper is to apply data mining technique for medical diagnostic system on web application supporting Thai language. This web would help users to reduce expense and time of visiting doctors. It is capable of giving preliminary diagnosis. The proposed system will discover the implication knowledge with association rules derived from formal concept analysis (FCA) to advice co-symptom of diagnosing to achieve more correctly. The association rules are built from its subconcept and superconcept relation from concept lattice. In addition, the proposed system supporting Thai language is challenged because this language is streaming string without the boundary delimiters. The proposed system is developed based on online web application to demonstrate real situation. The result show that the proposed system can suggest co-symptom to achieve more correctness for medical diagnostic.**

***Index Terms*—Medical Diagnostic System; Data Mining Technique; Web-based; Formal Concept Analysis.**

## I. INTRODUCTION

Data mining is important knowledge discovery in the information industry. The techniques of data mining are widely used to extract from big data [7, 9, 10]. One of the most popular techniques to perform data mining is discovering association rules [7, 11, 17]. Association rule is mostly used for discovering interesting relations between variables in large databases. In healthcare, the techniques of data mining has been used as well in various aspects to discover knowledge for healthcare information [6, 7, 15, 14, 20]. The relation of a large amount of information is generated to diagnosis, disease identification and treatment of an individual. Mostly, the diagnosis and treatment of disease from the clinical data set is used to extract the knowledge for providing scientific decision-making. To discover the related of medical data, association rules is applied to obtain interesting patterns. These rules may aid the diagnosis and prognosis of diseases to identify the relationship that occurs among several diseases. Moreover, association rules [11, 16, 17] have been applied to extract the knowledge from medical history, laboratory and demographic data [1, 13, 16, 20]. Thus, this paper proposes an application for medical diagnostic system based on association rules using formal concept analysis (FCA) in the general rules acquisition.

FCA [2, 12, 19] is widely used for data analysis in concept lattice form in information science [25]. This approach provide the hierarchical structure of information. The obtained structure in concept lattice is formed the relationship of generalization and specialization of information. For this reason, we can extract the feature dependency of information. The feature dependency derived from the relationship between subconccept and superconcept in concept lattice. In addition, the concept lattice prevent the redundancy inside a huge data [2, 12]. Thus, in this work we applied FCA to build knowledge structure using the previous medical history data. This knowledge base can generate strong association rules with attribute implication in order to find out the useful association relationship or pattern between the diagnosis and prognosis of several diseases. Thus, this work uses the extracted implication rules from the knowledge base to advice patients.

At the same time, the Internet has been widely used to promote many fields because of growing easy access anytime and anywhere. For this reason, many web applications for healthcare were developed. The online medical diagnostic is one aspect to serve the growth in number of patients. It has great potential to become a low-cost and effective source of basic healthcare promotion interventions. The patients can diagnose themselves in elementary case before they go to see the doctors. Thus, this work develops the medical diagnostic system based on web application using the mentioned attribute implication. The presented system supports Thai language. The challenge of diagnostic descriptions with this language is the difficulties in detecting Thai words and statements because it is a tie of context.

This article is organized as follows. Section 2 provides background of data mining technique, FCA, and Thai word segmentation. In Section 3, we present our proposed medical diagnostic system. Section 4 shows experiment results. Finally, Section 5 concludes the article.

## II. BACKGROUND

### *A. Data Mining*

In order to present the rationale of our system, it is essential to first present a brief description of data mining. The processes of discovery knowledge are recalled as follows in [10]. Data mining refers to extract a knowledge from large amounts of data. This paper divides process of data mining into 4 steps as follows.

#### *i. Data pre-processing*

This is an important step in the knowledge discovery process, because quality of knowledge depend on quality of data. In the real world data tend to be dirty, incomplete and inconsistent.

Thus, the step can help to improve the accuracy and efficiency of the subsequent mining process. The process of this step consists of data cleaning, data integration, and data transformation. This work is prepared for medical diagnostic in Thai language using Thai word segment. Moreover, we clean data and fill the missing value for incomplete data by doctor. The result is ready for next step in relational database form.

*ii. Data reduction*

This step can be applied to obtain a reduced representation of data set. The smaller volume is built to closely maintain the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results. For instance, the researchers apply rough set to reduce data prior extracting the knowledge [3].

*iii. Data modelling/ discovery*

This step extracts knowledge from the prepared data. Mostly, data modelling/ discovery apply intelligent methods to achieve data pattern. The discovery knowledge can be represented with the tasks of classification, clustering, association and so on. In this work, we use association rule to discovery knowledge by using FCA.

*iv. Solution analysis*

This step involves analysis of the results from data modelling/ discovery. The solution in our work is used to advice patients. To demonstrate the performance of our system, this work is assessed based on preference of users. Moreover, we compare the correctness of the prognosis between using associate rules and without.

*B. Formal Concept Analysis*

Formal concept analysis (FCA) [2,19] is a approach to data analysis with concept lattice form. The concept lattice structure indicate to hierarchical of data in explicit and implicit knowledge. The implicit knowledge is embedded in the data, and can be extracted in the form of dependency rules among the attributes describing the objects.

A formal context $\mathbb{K}:=(G,M,I)$include $G$, $M$, and $I$ where $G$ is called the objects, $M$ is called the attributes and $I$ is the relation between $G$ and $M$ of the context. $(g,m)\in I$ express that the object $g$ has the attribute $m$ . For a set $A\subseteq G$ of objects, $A'$ is defined as follows $A':=\{m\in M\mid gIm\ \ for\ all\ g\in A\}$. Correspondingly, for a set $B\subseteq M$ of attributes, $B'$ is defined as follows $B':=\{g\in G\mid gIm\ \ for\ all\ m\in B\}$.

A formal concept of $\mathbb{K}:=(G,M,I)$ is a pair $(A,B)$ with $A\subseteq G, B\subseteq M, A'=B$ and $B'=A$, where $A$ is the extent and $B$ is the intent of $(A,B)$. All formal concept is built to concept lattice. For more details, see [2]. This structure provides implication between attributes inside its hierarchical structure.

Attribute implication (over set of attributes $M$ ) is an expression $A\Rightarrow B$, where $A,B\subseteq M$ ($A$ and $B$ are sets of attributes). The implication can be read as: *if an object has all attributes from A, then it also has all attributes from B* and holds in the context $(G,M,I)$ if $A'\subseteq B'$. An implication can measure dependency by considering subconcepts and superconcepts from concept lattice. For more details, see [2].

This work applies FCA with attribute implication where set of objects is disease and remedy, and set of attributes is keywords of description of symptom. This knowledge base can generate attribute implication to find association relationship between symptoms of diagnosis in order to identify disease and remedy.

*C. Thai Word Segmentation*

This proposed system supports Thai query and stores medical history to suggest more correct prognosis. We applied Thai word segmentation approach to manage the Thai medical data. The feature of Thai word is streaming string led to more complex. Moreover, this language does not have boundary delimiters of character or word [5, 24]. For this reason, many researcher developed the approach for Thai language wrapping with using dictionary and non-dictionary to retrieve the keyword or word as discussed in [26]. To implement our system, we apply Thai Segmentation Program [28] to obtain keywords or word from Thai medical history data. At the same time, user's query use this method to obtain keywords or word. The process of word segmentation is showed in Figure 1.

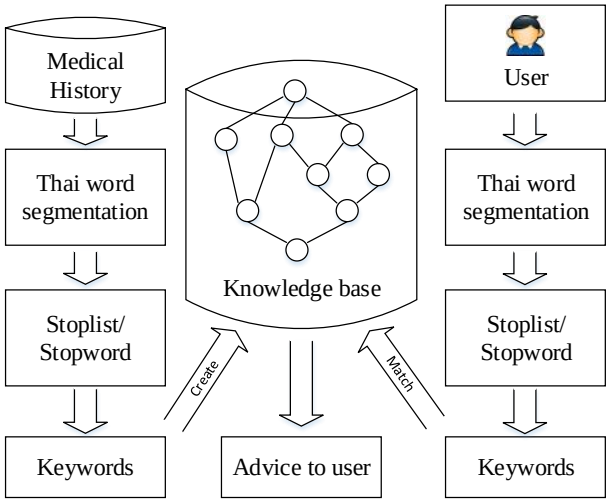



Figure 1: Thai word segmentation for retrieving knowledge to advice user

From Figure 1, the medical history consists of description symptom of disease and remedy in Thai free text form. It will be cropped with Thai Segmentation Program [28] in words in this work. Next, the stoplist (or stopwords) is used to achieve the indexing of word. At the same time, when the user input query to search name of disease with Thai free text in symptom form and then the joint symptom will be advised to enhance correctness of prognosis disease. The query will be also cropped into words and using a stoplist to obtain keywords for matching query and the prepared knowledge. The result shows the list of joint symptoms to provide user with more information.

## III. Related Works

Data mining applications of health have tremendous potential and usefulness [7,9,17]. In healthcare, data mining is used for the diagnosis and prognosis of diseases and to identify the relationship that occurs among several diseases. As healthcare

data are not limited to only quantitative data, it is also necessary to explore the use of data mining to expand the scope of what health care data mining can do.

The data mining techniques can used for supervised learning and unsupervised learning tasks. The supervised learning task will construct concise model of class labels for predictive feature [10]. The task of supervised learning is applied in healthcare application. For instance, Ruijuan [18] improved ID3 algorithm for breast-cancer prediction with 80% accuracy. Shweta [20] discussed various data mining approaches that had been utilized for breast cancer diagnosis and prognosis. The author reviews data mining techniques to apply in diagnosis and prognosis of cancer disease. The author found that decision tree is the best predictor with 93.62% accuracy of classification using UCI and SEER dataset, and the Bayesian network is a popular technique in medical prediction particularly. Sellappan et al. [21] developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using data mining techniques. They used decision trees, naïve bayes and neural network to compare classification accuracy. The results show that each technique has its unique strength in realizing the objectives of the defined mining goals. Moreover, the supervised learning in data mining techniques is discussed in much healthcare research surveyed in [1, 9, 14]. Thus, the supervised learning is successfully applied to enhance the prognosis and diagnosis.

On the other hand, the unsupervised learning task by using data mining techniques is regarded as a cluster analysis task. Cluster analysis or clustering aim to group a set of closest feature and divide the difference feature between groups. Each the closest group is called cluster [10]. In healthcare, cluster analysis is applied to divide numerous medical data into cluster due to small data. Each cluster can be identified with same feature that derive from the extraction of knowledge [22]. For instance, in [22], Bala et al. applied the K-Means clustering technique to predict heart disease diagnosis by using real and artificial datasets to experiment their system.

In addition, one of the most popular approaches in data mining is discovering association rules surveyed in [15, 17]. The researchers improve algorithm of association rule mining to enhance the performance of knowledge discovery [8, 11, 27]. Association rule mining is applied to improve decision making in medical diagnosis. Sridar et al. [14] applied association rule mining with artificial neural network (ANN) for the diagnosis of diabetes fully based online with real time input. It gives greater accuracy, more reliability and ease of use and maintenance. Rashid et al. [16] applied association rules mining for finding co-occurrences of diseases. The researchers implemented the system using clinical state correlation prediction (CSCP) with the healthcare repository that derived from patients. Sengupta et al. [6] applied Apriori association rule algorithm to discover associative rules among the clinical parameters in brain tumor data warehouse. The results indicate that the presented approach can follow the diagnostic procedure of brain tumor from large data volumes. Exarchos et al. [23] used association rule to extract rule-based classification model for the detection of ischemic beats in long duration electrocardiographic (ECG) recordings. Also, association rule mining is applied to give an efficient diagnosis in image form [1, 13]. For example, in [13], the authors develop a method based on association rule-mining to enhance the diagnosis of ultrasound kidney images. The method uses association rule mining to analyze the medical images and automatically generates suggestions of diagnosis.

In summary, the association rule mining is widely used to fulfill healthcare tasks. It is suitable for using in suggesting or predicting the diagnosis and prognosis in medical data warehouse. The association rules are derived from general knowledge that is representative of big data. However, the specific knowledge still require to response to user. Thus, a better way to achieve two types of knowledge at the same time is required. FCA is one approach to prepare both general and specific knowledge because it is represented through hierarchical conceptual structures called concept lattices. This structure provides specific knowledge through hierarchical conceptual and provides general knowledge through implication rules. For this reason, we applied FCA to extract the knowledge from the medical diagnosis system.

## IV. The Proposed System and Experiment Results

Figure 2 illustrated an overview of our system for users and admin. Firstly, administrator deals with previous medical data from clinics and local hospitals in data warehouse form. The admin prepares the medical data in Thai language that include symptom, disease, remedy and doctor using Thai word segmentation to obtain keywords. Afterwards, the set of keywords is reformed into knowledge base structure by using FCA. Secondly, when users or patients input detail of symptom, it will be matched with implication rules from knowledge base to identify co-symptom. Consequently, our system will interact with users and suggest the co-symptom to enhance the prognosis and remedy.

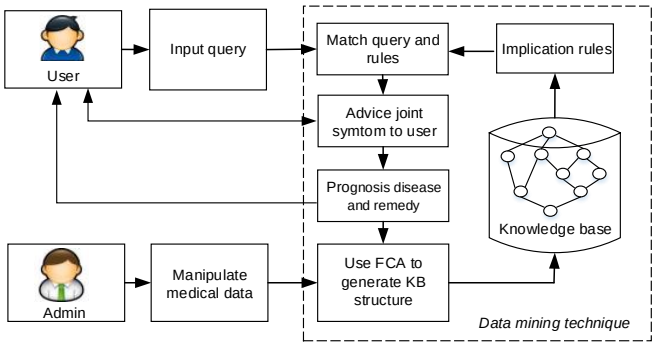
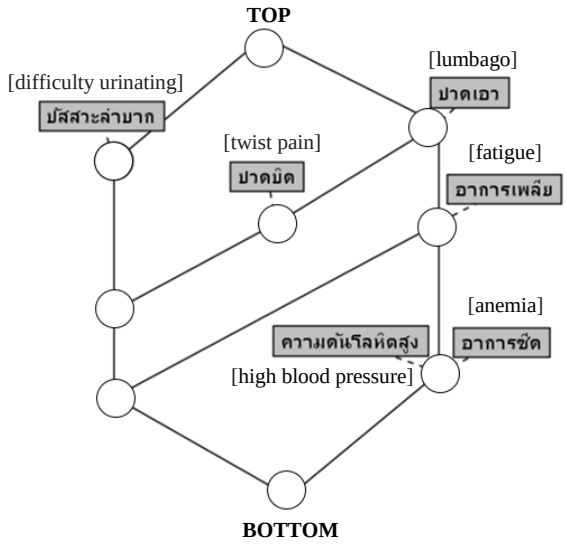



Figure 2: Overview our proposed system

Applying data mining technique, we use implication to achieve associate rules from the previous medical data to generate knowledge base. The set of keywords is represented in attribute form and name of disease is represented in object form. Both keyword and name of disease are derived from Thai word segmentation. There are used to generate associate rules for the extracted knowledge for suggestion to user. Table 1 show an example derived from Thai word segmentation. Consequently, the obtained data is used to construct knowledge base in concept lattice form showed in Figure 3 (left). Next, the concept lattice is extracted with implication rules showed in Figure 3 (right).

Table 1
An example of medical data after using Thai word segmentation

| | อาการซีด(anemia) | อาการเพลีย (fatigue) | ความดันโลหิตสูง (high blood pressure) | ปวดเอว (lumbago) | ปัสสาวะลำบาก (difficulty urinating) | ปวดบิด (twist pain) |
|---|---|---|---|---|---|---|
| กรวยไตอักเสบเรื้อรัง (Chronic pyelonephritis) | 1 | 1 | 1 | 1 | 0 | 0 |
| นิ่วไต (Kidney stones) | 0 | 1 | 0 | 1 | 1 | 1 |
| นิ่วท่อไต (Ureteral calculi) | 0 | 0 | 0 | 1 | 1 | 1 |
| นิ่วกระเพาะปัสสาวะ (Bladder stones) | 0 | 0 | 0 | 1 | 0 | 1 |
| กระเพาะปัสสาวะอักเสบ (Urinary infection) | 0 | 0 | 0 | 0 | 1 | 0 |
| ท่อปัสสาวะตีบ (Urethral stricture) | 0 | 0 | 0 | 0 | 1 | 0 |




(a)

**Rule 1:** อาการซีด ==> อาการเพลีย, ความดันโลหิตสูง, ปวดเอว
[anemia ==> fatigue, high blood pressure, lumbago ]

**Rule 2:** อาการเพลีย ==> ปวดเอว
[fatigue ==> lumbago]

**Rule 3:** ปวดบิด ==> ปวดเอว
[twist pain ==> lumbago]

**Rule 4:** ความดันโลหิตสูง ==> อาการซีด, อาการเพลีย, ปวดเอว
[high blood pressure ==> anemia, fatigue, lumbago]

**Rule 5:** อาการเพลีย ปวดเอว ปวดบิด ==> ปัสสวะลำบาก
[fatigue, lumbago, twist pain ==> difficulty urinating]

**Rule 6:** ปวดเอว ปัสสวะลำบาก ==> ปวดบิด
[lumbago, difficulty urinating ==> twist pain ]

(b)

Figure 3: Concept lattice from Table 1 (a) and implication rules from concept lattice (b)

The proposed overview system is developed as showed in Figure 4. This figure shows an example of application i.e. homepage and input interface to users. Next, the users click diagnosis button and our system presents co-symptom (to be selected by users) that is derived from implication rules by using FCA. The output shows detail of prognosis disease, remedy and suggested doctor. Moreover, our system can present the details of doctors who will be involved with treating advised prognosis.

We developed our system as online website. PHP language is used to implement system while MySQL database is used as database of detail of disease, remedy and doctor that is obtained from local hospitals in Surat Thani Province, Thailand. This data in Thai language consists of 100 records of symptom, diseases, remedy and doctor. We use 100 queries to test interaction and correctness of prognosis disease between our system and user's query. The result is showed as follows:

Table 2
The performance of our system to explain prognosis disease by using implication rules in FCA

| Detail of experiment | Amount of correctness (Times) |
|---|---|
| Ability to show the co-symptom | 89 |
| Not show the co-symptom | 11 |
| Ability of suggesting doctor to user | 65 |
| Correctness of disease prediction | 71 |


Figure 4: Some part of implementation for medical diagnostic

## V. CONCLUSION

The purpose of this paper is to apply data mining technique for medical diagnostic system on web application supporting Thai language. This web would help users to reduce expense and time of visiting doctors. It is capable of giving preliminary diagnosis. The proposed system will discover the implication knowledge with association rules derived from formal concept analysis (FCA) to suggest co-symptom for diagnosis more correctly. The association rules are built from its subconcept and superconcept relation from concept lattice. Moreover, a diagnostic system supporting Thai queries is challenging because of difficulty in detecting Thai words and statements. The proposed system supporting Thai language is challenged

because this language is streaming string without the boundary delimiters. The proposed system is developed based on online web application to demonstrate real situation. The result shows that the proposed system can suggest co-symptom to achieve better accuracy for medical diagnostic.

## ACKNOWLEDGMENTS

The authors are deeply grateful to the Faculty of Science and Industrial Technology, Prince of Songkla University, Surat Thani campus, Thailand. This research was financially supported by Prince of Songkla University, Surat Thani Campus 2015.

## REFERENCES

[1] Kavipriya, A., Gomathy, B.: Data Mining Applications in Medical Image Mining: An Analysis of Breast Cancer using Weighted Rule Mining and Classifiers. IOSR Journal of Computer Engineering, vol. 8, no. 4 (2013), pp. 18-23.

[2] Ganter, B., Wille, R.: Formal concept analysis: Mathematical Foundation. Springer, Heidelberg, New York, 1999.

[3] Tripathy, B. K., Acharjya, D. P., Cynthya, V.: A Framework for Intelligent Medical Diagnosis Using Rough Set with Formal Concept Analysis. International Journal of Artificial Intelligence & Applications (IJAIA), vol.2, no.5, (2011), pp. 45-66.

[4] Kumar, C. A.: Knowledge Discovery in Data Using Formal Concept Analysis and Random Projections. International Journal of Applied Mathematics and Computer Science, vol. 21, no. 4 (2011), pp.745-756.

[5] Haruechaiyasak, C., Kongyoung, S., Dailey, M.N.: A Comparative Study on Thai Word Segmentation Approaches. in: Processing of ECTI-CON. (2008). pp.1-4.

[6] Sengupta, D., Sood, M., Vijayvargia, P., Hota, S., Naik, P. K.: Association Rule Mining Based Study for Identification of Clinical Parameters Akin to Occurrence of Brain Tumor. Bioinformation, vol. 9, no. 11 (2013), pp.555-559.

[7] Tomar, D., Agarwal, S.: A Survey on Data Mining Approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, vol.5, no.5 (2013), pp. 241-266.

[8] AL-Zawaidah, F. H., Jbara, Y. H., Abu-Zanona, M. A.: An Improved Algorithm for Mining Association Rules in Large Databases. World of Computer Science and Information Technology Journal (WCSIT), vol.1, no. 7 (2011), pp. 311-316.

[9] Parvathi, I., Rautaray, S.: Survey on Data Mining Techniques for the Diagnosis of Diseases in Medical Domain. International Journal of Computer Science and Information Technologies, vol. 5, no. 1 (2014), pp. 838-846.

[10] Han, J., Kamber, M.: Data Mining Concepts and Techniques. The Second Edition, Morgan Kaufmann Publishers, March 2006.

[11] Hipp, J., Guntzer, U., Nakhaeizadeh, G.: Algorithms for Association Rule Mining A General Survey and Comparison. ACM SIGKDD, vol.2, no. 1(2000), pp. 58-64.

[12] Poelmans, J., Elzinga, P., Viaene, S., Dedene, R.: Formal Concept Analysis in Knowledge Discovery: A Survey. ICCS 2010, LNAI 6208, (2010), pp. 139153.

[13] Jose, J. S., Sivakami, R., Uma Maheswari N., Venkatesh, G.: An Efficient Diagnosis of Kidney Images Using Association Rules. International Journal of Computer Technology and Electronics Engineering, vol. 2, no. 2 (2012), pp. 14-20.

[14] Sridar, K., Shanthi, D.: Web Based Medical Diagnosis System Using ANN- ARM for the Diabetes Mellitus. International Journal of Computers and Distributed Systems, vol. 3, no. 3, (2013), pp.15-20.

[15] Devi, M.R., Sarojini, A.B.: Applications of Association Rule Mining in Different Databases. Journal of Global Research in Computer Science, vol.3, no. 8, (2012), pp. 30-34.

[16] Rashid, M.A., Hoque, M. T., Sattar, A.: Association Rules Mining Based Clinical Observations. arXiv:1401.2571, 2014.

[17] Zhao, Q., Bhowmick, S. S.: Association Rule Mining: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116, pp. 1-20.

[18] Hu, R.: Medical Data Mining Based on Decision Tree Algorithm. Computer and Information Science, vol. 4, no.5 (2011), pp. 14-19.

[19] Wille, R.: Formal concept analysis as mathematical theory of concepts and concept hierarchies. Lecture Notes Artificial Intelligent (LNAI-3626), Springer, (2005), pp.1-33.

[20] Kharya, S.: Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), vol.2, no.2 (2012), pp. 55-66.

[21] Palaniappan, S., Awang. R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. International Journal of Computer Science and Network Security, vol. 8, no. 8 (2008), pp.343-350.

[22] Bala, S. V., Devi, T., Saravanan, N.: Development of a Data Clustering Algorithm for Predicting Heart. International Journal of Computer Applications, vol. 48, no.7, (2012), pp.8-13.

[23] Exarchos, T. P., Papaloukas, C., Fotiadis, D. I., Michalis, L.K.: An Association Rule Mining-Based Methodology for Automated Detection of Ischemic ECG Beats. IEEE Transactions on Biomedical Engineering, vol. 53, no. 8 (2006), pp. 1531-1540.

[24] Theeramunkong, T., Usanavasin, S.: Non-Dictionary-Based Thai Word Segmentation Using Decision Trees. in: HLT 01 Proceedings of the first international conference on Human language technology research. (2001). pp. 1-5.

[25] Priss, U.: Formal Concept Analysis in Information Science. Annual Review of Information Science and Technology, vol. 40, no.1 (2006), pp.521-543.

[26] Aroonmanakun. W.: Collocation and Thai Word Segmentation. In: Proceedings Of SNLP-Oriental COCOSDA. (2002). pp. 68-75.

[27] Zhang, W., Ma, D., Yao, W.: Medical Diagnosis Data Mining Based on Improved Apriori Algorithm. Journal of Networks, vol. 9, no.5 (2014), pp. 1339-1345.

[28] Available at: http://www.arts.chula.ac.th/ling/wordseg