International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:11, 2008

Analysis of Textual Data based on multiple 2-class Classification Models

Shigeaki Sakurai and Ryohei Orihara

Abstract**—**This paper proposes a new method for analyzing textual data. The method deals with items of textual data, where each item is described based on various viewpoints. The method acquires 2-class classification models of the viewpoints by applying an inductive learning method to items with multiple viewpoints. The method infers whether the viewpoints are assigned to the new items or not by using the models. The method extracts expressions from the new items classified into the viewpoints and extracts characteristic expressions corresponding to the viewpoints by comparing the frequency of expressions among the viewpoints. This paper also applies the method to questionnaire data given by guests at a hotel and verifies its effect through numerical experiments.

Keywords**—**Text mining, Multiple viewpoints, Differential analysis, Questionnaire data

I. Introduction

As computers and network environments are becoming ubiquitous, many kinds of questionnaires are now conducted on the Web. A simple method 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. Text mining techniques may be used as the analysis method.

Even though many text mining techniques [2] [3] [9] [10] [12] 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, in order to classify textual responses to questionnaires into various viewpoints, and to discover characteristic expressions corresponding to 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.

Shigeaki Sakurai is with the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, e-mail: shigeaki.sakurai@toshiba.co.jp

Ryohei Orihara is with the HumanCentric Laboratory, Corporate Research & Development Center, Toshiba Corporation, Kawasaki, e-mail: ryohei.orihara@toshiba.co.jp

II. Analysis of textual responses

A. Analysis targets

Many kinds of comments are expressed in textual responses to questionnaires. It is important to investigate all textual responses in detail and to implement 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 impractical to implement all required measures due to constraints regarding cost, time, etc. 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.

B. Analysis policy

Respondents to a questionnaire can freely describe their opinions in textual responses and 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: bad aspects of the hotel, good aspects of the hotel, and requests to the hotel. That is, the guest may complain about a small room, admire a delicious dinner, and request internet access. It is necessary for analysts to analyze the responses according to three different viewpoints. If the respondents were willing to classify their responses into the viewpoints and put them into columns, the analysis task of a questionnaire would be easy. However, since the respondents would be likely to find such an additional task troublesome, the respondents would not be willing to answer the questionnaire if it also involved the additional task. A low response rate of the questionnaire would be likely. It is necessary to deal with textual responses that include multiple viewpoints and are freely described in order to ease the burden on respondents.

We first considered a method that uses passage extraction techniques [6] [13] for that purpose. The techniques can extract specific parts of textual data and are used effectively in a question answering task. However, many passage extraction techniques are required in order 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 cannot extract the specific parts using the techniques.

Next, we considered classifying each textual response by using a classification model. It is necessary for the model to deal textual responses that include a single viewpoint and textual responses that include multiple viewpoints. On the other hand, the model is inductively acquired from training examples. The model requires a large number of training examples in order to distinguishes between the former responses and the latter responses, because the latter responses are composed of combinations of the former responses and there are various combinations. Therefore, the method based on a classification model is not always appropriate.

We note the acquisition of 2-class classification models corresponding to each viewpoint. The classification models can identify whether or not a textual response corresponds to each viewpoint. The models are more easily acquired from training example sets than the classification model of multiple viewpoints, because the acquisition of the former models does not require the combinations of multiple viewpoints. We can identify viewpoints that correspond to each textual response by using the models. We can grasp rough trends of the textual response by checking the number of the responses included in viewpoints. On the other hand, we also note the extraction of expressions from textual responses included in a specific viewpoint. Then, it should be noted that the extracted expressions are not always related to the specific viewpoints, because they can be related to other viewpoints that simultaneously occur with the specific viewpoint. However, the frequencies related to the other viewpoints are much smaller than the frequency of expressions related to the specific viewpoint. We can extract expressions that correspond to each viewpoint by comparing the frequency of expressions extracted in each viewpoint. We can grasp important topics corresponding to each viewpoint. Based on the above discussion, we propose concrete methods of classification and extraction in the following section.

C. Analysis method

A new analysis method acquires 2-class classification models for viewpoints in a learning phase. Also, in an inference phase, the method classifies textual responses into viewpoints based on the models and extracts expressions corresponding to viewpoints. In the following, we explain each phase in detail.

1) Learning phase: The learning phase is composed of three processes: the feature extraction process for learning, the generation process of training examples, and the inductive learning process as shown in Fig.1. The method deals with a language without word segmentation, such as Japanese.

The first process decides attributes for textual responses and generates attribute vectors for each textual response. At first, the process decomposes each textual response into words with corresponding parts of speech by using morphological analysis [5]. The process calculates their tf-idf values [11] by using Equation (1). If their tf-idf values are bigger than or equal to a threshold and their parts of speech are included in a designated

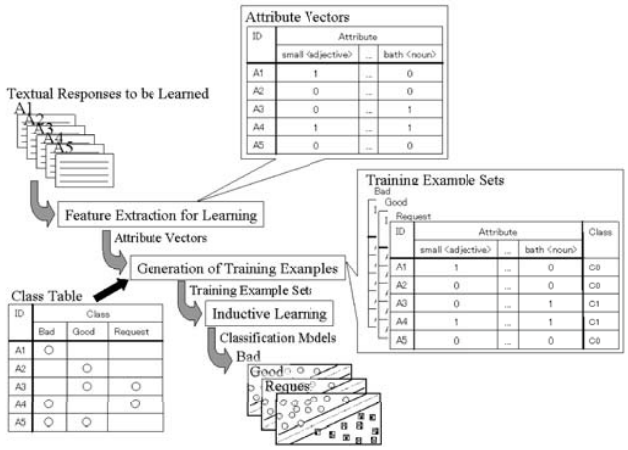

Fig. 1. Learning phase

set of parts of speech, words are selected as attributes.

$$\text{tf-idf}_i = \frac{1}{D} \cdot \log_2 \left(\frac{D}{d_i}\right) \cdot \sum_j \frac{\log_2 (t_{ij} + 1)}{\log_2 w_j} \qquad (1)$$

Here, D is the total number of textual responses, d_i is the number of textual responses that include the i-th word, w_j is the number of words included in the j-th textual response, and t_{ij} is the number of i-th words included in the j-th textual response.

Next, the process evaluates whether or not the attributes are included in a textual response. If the attributes are included, the process gives 1s to the corresponding attribute values. Otherwise, the process gives 0s to them. An attribute vector characterized by 1s and 0s is generated for the textual response. The upper right side in Fig.1 shows attribute vectors corresponding to textual responses A1 $\sim$ A5. In this figure, the response A1 includes the word "small" but does not include the word "bath". Their tf-idf values are bigger than or equal to the threshold.

The second process generates training example sets corresponding to each viewpoint. At first, the process selects a viewpoint. The process checks whether the viewpoint is assigned in a textual response or not by referring to a class table. Here, the class table is given by analysts and describes relationships between responses and viewpoints. An example of a class table is shown at lower left in Fig.1. In this figure, the response A3 includes viewpoints "Good" and "Request", but does not include a viewpoint "Bad". For each viewpoint, if the viewpoint is assigned, the process assigns the class c_1 with respect to the viewpoint to the textual response. Otherwise, the process assigns the class c_0 to it. The process integrates an attribute vector of the textual response with the class and generates a training example of the response. The generation is performed for combinations of responses and viewpoints. An example of a training example set corresponding to the viewpoint "Request" is shown at lower right in Fig.1.

The third process acquires 2-class classification models from each training example set. Each model is a model for a viewpoint. In this paper, the process uses a support vector

machine (SVM) [14] to acquire the models, because many papers [1] [7] [8] [15] have reported that an SVM gives high accuracy ratios for text classification. The process acquires 2-class classification models described with hyperplanes by using an SVM.

2) Inference phase: The inference phase is composed of three processes: the feature extraction process for inference, the class inference process, and the expression extraction process as shown in Fig.2.

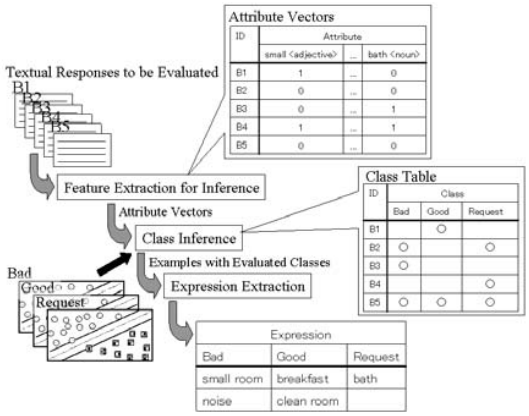

Fig. 2. Inference phase

The first process generates attribute vectors for each textual response to be evaluated. At first, the process evaluates whether or not the attributes selected by the learning phase are included in a textual response. But, the textual response is previously decomposed into words with corresponding parts of speech by using the morphological analysis. If the attributes are included in the textual response, the process gives 1s to the corresponding attribute values. Otherwise, the process gives 0s to them. The evaluation is performed for all textual responses.

The second process infers viewpoints corresponding to each textual response to be evaluated. At first, the process applies the textual responses to each 2-class classification model. The process evaluates whether classes of the textual responses corresponding to the models are c_0s or c_1s. If the textual responses are classified into c_1s, corresponding viewpoints are assigned to the textual responses. The process can assign multiple viewpoints to some textual responses. An example of an inferred class table corresponding to the responses B1 $\sim$ B5 is shown at lower right in Fig.2. In this figure, the viewpoint "Good" is assigned to the response B1 and the viewpoints "Bad" and "Request" are assigned to the response B2.

The third process extracts characteristic expressions corresponding to each viewpoint. At first, the process extracts expressions from the textual responses. Here, the expressions are words to which specific parts of speech are assigned or phrases to which specific sequences of parts of speech such as <adjective> and <noun> are assigned. The specific parts of speech and the specific sequences are designated by analysts. The process calculates the frequency of the expressions in each viewpoint. Also, the process extracts expressions which satisfy Equation (2) and Equation (3).

$$F_1 - F_2 \geq \text{minimum frequency difference} \quad (2)$$

$$\frac{F_1 - F_2}{F_1} \geq \text{minimum occupation ratio} \quad (3)$$

Here, F_1 is the maximum frequency of expressions in the viewpoints. F_2 is the next one. Equation (2) evaluates whether the expression is frequent or not for viewpoints. The equation can get rid of common expressions to some extent. On the other hand, Equation (3) evaluates whether the expression is characteristic for a specific viewpoint and the expression is not characteristic for other viewpoints. If these values are bigger than or equal to the predefined thresholds, that is, the minimum frequency difference and the minimum occupation ratio, the expression is regarded as a characteristic expression included in the viewpoint with the maximum frequency. In the following, the left term in Equation (2) is called the frequency difference and the left term in Equation (3) is called the occupation ratio. We can anticipate that the process extracts frequent and characteristic expressions corresponding to each viewpoint, because Equation (2) evaluates frequency of expressions and Equation (3) evaluates specialty of expressions for the viewpoints.

For example, five textual responses C1 $\sim$ C5 are given. Here, C1 and C2 have the expression "small room" and have the viewpoint "Bad". C3 has the expressions "small room" and "clean room", and has the viewpoints "Bad" and "Good". C4 and C5 have the expression "clean room" and have the viewpoint "Good". Also, the minimum frequency difference and the minimum occupation ratio are 2 and 0.5, respectively. Then, "clean room" occurs 3 times in the case of "Good" and occurs once in the case of "Bad". Its frequency difference is equal to 2 (=3 $-$ 1) and its occupation ratio is equal to 0.667 ($= \frac{2}{3}$). Therefore, the difference and the ratio are bigger than or equal to the thresholds. "clean room" is regarded as a characteristic expression of "Good". Similarly, "small room" is regarded as a characteristic expression of "Bad".

III. Numerical experiments

A. 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. 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, D4 = D1 $\cup$ D3, and D1 $\cap$ D3 = ϕ. The frequency of the textual responses in the data set is shown in Table I. In Table I, "Yes" indicates the number that includes a viewpoint and "No" indicates the number that does not include a viewpoint. Also, we used SVM software [4] with

TABLE I
DISTRIBUTION OF TEXTUAL RESPONSES

	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	

linear kernel, because the kernel gives comparatively high classification efficiency without adjusting its parameters.

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

TABLE II
LEXICAL FILTER

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

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-fold cross-validation experiments were applied to textual responses with attribute values and a single viewpoint. Also, the 10-fold cross-validation experiments were performed for three viewpoints. Moreover, these numerical experiments were performed for each lexical filter and each threshold. We calculated the accuracy ratio defined by Equation (4) for each viewpoint, each filter, and each threshold.

$$\text{accuracy ratio} = \frac{N_r}{N_t} \tag{4}$$

Here N_r is number of correctly classified textual responses and N_t is number of textual responses.

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 in accordance with the results of the former experiments. The number of attributes was about 1,400. We also performed 10-fold cross-validation experiments for each viewpoint and each data set, and calculated accuracy ratios.

Lastly, we extracted expressions from textual responses in D3 by using two expression methods. One is the proposed extraction method and the other is the method based on the maximum frequency. The latter method extracts expressions which have designated parts of speech and their designated sequences, and assigns viewpoints with the maximum frequency to the expressions. Here, the former method uses 3 as the minimum frequency difference and uses 0.6 as the minimum occupation ratio. The latter method extracts expressions whose number is equal to the number of expressions extracted by the former method in each viewpoint. Also, these methods extract nouns as expressions. We classified extracted expressions into three categories: valid expressions, invalid expressions, and irrelevant expressions. That is, we evaluated whether the extracted expressions are relevant to viewpoints or not by looking through them and evaluated whether the relevant expressions are valid for their viewpoints or not by reading textual responses including them. Suppose that, the expression "service" is evaluated as "Bad". "Service" is regarded as a relevant expression, because the expression is directly tied to the viewpoints "Bad", "Good", and "Request". If most textual responses including "service" describe bad service, "service" is classified into valid expressions. Otherwise, "service" is classified into invalid expressions. On the other hand, the expression "sudden" is evaluated as "Bad". "sudden" is classified into irrelevant expressions, because the expression is not directly tied to the viewpoints "Bad", "Good", and "Request".

B. Experimental Results

Table III shows results for changing lexical filters and thresholds. Each cell shows average accuracy ratios for 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.

TABLE III
ACCURACY RATIO FOR THRESHOLDS AND FILTERS IN D1

	Threshold					
Ft.	0.000	0.005	0.010	0.015	0.020	Avg.
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
Avg.	0.646	0.646	0.644	0.630	0.637	0.641

Fig.3 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 accuracy ratio, as shown in Table III.

Lastly, Table IV shows experimental results of the expression extraction. In this table, the ratios of valid expressions, defined by Equation (5), are shown.

$$\text{validity ratio} = \frac{N_v}{N_v + N_c} \tag{5}$$

Here, N_v is number of valid expressions and N_c is number of invalid expressions.

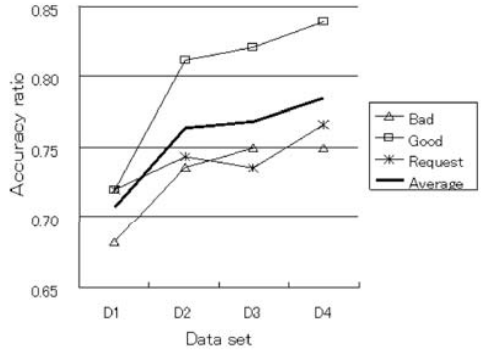

Fig. 3. Accuracy ratio for data sets

TABLE IV
VALIDATION OF EXPRESSIONS

	Bad	Good	Request	Avg.
Proposed	0.929	0.700	0.667	0.815
Maximum	0.375	0.250	0.500	0.333

C. Discussion

Setting of viewpoints: In this analysis task, we used three viewpoints. The viewpoints are not always applicable to all analysis tasks. However, the viewpoints are applicable to analysis of the voice of customer in most fields of service industry. For most fields of service industry, large amounts of data are available. 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. This is why the lexical filters including adjectives and nouns provided relatively high accuracy 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 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 deals with all parts of speech. Therefore, the L9 filter gives the highest accuracy ratio. However, the filter causes an increase in the number of attributes. The L5 or L7 filters should be used if calculation speed and memory size are important considerations. This is why the numbers of their attributes are comparatively small and their average accuracy ratios are almost equal to those for 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 2-class 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 accuracy ratios. The thresholds are not particularly sensitive. The reason for 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 accuracy ratio becomes higher as the number of textual responses increases. The case of D4 is about 8% higher than the case of D1. This is why a more appropriate 2-class classification model is acquired by using many textual responses. On the other hand, the accuracy curves have not converged. If more training examples are used, the proposed method may give a higher accuracy ratio.

Validity of extracted expressions: The proposed method gives a comparatively high validity ratio. Also, the method extracts more valid expressions than the method based on the maximum frequency. In the case of the latter method, the extracted expressions tend to be common expressions and they are not always characteristic expressions for specific viewpoints. The proposed method extracts characteristic expressions of viewpoints by referring to the frequency difference and the minimum occupation ratio.

On the other hand, the proposed method extracts comparatively many irrelevant expressions. However, we can easily judge whether the expressions are relevant or not. We can grasp important expressions corresponding to viewpoints. Also, it is not always necessary to check all extracted expressions because it is impractical to implement all measures related to the expressions. Therefore, we do not care even if there are many irrelevant expressions.

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

IV. SUMMARY AND FUTURE WORK

This paper proposed 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 showed that accuracy ratios based on 2-class classification models were improved by an increase in training examples. We also showed that the method extracted valid expressions of textual responses. We think the method is efficient for analyzing textual responses to questionnaires and the acquired 2-class classification models can be used to analyze similar textual responses.

In the future, we intend to develop a system in which the method is applied via a graphical user interface. The system provides an environment in which many analysts can easily analyze the textual responses to questionnaires. Also, we will 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," *Proc. of the 10th Intl. Sympo. on String Processing and Information Retrieval*, 2003, Manaus, Brazil, pp. 183-196.

[2] R. Feldman and H. Hirsh, "Mining Text using Keyword Distributions," *Journal of Intelligent Information Systems*, vol. 10, no. 3, pp. 281-300, 1998.
[3] M. A. Hearst, "Untangling Text Data Mining," *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, Montreal, Canada, pp. 20-26.
[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/papers/guide/guide.pdf, 2003.
[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," *Proc. of the Pacific Association for Computational Linguistics 2001*, 2001, Kitakyushu, Japan, pp. 127-135.
[6] A. Ittycheriah, M. Franz, W. -J. Zhu, and A. Ratnaparkhi, "IBM's Statistical Question Answering System," *Proc. of the 9th Text Retrieval Conf*. 2000, Gaithersburg, Maryland, USA, pp. 229-234.
[7] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proc. of the 10th European Conf. on Machine Learning*, 1998, Dorint-Parkhotel, Chemnitz, Germany, pp. 137-142.
[8] T. Joachims, "Transductive Inference for Text Classification using Support Vector Machines," *Proc. of the 16th Intl. Conf. on Machine Learning*, 1999, Bled, Slovenia, pp. 27-30.
[9] S. Sakurai, Y. Ichimura, A. Suyama, and R. Orihara, "Acquisition of a Knowledge Dictionary for a Text Mining System using an Inductive Learning Method," *Proc. of the IJCAI 2001 Workshop on Text Learning: Beyond Supervision*, 2001, Seattle, Washington, USA, pp. 45-52.
[10] S. Sakurai and A. Suyama, "An E-mail Analysis Method based on Text Mining Techniques," *Applied Soft Computing*, vol. 6, no. 1, pp. 62-71, 2005.
[11] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval," Mcgraw-Hill, New York, USA, 1983.
[12] P. -N. Tan, H. Blau, S. Harp, and R. Goldman, "Data Mining of Service Center Call Records," *Proc. of the 6th Intl. Conf. on Knowledge Discovery and Data Mining*, 2000, Boston, Massachusetts, USA, pp. 417-423.
[13] S. Tellex, B. Katz, J. Lin , and A. Fernandes, "Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering," *Proc. of the 26th Intl. Conf. on Research and Development in Information Retrieval*, 2003, Toronto, Canada, pp. 41-47.
[14] V. N. Vapnik, "The Nature of Statistical Learning Theory," Springer, New York, USA, 1995.
[15] Y. Yang and X. Liu, "A Re-examination of Text Categorization Methods," *Proc. of the 22nd Intl. Conf. on Research and Development in Information Retrieval*, 1999, Berkeley, California, USA, pp. 15-19.

Shigeaki Sakurai received an MS degree in mathematics and a Ph.D. degree in industrial administration from Tokyo University of Science, Japan, in 1991 and 2001, respectively. He was a Professional Engineer of Japan in the field of information engineering in 2004.

He is a research scientist at the System Engineering Laboratory, Corporate Research & Development Center, Toshiba Corporation. His research interests include data mining, soft computing, and web technology.

Dr. Sakurai is a member of IEICE, SOFT, and JSAI.

Ryohei Orihara received a BS degree, an MS degree, and a Ph.D. degree in engineering from the University of Tsukuba, Japan, in 1986, 1988 and 1999, respectively.

He is the laboratory leader at the HumanCentric Laboratory, Corporate Research & Development Center, Toshiba Corporation. He is also a part-time associate professor at Tokyo Institute of Technology, Japan. His research interests include machine learning, creativity support systems, analogical reasoning, metaphor understanding, data mining and text mining.

Dr. Orihara is a member of IPSJ, JSAI, and JSSST.