Evaluation of Products by Analysis of User-Review using HK Graph

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

Tomohiro Yoshikawa
Nagoya University

Takeshi Furuhashi
Nagoya University

Eiji Hirao
NEC corporation

Hiroto Iguchi
NEC corporation

Abstract—**Recently, the sites in internet on which users can write private ideas and opinions are increasing. In addition, the number of people who want to know other's opinions about the interested products is also increasing. However, it is very difficult for people to read whole reviews on internet. This study tries to develop a new review analysis system which shows evaluation information about products using graph structure of evaluation keywords. This paper focuses on the part of extraction of evaluation keywords from reviews on internet. This paper shows the extraction method for evaluation keywords and the result of keyword graph structure with extracted evaluation words. It employs HK Graph (Hierarchical Keyword 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, new styles of the Web sites such as weblogs are rapidly spreading because people can write their ideas or opinions in weblogs much easier than in traditional websites. Also, the number of the websites which contain a lot of consumers ’ opinions or reviews about some products like electrical appliances or cars is increasing. People often want to know the reviews before they buy something. There are some websites where people rate some products on a scale of 1 to 5. However, it is hard to understand the reasons why they are evaluated as good or bad. Moreover, it is uncertain whether users ’true opinions hidden deeply inside their minds appear on such quantitative information.

In Japan, “kakaku.com”[1] is well known as a major website which collects and shows not only a lot of consumer's evaluation rates but also their free-written reviews for various products. We think the users ’ true opinions come out on the reviews not on rates. However, there are more than 100 reviews for each product in this website and huge number of pages including the reviews for that product. Then it is difficult for us to read all reviews on internet especially when we want to compare the interested one with others. Therefore, the demands for the analysis system of evaluation information of reviews for products are strongly growing.

HK Graph(Hierarchical Keyword Graph)[2] 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 study tries to develop an evaluation support system of products using the text contents on Web consumers ’reviews for them based on HK Graph. HK Graph divides text data into words using morpheme and phrase analysis tool, and then it shows the hierarchical relationship among words. However, the criterion of extraction of words in HK Graph is the co-occurrence, then most of them are general words rather than evaluation keywords. The general words often prevent us from the effective analysis or evaluation. It is needed to extract the words which are related to evaluation of the products.

This paper proposes the extraction method of the words related to the evaluation of products from reviews on internet. In the proposed method, the evaluation keywords are extracted using modification relation of words and retrieved frequency on internet. It is expected that the extracted keyword graph can support the effective evaluation of products.

This paper employs “kakaku.com” as the website of reviews and applies the proposed method to the review texts for mobile phones. This paper investigates the performance of extraction of evaluation keywords based on the collect evaluation keywords defined by hand. Finally, it shows the result of keyword graph with extracted evaluation words by the proposed method using HK Graph.

II. HK Graph

HK Graph extracts the words which have high co-occurrence with the words selected by the user from the target sentences and shows them as a hierarchical keyword graph. Then we can grasp the abstract of the text. The features of HK Graph are the hierarchical structure and interactive search, in which it starts from the selection of words by a user with he/she interest and he/she can proceed into deeper layer of interested words shown as the keyword graph. The algorithm of HK Graph is shown as follows.

A. Division into Words

The first step in the algorithm of HK Graph is to divide text into words by applying Cabocha[3] to the text. Cabocha is a Japanese language morphological and paragraphic analysis tool. Unlike English which has spaces between every word, it is difficult to divide Japanese text without tools like Cabocha. Applying Cabocha to the target texts, particles, symbols (punctuation, parentheses), pronouns, conjunctions and adnominal words 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 to select some keywords which he/she is interested in out of the divided words in 2.1. As the selected words here are the bases to start analysis, the selected 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. As for analysis of consumers ’ reviews, Bases will be the name of products.

C. Extraction of Main-node

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

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

B_i is Base, W_j is each divided word in 2.1 and $N(X)$ is the number of texts 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 called “ 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, $B_1 \sim B_3$ are Base, A_1 is All connected Main-node, $M_{12} \sim M_{23}$ are Multi connected Main-node, $S_{11} \sim S_{32}$ are Single connected Main-node and Sub_1, Sub_2 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.

III. Extraction of Evaluation Keywords

A. Evaluation Keywords

For the review analysis, we want to create a graph which contains evaluation information as Main-nodes and Sub-nodes. Using reviewing website for products as the target texts, the ratio of the words for evaluation to the extracted words increases. However, there are still a lot of words with no relation to evaluation. Because HK Graph extracts the words based on the co-occurrence, and the words with no relation are also extracted as long as their co-occurrence is high. Evaluation information consists of evaluation keywords. This paper defines evaluation keywords as Evaluated words and Evaluating words and extracts them. This paper focuses on such an extraction of them from reviews.

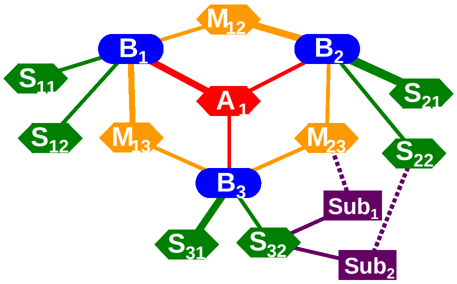

Fig. 1. Image of HK Graph

Fig.2 shows an example of a sentence in a review for a mobile phone. In this sentence, evaluation keywords correspond to “ design ” (Evaluated word) and “ cool ” (Evaluating word). HK Graph shows the name of products as Base, Evaluated words as Main-nodes and Evaluating words as Sub-nodes. The following subsections describe the extraction method of these evaluation keywords.

B. Extraction of Evaluated Words

Evaluated words are defined as the words which represent the features of products and are the focused points in the evaluation such as “ design ”, “ price ”,“ size ” and so on. Based on the knowledge of modification relationships obtained by Cabocha, this method extracts the words that are modifying adjective or adjectival verb as the candidates of Evaluated words. Finally, Evaluated words are extracted based on the threshold of modifying times.

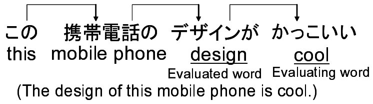

Fig. 2. Example of Evaluation Sentence for Mobile Phone

C. Extraction of Evaluating Words

Evaluating words are defined as the words which describe the features of evaluated words such as “ cool ”, “ expensive ”, “ small ” and so on. As for evaluating words, the words which are modified by Evaluated words defined in 3.2 are chosen. When all modified words by Evaluated words are extracted as the Evaluating words, there are still a lot of improper words as Evaluating words, therefore these words are considered as candidates of Evaluating words. The proposed method employs $Dice$ coefficient which is shown below as a criteria for co-occurrence.

$$Dice(P_i, W_j) = \frac{N(P_i, \gamma, W_j)}{N(W_j)} \qquad (2)$$

Where Pi is Evaluated word, Wj is a candidate of Evaluating words and γ is a particle. $N(X)$ is the retrieval number of the word X, which is counted as the retrieval frequency for the internet search using Yahoo API[4]. The numerator in eq.(2) is the retrieval frequency for the search e.g., "dezain(Pi) ga(γ) kakkoi(Wj) (design is cool)". According to the investigation of examination shown in 4.3, "ga" is employed as the most appropriate particle. The threshold for $Dice$ coefficient is decided and the candidates which have higher $Dice$ coefficient are extracted as Evaluating words.

IV. Experiment

A. Extraction of Collect Evaluating Words

In this paper, an experiment is done to analyze reviews about mobile phones obtained from a website called as "kakaku.com" which is a major website holding a lot of kinds of reviews in Japan. Here, 172 reviews for 8 kinds of mobile phones in "kakaku.com" were used. First, Cabocha was applied to the reviews and 4167 words were divided. Then 179 words were deleted as noise words described in 2.1, and 15 (The candidates of Evaluated words were 209.) words were extracted as Evaluated words using the extraction method shown in 3.2. Then, 668 candidates of Evaluating words were obtained by the proposed method. These candidates were the words which were modified by the extracted Evaluated Words.

This paper investigates the appropriate threshold of $Dice$ coefficient for Evaluating words described in 3.3. If the threshold is very low level, it will extract a lot of Evaluating words and improper words at the same time. If high level, we would get few improper words, but few Evaluating words instead. In this experiment, for quantitative investigation of the proposed method, the words which express evaluation in these candidates of 668 words were defined by ourselves not automatically. Then 210 words were defined as the collect Evaluating words.

B. Accuracy and Coverage Rate

Fig.3 is a Venn diagram that shows the sets of all words (668 words), collect Evaluating words (210 words) and Extracted words. Eq.(3) shows accuracy which represents how many Extracted words are proper or accurate in the extracted words and eq.(4) shows coverage rate which represents how many Evaluating words are extracted in the collect Evaluating words.

$$Accuracy(\%) \quad = \quad \frac{N(E_v \cap E_x)}{N(E_x)} \times 100 \tag{3}$$

$$CoverageRate(\%) \quad = \quad \frac{N(E_v \cap E_x)}{N(E_v)} \times 100 \tag{4}$$

In these expressions, Ev is the collect Evaluating words, Ex is the extracted words, and $N(X)$ is the number of words X.

Usually, there is a trade-off between accuracy and coverage rate. It is necessary to decide the appropriate threshold that balances both accuracy and coverage rate. Higher accuracy and coverage rate mean larger $Ev \cap Ex$ part and smaller the other part of Ev and Ex in Fig.3. Then this paper employs mutual information between two conditions, one is of whether a word is the Evaluating word and the other is of whether the word is the Extracted word, to decide appropriate threshold. When the set of Ev is equal to the set of Ex, the mutual information is maximal and both the accuracy and the coverage rate are 100 % at the same time. Therefore, it would be better to decide the threshold to maximize the mutual information. The expression of mutual information is shown below.

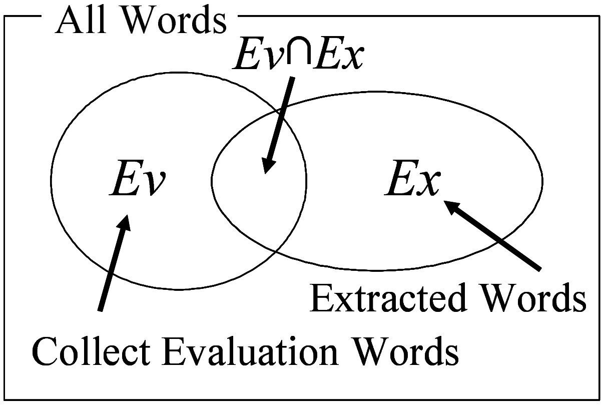

Fig. 3. Image of Extracted Words

$$I(Ev;Ex) = H(Ev) - H(Ev|Ex) \tag{5}$$

In this expression, $H(Ev)$ is the entropy of Ev (collect Evaluating words), and $H(Ev|Ex)$ is the entropy of the conditional probability of E_v, given Ex (extracted words). Each entropy is given as the expression (6) and (7), respectively.

$$H(Ev) = -\sum_{i=1}^{n} p(Ev_i) log_2 p(Ev_i) \tag{6}$$

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 ($p(Ev_1) = \frac{N(Ev)}{N(AllWords)}$, $p(Ev_2) = \frac{N(\overline{Ev})}{N(AllWords)}$).

$$H(Ev|Ex) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(Ev_i, Ex_j) log_2 p(Ev_i|Ex_j) \tag{7}$$

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 ($p(Ev_1|Ex_1) = \frac{N(Ev \cap Ex)}{N(Ex)}$, $p(Ev_2|Ex_2) = \frac{N(\overline{Ev} \cap \overline{Ex})}{N(\overline{Ex})}$).

C. Selection Particles for Extracting Evaluating Words

First, the appropriate particles were investigated based on the comparison of the maximal mutual information. The following 6 kinds of particle in Japanese was compared, "ga","no","wo","ni","wa" and "mo". These particles express

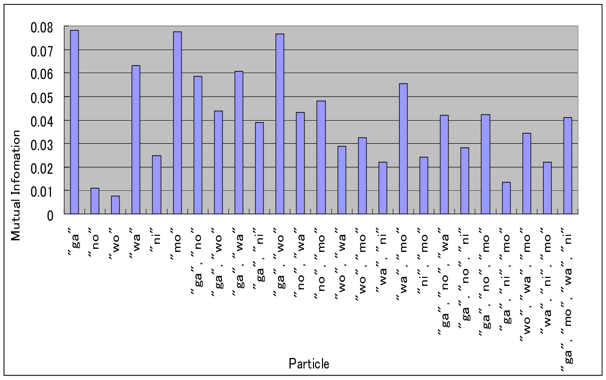

Fig. 4. Maximal Mutual Informations of Particles

relation between Evaluated words and candidates of Evaluating words. “ga” and “wa” indicate the relation of subjects and predicates ($\simeq$ be), “wo” and “ni” indicate that of objects and predicates ($\simeq$ at, on, to, etc), “mo” has parallel mean in words ($\simeq$ too, and) and “no” is the meaning of possessive ($\simeq$ of). These particles connect from a substantive (Evaluated word) and express relationship of phrase.

Mutual information was calculated for each particle and the combinations of them. Fig.4 shows the comparison result of their maximal mutual information. For example, when using two particles, numerator of $Dice$ coefficient (eq.(2)) was the sum of each retrieval number. Using the particle “ga” shows the biggest maximal mutual information in Fig.5, then this paper employs “ga” for the most appropriate particle to calculate $Dice$ coefficient.

Fig.5 shows the accuracy and the 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 30.5 % 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 eq.(2) was appropriate to extract Evaluating words. The maximum accuracy was approximately 80 % while the coverage rate declined to 20 %-30 %. In this experiment, the threshold with the maximal mutual information was 0.00004 in which the accuracy was 68.1 % and the coverage rate was 60.0 % (Fig.5).

D. Generated HK Graph

HK Graph was generated using the evaluation keywords extracted by the proposed method. The result is shown in Fig.6. In the figure, “ SH904i ” and “ N904i ” is the name of mobile phones. It shows that the keyword graph contains Evaluated words such as “ design ”, “ response ” and “ button ” as Main-nodes and Evaluating words such as “ fast ”, “ like ” and “ small ”as Sub-nodes, which are appropriate evaluation keywords. The results show that the proposed method is suitable for the extraction of evaluation information from reviews.

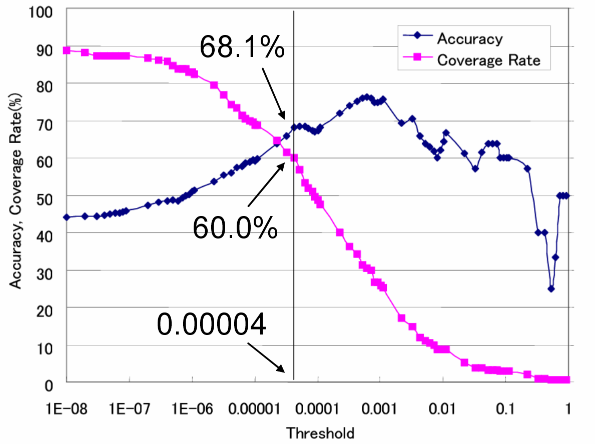

Fig. 5. Accuracy and Coverage Rate

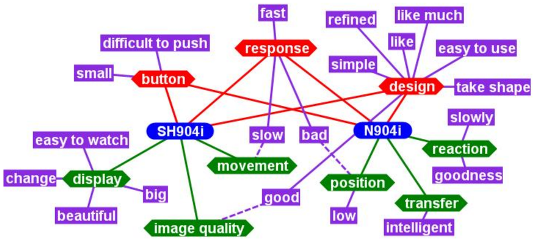

Fig. 6. Result of HK Graph with Extracted Evaluation Keywords

V. Conclusion

This paper proposed an evaluation support system of products using the text contents on Web consumers’ reviews for them based on HK Graph. This paper focused on the extraction method of the words related to the evaluation of products from reviews on internet. This paper employed “kakaku.com” as the website of reviews and applied the proposed method to the review texts for mobile phones. This paper investigated the performance of extraction of evaluation keywords based on the collect evaluation keywords defined by hand. It showed the result of keyword graph with extracted evaluation keywords. For the further works, more investigation to extract appropriate evaluation keywords will be needed, and we will apply the proposed method to other reviews or texts.

References

[1] Kakaku.com, http://bbs.kakaku.com/bbs/

[2] 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.

[3] 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

[4] Yahoo API, http://developer.yahoo.co.jp/