Extraction of Important Keywords in Free Text of Questionnaire Data and Visualization of relationship among sentences

Yuki Uchida
Nagoya University
uchida@cmplx.nagoya-u.ac.jp

Tomohiro Yoshikawa
Nagoya University

Takeshi Furuhashi
Nagoya University

Eiji Hirao
NEC corporation

Hiroto Iguchi
NEC corporation

***Abstract*—Recently, companies often carry out questionnaire(s) and develop marketing strategies. There are usually two types of forms for the answer of a questionnaire. One is the form to select prepared answers and the other is free text form. The true message might be in the text form rather than the numerical part, then the analysis of free text form is needed. The amount of text in a questionnaire is, however, usually large and difficult to read whole text data for analysis. This study tries to develop a free text analysis support system which visualizes relationships among respondents based on their texts and shows their opinions using graph structure of keywords. First, this paper proposes the extraction method of important keywords in their opinions based on the modification relationships. Next, it clusters the respondents interactively on visible space using MDS. Finally, it shows their opinions using HK Graph which can visualize the relationship among words with hierarchical network structure based on the co-occurrence information for the keyword graph.**

I. Introduction

Recently, companies often carry out a questionnaire(s) and develop marketing strategies such as the discovery of buying groups for the target products, the prediction of market size, the services for customer satisfaction and the design of products. Therefore, the analysis of questionnaire data is important for them. There are usually two types of forms for the answer of a questionnaire. One is the form to select prepared answers, which can be converted into numerical value, and the other is the free text form, in which the idea of a respondent is described freely. In the select form, respondents are asked to answer to each question by grading their impression about the evaluation objects from multiple grade scales, and it can quantify people's impressions about evaluation subjects. The true idea might be in the text form rather than the numerical part, because respondents can describe their ideas freely while they can only select the grade in the select form. Then the analysis of free text form is needed for companies. It is, however, difficult to analyze the free text because it takes much time to grasp whole text data and analyze the contents.

There are some conventional text mining methods for the classification of text, in which the method based on frequencies of words in text[1] is one of the most popular one. It is important for the classification of text to consider sentence structures. Some classification methods based on the similarities of nouns and predicates among sentences[2] and that of expressions in the end of sentences[3] have been reported. These are, however, not appropriate to be extracted as the keywords from free text in questionnaires, because respondents are not always write them in correct grammar with nominative/predicative. It is also difficult to decide the number of clusters in these classification methods. In addition, analyzers have to read all contents in each cluster after the classification. It is also needed to support this analysis part.

This paper proposes the extraction method of important keywords from free text to consider sentence structures. In the proposed method, the keywords are extracted using the modification relation of words and retrieved frequency on the Internet. It also proposes an interactive clustering method of the respondents on visible space which is the result of Multi Dimensional Scaling (MDS)[4] based on the similarity of extracted keywords. It enable a user to grasp relationships among contents based on their keywords and features/tendencies of the respondents' opinion by the interactive clustering on visible space.

HK Graph (Hierarchical Keyword Graph)[5][6] is a powerful text mining method that can visualize the relationships among words in text using a hierarchical graph structure. The relationship is based on their co-occurrence, i:e: how often the words appear together in the text. This paper applies HK Graph to the texts in the clusters to support the detail analysis of their contents without actually reading all of them.

This paper applies the proposed method to a web questionnaire data which is about natural disasters. It extracts the keywords for the consideration of disasters from them, and then it applies MDS based on the extracted keywords to cluster the respondents on visible space. In addition, it visualizes the relationships among keywords with hierarchical network structure in each cluster using HK Graph.

II. HK Graph

HK Graph extracts the words which have high co-occurrence with the words selected by a user from the target sentences and shows them as a hierarchical keyword graph. Then we can grasp the essence of the text. The features of HK Graph are the hierarchical structure and interactive search, with which users can start to analyze the text related to the items they are interested in, and they can proceed into deeper

978-1-4244-3597-5/09/$25.00 ©2009 IEEE

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[7]. 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. After applying Cabocha to the target texts, particles, symbols (punctuation, parentheses), pronouns, conjunctions and adnouns 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 some keywords of interest out of the words resulting from the process described in Sec. Ⅱ.A. As the words selected here are the bases of the analysis, they are called a "Base." In the next step, the co-occurrence between each Base and other words in the texts are calculated, and high co-occurrence words are extracted. Therefore, if other bases are selected, extracted words and the keyword graph created from them are also different. When consumers' reviews are analyzed, Bases will be the names of products.

C. Extraction of Main-node

In the next step, the words which have high co-occurrence with 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.

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

B_i is the Base, W_j is each word divided out in Sec. Ⅱ.A, and $N_s(X)$ is the number of texts including the word X. Using eq.(1), All connected Main-nodes which have high co-occurrence to all Bases, Multi connected Main-nodes to plural Bases and Single connected Main-nodes to single Base are extracted as Main-nodes.

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. The words, called "Sub-nodes," are extracted using eq.(1) where the Base is replaced with the selected Main-node. A 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

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

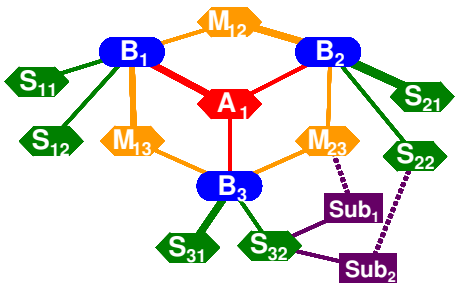

Fig. 1. Image of HK Graph

III. Proposed Method

A. Extraction of Evaluation Keywords

This paper defines evaluation keywords, which are regarded as the evaluation words for the target object(s) in a questionnaire, as Evaluated words and Evaluating words and extracts them. HK Graph shows each group (cluster) as Base, Evaluated words as Main-nodes and Evaluating words as Sub-nodes. The following subsections describe the extraction method of these evaluation keywords.

1) Extraction of Evaluated Words: Evaluated words are defined as the words which represent the target objects themselves or related words and are the focused opinions in the text such as "earthquake," "typhoon," "fire" and so on. When respondents describe their opinions in detail, it is thought that the straightforward expressions could be done as adjectives or adjectival verbs, e.g. "fear," "dangerous," "rare" and so on. Based on the knowledge of modification relationships obtained by Cabocha, this method extracts the words, nouns an unknown words which are not registered in Cabocha, that are modifying adjective or adjectival verb as the candidates of the Evaluated words. Finally, Evaluated words are extracted based on the threshold of modifying times.

These three rules are additionally applied to the extraction of evaluated words.

(a) The words which contain Hiragana (Japanese) are excepted.
(b) The words which contain Evaluated words are extracted. If the word "earthquake" is extracted as Evaluated word, "Tokai earthquake" is also extracted.
(c) The words which modify Evaluated words are extracted. In the sentence "I am afraid fire, typhoon and earthquake," Cabocha determines that "fire" modifies "typhoon" and "typhoon" does "earthquake." If the word "earthquake" is extracted as Evaluated word, "fire" and "typhoon" are also extracted.

2) Extraction of Evaluating Words: Evaluating words are defined as the words which describe the opinions for the Evaluated words such as the adjectives or adjectival verbs describe above. As for Evaluating words, the words which are modified by Evaluated words defined in Sec. Ⅲ.A.1) are chosen. When all modified words by Evaluated words are extracted as Evaluating words, there are still a lot of

improper words as Evaluating words, therefore these words are considered as the candidates of Evaluating words. The proposed method employs $Dice$ coefficient which is shown below as a criteria for co-occurrence.

$$Dice(P_i, W_j) = \frac{N_h(P_i, \gamma, W_j)}{N_h(W_j)} \quad (2)$$

In the above equation, P_i is an Evaluated word, W_j is a candidate Evaluating word, and γ is a particle. $N_h(X)$ is the frequency of the word X, which is counted as the frequency for the Internet search using Yahoo API[8]. The numerator in eq.(2) is the retrieval frequency for the search, e.g. "earthquake(Pi) ga(γ) scary(Wj) (earthquake is scary)." According to the investigation of examination shown in [9], "ga" is employed as the most appropriate particle. The threshold for $Dice$ coefficient is decided based on the mutual information[8] and the candidates which have higher $Dice$ coefficient are extracted as Evaluating words.

B. Definition of the Similarity between texts

The similarity between two texts are defined. The frequency of Evaluated words in each text is counted. The equation of frequency vector $\mathbf{A}$ in text A to calculate the similarity between text A and B is shown below.

$$\mathbf{A} = (x_{Ai} | X_i \in A \cup B) \quad (3)$$

In this equation, X_i is Evaluated words extracted in text A or B, x_{Ai}, the elements of the frequency vector, is the frequency of X_i in text A. The number of elements is that of Evaluated words in text A and B. The similarity between text A and B are defined in the following equation.

$$Sim(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{|\mathbf{A}|\,|\mathbf{B}|} = \frac{\sum x_{Ai} x_{Bi}}{\sqrt{\sum x_{Ai}^2}\sqrt{\sum x_{Bi}^2}} \quad (4)$$

The relationships among the texts are visualized using MDS based on each similarity between two texts.

C. Free Text Analysis Support System

Free text analysis support system is implemented using the method described above. The interface of the proposed system is shown in Fig.2. In this figure, the right part is the result of MDS and the left one is that of HK Graph which presents Evaluated words and Evaluating words. It enables a user to cluster interactively on the result of MDS by circle or polygon with arbitrary size. Then the clusters are selected as Bases for HK Graph and the extracted evaluation keywords, a user can select the number of presented evaluation words, are shown with connected links.

IV. Experiment

A. Extraction of Correct Evaluated Words

An experiment was done to analyze 1000 free texts acquired by web questionnaire about natural disasters. The questions was "Please describe your opinions about natural disasters freely." First, Cabocha was applied to the texts and 2835 words were divided. 32 words were deleted as noise words described in Sec. Ⅱ.A, and 245 candidates of Evaluated words were extracted using the extraction method shown in Sec. Ⅲ.A.1). This paper investigated the appropriate threshold of modifying times. In this experiment, for quantitative investigation of the proposed method, the words which could be thought appropriate keywords in 1486 nouns and unknown words were defined by ourselves not automatically. Then 479 words were defined as the correct Evaluated words.

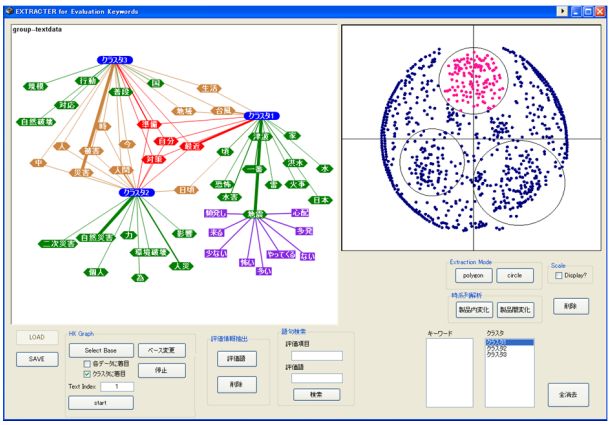

Fig. 2. Image of Free Text Analysis Support System

B. Accuracy and Coverage Rate

Fig.3 is a Venn diagram that shows the sets of all words (1486 nouns and unknown words), correct Evaluated words (479 words) and extracted Evaluated words. Note that Evaluated words are extracted based on the threshold of modifying times as the first step, then additional rules shown in Sec. Ⅲ.A. are applied and finally Evaluated words are extracted. Eq.(5) shows the accuracy which represents how many extracted words are proper or accurate in the extracted words and eq.(6) shows the coverage rate which represents how many Evaluated words are extracted in the correct Evaluated words.

$$Accuracy(\%) = \frac{N(E_v \cap E_x)}{N(E_x)} \times 100 \quad (5)$$

$$CoverageRate(\%) = \frac{N(E_v \cap E_x)}{N(E_v)} \times 100 \quad (6)$$

In these equations, E_v is the set of correct Evaluating words, E_x is that of the extracted words, and $N(X)$ is the number of words in a word set X.

Usually, there is a trade-off between accuracy and coverage rate. It is necessary to decide the appropriate threshold, the one that balances both accuracy and coverage rate. Higher accuracy and coverage rate mean a larger $E_v \cap E_x$ part and smaller $E_v \cap \overline{E_x}$ and $\overline{E_x} \cap E_v$ parts. This paper employs mutual information between Evaluating words and the extracted words to decide the appropriate threshold. When all extracted words are the

correct words and all correct words are extracted, the mutual information is maximal and both the accuracy and the coverage rate are concurrently 100%. Therefore, it would be better to decide the threshold to maximize the mutual information. The equation of mutual information is shown below.

$$I(Ev; Ex) = H(Ev) - H(Ev|Ex) \tag{7}$$

In this equation, $H(Ev)$ is the entropy of Ev (correct Evaluating words), and $H(Ev|Ex)$ is that of the conditional probability of E_v, given Ex (extracted words). Each entropy is given as the eq.(8) and (9), respectively.

$$H(Ev) = -\sum_{i=1}^{n} p(Ev_i) log_2 p(Ev_i) \tag{8}$$

Ev_i is the set of Evaluating words or that of the others, then n becomes 2. $p(Ev_i)$ is the probability of Ev_i ($p(Ev_1) = \frac{N(Ev)}{N(AllWords)}$, $p(Ev_2) = \frac{N(\overline{Ev})}{N(AllWords)}$).

$$H(Ev|Ex) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(Ev_i, Ex_j) log_2 p(Ev_i|Ex_j) \tag{9}$$

Ex_j is the set of extracted words or that of the others, then m also becomes 2. $p(Ev_i|Ex_j)$ is the conditional probability of Ev_i, given Ex_j ($p(Ev_1|Ex_1) = \frac{N(Ev \cap Ex)}{N(Ex)}$, $p(Ev_2|Ex_2) = \frac{N(\overline{Ev} \cap \overline{Ex})}{N(\overline{Ex})}$).

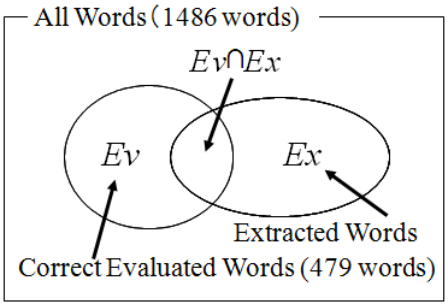

Fig. 3. Image of Extracted Words

Fig.4 shows the accuracy and the coverage rate with changing the threshold of modifying times. Though they were not the candidates of Evaluated words in the proposed method, "modifying times was 0" means that extracted all nouns and unknown words as Evaluated words. In this figure, we can see the accuracy increases when the threshold becomes higher. The maximum accuracy was approximately 60% while the coverage rate declined to 0% - 10%. In this experiment, the threshold of modifying times with the maximal mutual information was 1, i.e. the words which modify adjectives or adjectival verbs at least once were extracted as Evaluated words. Then the accuracy was 45.0% and the coverage rate was 68.0%, respectively. Higher values of these performance will be better, and more improvement of the extraction method is needed. However, this extraction of important keywords will support appropriate calculation of similarity between texts and presentation of keywords by HK Graph to grasp the outline or tendency in the texts comparing with the case of no extraction and using all words. Then this paper employs the extracted Evaluated words for the experiment.

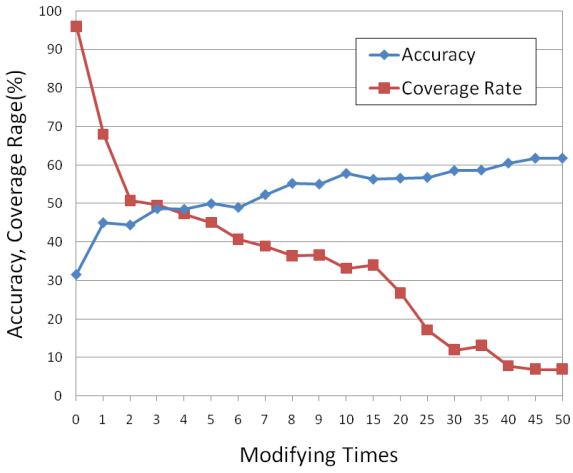

Fig. 4. Accuracy and Coverage Rate

C. Visualization using MDS based on Similarities between texts

650 Evaluated words were extracted from 1486 nouns and unknown words by the threshold shown in Sec. IV.C. Similarities between every two texts were calculated using eq.(4). The result of MDS is shown in Fig.5. In the figure, each dot corresponds to one text data by a respondent. In Fig.5, close dots mean that they have similar frequency vector one another, which is expected to be similar description in meaning, and vice versa.

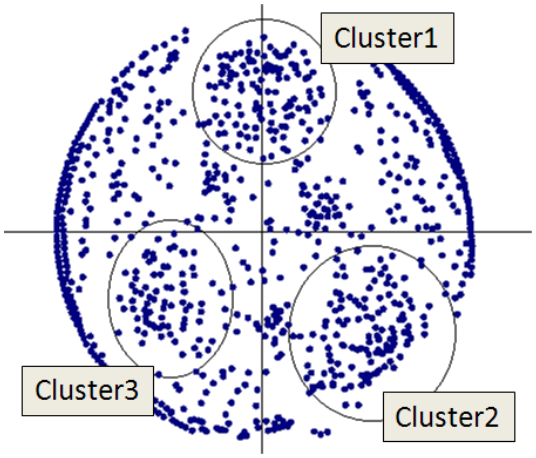

Fig. 5. Result of MDS

D. Interactive Clustering and Generation of HK Graph

First, this paper clusters the respondents on visible space. In this experiment, three groups in Fig.5 were clustered. Each circle in Fig.5 corresponds to each cluster. Fig.6 shows generated HK Graph using the evaluation keywords extracted from each cluster. In this figure, Bases named "Cluster1," "Cluster2" and "Cluster3" are those in Fig.5. Fig.6 shows that the respondents in Cluster1 have strong relationship to "earthquake," Cluster2 is to "natural disaster," Cluster3 is to "disaster," respectively. Fig.6 also shows Evaluating words connected to "earthquake." We can see that the respondents in Cluster1 are afraid of earthquake, especially that it often occurs. We can also see Cluster1 describes concrete disaster such as "flood," "typhoon," "tunami," "fire" and so on while the others use abstract expression. All connected Main-node shows every cluster describe measures and preparation for disasters.

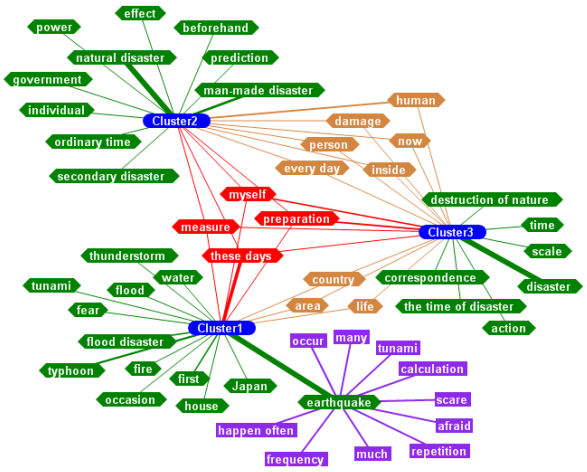

Fig. 6. Result of HK Graph among Three Clusters

V. Conclusion

This paper proposed free text analysis support system which visualized the relationships among text contents of respondents using MDS and showed their opinions using HK Graph. This paper showed the extraction method of important keywords from free text and the interactive clustering method of respondents on visible space. This paper applied the proposed method to a questionnaire data about natural disasters. This paper investigated the performance of extraction of keywords based on the correct keywords defined by hand. It showed the result of MDS based on the extracted keywords and clustered the respondents interactively on visible space. It also showed the generated HK Graph with the extracted evaluation words from each cluster. For the further works, more investigation to extract appropriate keywords will be needed, and we will apply the proposed method to other free texts.

References

[1] Takenobu Tokunaga, "johokensaku to gengoshori," (in Japanese) University of Tokyo Press, 1999
[2] Yuto Murakami, Yoshikazu Tanizawa, Dongli Han, Minoru Harada, "Automatic classification of Open-Ended Questionnaires based on semantic analysis," (in Japanese) The 66th National Convention of IPSJ, pp.171-172, 2005
[3] Hiroko Inui, Kiyotaka Uchimoto, Masaki Murata, Hitoshi Isahara, "Classification of Open-Ended Questionnaires based on Predicative," (in Japanese) NLP-99, pp.181-188, 1998
[4] Norimasa Hayshida, Hideyuki Takagi, "Visualized IEC: Interactive Evolutionary Computation with Multidimensional Data Visualization," IEEE InternationalConference on Industrial Electronics, Control and Instrumentation (IECON2000), Vol.4, 2738-2743, 2000
[5] Takahiro Okabe, Tomohiro Yoshikawa, Takeshi Furuhashi, "Proposal of Multi-Connected Hierarchical Text Mining Method for Medical Incident Reports," (in Japanese) The 22nd Fuzzy System Symposium, pp.211-214, 2006
[6] Bo Hao, Tomohiro Yoshikawa, Takeshi Furuhashi, Sin-ichiro Sugiura, "A Development of Early Diagnosis and Hospital Search Support System for Integrate Medical Support System," Proc. the 2nd International Conference on Knowledge Generation, Communication and Management (KGCM 2008), pp.55-60, 2008
[7] Taku Kudo, Yuji Matsumoto, "Japanese Dependency Analysis using Cascaded Chunking," 6th Conference on Natural Language Learning, pp.63-69, 2002
[8] Yahoo API, http://developer.yahoo.co.jp/
[9] Yuki Uchida, Tomohiro Yoshikawa, Takeshi Furuhashi, Eiji Hirao, Hiroto Iguchi, "Evaluation of Products by Analysis of User-Review using HK Graph," Proc. of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems(SCIS & ISIS 2008), pp.376-379, 2008