A Text Mining Approach to Find Patterns Associated with Diseases and Herbal Materials in Oriental Medicine

Ji Hoon Kang, Dong Hoon Yang, Young Bae Park, and Seoung Bum Kim

***Abstract*—Currently, in Oriental Medicine, obtaining the statistical backgrounds is a very important and challenging issue to be authorized as the evidence-based medical science. Medicinal treatment is one of main part in Oriental Medicine, and it has some difficulties to specify the effectiveness of herbs because of its own theoretical basis and practical experience. The main purpose of this study is to clarify the relationship between herbal materials and efficacy of remedies in Bangyakhappyeon which is one of the representative textbook in Oriental Medicine with a text mining approach. We adopted an association rule induction combined with the network analysis and found useful and informative relationships between the symptoms and herbal materials.**

***Index Terms*—Oriental medicine database; bangyakhappyeon; herbal materials; association rules**

I. Introduction

Based on traditional theory compiled through thousands of years of practice and research by Oriental Medicine experts, a large amount of knowledge has accumulated in the form of ancient books and modern literature. At the same time, it is becoming harder to understand the interrelated roles of herbal materials in complex prescriptions. These situations create an urgent need for methods that will allow researchers to use Oriental Medicine data thoroughly, effectively, and efficiently from a knowledge discovery database (KDD). Because accumulated clinical experience from the books has great value, a KDD database is necessary to take advantage of this valuable resource (Bath, 2004).

The most prominent definition of a KDD was proposed by Fayyad et al. (1996). In that study, a KDD was defined as “the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.” The KDD is a growing research field consisting of many methods, including visualization, clustering, classification, and association rules (Witten et al., 2011). KDD is useful for analysing and understanding large quantities of data. The subject of KDD has attracted considerable interests in many disciplines, and their widespread applications range from finance, marketing, and telecommunication to scientific analysis (Hatonen et al., 1996). KDD methodology is particularly suitable for the complex organisms, processes, and relationships found in medicine.

Recently, the techniques in KDD are utilized to extract meaningful patterns from unstructured databases (e.g., text, image, etc) as well as structured databases. Text mining is the process of extracting useful information from text. Compared with the intensive research work and the immense publicly accessible bibliographic literature data of modern biomedical science, the development of text mining for Oriental Medicine is still in its early stages. Interestingly, text mining research in Oriental Medicine to date has been focused on literature-based discovery.

The relationships between disease patterns and prescriptions in Shanghanrun, Dongeuibogam, and Bangyakhappyeon, all regarded as typical ancient books, are limited to frequency analysis. Hong et al.(2011) proposed network analysis to identify relationships between disease patterns and prescriptions.

Obviously, text mining of Oriental Medicine is still in its early stages. Medical disciplines have many information and data sources, including data generated during practical clinical processes and research activities. Clinical practice in Oriental Medicine is a kind of complex clinical experiment of trying to effectively apply a vast amount of largely uncategorized information and data sources concerning symptoms, herbal treatments, and drugs. It is important to develop new clinical research measures that will help Oriental Medicine practitioners to efficiently use data collected from clinical practice. At the same time, this will promote the development of TCM (Traditional Chinese Medicine) and TKM (Traditional Korean Medicine) from their status as the collected experience of individuals a system of into evidence-based medicine.

In Korea, Bangyakhappyeon is a simplified prescription book based on the information in the Donguibogam, which was published in 1885 (Kim and Yoon, 1991). This book has long been used by clinicians. The main purpose of this present study is to construct the entire Bangyakhappyeon disease-prescription-drug pattern and analyze the relationships between disease patterns and drugs. We obtained specific medical material with a strong association for internal damage and visualized the network for herbal materials and symptoms. High-frequency drugs and high-specificity drugs as well as frequently used paired-drugs for internal damage were researched.

We found that the results are clinically useful resources that can be potentially utilized. The results of this study can be used as basic information for future drug development and for a clinical decision support system for Oriental Medicine clinicians.

II. Data

Bangyakhappyeon consists of 54 categories including

Manuscript received March 15, 2012; revised April 20, 2012.

J. H. Kang and S. B. Kim are with the School of Industrial Management Engineering, Korea University, Seoul, Korea

D. H. Yang, Y. B. Park are with the Department of Biofunctional Medicine & Diagnostics, Kyung Hee University, Seoul, Korea

prescriptions for the main disease patterns. Categories 1 through 18 contain miscellaneous frequent diseases, categories 19 through 30 consider the internal parts of the body, including essence, spirit, Qi, and blood. In addition, categories 31 through 52 are devoted to physical maladies such as those of the eyes, ears, and mouth. The remaining two categories (53 through 54) consider gynecology, obstetrics, and pediatrics. Disease patterns are summarized based on the cause, nature, and location of the pathological changes at particular stages of diseases.

Bangyakhappyeon contains 521 prescriptions and 305 herbal materials. Most prescriptions (formulas and recipes) comprise several herbal ingredients per prescription. We analyzed the formulas without considering dosages because dosage information in Bangyakhappyeon is highly variable. As shown in Table I, we constructed a binary matrix in which columns are herbal materials, rows represent prescriptions, and each cell has either a 0 or a 1.

Many kinds of prescriptions and herbal materials account for the 52 types of diseases in the book. If the variety of prescriptions were more extensive, the number of materials would grow in a geometric progression. Therefore, the important medicines for certain disease patterns are difficult to discern. The dataset contains several combinations of prescriptions and herbal materials used, including duplications, in 52 disease categories. For example, 39 prescriptions are available for internal damage, 91 herbal materials can be used, and 340 materials, including duplications, are available.

III. Method

A. Association Rules

Association rules have been widely used to find out relationships between item sets in large databases. Association rules are also called a market basket analysis because they have been used to discover which groups of products tend to be purchased together in markets (Borgelt and Kruze, 2002). Association rules can be expressed in the form of "if-then" statements in a probabilistic sense. For example, "if A-event occurs, then B-event also ensues." Here, the event, which occurs first (Item A) is called an antecedent event and the event that follows the first event (Item B) is called a consequent event.

In general, association rules are generated in two stages. First, a set of frequent rules is generated. Second, the strength of the rules, obtained from the first stage is evaluated. For the first stage to generate the rules, an Apriori algorithm has been widely used (Agrawal et al., 1993). The main idea of the Apriori algorithm is to generate frequent one-item sets and uses this information to generate two-item sets, and then three-item sets until frequent item sets of all sizes are generated. The concept of support is used to determine the frequent item sets. The support of a rule with an antecedent Item set A and a consequent Item set B (A→B) is simply defined as the proportion of transactions that include all antecedent and consequent item sets and can be calculated by the following probability (Agrawal et al., 1994):

$$\text{Support} : (A \rightarrow B) = P\,(A \cup B), \qquad (1)$$

where N represents the total number of transactions in the database.

Having found a number of candidate rules from the Apriori algorithm, the goal is now to assess the strength of the rules. Two main measures to achieve this goal are confidence and lift. Confidence is defined as the ratio of support value to the number of transactions of all the antecedent items sets. Therefore, the confidence of the rule (A→B) can be represented by the following conditional probability:

$$\text{Confidence} : (A \rightarrow B) = P\,(B \mid A). \qquad (2)$$

The lift value of a rule is the ratio of the number of transactions of consequent item sets given that antecedent item set has occurred to the number of transactions of consequent item sets in all transactions. In other words, the lift is the ratio of one confidence to another confidence, assuming that consequent transactions are independent with antecedent items (Tan et al., 2006). The lift value of the rule (A→B) can be calculated as follows:

$$\text{Lift} : (A \rightarrow B) = P\,(B \mid A) / P(B). \qquad (3)$$

A lift value greater than 1 implies that the degree of association between the antecedent and consequent item sets is higher than in a situation in which the antecedent and consequent item sets are independent.

IV. Results

We used association rules to characterize the relationships between diseases and herbal materials. Table I shows the relationship between internal damage and the associated herbal materials and their corresponding confidence and lift. Association rules of other diseases (other than internal damage) were obtained. However, the results are not shown in this paper due to page limits. Here, internal damage can be considered as an antecedent item set, and the herbal materials can be considered as a consequent item set. Table I shows the 11 association rules between internal damage (antecedent) and their associated herbal materials (consequent) that have a support value of at least 20%. The minimum support value is usually determined by the users.

Table I also shows the values of confidence and lift values that can be used to judge the strength of rules. Herbal materials with high confidence and lift values have strong relationships with the disease, internal damage. For example, the rule (Internal damage→ Atractylodis Rhizoma Alba) has the highest confidence(50%), meaning that Atractylodis Rhizoma Alba is the most frequently used herbal material for treating internal damage. However, it is interesting to note that the rule (internal damage→ Atractylodis Rhizoma Alba) has a relatively low lift value (1.86). This implies that Atractylodis Rhizoma Alba is frequently used to treat other diseases as well as internal damage, and thus, can be considered as a generally used material. Conversely, Crataegii Fructus, despite its relatively lower confidence (20.59%), has a high lift value (9.75). This implies that Crataegii Fructus is an herbal material specifically used for treating internal damage. It can be seen that Crataegii Fructus and Massa Medicata Fermentata have larger lift values compared with their confidence values; the opposite is true for Atractylodis Rhizoma Alba and Poria(white). As mentioned earlier, the larger lift values (compared with their confidence values) of Crataegii Fructus and Massa Medicata Fermentata imply that these medicines are preferred for treating internal damage. On the other hand, the low lift (compared with confidence values) values of Atractylodis Rhizoma Alba and Poria(white) imply that they are globally used medicines for general diseases, including internal damage.

TABLE I: CONFIDENCE AND LIFT BETWEEN DISEASE (INTERNAL DAMAGE) AND HERBAL MATERIALS

Antecedent (Disease)	Consequent (herbal materials)	Confidence (%)	ift
Internal damage	*Atractylodis Rhizoma Alba*	50.00	1.86
	Poria(White)	44.12	2.28
	Magnoliae Cortex	41.18	3.20
	Pinelliae Rhizoma	38.24	2.08
	Atractylodis Rhizoma	32.35	2.59
	Aaurantii Immaturus Fructus	29.41	5.28
	Massa Medicata Fermentata	29.41	5.89
	Amomi Fuctus	23.53	5.33
	Aucklandiae Radix	23.53	2.31
	Paeoniae Radix Alba	23.53	1.29
	Crataegii Fructus	20.59	9.75

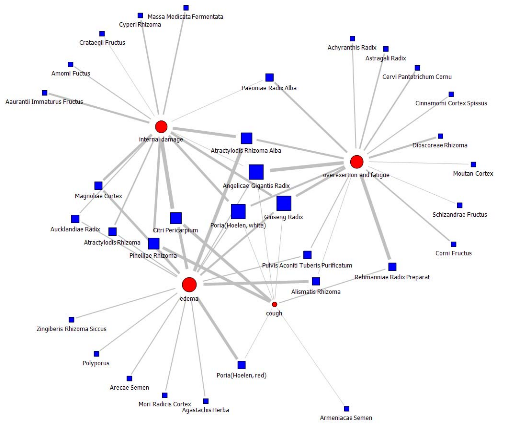

Fig. 1. A network graph to visualize association rules between diseases (internal damage, edema, coughs, and overexertion and fatigue) and their associated herbal materials.

Fig. 1 displays the result of network analysis to illustrate the relationships between diseases and their associated herbal materials. The size of circles (diseases) and squares (herbal materials) represents the frequency of elements with the prescriptions, and the thickness of lines indicates the strength of association rules. For example, Angelicae Ginseng Radix, represented by a large rectangle, has a high frequency as treatment for all four diseases considered here. It has been recognized in Oriental Medicine that *Angelicae Ginseng Radix* is a commonly used and multipurpose herbal material. Through network analysis, we can readily see the list of herbal materials that were used together to treat a certain disease. Overall, network analysis is very helpful in understanding the fundamental principles of prescriptions and the effects of medicines.

V. CONCLUSIONS

This paper aims at extracting useful information from Oriental Medicine databases. We have adopted some text mining approaches such as association rule induction and a couple of graphical expressions to reveal the patterns associated with diseases and the related herbal materials used in Oriental Medicine. Association rule algorithm can help us deduce meaningful rules on associations among item sets. Support, confidence, and lift, calculated from the association rules can be used to assess the strength of these rules. In addition, we used a network analysis to effectively visualize the association rules.

Useful knowledge can be extracted from the text mining method conducted in this study. Association rules between disease patterns and herbal materials can be useful for the development of new medicines in Oriental Medicine because they help identify the important herbs which correspond to certain prescriptions. Moreover, when a clinician diagnoses and treats patients, full use can be made of the information to arrive at an accurate diagnosis and effective treatment.

We hope that the systematic procedure presented in this paper can be expanded to other publications or fields in Oriental Medicine to summarize the documents and to uncover useful knowledge on human health.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea Grant 0003811, Brain Korea 21 (Networked Enterprise).

REFERENCES

[1] P. A. Bath, "Data mining in health and medical information," *Annual Review Information Science Technology*, vol. 38, no. 1, pp. 331-369.

[2] U. Fayyad, G. P. Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag, vol. 17, no. 3, pp.37-54, 1996.

[3] I. H. Witten, E. Frank, and M. A. Hall, "Data Mining, Morgan Kaufmann," Burlington, MA, 2011.

[4] K. Hatonen, M. Klemettinen, H. Mannila, P. Ronkainen, and H. Toivonen, "Knowledge discovery from telecommunication network alarm databases," *In Proc. of the 12th International Conference on Data Engineering ICDE'96 (New Orleans)*, pp. 115–122, 1996.

[5] D. K. Hong, S. H. Yook, M. Y. Kim, Y. J. Park, H. S. Oh, D. H. Nam, and Y. B. Park, "A Structural Analysis of Sanghanron by Network Model- Centered on Symptoms and Herbs of Taeyangbyung Compilation in Sanghanron," Korean Oriental Med, vol. 32. no. 1, pp.56-66, 2011.

[6] H. T. Kim and C. Y. Yoon, "A study on Bang-Yak-Hap-Pyun," *Journal of Oriental Medical Classics*, vol. 5, pp.151-199,1991.

[7] C. Borgelt and R. Kruze. Induction of association rules: Apriori implementation. 5th Conference on Computational Statistics (Compstat 2002, Berlin, Germany), pp.3 95-400, Physica Verlag, Heidelberg, Germany,2002.

[8] R. Agrawal, T. Imielienski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," In *Proc. Conf. on Management of Data*, pp. 207–216. New York: ACM Press,1993.

[9] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB, pp.487–99, 1994.

[10] P. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining, Pearson, Boston, 2006.