A Study on Disease Search Support System using HK Graph

Bo Hao, Tomohiro Yoshikawa, Takeshi Furuhashi, Shin-ichi Sugiura

Abstract—**Recently, public interest in health is growing. In addition, the rapid development of information technology gives us much medical information, that places increased demands on the medical support systems using the information. The aim of this study is to develop a medical support system by using Hierarchical Keyword Graph (HK Graph). This system infers and shows the candidates of the name of disease with graph structure after a user inputs his/her symptoms. Then it shows appropriate hospitals considering the disease, specialty of each hospital, the distance from home, and so on. This paper focuses on the part of the inference of disease from inputted symptoms. It proposes interactive disease search support system using HK Graph which can visualize the relationship among words with network structure based on the co-occurrence information.**

I. Introduction

Recentry, the public interest in health is growing, and TV programs dealing with medical science and medical books for non-special are becoming popular. They are part of a trend in which the medical information is getting disclosed in various ways. In particular, the development of systems which provide medical information through internet is advancing rapidly. The demands on needs for medical support systems will continue to grow.

However, there are still many technical terms in the medical field which are difficult to understand for non-specialists. For example, though the Web pagecite [1] that explains a “facility criterion” enacted by Japanese Health Minister tries to make it easy for a lay audience to understand, many technical terms are used in the page. Importantly, the diagnosis and medical treatment are performed based on the facility criterion, and the fee is also calculated based on it. Therefore, some problems are reported. For example, patients can not get a diagnosis or treatment at the hospital to which they go for attention and must find another one; the payment becomes expensive because the hospital does not have the appropriate facility criterion. To avoid these problems, each patient has to understand the facility criterion first and select a hospital considering his/her disease and the facility criterion of the hospital comprehensively. However, it is not realistic to entrust such decisions to the user. Therefore, when a user gets sick and reports “headache” and “stomachache” vaguely, it is helpful for the user to be given the names of diseases and symptoms which are related to his/her condition in response to inputting the symptoms. Moreover, by showing appropriate hospitals with the appropriate facility criterion, with geographic and other relevant information, the aforementioned problems without demanding users to have medical knowledge might be solved.

Bo Hao, Tomohiro Yoshikawa and Takeshi Furuhashi are with the Department of Computational Science and Engineering, Shin-ichi Sugiura is with the MEXT Innovate Research Center for Preventive Medical Engineering, Nagoya University, Nagoya, Japan (email: haobo@cmplx.cse.nagoya-u.ac.jp {yoshikawa, furuhashi}@cse.nagoya-u.ac.jp ssugiura@med.nagoya-u.ac.jp).

Yahoo Health Care [2] provides a system which infers the name of a disease from symptoms. However, this system performs only a simple matching search based on inputted keywords. When a user inputs a symptom which often appears, the number of the search result will be large and it is difficult for user to select the right disease from them. Moreover, this system can not consider the severity of symptoms and it is difficult to prompt a user to generate symptoms other than inputted.

The aim of this study is to develop a medical support system described above. It employs Hierarchical Keyword Graph (HK Graph) [3], an effective text mining method that can visualize the relation among words in texts with a graph structure based on their co-occurrence for the disese search support system of name of diseases which is a part of the medical support system. The disease search support system can infer the name of disease from the symptoms inputted by user. Applying HK Graph to the inputted symptoms, the system can show symptoms which have high relation to inputted ones with the hierarchical keyword graph structure. One of the features of this system is its interactive support to diagnose the likely disease state. For example, a user can input not only the nature but also the severity of symptoms, such as strong dizziness, a little nausea and so on, or can add more symptoms according to the visualized graph. Then the system can reflect their information to the search result.

Osawa et al. [4] proposed a text mining method called “KeyGraph” which can visualize the relation among words in texts with graph structure. This method enables us to grasp important relations among words and phenomena by extracting words in texts based on the appearance and co-occurrence and by visualizing them in a graph structure. Though KeyGraph can present the relation among words effectively, it is difficult for a user to actively identify certain words that would be most useful for establishing the appropriate relationship. The interactive support is one of the most important features of the proposed system, because the search is carried out through the interaction between inputs such as the nature and the severity (weight) of symptoms from a user and the visualized symptoms within the hierarchical graph structure. Therefore, this study employs HK Graph for the disease search support system in which a user can actively give and get information.

This paper describes the disease search support system (DSSS) which is a part of a broader medical support system, the aim of this study. In this DSSS, symptom words that highly related to the inputted ones are shown with hierarchical graph structure using HK Graph. The search target to extract words is the texts of medical dictionaries or medical Web pages. Therefore, most of the extracted words are not related to name of diseases or symptoms but general words when these sentences in the texts are divided into words. Then this paper tries to extract symptom words automatically from the words in texts on Web pages or dictionaries first. It applies the proposed extraction method using the Web texts of disease-symptom pages in Yahoo Health Care. And then

it shows disease ranking method based on the information of symptoms on HK Graph considering inputted severity of symptoms by user. It studies the effectiveness of the proposed method based on the accuracy and coverage rate of symptom words and the transition of ranking of diseases.

II. HK Graph

HK Graph is a text mining method that extracts words based on the co-occurrence between selected attributes by a user and words in sentences. It then generates a hierarchical keyword graph. In HK Graph, a user can analyze and search the information of texts actively and interactively. The algorithm of HK Graph is described as follows.

A. Division into Words

First, sentences in the target text data are divided into words by applying Cabocha [5] which is Japanese language morphological and paragraphic analysis tool. Particles, symbols (punctuation, parentheses), pronouns, conjunctions and adnominal words given by morphological information are regarded as noise words which are not needed for disease search, and they are deleted.

B. Selection of Base

As the next step, a user selects some keywords which represent his/her symptoms from the extracted words in II.A. Though this system will enable a user to input the symptoms freely, he/she "selects" them in this paper, because more study is needed to understand the implication of vagueness and synonymy of description. In this system, the inputted (selected) symptom words are used as attributes, and the texts written about symptoms for diseases are used as the target texts. As the selected words here are the bases to start analysis, these words are called "Base". In the next step, the co-occurrence between Base and words in texts are calculated and high co-occurrence words are extracted. Then if other bases are selected, extracted words and the keyword graph with them are also different. That is one of the most different points from KeyGraph which generates keyword graph structure automatically with the calculation of co-occurrence among words.

C. Extraction of Main-node

In the next step, words are extracted which have high co-occurrence with Base. $Jaccard$ coefficient is used as the co-occurrence value. The expression of co-occurrence is shown below.

$$Jaccord(B_i, W_j) = \frac{N(B_i \cap W_j)}{N(B_i \cup W_j)} \quad (1)$$

B_i is Base, W_j is each divided word in II.A and $N(X)$ is the number of sentences including the word X. Using expression (1), the words with high co-occurrence to all Base (All connected Main-node), those to plural Base (Multi connected Main-node) and those to single Base (Single connected Main-node) are extracted. The extracted words are called "Main-node".

D. Extraction of Sub-node

When a user wants to know more about a certain Main-node, HK Graph can extract the words which have high relation to the selected Main-node. The words "Sub-node" are extracted based on the calculation replacing Base with the selected Main-node in the expression (1). Sub-node is shown when the user clicks a Main-node. Each Sub-node is also connected with another highly related Main-node.

E. Presentation of Hierarchical Keyword Graph

The image of output of HK Graph is shown in Fig.1. In this figure, B_1〜B_3 are Base, A_1 is All connected Main-node, M_{12}〜M_{23} are Multi connected Main-node, S_{11}〜S_{32} are Single connected Main-node and Sub_1, Sub_2, S_{32} are Sub-node of S_{32}. Base and Main-node are connected with their links and the value of co-occurrence is expressed in the thickness of each link.

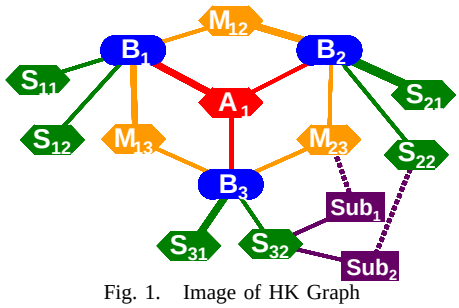

Fig. 1. Image of HK Graph

III. Proposed Method

A. Extraction of Symptom Words

Using medical dictionaries or medical Web pages as the target texts, the ratio of the name of diseases and symptoms to the extracted words increases. However, there are still many words with no relation to symptoms. HK Graph extracts the words based on the co-occurrence, then the words with no relation are also extracted as long as their co-occurrence is high. Fig.2 shows the result of HK Graph to medical Web texts with "fever" and "snivel" as Base. In this figure, the words such as "in a few months", "a few days", "falling" and "case" are not appropriate for supporting a search or prompt of other symptoms. Therefore, this paper proposes the extraction method to extract symptom words which are symptoms themselves or directly related to them using Web information from the words obtained in II.A.

First, as described in II.A, noise words are deleted after applying Cabocha to divide sentences. Then the retrieval frequency of single search result and AND search result are counted between each obtained word with "symptom" using Yahoo API [6]. $Dice$ coefficient is used as the co-occurrence here instead of $Jaccord$. The search "symptom (word)" and $Dice$ coefficient can extract the words often used as symptoms. The expression of $Dice$ coefficient is shown below.

$$Dice(Symp, W_i) = \frac{N(Symp \cap W_i)}{N(W_i)} \quad (2)$$

$Symp$ is the word "symptom", W_i is each word obtained in II.A, $N(X)$ is the retrieval frequency of the word X . In the proposed method, a threshold is employed for the co-occurrence above and the words with higher co-occurrence than the threshold are regard as symptom words.

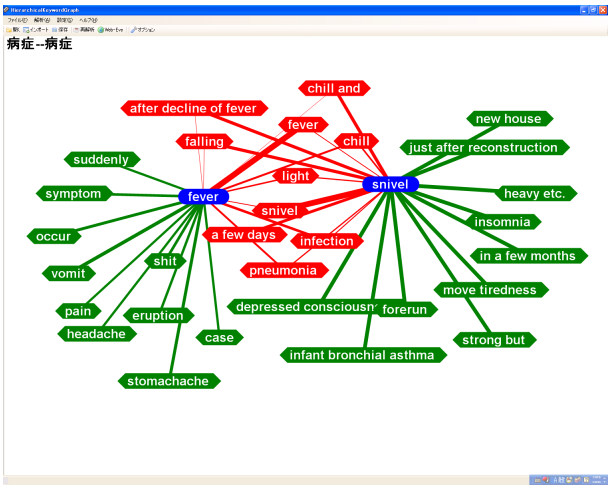

Fig. 2. Generated Symptom Graph (Base: “fever”, “snivel”)

B. Disease Ranking Method

DSSS infers the candidates of diseases using symptoms on HK Graph and severity for them. In this paper, the description of symptoms for a disease in Yahoo Health Care, in which the number of diseases is 502, is used to calculate the relationship between a candidate disease and symptoms on symptom graph (HK Graph). Here, the description of symptoms for diseases is called symptom sentence. First, the diseases which include at least one inputted symptom in symptom sentence are extracted. Then, the co-occurrence between words in symptom sentence of a disease and the symptoms on symptom graph is calculated as the relationship between them. A user gives the severity as 10 levels to the input symptoms and newly prompted ones by related symptom words on symptom graph. A disease which has relationship with high severity symptoms given by the user will have more possibility to be the actual disease. Then this system uses the severity information for the above calculation of relationships and ranks the diseases based on them. Here, the severity of a symptom is defined as the weight of the node, and the relationship R for each disease is calculated by the following function.

$$R(D_i) = \sum_{j}^{n_n} \sum_{k}^{n_{si}} \frac{Weight(N_j) \times Jaccord(N_j, W_{ik})}{n_{si}} \quad (3)$$

D_i is i-th disease, n_n is the number of nodes on HK Graph, n_{si} is the number of words in the symptom sentence of i-th disease, $Weight(N_j)$ is the weight for node N_j, N_j is j-th node on HK Graph, W_{ik} is k-th word in symptom sentence of i-th disease.

IV. Experiment

A. Extraction of Symptom Words

This paper applies the proposed method to the text data of Yahoo Health Care, and it investigates the performance for the extraction of symptom words. The text data were 502 diseases and the description of symptoms for them, which were provided by Yahoo Health Care on the internet. First, Cabocha was applied to these text data, and 37,200 words were divided. Then 11,018 words were extracted after the noise words were deleted described in II.A. In this paper, the words which express symptoms or related with symptoms were selected by hand from these 11,018 words to investigate the proposed method quantitatively. Then 4,670 words were defined as symptom words. Defined symptom words are shown below.

- symptom (fever, vomit, etc.)
- a part of body (arm, skin, etc.)
- attribute (lady, child, etc.)
- name of disease (melancholy, hey fever, etc.)

B. Accuracy and Coverage Rate

Words were extracted by changing the threshold of co-occurrence defined in expression (2). The image of all words (11,018), the symptom words (4,670) and extracted words is shown in Fig.3.

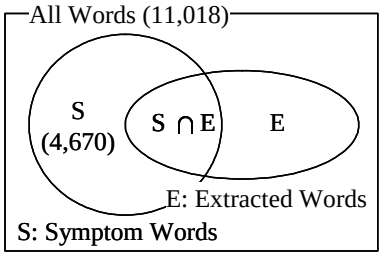

Fig. 3. Image of Extracted Words

To investigate how many symptom words defined in IV.A were included in the extracted words by the proposed method, the accuracy is defined as the expression (3).

$$Accuracy(\%) = \frac{N(S \cap E)}{N(E)} \times 100 \quad (4)$$

In this expression, S is the symptom words, E is the extracted words and $N(X)$ is the number of word X. And the coverage rate of the symptom words is defined below.

$$Caverage(\%) = \frac{N(S \cap E)}{N(S)} \times 100 \quad (5)$$

Fig.4 shows the accuracy and coverage rate with changing the threshold of co-occurrence. Though it does not show in the graph, the basic accuracy and the coverage rate without the threshold were 42.4% and 100%, respectively. In this figure, we can see the accuracy increases when the threshold becomes higher. This relation shows that the co-occurrence defined in expression (2) was appropriate to extract symptom words. The maximum accuracy was about 80% while the coverage rate declined to about 10%-20%. It is necessary to decide the appropriate threshold that balances both the accuracy and the coverage rate. Higher accuracy and coverage rate mean larger $S \cap E$ part and smaller the other part of S and E in Fig.3. Then this paper employs mutual information to decide appropriate threshold. The threshold with the maximal mutual information is taken. The expression of mutual information is shown below.

$$I(S;E) = H(S) - H(S|E) \quad (6)$$

In this expression, $H(S)$ is the entropy of S (Symptom Words), and $H(S|E)$ is the entropy of the conditional

probability of S, given E (Extracted Words). Each entropy is given as the expression (6) and (7), respectively.

$$H(S) = -\sum_{i=1}^{n} p(s_i)\log_2 p(s_i) \quad (7)$$

s_i is the set of symptom words or that of the others, then n becomes 2. $p(s_i)$ is the probability of s_i ($p(s_1) = \frac{N(S)}{N(AllWords)}$, $p(s_2) = \frac{N(\bar{S})}{N(AllWords)}$).

$$H(S|E) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(s_i, e_j)\log_2 p(s_i|e_j) \quad (8)$$

e_j is the set of extracted words or thaf the others, then m also becomes 2. $p(s_i|e_j)$ is the conditional probability of s_i, given e_j ($p(s_1|e_1) = \frac{N(S\cap E)}{N(E)}$, $p(s_2|e_2) = \frac{N(\bar{S}\cap\bar{E})}{N(\bar{E})}$).
In this experiment, the threshold with the maximal mutual information was 0.2 in which the accuracy was 72.3% and the coverage rate was 70.7% (Fig.4). The required accuracy can be 60%-80% to support a search or prompt of other symptoms. Therefore, the performance in this experiment is good, and symptom words were well extracted.

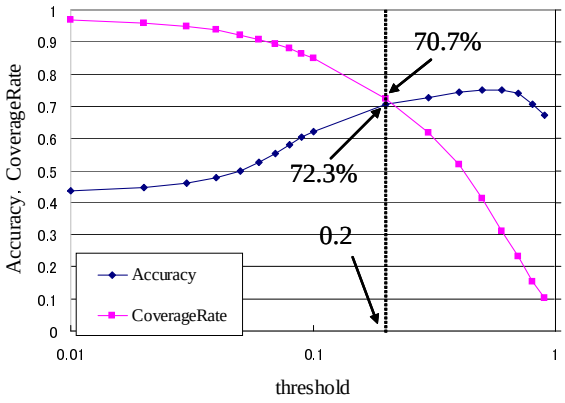

Fig. 4. Accuracy and Coverage Rate

C. Generated Symptom Graph

Symptom graph was generated using the extracted words by the proposed method. The result is shown in Fig.5. In this example, the threshold was 0.2 in which the mutual information was largest. Compared with Fig.2, the unrelated words to symptoms are fewer, and appropriate keyword graph of symptom words is acquired.

D. Transition of Disease Ranking

The ranking of diseases is determined based on the relationship R shown in (3) using the symptoms on the graph and the given severity for them. Here, the effect of changing weight for symptoms was investigated. The symptoms "fever" and "snivel" were inputted, and symptom graph shown in Fig.5 was generated. Fig.6 shows the transition of disease ranking when the weight for the symptom "eruption" in Fig.5 was given from the default value to 10 times of it. The ranks of the diseases which include "eruption" in the symptom sentences become higher in proportion to the weight, and it shows changing weight is reflected to the disease ranking. It is expected to support a user to find his/her possible disease quickly.

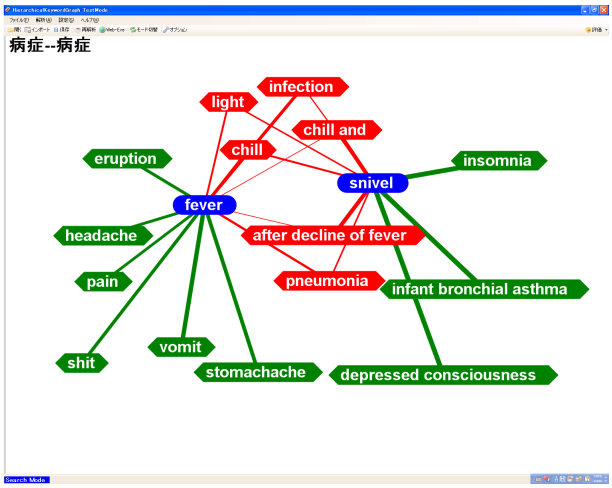

Fig. 5. Generated Symptom Graph after the Extraction of Symptom Words

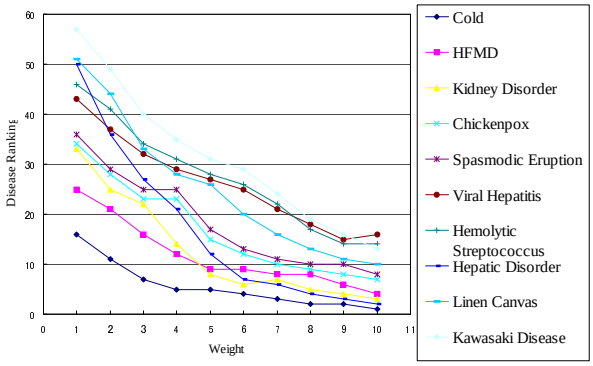

Fig. 6. Transition of Disease Ranking

V. Conclusions

This paper described the disease search support system which was a part of a broader medical support system. It proposed the extraction method of symptom words from the words in medical Web texts or dictionaries. It also proposed the disease ranking method based on the symptoms and their severity. This paper applied the proposed method to the Web texts of disease-symptom in Yahoo Health Care and studied the effectiveness of the proposed method based on the accuracy, coverage rate of symptom words and the transition of disease ranking. Moreover, an example of a hierarchical keyword graph using the proposed method was shown. For the future work, we will develop the total medical support system including the disease search support system.

References

[1] WAM NWT : http://www.wam.go.jp/iryo/
[2] Yahoo Health Care: http://health.yahoo.co.jp/index.html
[3] Takahiro Okabe, Tomohiro Yoshikawa, Takeshi Furuhashi, "Proposal of Multi-Connected Hierarchical Text Mining Method for Medical Incident Reports" (in Japanese), The 0th Fuzzy System Symposium, pp.211-214, 2006
[4] Yukio Ohsawa, Benson Nels E, Masahiko Yachida, "KeyGraph: Automatic Indexing by Segmenting and Unifing Co-occurrence Graphs" (in Japanese), The transactions of the Institute of Electronics, Vol.J82-D-I, No.2(19990225), pp.391-400, 1999
[5] Taku Kudo, Yuji Matsumoto, "Japanese Dependency Analysis Using Cascaded Chunking" (in Japanese), The journal of Information Processing Society of Japan, Vol.43, No.6, pp.1834-1842, 2002
[6] Yahoo Web API: http://developer.yahoo.co.jp/