A Classification Method of Open-ended Questionnaires using Category-based Dictionary from Sampled Documents

Keiichi Hamada[1], Masanori Akiyoshi[2], Masaki Samejima[1] and Hiroaki Oiso[3]
[1]*Department of Information Science and Technology, Osaka University, 2-1 Yamadaoka, Suita, Osaka, Japan*
[2]*Faculty of Applied Information Science, Hiroshima Institute of Technology, 2-1-1 Miyake saeki-ku, Hiroshima, Japan*
[3]*Codetoys K. K, 8F Dojima building, 2-6-8 Nishitenma, Kita-ku, Osaka, Japan*

Keywords: Open-ended Questionnaires, Typical Words Involvement Degrees, Co-occurrence Pattern.

Abstract: This paper addresses a classification method of open-ended questionnaires using a category-based dictionary. Different from other classification methods, our proposed method introduces a category-based dictionary which is generated from a small set of categorized samples. This category-based dictionary is used to judge questionnaire categories with tf-idf(term frequency inverted document frequency) and co_tf-idf(co-occurrence tf-idf). Experimental questionnaires about a university lecture show that 71% of these questionnaires are classified accurately.

1 INTRODUCTION

Recently, various types of questionnaires (*e.g.* close-ended, open-ended) are collected to improve contents or services. In open-ended questionnaires, people write opinions in their own words that are expected to involve significant information. Analysts classify the questionnaires into categories for grasping which types of opinions are useful. However, analysts spend a lot of time for reading the questionnaires in order to classify them. In order to classify them into categories efficiently, this paper addresses an efficient classification of open-ended questionnaires.

One of the document classification methods is text mining(Berry, 2003) which analyzes and classifies large amounts of text data (*e.g.* news articles(Atkinson and Van der Goot, 2009), patent documents(Tseng et al., 2007)). SVM(Support Vector Machine) and clustering are also the popular machine learning techniques that are useful for questionnaire classification by a number of questionnaires(Zhang and Lee, 2003)(Chim and Deng, 2008). However, the questionnaires include grammatical errors and typos, and are not accumulated for the classification, while make it difficult to apply the text mining and the machine learning classifier.

tf-idf(term frequency inverted document frequency) (Salton and Buckley, 1988) that indicates characteristics of words is a successful approach to document classification(Ramos, 2002). It is possible to find characteristic words in each category by tf-idf. Documents can be classified by emphasizing the characteristic words in comparing documents. But some classification methods using tf-idf(Trieschnigg et al., 2009) need tuning parameters to be determined manually for every targets, which are difficult to decide based on questionnaires that are not accumulated. In addition, text mining techniques using co-occurrence patterns have been proposed for supporting document classification (*e.g.* a keyword extraction algorithm using a set of co-occurrence between each term and frequent terms(Matsuo and Ishizuka, 2004)).

Based on our investigation on contents of questionnaires in each category, we found that there are characteristic words and co-occurrence patterns. So, we propose the classification method using tf-idf which considers words and co-occurrence patterns. In order to reflect the characteristics of the categories to tf-idf, the proposed method uses samples of questionnaires categorized by analysts in advance and calculates "typical words involvement degree" between an inputted questionnaire and each category based on the samples. The questionnaire is classified into some categories that have high typical words involvement degrees with the questionnaire.

Hamada K., Akiyoshi M., Samejima M. and Oiso H..
A Classification Method of Open-ended Questionnaires using Category-based Dictionary from Sampled Documents.
DOI: 10.5220/0003971601930198
In *Proceedings of the 14th International Conference on Enterprise Information Systems* (ICEIS-2012), pages 193-198
ISBN: 978-989-8565-12-9

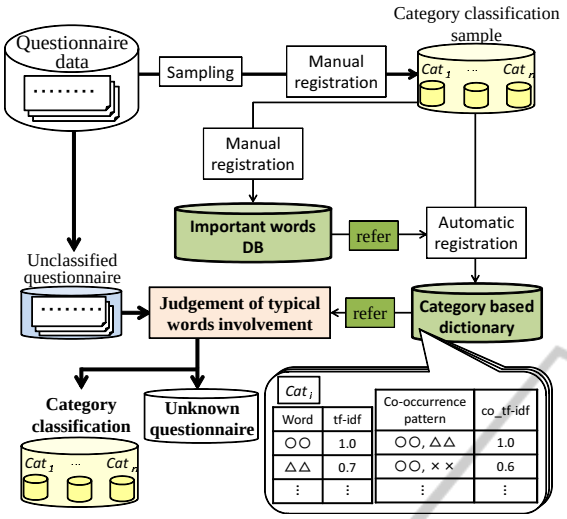

Figure 1: Outline of analysis support system.

2 CLASSIFICATION SUPPORT SYSTEM

2.1 System Overview

Figure 1 shows the outline of the classification support system. In advance, analysts pick up parts of questionnaire data, and classify the sentences of the questionnaires into categories Cat_i. These sentences are defined as "category classification samples". The goal of this research is to classify questionnaires using as few "category classification samples" as possible, because there are not many questionnaires and a few "category classification samples" make analysts easy to classify to categories they expect.

For the classification, this system uses only noun, verb, and adjective that indicate the contents of the questionnaires. The questionnaires are classified by using similarity between "category classification samples" and the questionnaires.

Because each sentence may have different contents in the questionnaire, this system does not classify the questionnaire, but a sentence in the questionnaire. In case that sentences in a questionnaire have contents of different categories(*e.g.* Cat_1 and Cat_2), the questionnaire is classified into the categories(*e.g.* Cat_1 and Cat_2).

Being inputted to the system, a questionnaire is separated to sentences for the above reason. And the system removes words in "general DB" which is a database of stop words(*e.g.* do, be).

In order to classify questionnaires with high accuracy, we need to define how to decide the similarity.

2.2 Approach

The typical method to decide similarities is to find common words in "category classification samples". But, the common words in these samples do not indicate characteristics of words. So, questionnaires are often classified wrongly.

Because analysts classify questionnaires based on the meaning of the categories, the questionnaires are similar to each other in the same category. According to investigate questionnaires, it is considered that there are some features of questionnaires as below.

- A questionnaire includes typical words which are words or synonyms included in other questionnaires in the same category. And, co-occurrence patterns consisting of the typical words appear in a questionnaire.
- Some categories have important words that characterize the categories.

We define "typical words involvement degree" as the similarity to classify questionnaires based on the above features. This degree is an index based on how many typical words and co-occurrence patterns appears in a questionnaire. Also, this degree is calculated by "category-based dictionary" that consists of typical degrees of words and co-occurrence patterns. And "important words DB" which is a database of words that analysts consider to characterize a category. "Category-based dictionary" is constructed by category classification samples, and "important words DB" is manually constructed by analysts.

3 JUDGMENT OF TYPICAL WORDS INVOLVEMENT DEGREES

3.1 Construction of "Category-based Dictionary"

As we mentioned before, a sentence in a questionnaire often includes typical words and co-occurrence patterns. It is necessary to calculate typical degrees of words and co-occurrence patterns. As for the construction of "category-based dictionary", $tf-idf$ and $co\_tf\text{-}idf$(co-occurrence $tf\text{-}idf$) are used as these degrees as shown in (1), (2), and Table 1.

$$tf\text{-}idf(w_j) = tf(w_j) \times \log \frac{N}{df(w_j)} \quad (1)$$

$$co\_tf\text{-}idf(w_k, w_l) = tf_{co}(w_k, w_l) \times \log \frac{N}{df_{co}(w_k, w_l)} \quad (2)$$

Table 1: Definition and condition of symbols.

Symbol	Definition/Condition
j,k,l	$1 \leq j,k,l \leq Const$ (*Const* is the number of words in "category classification samples")
w_j	a word
$tf(w_j)$	the number of occurrences for w_j in a category
N	the number of all categories
$df(w_j)$	the number of categories w_j appears
$tf_{co}(w_k,w_l)$	the number of occurrences which w_k and w_l appear together in a category
$df_{co}(w_k,w_l)$	the number of categories w_k and w_l appears

"Important words DB" is used for reflecting important words in typical words to typical degrees. When a word is in a set "X" of words in "important words DB", importance degrees of typical words($C_1, C_2 \geq 1$) are weighted to $tf-idf$ and $co\text{-}tf\text{-}idf$ as shown in (3), (4).

$$tf\text{-}idf(w_j) = C_1 \times tf(w_j) \times \log \frac{N}{df(w_j)} \quad (3)$$

$$co\text{-}tf\text{-}idf(w_k,w_l) = C_2 \times tf_{co}(w_k,w_l) \times \log \frac{N}{df_{co}(w_k,w_l)} \quad (4)$$

$$s.t. \quad w_j \in X, w_k \text{ or } w_l \in X$$

Finally, because $tf-idf$ and $co\text{-}tf\text{-}idf$ in "category-based dictionary" may not be normalized, we make them normalize into a range of 0 and 1 for each category.

3.2 Calculation of Typical Words Involvement Degrees

"Typical words involvement degree" should be based on both typical words and co-occurrence patterns. So, typical words involvement degree R for a category is decided by an average of R_t and R_{cot} as shown in (5) - (9), and Table 2.

$$R = \frac{R_t + R_{cot}}{2} \quad (5)$$

$$R_t = \frac{R'_t}{n_t} \quad (6)$$

$$R_{cot} = \frac{R'_{cot}}{n_{cot}} \quad (7)$$

$$R'_t = \sum_{1 \leq m \leq n_t} tf\text{-}idf(w_m) \quad (8)$$

$$R'_{cot} = \sum_{1 \leq p < q \leq n_t} co\text{-}tf\text{-}idf(w_p,w_q) \quad (9)$$

Table 2: Definition of symbols.

Symbol	Definition
n_t and n_{cot}	the number of common words and common co-occurrence patterns between "category-based dictionary" and the sentence

3.3 Judgment by Typical Words Involvement Degree and It's Example

A questionnaire is classified into the target category by judging whether typical words involvement degree R is higher than the threshold value S_i as shown in (10), (11), and Table 3.

$$S_i = \frac{sum_i}{num_i} \quad (10)$$

$$sum_i = \sum_{O_r \in Cat_i} R_{O_r} \quad (11)$$

Table 3: Definition of symbols.

Symbol	Definition
num_i	the number of sentences in Cat_i
O_r	the sentence r
R_{O_r}	typical words involvement degree R for the sentence r

Figure 2 shows an example of typical words involvement judgment. This target sentence includes word "History" and co-occurence pattern "Calculator, History" and "Calculator, Knowledge". The typical words involvement degree R for Cat_i is 0.6965 which is higher than $S_i = 0.5$. Thus, the questionnaire including this target sentence is classified to Cat_i.

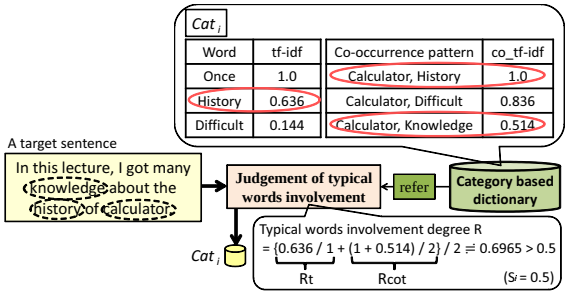

Figure 2: An example of typical words involvement judgment.

3.4 Automatic Acquisition Method of Importance Degrees C_1, C_2

The system needs to determine C_1 and C_2 automatically based on relations between category classifi-

cation samples and "category-based dictionary", because it is difficult for analysts to determine C_1 and C_2 in advance.

A correct sentence which is classified into a target category in category classification samples should have a high typical words involvement degree because the sentence has common words with the target category. However, an incorrect sentence which is not classified into the target category should have a low typical words involvement degree because the sentence lacks many related words to the topic. In addition, it should have low standard deviation of typical words involvement degrees for correct sentences because typical words involvement degrees of the correct sentences are similar each other.

This system determines suitable C_1, C_2 under the conditions that (1) a correct sentence has higher typical words involvement degree than that every incorrect sentence has, and (2) the standard deviation of typical words involvement degrees for correct sentences is as low as possible as shown in Figure 3.

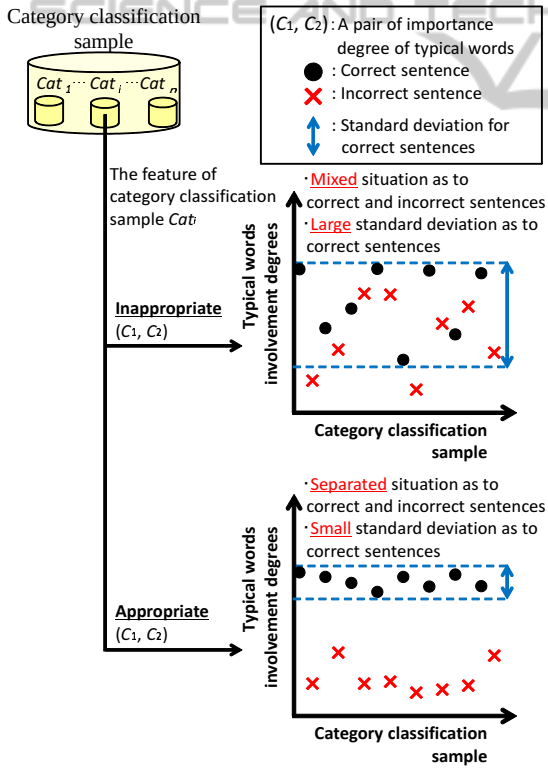

Figure 3: Desirable situation as to correct and incorrect opinions in category classification sample.

Based on this, the automatic acquisition algorithm is as follows.

1. For each combination C_1 and C_2, calculate the number of incorrect sentences as *number*, that have higher typical words involvement degree than the minimum of correct sentences in the condition of $1.0 \leq C_1 \leq C_{max} C1.0 \leq C_2 \leq C_{max}$.
2. Create class P including the combination C_1 and C_2 which has lowest *number*.
3. For each combination of C_1 and C_2 in P, calculate the standard deviation as *sd*, of typical words involvement for correct sentences.
4. Determine the combination of C_1 and C_2 that lead the lowest *sd*.

4 EVALUATION

4.1 Results of Experiment

We executed an experiment to evaluate effectiveness of our proposed method. The target questionnaires data are questionnaires on "Give what you learned or what you feel about calculators' history" in a university lecture. The number of questionnaires is 165 with an average of 3.0 sentences per a questionnaire and with an average of 8.3 words per a sentence. We provided 10 categories in advance, and each extraction has 20 questionnaires as "category classification samples" that each category has at least two sentences. An average of sentences including category classification samples per an extraction is 67.0. In this experiment, we calculated the average of results for 5 times of extractions. Evaluation criteria are the recall rate, precision rate, and F measure as shown in Table4.

For separating a sentence to words, we used morphological analysis "Japanese morphological analysis" in Yahoo!Japan Developer Network. In addition, we defined $C_{max} = 10.0$ in automatic acquisition of importance degrees C_1 and C_2.

We compared the effectiveness by the proposed method to ones by other methods: the method of clustering and the method of SVM. In the SVM, the numbers of occurrences for the top five frequently-used words are used as vector elements.

Table 4: Definition of evaluation criteria.

Criteria	Definition
recall rate	*recall* = [$\sum_i$ the number of classified questionnaires correctly using a method for Cat_i] / [$\sum_i$ the number of classified questionnaires manually for Cat_i]
precision rate	*precision* = [$\sum_i$ the number of classified questionnaires correctly using a method for Cat_i] / [$\sum_i$ the number of classified questionnaires manually for Cat_i]
F measure	[2× *recall* × *precision*] / [*recall* + *precision*]

Figure 4 shows that our proposed method is the best classification accuracy of the three methods. The accuracy allows analysts to understand the number of questionnaires in each category and the contents of categories reading "category classification samples".

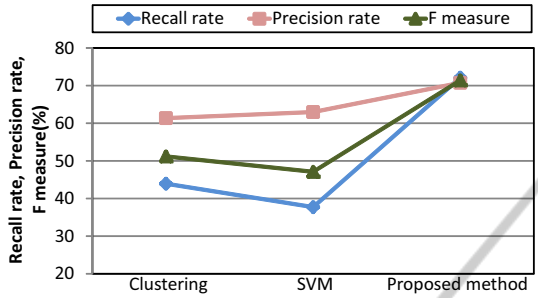

Figure 4: Result of classification experiment.

4.2 Discussion

Figure 5 shows F measures in increasing the number of category classification when the number of samples is changed from 20 to 100.

Even if category classification samples are increased, the accuracy of the proposed method is better than the method of SVM. So, it is confirmed that the proposed method does not depend on the number of category classification samples. Thus, our proposed method can classify questionnaires at a reduced cost.

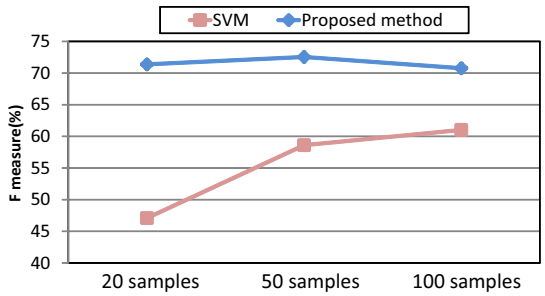

Figure 5: Result of classification experiment by using increased category classification sample.

In order to verify the adequacy of important words, we compared our proposed method to the method without important words as $C_1 = C_2 = 1.0$. Figure 6 shows the result for Data Set "A"(F measure = 72.4%) which has the best F measure in 5 data sets using our proposed method. Figure 6 shows that recall rate is increased by 4%, precision rate is increased by 11%, F measure is increased by 8%.

Table 5 shows C_1 and C_2 for each category including more than 20 questionnaires and F measure by our proposed method. C_1 and C_2 are decided to 1.0

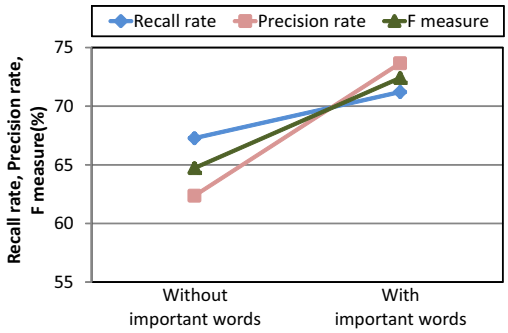

Figure 6: Result of comparison with/without important words.

in Category 4 because it does not have any important words. Table 6 shows the result of sensitivity analysis of changes in C_1 and C_2 for Category 6.

Table 6 shows that the F measure by our proposed method is 5% lower than that one by using the best C_1 and C_2 in Category 6. Thus, we can prove important words are effective, and it is possible to improve classification accuracy by the improvement of the automatic acquisition method of C_1 and C_2.

Table 5: C_1 and C_2 for each category and F measure.

Category Number	1	2	3	4	5	6
C_1	6.8	2.0	1.0	1.0	10.0	3.0
C_2	1.0	3.4	1.2	1.0	10.0	1.0
F measure(%)	76.6	67.9	82.5	55.4	85.1	75.0

Table 6: F measure(%) in category 6.

C_1 \ C_2	1	3	5	7	9	10
1	68.1	76.0	76.0	76.0	76.0	76.0
3	75.0	80.0	78.6	77.2	77.2	77.2
5	75.0	78.6	77.2	77.2	77.2	77.2
7	75.0	78.6	77.2	77.2	77.2	77.2
9	75.0	77.2	77.2	77.2	76.7	76.7
10	75.0	77.2	77.2	77.2	76.7	76.7

Figure 7 shows the results for each category in Data Set "A" and Data Set "B"(F measure = 68.6%). Table 7 shows the number of manually classified questionnaires for each category that includes more than 20 questionnaires. Both of these results show that the classification accuracy differs in each category. The F measure for Category 5 in both data sets is about 85%, but Category 4 is about 50%. This difference depends on whether the category's content is clear or not. Table 8 and 9 show the sentences included in Category 4 and 5, respectively. The sentences in categories which have clear contents *e.g.* Category 5 includes clear words that characterize the category, and these clear words have a high value of

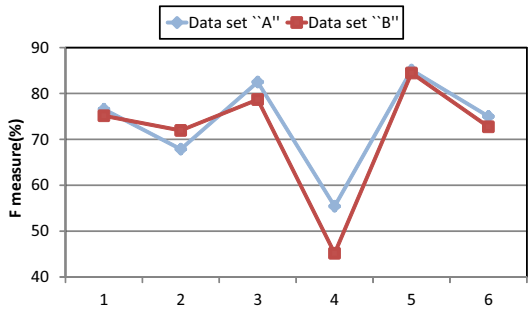

Figure 7: Result of classification experiment for 2 category classification sample patterns.

tf-idf and co_tf-idf. So, the sentences with clear contents can be classified accurately. On the other hand, in categories which is confused contents such as Category 4, the system can not identify words that characterize the category, and the words appear in other categories. Thus it is difficult to classify in these categories.

Table 7: The number of correctly categorized opinions for each category.

Category Number	1	2	3	4	5	6
Data set “A”	89	53	46	39	25	22
Data set “B”	89	60	36	43	21	30

Table 8: Examples of category 4 “Dangerousness and how to deal with information society”.

I would like to learn how to deal with overflooding information.
But such information is not always right.
But I do not know whether it is good to depend on information in the web.

Table 9: Example of category 5 “Electronic tag technology”.

And it is nice to know electronic tag is used in book stores’ security system.
I understood that electronic tags are used everywhere.
I think there will be no more cash registers in the future because electronic tags are used for in all goods.

5 CONCLUSIONS

This paper addressed the classification method of open-ended questionnaire using category-based dictionary from category classification samples. Our proposed method uses typical words involvement degree which is an index that measures the number of typical words and co-occurrence patterns that characterize a category. By applying our proposed method to questionnaires about a university lecture, 71% of these questionnaires are classified accurately. As a result of experiments, the clearer the contents are, the more accurate the proposed method can classify the questionnaires.

REFERENCES

Atkinson, M. and Van der Goot, E. (2009). Near real time information mining in multilingual news. In *Proceedings of the 18th international conference on World wide web*, WWW ’09, pages 1153–1154, New York, NY, USA. ACM.

Berry, M. (2003). *Survey of Text Mining : Clustering, Classification, and Retrieval*. Springer.

Chim, H. and Deng, X. (2008). Efficient phrase-based document similarity for clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20:1217–1229.

Matsuo, Y. and Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169.

Ramos, J. (2002). Using TF-IDF to Determine Word Relevance in Document Queries. Technical report, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855e.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing and management*, 24(5):513–523.

Trieschnigg, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., and Rebholz-Schuhmann, D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics (Oxford, England)*, 25(11):1412–1418.

Tseng, Y.-H., Lin, C.-J., and Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Inf. Process. Manage.*, 43:1216–1247.

Zhang, D. and Lee, W. S. (2003). Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR ’03, pages 26–32, New York, NY, USA. ACM.