Analysis of Textual Data with Multiple Classes

Shigeaki Sakurai[1], Chong Goh[1,2], and Ryohei Orihara[1]

[1] Corporate Research & Development Center, Toshiba Corporation
[2] Carnegie Mellon University

Abstract. This paper proposes a new method for analyzing textual data. The method deals with items of textual data, where each item includes various viewpoints and each viewpoint is regarded as a class. The method inductively acquires classification models for 2-class classification tasks from items labeled by multiple classes. The method infers classes of new items by using these models. Lastly, the method extracts important expressions from new items in each class and extracts characteristic expressions by comparing the frequency of expressions. This paper applies the method to questionnaire data described by guests at a hotel and verifies its effect through numerical experiments.

1 Introduction

As computers and network environments have become ubiquitous, many kinds of questionnaires are now conducted on the Web. A simple means of analyzing responses to questionnaires is required. The responses are usually composed of selective responses and textual responses. In the case of the selective responses, the responses can be analyzed relatively easily using statistical techniques and data mining techniques. However, the responses may not correspond to the opinions of respondents because the respondents have to select appropriate responses from among those given by the designers of the questionnaires. Also, the designers are unable to receive unexpected responses because only those expected by the designers are available. On the other hand, in the case of the textual responses, the respondents can freely describe their opinions. The designers are able to receive more appropriate responses that reflect the opinions of the respondents and may be able to receive unexpected responses. Therefore, textual responses are expected to be analyzed using text mining techniques. Even though many text mining techniques [2] [3] [8] [9] have previously been studied, textual data has not always been analyzed sufficiently. Since analysis may be undertaken for various purposes and there are various types of textual data, it is difficult to construct a definitive text mining technique. The text mining technique must reflect the features of the textual data. In this paper, we propose a new analysis method that deals with textual data that includes multiple viewpoints. The method is designed to deal with free-form textual responses, to classify textual responses to questionnaires according to various viewpoints, and to discover characteristic expressions corresponding to

M.-S. Hacid et al. (Eds.): ISMIS 2005, LNAI 3488, pp. 112–120, 2005.

each viewpoint. We apply the proposed method to the analysis of textual responses given by guests at a hotel and verify its effect through numerical experiments.

2 Analysis of Textual Responses

2.1 Analysis Targets

Many kinds of comments may be expressed in textual responses to questionnaires. It is important to investigate all textual responses in detail and to include measures that resolve problems in the responses. However, the amount of textual responses that analysts can investigate is limited. Even if they could investigate all textual responses, it would be impracticable to include all required measures due to constraints regarding cost, time, etc. Therefore, it is necessary to show rough trends for the textual responses and to extract the important topics from them. Thus, we propose a method that classifies textual data into various viewpoints and extracts important expressions corresponding to each viewpoint.

2.2 Analysis Policy

Respondents to a questionnaire can freely describe their opinions in textual responses, and the respondents can provide responses that include multiple viewpoints. For example, in the case of a questionnaire for guests at a hotel, a guest may provide a textual response that includes three viewpoints, e.g. bad aspects of the hotel, good aspects of the hotel, and requests to the hotel. If the respondents classified their responses according to the viewpoints and put them into columns, the analyst's task would be easy. However, since the respondents would be likely to find such a questionnaire troublesome, a low response rate would be likely. It is necessary to allow textual responses that include multiple viewpoints in order to ease the burden on respondents.

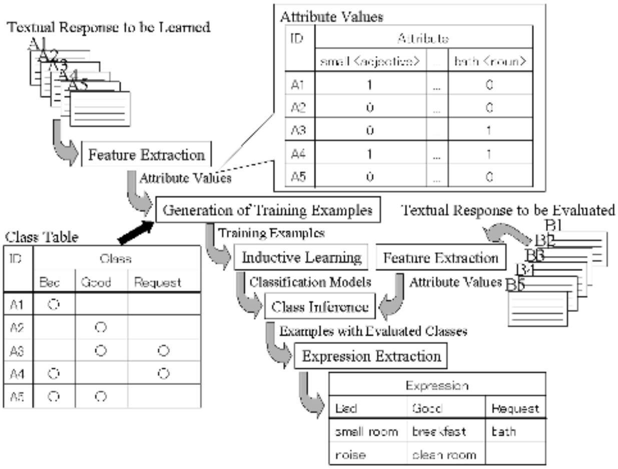
We first consider a method that uses passage extraction techniques [6] [10] for that purpose. The techniques can extract specific parts of textual data and are used effectively in a question answering task. However, it is necessary for many passage extraction techniques to measure the distance between a standard sentence and parts of the textual data. In the case of analysis of textual responses to questionnaires, it is difficult to decide what corresponds to a standard sentence. Therefore, we can not extract the specific parts using the techniques.

Next, we consider classifying each textual response by using a classification model. The classification model has to identify textual responses that include a single viewpoint and textual responses that include multiple viewpoints. It is difficult for the model to identify the former responses with the latter responses, because the former responses may be a part of the latter responses. A large number of training examples are required to inductively learn the model, because the latter responses are composed of combinations of the former responses. Therefore, a method based on a classification model is not always appropriate.

Thus, we try to acquire classification models corresponding to viewpoints. Here, the classification models can identify whether or not a textual response corresponds to a viewpoint. The models are acquired from training example sets that correspond to viewpoints. We can identify viewpoints that correspond to each textual response by using the models and can also extract expressions from textual responses included in a specific viewpoint. Here, it should be noted that the expressions are not always related to the specific viewpoints. This is the reason they can be related to other viewpoints that simultaneously occur with the specific viewpoint. However, the number is much smaller than the number of expressions related to the specific viewpoint. Therefore, we can extract expressions that correspond to each viewpoint by comparing the number of expressions extracted in each viewpoint. We consider that the classification and the extraction can analyze textual responses to questionnaires.

2.3 Analysis Flow

We constructed a new analysis method based on the policy described in subsection 2.2. The method is composed of five processes as shown in Figure 1. Here, the method deals with a language without word segmentation, such as Japanese. In the following, the processes are explained.



ID	Attribute		
	small <adjective>	...	bath <noun>
A1	1	...	0
A2	0	...	0
A3	0	...	1
A4	1	...	1
A5	0	...	0

ID	Class		
	Bad	Good	Request
A1	○		
A2		○	
A3		○	○
A4	○		○
A5	○	○	

Expression		
Bad	Good	Request
small room	breakfast	bath
noise	clean room	

Fig. 1. Analysis flow

Feature Extraction Process: The process decomposes each textual response into words with corresponding parts of speech by using morphological analysis [5]. The process extracts words, if their tf-idf values are bigger than or equal to a threshold and their parts of speech are included in a designated set of parts of speech. The process regards the extracted words as attributes. The process also evaluates whether or not the words are included in a textual response. If the words are included, the process gives 1s to the corresponding attributes.

Table 1. Training examples corresponding to the viewpoint "Bad"

ID	Attribute			Class
	small <adjective>	$\cdots$	bath <noun>	
A1	1	$\cdots$	0	c_1
A2	0	$\cdots$	0	c_0
A3	0	$\cdots$	1	c_0
A4	1	$\cdots$	1	c_1
A5	0	$\cdots$	0	c_1

Otherwise, the process gives 0s to them. Therefore, a column vector as shown at the upper-right side in Figure 1 is assigned to each textual response.

Generation Process of Training Examples: The process selects a viewpoint in the class table. The process evaluates whether the viewpoint is assigned in a textual response or not. If the viewpoint is assigned, the process assigns the class c_1 to the textual response. Otherwise, the process assigns the class c_0 to it. The process generates the training example set corresponding to the viewpoint by integrating attribute values with the classes. Table 1 shows an example of training examples corresponding to the viewpoint "Bad".

Inductive Learning Process: The process acquires classification models from each training example set by solving 2-class classification tasks. Each model corresponds to a viewpoint. In this paper, the process uses a support vector machine (SVM) [4] to acquire the models, because many papers [1] [7] have reported that an SVM gives high precision ratios for text classification. The process acquires classification models described with hyperplanes by using an SVM.

Class Inference Process: The process applies textual responses to be evaluated to each classification model. Here, each textual response is characterized by words extracted from the textual responses to be learned. The process infers classes, c_0s or c_1s, corresponding to the textual responses to be evaluated for each viewpoint. In Figure 1, three kinds of classes corresponding to the viewpoints "Bad", "Good", and "Request" are assigned to the responses B1 $\sim$ B5.

Expression Extraction Process: The process extracts expressions from the textual responses with class c_1. Here, the expressions are words that are specific parts of speech or phrases that are specific sequences of parts of speech such as <adjective> and <noun>. The words and the sequences are designated by analysts. The process calculates the frequency of the expressions in each viewpoint and assigns the viewpoint with the maximum frequency to the expressions. Lastly, the process extracts expressions that are bigger than or equal to a threshold. The expressions are regarded as characteristic expressions in the viewpoints.

For example, assume that five textual responses are given. Here, two textual responses include the expression "small room", two textual responses include the expression "clean room", and one textual response includes the expressions "small room" and "clean room". Also, assume that textual responses with "small room" are classified into "Bad" and textual responses with "clean room" are classified into "Good". In the case of "Good", "clean room" occurs 3 times and "small room" occurs once. Similarly, in the case of "Bad", "clean room" occurs once and "small room" occurs 3 times. Therefore, we can extract "clean room" as an expression relating to "Good" and "small room" as an expression relating to "Bad".

3 Numerical Experiments

3.1 Experimental Method

We used textual responses to a questionnaire collected from guests at a hotel. Each textual response contains comments on the hotel. The comments have three viewpoints: bad aspects of the hotel, good aspects of the hotel, and requests to the hotel. Analysts read each textual response and assigned three viewpoints to each textual response. Thus, some textual responses have multiple viewpoints and other textual responses have a single viewpoint. We collected a total of 1,643 textual responses with viewpoints assigned by analysts not as a single set but as the result of 4 separate attempts. The data sets D1, D2, D3, and D4 corresponding to 4 attempts are related such that D2 $\subseteq$ D3 and D4 = D1 + D3. The frequency of the textual responses in the data set is shown in Table 2. In Table 2, "Yes" indicates the number that includes a viewpoint and "No" indicates the number that does not include a viewpoint.

Table 2. Distribution of comments

	D1		D2		D3		D4	
	Yes	No	Yes	No	Yes	No	Yes	No
Bad	48	59	603	714	693	843	741	902
Good	62	45	707	610	823	713	885	758
Request	51	56	457	860	506	1,030	557	1,086
Total	107		1,317		1,536		1,643	

In order to evaluate the difference in the feature extraction process, we used 9 lexical filters and 5 thresholds of tf-idf values. Each filter extracts the part of speech designated by Table 3. That is, a filter L1 extracts adjectives and a filter L9 extracts all words. Also, the thresholds are changed in the range $0.000 \sim 0.020$.

At first, we performed numerical experiments by using D1. We extracted attributes from textual responses included in D1 by using a lexical filter and a threshold. 10-Cross validation experiments were applied to textual responses with attribute values and a single viewpoint. Also, the 10-Cross validation experiments were performed for three viewpoints. Moreover, these numerical experiments were performed for each lexical filter and each threshold. We calculated

Table 3. Lexical filter

Filter	Part of speech	Filter	Part of speech	Filter	Part of speech
L1	adjective	L4	adjective, verb	L7	adjective, verb, noun
L2	verb	L5	adjective, noun	L8	L7, numeral, symbol, alphabet, desinence, interjection, unknown
L3	noun	L6	verb, noun	L9	All parts of speech

the precision ratio defined by Formula (1) for each viewpoint, each filter, and each threshold.

$$\text{precision ratio} = \frac{\text{Number of correctly classified textual responses}}{\text{Number of textual responses}} \tag{1}$$

Next, we performed numerical experiments by using D2, D3, and D4. We extracted attributes from textual data included in each data set, where we used a lexical filter and a threshold selected by the former experiments. The number of attributes was about 1,400. We also performed 10-Cross validation experiments for each viewpoint and each data set, and calculated precision ratios.

Lastly, we extracted expressions from textual responses in D1. We selected lexical filters and thresholds that corresponded to maximum precision ratios and acquired classification models for each viewpoint. We set 2 as the threshold of the expression extraction process and extracted nouns. We investigated which textual responses included extracted words and classified extracted words into four categories: correct category, wrong category, mixed category, and neutral category. Here, correct category indicates that an extracted word corresponds to its viewpoint, wrong category indicates it does not correspond to its viewpoint, mixed category indicates it corresponds to its viewpoint and also corresponds to other viewpoints, and neutral category indicates it is not related to all viewpoints. We calculated frequency ratios defined by Formula (2) for the correct category, the wrong category, and the mixed category.

$$\text{frequency ratio} = \frac{\text{Number of expressions of each category}}{\text{Number of expressions except the neutral category}} \tag{2}$$

3.2 Experimental Results

Table 4 shows results for changing lexical filters and thresholds. Each cell shows average precision ratios in three viewpoints. The last row shows average values when using the same threshold and the last column shows average values when using the same lexical filter.

Figure 2 shows results for changing data sets. Solid lines in the figures indicate results for “Bad”, “Good”, and “Request”. A solid heavy line indicates average values for three viewpoints. Here, we used a lexical filter L9 and a threshold 0.005, because the filters and the threshold give a model with a stable precision ratio, as shown in Table 4.

Lastly, Table 5 shows frequency ratios for D1.

Table 4. Precision ratio for thresholds and filters in D1

	Threshold					
Filter	0.000	0.005	0.010	0.015	0.020	Average
L1	0.667	0.667	0.664	0.673	0.660	0.666
L2	0.508	0.508	0.514	0.520	0.539	0.518
L3	0.586	0.586	0.592	0.579	0.583	0.585
L4	0.660	0.660	0.629	0.617	0.648	0.643
L5	0.676	0.676	0.695	0.664	0.695	0.681
L6	0.651	0.651	0.626	0.617	0.611	0.631
L7	0.682	0.682	0.707	0.657	0.664	0.679
L8	0.688	0.688	0.688	0.654	0.667	0.677
L9	0.698	0.698	0.682	0.685	0.667	0.686
Average	0.646	0.646	0.644	0.630	0.637	0.641

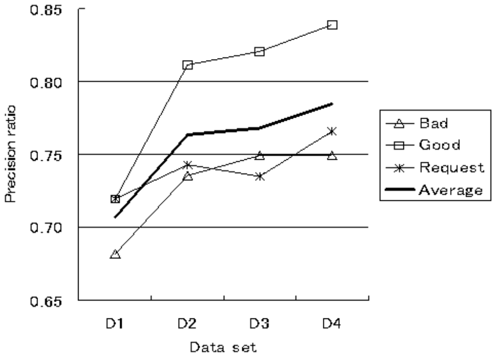


Fig. 2. Precision ratio for data sets

3.3 Discussion

Setting of Viewpoints: In this analysis task, we used three viewpoints. The viewpoints can not always apply to all analysis tasks. However, the viewpoints can apply to analysis of the customer voice in the service field. We have large amounts of data in the field. Therefore, we consider that the viewpoints have a wide range of application tasks.

Influence of Lexical Filters: The textual responses describe the impressions of the guests. Expressions that include adjectives and nouns are important. They lead to the correct viewpoint classification. We believe this is the reason the lexical filters that included adjectives and nouns provided comparatively high precision ratios. On the other hand, the morphological analysis engine sometimes leads to incorrect word segmentation. In particular, the engine tends to fail in the case of word segmentation for text that includes new words and proper nouns. This causes the engine to identify the words as unknown words, or to segment

Table 5. Frequency ratio of expressions

	Bad	Good	Request
Correct	0.887	0.512	0.800
Wrong	0.065	0.198	0.086
Mixed	0.048	0.291	0.114

the words at wrong positions and assign wrong parts of speech to the words. The L9 lexical filter is able to deal with new words and proper nouns because the filter extracts all parts of speech. Therefore, the L9 filter gives the highest precision ratio. However, the filter causes an increase in attributes. The L5 or L7 filters should be used, if calculation speed and memory size are important considerations. This is the reason the numbers of their attributes are comparatively small and their average precision ratios are almost equal to the L9 filter.

Influence of the Thresholds: The number of attributes increases as the threshold of the feature extraction process becomes low. When an inductive learning method uses large amounts of attributes, the method tends to acquire a classification model which excessively depends on training examples. It is necessary to select an appropriate threshold. However, in these textual responses, the difference in the thresholds does not lead to a big difference in precision ratios. The thresholds are not relatively sensitive. We believe the reason behind this result is the low number of irrelevant words, because each textual response deals with limited topics and is described in a comparatively short sentence.

Influence of Increase in Textual Responses: The precision ratio becomes higher as the number of textual responses increases. The case of D4 is about 8% higher than the case of D1. We believe this is why a more appropriate classification model is acquired by using many textual responses. On the other hand, the precision curves do not always converge. If more training examples are used, the proposed method may give a higher precision ratio.

Validation of Extracted Expressions: In the case of "Bad" and "Request", frequency ratios of "Correct" are comparatively high and the proposed method extracts valid expressions. On the other hand, in the case of "Good", the frequency ratio of "Correct" is low. We believe this is the reason many "Good" textual responses are accompanied by topics of other classes. That is, the classification model for "Good" tends to classify a textual response as "Good". If some guests assign topics to "Good" and other guests assign the topics to "Bad" or "Request", the expressions corresponding to the topics tend to acquire the maximum frequency in the case of "Good". Therefore, the frequency ratio of "Mixed" becomes high and the frequency ratio of "Correct" becomes low in the case of "Good". In the future, it will be necessary to devise a method that identifies topics with multiple viewpoints.

According to the above discussion, we believe that the proposed method is able to classify textual responses to questionnaires and extract valid expressions to some extent. The method therefore makes it possible for analysts to easily acquire new knowledge from textual responses.

4 Summary and Future Work

This paper proposes a new analysis method in order to analyze textual responses to questionnaires. The method was applied to questionnaire data collected from guests at a hotel. We show that precision ratios based on classification models were improved by an increase in training examples. We also show that the method extracted valid expressions of textual responses. We believe the method is efficient for analyzing textual responses to questionnaires.

In the future, we hope to develop a method that extracts more characteristic expressions. In this paper, we adopted a simple method based on frequency, but the method extracts many words included in the neutral category. The frequency of words included in the neutral category must be reduced. We also hope to develop a system in which the method is applied via a graphical user interface, and will also attempt to apply the method to other types of questionnaire data.

References

1. A. Cardoso-Cachopo and A. L. Oliveira, “An Empirical Comparison of Text Categorization Methods,” Proceedings of the 10th International Symposium on String Processing and Information Retrieval, 183-196, 2003.
2. R. Feldman and H. Hirsh, “Mining Text using Keyword Distributions,” Journal of Intelligent Information Systems, 10:281-300, 1998.
3. Marti A. Hearst, M. A. Hearst, “Untangling Text Data Mining,” Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
4. C. -W. Hsu, C. -C. Chang, and C. -J. Lin, “A Practical Guide to Support Vector Classification,” http://www.csie.ntu.edu.tw/˜ cjlin/libsvm/.
5. Y. Ichimura, Y. Nakayama, M. Miyoshi, T. Akahane, T. Sekiguchi, and Y. Fujiwara, “Text Mining System for Analysis of a Salesperson’s Daily Reports,” Proceedings of Pacific Association for Computational Linguistics 2001, 127-135, 2001.
6. A. Ittycheriah, M. Franz, W. -J. Zhu, and A. Ratnaparkhi, “IBM’s Statistical Question Answering System,” Proceedings of the 8th Text Retrieval Conference, 2000.
7. L. M. Manevitz and M. Yousef, “One-Class SVMs for Document Classification,” Journal of Machine Learning Research, 2:139-154, 2001.
8. S. Sakurai, Y. Ichimura, A. Suyama, and R. Orihara, “Acquisition of a Knowledge Dictionary for a Text Mining System using an Inductive Learning Method,” IJCAI 2001 Workshop on Text Learning: Beyond Supervision, 45-52, 2001.
9. P. -N. Tan, H. Blau, S. Harp, and R. Goldman, “Textual Data Mining of Service Center Call Records,” Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, 417-423, 2000.
10. S. Tellex, B. Katz, J. Lin, and A. Fernandes, “Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering,” Proceedings of the 26th Annual Conference ACM SIGIR Conference on Research and Development in Information Retrieval, 2003.