A study of visualization method with HK Graph for free text using grouping words based on their concept

Daisuke Kobayashi*, Tomohiro Yoshikawa†, Takeshi Furuhashi‡ and Eiji Hirao§
*Nagoya University
Email: kobayashi@cmplx.cse.nagoya-u.ac.jp
†Nagoya University
Email: yoshikawa@cmplx.cse.nagoya-u.ac.jp
‡Nagoya University
Email: furuhashi@cmplx.cse.nagoya-u.ac.jp
§NEC Corporation

***Abstract*—Many companies carry out questionnaires. These questionnaires often have questions which need respondents to answer by free description. It is, however, inefficient for an analyzer to read whole data for getting outlines or classifying them. The authors have proposed the HK Graph (Hierarchical Keyword Graph) which is a support tool for text mining. HK Graph can visualize the relationships among attributes and words with hierarchical graph structure based on frequency of co-occurrence. However, the result of HK Graph is not helpful enough for the analyzer to grasp the outlines of the texts and extract opinions from them. This paper presents a new visualization method for the HK Graph incorporating a grouping method based on concepts of words. An experiment is carried out by applying the proposed method to actual questionnaire data on disasters and studies the effectiveness of the proposed method.**

I. Introduction

Recently, companies often carry out questionnaires for planning marketing strategies. There are a lot of objectives of questionnaires, e.g., finding needs of products or services, predicting demand or market scale, investigating images of brands or levels of customer satisfaction, and so on. Thus, questionnaires are important information sources for companies, and effective analysis methods for them are strongly needed.

There are two types of formats in questionnaire. One is to answer from some choices for a question or directly answer a question by marking a numerical value, the other is to answer by free description. In the former case, it is easy to quantify the responses, and to apply multivariate analysis methods. However, it is difficult to extract real opinions or impressions of respondents unless good setting of questions is done, because the respondents can respond only to the prepared questions. On the other hand, in the latter case, it is expected to extract real opinions of respondents because they can write their responses freely. However the analysis of free text data is difficult because numerical analysis methods can not be applied and it takes long time to aggregate and analyze them. In many analyses of questionnaires, the analysts have to read whole text data for getting outlines or manually classifying them. Therefore, the demand for the support system by text mining methods for the analysis of free texts in questionnaire has been growing.

A lot of text mining methods to support the analysis of texts have been reported [1], [2], [3]. This study focuses on the approach by visualization. Analysis of free texts in questionnaires needs interactivity because a lot of grammar or syntax errors are contained in responses, and the visualization is expected to be effective for interactive analyses. There are some conventional systems to support the analysis of texts with visualization such as Key Graph [4], ACCENT [5], $KOTOBA$ network [6], and so on. Key Graph can show the relationships among words in texts based on their frequencies of co-occurrence; however, Key Graph is not designed for interactive word search. The displayed result is fixed. ACCENT and $KOTOBA$ network shows relationships of words in a form of graph structure. However, the interactivity of these methods is low because the graph has to be restructured every time a user needs to analyze further into the details.

The authors have developed Hierarchical Keyword Graph (HK Graph) [7], [8] as the support system for the analysis of texts with visualization of relationships among words. HK Graph can extract relevant words based on their co-occurrence for the words input by a user or for the groups, e.g., men and women, and visualize them as a graph structure. In addition, HK Graph can show the words as a hierarchical structure, and then a user can analyze the focused word in depth by seeing the lower layer of the word.

In HK Graph, divided words are regarded as different words except for the perfect matching ones. Therefore, synonyms such as different representation and different words are regarded as different words even if they were used as the same meaning. Then these words are shown separately or often not presented because the frequency of appearance becomes low. In free texts of questionnaires, the representation of words

(Hiragana/Kanji etc.) and using words are often different in individuals. Consequently, aggregating these synonyms has been an important issue.

This paper considers the similarities of words in HK Graph based on their "concept" of words which are defined by the thesaurus, and proposes a method to feed back them into the visualization result with HK Graph by grouping the words based on their concepts. The proposed method can aggregate the words which were separately shown in the conventional HK Graph, and it is expected that the visualization of aggregated words makes it easier to grasp the outlines of the texts and extract proper features of them. This paper applies the proposed method to actual questionnaire data on disasters and studies the effectiveness of the proposed method.

II. HK Graph

In HK Graph, users can interactively analyze text data by starting to analyze the text related to the items they are interested in, and proceeding into deeper layers of words of interest shown as a keyword graph. The algorithm of HK Graph is as follows.

A. Division into Words

The first step in the algorithm of HK Graph is to divide the target texts into words by applying Cabocha[5]. Cabocha is a Japanese language morphological and paragraphic analysis tool. Unlike English, which separates words with spaces, it is difficult to divide Japanese text without tools like Cabocha. Applying Cabocha to the target texts, particles, symbols (punctuation, parentheses), pronouns, conjunctions and adnouns given by morphological information are regarded as noise words which are not needed for analysis, and they are deleted.

B. Selection of Base

The second step is for users to select contents to be analyzed, e.g., sex, income, age, the texts containing a certain word or sentence, the target group, and so on. If they want to grasp the outlines or the tendency of texts, they can select all texts. The selected group is called "Base." It is one of the features of HK Graph to let users give the start nodes of their interest which create the graph structure.

C. Extraction of Main-node

High co-occurrence words in the texts with the selected Base are extracted. The extracted words are called "Main-nodes." $Jaccard's$ coefficient is used as the co-occurrence value. The equation of co-occurrence is shown below.

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

B_i is the Base i, W_j is each word divided out in II.A., $S(X)$ is the texts including the Base/word X, and $N(S)$ is the number of texts of S. Using eq.(1), the words with high co-occurrence to all Bases (All connected Main-node), those to plural Bases (Multi connected Main-node), and those to single Base (Single connected Main-node) are extracted.

D. Extraction of Sub-node

When a user wants to know more about a certain Main-node, HK Graph can extract the words which are closely related to the selected Main-node as a lower layer. The words, called "Sub-nodes," are extracted using eq.(1) where the Base is replaced with the selected Main-node. Thus in HK Graph, the words which have high co-occurrence with the focused words can be expanded as the lower layer, then a user can analyze the texts by proceeding into deeper layers of words of his/her interest.

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 Bases, M_{a1} is All connected Main-node, M_{m121}-M_{m231} are Multi connected Main-nodes, M_{s11}-M_{s32} are Single connected Main-nodes, and S_{211}, S_{222}, etc. are Sub-nodes. A Base and Main-nodes, a Main-node and Sub-nodes are connected with their links, and the value of co-occurrence is expressed in the thickness of each link.

In this way, the relationships of the words are hierarchically presented in HK Graph. In addition, HK Graph can show the new graph starting from the presented words, either Main-nodes or Sub-nodes, as new Bases.

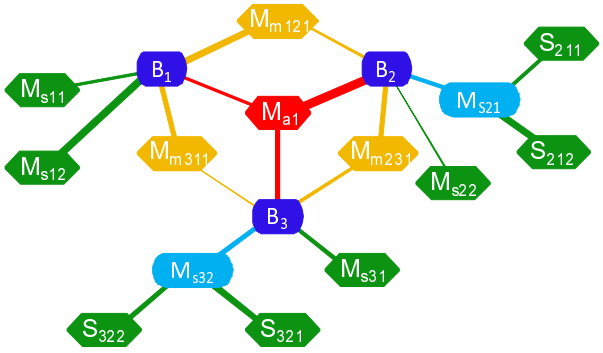

Fig. 1. Image of HK Graph

III. Proposed Method

The objective of this study is to support the analysis of free texts in questionnaires free texts using HK Graph. However, synonyms are regarded as different words in HK Graph as described in I. This paper considers the similarities of words in HK Graph based on their "concept" of words which are defined by the thesaurus, and proposes a method to feed back them into the visualization result with HK Graph by grouping the words based on their concepts. The details of the proposed method are as follows.

A. Definition of Concept

This paper uses "Nihongo Dai-Thesaurus [9]" to define "concept." This thesaurus consists of 6 chapters. Each chapter consists of 67 sections, and each section has 1044 categories. Each category has 1-30 groups of belonging words. The small

word group is the minimum unit of synonyms to be grouped. Consequently in this paper, the sections in this thesaurus are defined as "large concepts," the categories are "middle concepts," and the small word groups are "small concepts," respectively. Based on these concepts, the words in texts are grouped.

B. Generation of Concept Group

After the division of texts into words, it gives the label(s) of large concept, middle concept, and small concept to each word. Concretely, each word is searched in belonging words to small word groups in the thesaurus, and the label of corresponding large concept, middle concept, and small concept are given to the word when the word is found. Plural concepts can be given to one word because a word often has some meanings and appears in some small word groups in the thesaurus. Large concept groups, middle concept groups, small concept groups, in which same belonging words are, are generated based on the concept labels of words. Then, in HK Graph shown in II., the nodes are replaced to these concept groups. Each concept group is shown with the corresponding label in the thesaurus.

C. Application to HK Graph

As described II.C., the high co-occurrence words in texts with the Bases are extracted as Main-nodes in HK Graph. Instead of the words, the relative concept groups (the label is the word used in the thesaurus) to the Bases are extracted and connected as Main-nodes. The algorithm of HK Graph with the concept groups is shown as follows.

1) Extraction of Concept Group: The concept groups which have high co-occurrences with the selected Base are extracted. A user can select the size of the concept groups (large, middle, small concept groups) and the word class to be extracted. In a concept group, these are nouns, verbs, adjectives, and so on. For example, when noun is selected, only nouns in each concept group are extracted and shown. The co-occurrence between Base B_i and concept group G_i is calculated by the $Jaccard's$ coefficient (eq.(2)) similar to eq.(1). G_j in eq.(2), which is the difference from eq.(1), is each word in the texts which belongs to the selected size of the concept group j by the user.

$$Jaccard(B_i, G_j) = \frac{N(S(B_i) \cap S(G_j))}{N(S(B_i) \cup S(G_j))} \quad (2)$$

2) Extraction of Lower Layer Concept Group: Smaller concept groups are extracted as lower layer nodes from the concept groups in III.C.1). Concretely, middle concept groups are extracted as the lower layer of large concept groups, small concept groups are extracted as the lower layer of them. In addition, from small concept groups, the words in the texts which belong to the concept word in the thesaurus are extracted. Furthermore, the lower layers can be hierarchically extracted in the same way with II. In this way, a user can analyze the texts from the outlines to the detail by the hierarchically visualized concepts and words.

3) Extraction of Co-occurring Words: The proposed method can also extract the words with high co-occurrences to a certain concept group. The co-occurrence between a word and a concept group is calculated by that between the word and the belonging words to the concept group. The equation of co-occurrence is shown below.

$$Jaccard(G_i, W_j) = \frac{N(S(G_i) \cap S(W_j))}{N(S(G_i) \cup S(W_j))} \quad (3)$$

IV. Experiment

A. Questionnaire Data and Division into Words

In this experiment, the proposed method was applied to 500 free texts of a questionnaire on disasters. The question of the questionnaire was "Please write your opinions or thoughts on natural disasters over 50 characters (in Japanese)." Actually, including "I don't have comments in particular" and so on, 125 respondents out of 500 answered under 50 characters. Firstly, the text data were divided into words as described in II.A., and the words and concepts not to be needed for the analysis ('thing,' 'is,' 'there,' 'that,' and a concept "point" coming from a word 'thing') were excluded by hand. The following figures and descriptions of the results are originally written in Japanese and translated into English.

B. Study of Grouping Words

First, the effect of grouping words based on concepts was studied. The proposed method, HK Graph with concept, was applied to the above questionnaire data, and the co-occurring words described in III.C.3) to "flood" which was one of the small concept groups were shown in Fig.2. In Fig.2, noun was selected as the presenting word class. A lot of related words such as 'typhoon,' 'heavy rain,' 'river,' and so on to the small concept group "flood" can be seen in Fig.2. Each of them was not co-occurred with the word 'flood' itself but co-occurred with the words belonging to "flood" in Nihongo Dai-Thesaurus [9]. This small concept group "flood" consisted of 7 words in the thesaurus, which were 'flood,' 'water damage, ' 'overflow, ' 'flash flood, ' 'freshet, ' 'inundation above floor level, ' and 'immersion ' (originally in Japanese). In Fig.2, all connected words to "flood" node were extracted over two times, i.e., they were used more than two times with other seven words above in the texts. On the other hand, conventional HK Graph was applied to the same texts and the result that the words co-occurring with 'flood' in noun was shown in Fig.3. The number of the sentences having the connected words in Fig.3 with 'flood' was one each, except for 'thunder.' The numbers of the sentences belonging to each "flood" node were 29 in Fig.2 and 10 in Fig.3, respectively. Therefore, about three times sentences were linked to the concept group.

In Fig.3, the number of sentences with each node extracted from the word 'flood' was one except that of 'thunder' was two. That is to say, the analysis on the co-occurring words with the word 'flood' does not give the outlines of the texts but a part of them. On the other hand, the words whose co-occurrences with "flood" as a concept group were high because of the overlaps of co-occurring one another while those with

'flood' as a word were low were extracted such as 'heavy rain' (with 8 sentences), 'stream' (with 6 sentences), 'river' (with 3 sentences), and so on in Fig.2. By grouping words into the concepts, the words related to "flood" could be grasped. In fact, when the result of Fig.2 is compared to Fig.3, the easiness to grasp the tendency of sentences is obviously different.

In addition, 'earthquake' strongly co-occurs with "flood" in Fig.2. In the texts, 'flood' and 'earthquake' were often used together in the same sentences. However, when co-occurrence between them is calculated, it becomes low on eq.(1) because 'earthquake' was used overwhelmingly in all texts. Grouping words based on concepts overcame one of the weakness of the $Jaccard's$ coefficient, the words with low appearance frequency tend to have high co-occurrence and vice versa, and appropriate co-occurrences were acquired. Note that in the case of using a function for the co-occurrence which is valued on the frequency instead of the $Jaccard's$ coefficient, only high frequency words which are sometimes not important for analysis are extracted.

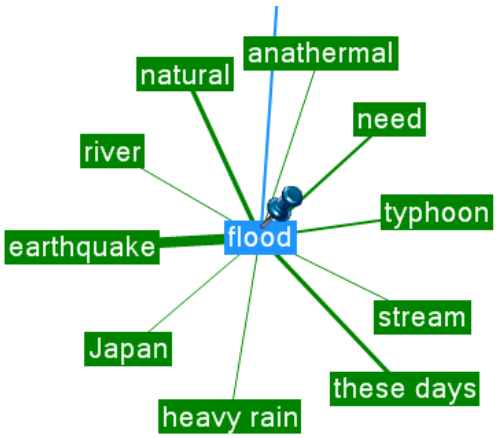

Fig. 2. Co-occurring words to concept group "flood"

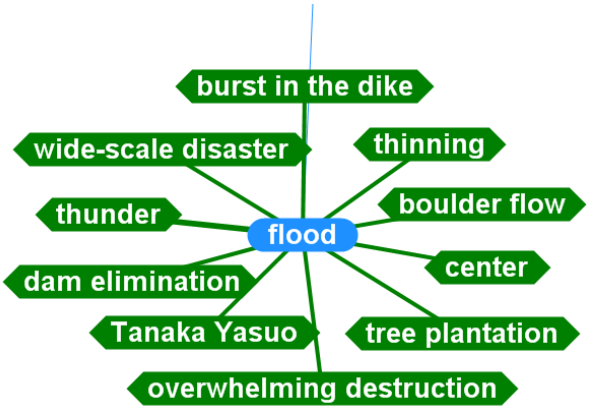

Fig. 3. Co-occurring words with word 'flood'

C. Study of Grasping Outline of Text

Next, the easiness to grasp the outlines of texts was studied comparing with the conventional HK Graph. Fig.4 shows the result of the conventional HK Graph when the extracted word class was noun and the number of presented Main-node was 50, and Fig.5 shows that of the proposed method when the extracted word class was noun and the number of Main-node was 10 small concept groups in the way described in III.C.2), respectively. In Fig.4, it is difficult to grasp the outlines of texts, because synonyms are not grouped and they are discretely positioned. On the other hand, Fig.5 shows the hierarchical structure of the form "small concepts → words" with semasiological gatherings, which enables us to grasp the responses of the questionnaire , e.g., the respondents wholly wrote about "earthquake", "environment," and so on in this questionnaire.

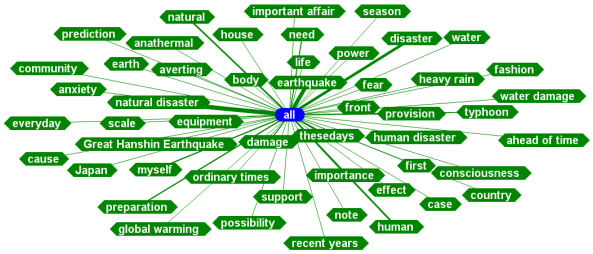

Fig. 4. Conventional HK Graph

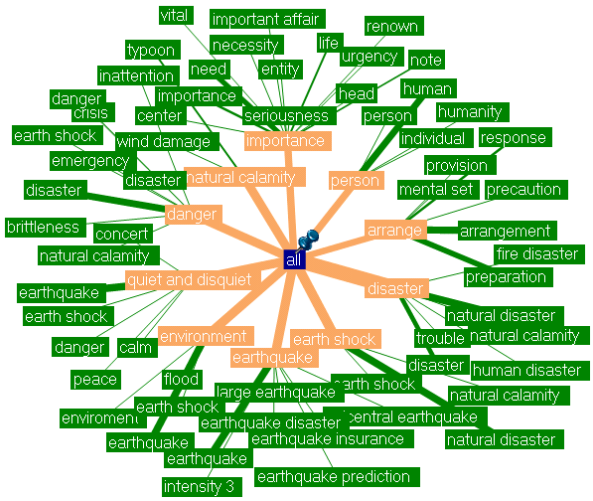

Fig. 5. Proposed method (Main-node : small concept group)

Fig.6 shows the result based on the expansion way shown in III.C.2). As the questionnaire was on the natural disaster, "meteorological phenomenon." was selected out of the 10 large concept groups, and "convulsion of nature" was selected as the middle concept groups extracted from "meteorological phenomenon". The small concept groups extracted from

"convulsion of nature" and the words belonging to each small concept group were presented. Some words such as "earth shock", "flood", "earthquake", and so on are included in two levels in Fig.6, because these words are also plurally represented in Nihongo Dai-Thesaurus [9]. By proceeding into deeper layers grouping the hierarchical structures of concepts and words interactively, a user can analyze the texts from abstract concepts to concrete words while he/she is classifying the texts by the concepts.

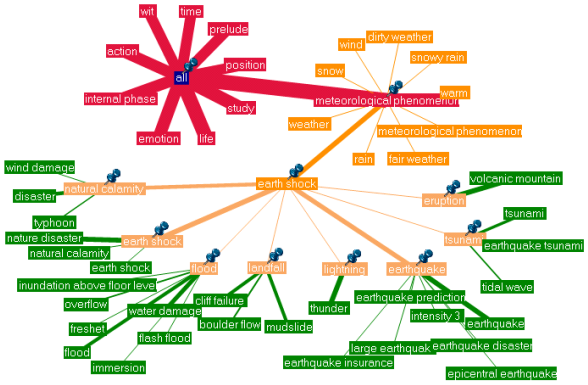

Fig. 6. Proposed method (large concept → middle concept → small concept)

V. Conclusion

This paper considered the similarities of words in HK Graph based on their "concept" of words which were defined by the thesaurus, and proposed a method to feed back them into the visualization result with HK Graph by grouping the words based on their concepts. This paper applied the proposed method to actual questionnaire data on disasters and studied the effectiveness of grouping words by the concepts and easiness to grasp the outlines of texts comparing with the conventional HK Graph. Future work is the study on the extraction of the proper concept for the word belonging to some concepts considering the meaning in the texts.

Acknowledgment

This work was supported by the Grant-in-Aid for Scientific Research (C) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan, Grant number: 22500088.

References

[1] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1999, pp. 16–22.

[2] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, and O. Zamir, "Text mining at the term level," *Principles of Data Mining and Knowledge Discovery*, pp. 65–73, 1998.

[3] Y. Ichimura, T. Hasegawa, I. Watanabe, and M. Sato, "Text mining : Case studies(special issue :," *Journal of Japanese Society for Artificial Intelligence*, vol. 16, no. 2, pp. 192–200, 2001-03-01. [Online]. Available: http://ci.nii.ac.jp/naid/110002808468/en/

[4] Y. Ohsawa, N. Benson, and M. Yachida, "KeyGraph: Automatic indexing by co-occurrence graph based on building construction metaphor," in *Proceedings of the Advances in Digital Libraries Conference*. IEEE Computer Society, 1998, p. 12.

[5] K. Misue and I. Watanabe, "Visualization of keyword association for text mining," *IPSJ SIG Notes*, vol. 99, no. 57, pp. 65–72, 1999.

[6] M. Masaki, O. Shinobu, T. Yashuhiro, S. Masanari, I. Makoto, and F. Miki, *Jireidemanabu text mining (in Japanese)*. Kyoritsu Shuppan, Tokyo, 2008.

[7] T. Okabe, T. Yoshikawa, and T. Furuhashi, "A proposal of analysis system for medical incident reports using metadata and co-occurrence information," *Journal of Japan Society for Fuzzy Theory and Intelligent Informatics*, vol. 18, no. 5, pp. 689–700, 2006.

[8] T. Yoshikawa, Y. Uchida, T. Furuhashi, E. Hirao, and H. Iguchi, "Extraction of evaluation keywords for analyzing product evaluation in user-reviews using hierarchical keyword graph," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 4, pp. 457–462, 2009.

[9] T. Yamaguchi, *Nihongo Dai-Thesaurus CD-ROM*. Taishukan Shoten, Tokyo, 2006.