IADIS International Journal on WWW/Internet
Vol. 12, No. 2, pp. 102-114
ISSN: 1645-7641

# ESTIMATION OF REVIEW HELPFULNESS BY CONTENT COVERAGE AND WRITING STYLE

Akihide Bessho. *University of Hyogo. 2167 Shosha, Himeji, Hyogo 671-2280, Japan*

Takayuki Yumoto. *University of Hyogo. 2167 Shosha, Himeji, Hyogo 671-2280, Japan*

Manabu Nii. *University of Hyogo. 2167 Shosha, Himeji, Hyogo 671-2280, Japan*

Kunihiro Sato. *University of Hyogo. 2167 Shosha, Himeji, Hyogo 671-2280, Japan*

### ABSTRACT

Customer reviews are helpful information to decide what to buy in EC sites. However, some customer reviews are not helpful. Therefore, users must choose helpful ones. This is difficult especially for users who do not have enough knowledge about the products. In this paper, we propose methods to classify customer reviews into helpful and unhelpful. To classify them, we focus on content coverage and writing style of reviews. Coverage expresses how many important words are contained by the reviews. Writing style is expressed as “formal” or “informal”, and it is classified by a machine learning technique. We made test data from user votes in Amazon.co.jp, and evaluated our methods. The accuracy of our rule-based method was 0.69. However, the method using coverage does not work well when many reviews have not been written yet. We analyzed the effects of the number of reviews, and confirmed that our rule-based method is less affected and achieves better accuracy than the method using only coverage.

### KEYWORDS

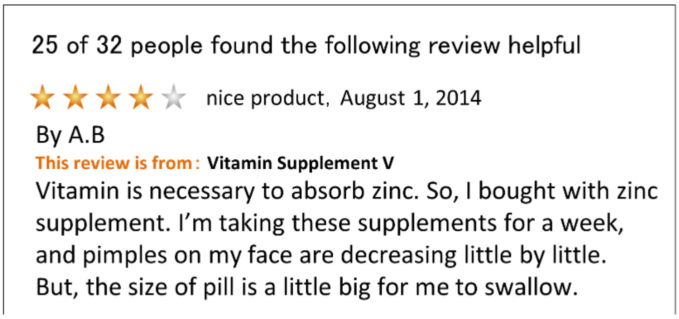
Customer review, Helpfulness, Coverage, Writing style, Support vector machine

## 1. INTRODUCTION

Recently, many people use e-commerce sites to purchase products. Customer reviews play an important role to help customers obtain information about products. For example, on Amazon.com, users can write reviews of products and rate them with 1-5 stars. Figure 1 shows an example of a customer review. Amazon.com summarizes the results and shows the distribution of the number of stars. The users can roughly know whether the product is good whether from it. If they want to know more details, they must read the reviews.

For efficient understanding of reviews, there are many methods such as estimating the sentiment of reviews (Turkey et al. 2002, Gamon 2004), estimating ratings for products (Pang et al. 2005), and summarizing reviews (Hu et al. 2004, Zhuang et al. 2006). However, we often find reviews that are not helpful. Furthermore, there are review spams (Liu et al. 2007, Jindal et al. 2008). Against these problems, some sites provide voting facilities for reviews. For example, users can vote on whether the review is helpful or not. In Figure 1, 25 users voted the review as helpful, and 7 users voted as unhelpful.

Although such facilities help users to find helpful or unhelpful reviews, some sites do not provide these facilities, and newly posted reviews have no votes. To solve these problems, some researchers analyze reviews from the aspect of helpfulness (Kim et al. 2006, Danescu-Niculescu-Mizil et al. 2009, Lu et al. 2010, and Mudambi et al. 2010). We focus on the coverage and writing style of reviews. Coverage is a measure expressing how many important words are contained in a review in comparison with other reviews for the same product. We define coverage as a frequency-based measure. Writing style expresses whether the review text is described in formal style or informal style. We built a writing style classifier by support vector machines. We assume that a review whose coverage is high and is described in formal style is helpful. From this idea, we propose methods to estimate the helpfulness of reviews.



25 of 32 people found the following review helpful

★★★★☆ nice product, August 1, 2014

By A.B

This review is from: **Vitamin Supplement V**

Vitamin is necessary to absorb zinc. So, I bought with zinc supplement. I'm taking these supplements for a week, and pimples on my face are decreasing little by little. But, the size of pill is a little big for me to swallow.

Figure 1. Example of customer review

## 2. RELATED WORK

To obtain information from reviews efficiently, there are two approaches. The first approach is summarization. In the methods of Hu et al. (2004) and Zhuang et al. (2006), they estimate polarity (positive or negative) of reviews and count the number of positive reviews and negative ones for each feature (e.g. picture quality and size), and make text summaries. Liu et al. (2005) also used a similar number to their work, but they visualized the results as bar charts for comparing competitive products. Their approach is good for grasping the overview of all reviews. In this approach, however, the summarization quality may be reduced when there are low-quality reviews or review spams. Liu et al. (2007) proposed a method to find low quality reviews, and Jindal et al. (2008) proposed a method to detect review spams.

The second approach is to find helpful reviews. Mudambi et al. (2010) analyzed the reviews of six products on Amazon.com and pointed out the relation between review helpfulness and ratings for the products. Furthermore, Danescu-Niculescu-Mizil et al. (2009) focused on the differences in the average ratings. Kim et al. (2006) and Lu et al. (2010) proposed methods to estimate helpfulness by machine learning techniques. They focused on basic features on text such as n-gram, length of reviews and HTML structure. Furthermore, Lu et al. focus on information related to the reviewers. We also estimate helpfulness and we focus on the textual features of reviews. We especially focus on comprehensive measures for users. The coverage and writing style are persuasive measures to explain why a review is helpful.

## 3. CUSTOMER REVIEW

Some EC sites have review systems. Users can write their impressions as customer reviews on each product, and they can rate the product. Some sites also provide voting systems for reviews. For example, users can vote each review as helpful or not on Amazon.com. By using the number of votes, we define the support score of a review. $\sup(r_i)$ is the support score of review $r_i$ and is defined as:

$$\sup(r_i) = vote_{good}(r_i) - \left(vote_{all}(r_i) - vote_{good}(r_i)\right)$$

$vote_{all}(r_i)$ is the number of all votes for review $r_i$, and $vote_{good}(r_i)$ is the number of helpful votes to $r_i$. A higher support score means the review is more helpful. In Figure 1, $vote_{good}(r_i) = 25$, $vote_{all}(r_i) = 32$, and $\sup(r_i) = 18$.

For our research, we collected customer reviews of 51 products that have at least 80 reviews from Amazon.co.jp. We show the overview of the collected data in Table 1. We automatically labeled some reviews as helpful or unhelpful as follows:

1. If $\sup(r_i) \geq 10$, review $r_i$ is labeled as helpful.
2. If $\sup(r_i) \leq -10$, $r_i$ is labeled as unhelpful.

In this paper, we try to estimate whether a given review is helpful or unhelpful, and the labels are regarded as correct answers.

From Table 1, you can see that the number of helpful reviews is more than double that of the unhelpful reviews. To avoid bias, we randomly selected helpful reviews and removed the labels until the number of the helpful reviews and the unhelpful reviews became the same. We used the 492 labeled reviews for testing our methods. We show the number of all reviews (All) and the labeled reviews (Labeled) per product in Table 2. All reviews are used for computing coverage and the labeled reviews are used for evaluation. Table 3 shows the distribution of stars for the labeled reviews. This means that the number of good reputations and the number of bad reputations are almost same.

Table 1. Overview of collected reviews

| | |
|---|---:|
| Date | June 27, 2014 |
| Number of products | 51 |
| Number of reviews | 18617 |
| Number of helpful reviews | 712 |
| Number of unhelpful reviews | 246 |

# 4. COVERAGE

## 4.1 Computing Coverage

We model a customer review as a set of keywords. $V_{r_i}$ denotes a set of keywords of review $r_i$. We regard nouns and noun phrases as keywords. To find keywords, we first find independent nouns from each review text with a morphological analyzer. We used a Japanese morphological analyzer called MeCab[1](Kudo et al., 2004). Next, we find a sequence of independent nouns, and regarded it as a noun phrase. These noun phrases are keywords. Independent nouns that any noun phrases do not contain are keywords.

We consider that each keyword has different importance. We assume that the more reviews contain a keyword, the more important the keyword is. To express this importance, we use the document frequency (DF), which is often used for information retrieval. Here, we define $DF(w,S)$ as the number of reviews containing a keyword $w$ in a set of reviews $S$.

We define coverage as follows.

$$Cov(r_i) = \frac{\sum_{w \in V_{r_i}} DF(w, R \setminus \{r_i\})}{\sum_{w \in V_R} DF(w, R \setminus \{r_i\})} \quad (1)$$

$V_R$ is a set of keywords that appear in the set of the reviews $R$. $V_R$ satisfies the following equation:

Table 2. Number of reviews per product

| Product | All | Labeled | Product | All | Labeled |
|---|---|---|---|---|---|
| Air cleaner1 | 594 | 8 | Medicine1 | 162 | 7 |
| Air cleaner2 | 184 | 5 | Medicine2 | 460 | 34 |
| Bath scales | 300 | 1 | Microphone | 349 | 10 |
| Blanket | 381 | 3 | Mobile battery | 1836 | 27 |
| Car navigation system1 | 320 | 12 | Mouse1 | 727 | 20 |
| Car navigation system2 | 111 | 2 | Mouse2 | 1433 | 4 |
| Cleanser | 224 | 2 | Mouse3 | 397 | 2 |
| Digital camera | 168 | 7 | Mouse4 | 194 | 5 |
| Earphone | 144 | 6 | MP3 player | 130 | 17 |
| Electric pot | 152 | 5 | Office chair | 132 | 4 |
| Electric shaver | 422 | 5 | Pressure cooker | 176 | 9 |
| Electric toothbrush | 474 | 7 | Printer | 167 | 4 |
| Ethanol | 194 | 6 | Roomba | 179 | 3 |
| Game | 93 | 14 | Soap1 | 106 | 1 |
| HDD1 | 774 | 20 | Soap2 | 95 | 3 |
| HDD2 | 299 | 10 | Supplement1 | 337 | 8 |
| HDMI cable | 817 | 29 | Supplement2 | 597 | 28 |
| Heater | 265 | 2 | Supplement3 | 335 | 14 |

[1] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

| | | |
|---|---|---|
| IC recorder | 359 | 4 |
| iPod classic | 241 | 28 |
| iPod nano | 217 | 9 |
| iPod shuffle | 187 | 3 |
| iPod touch | 394 | 29 |
| Keyboard | 799 | 10 |
| Mechanical pencil | 93 | 11 |
| Sweeper1 | 84 | 5 |
| Sweeper2 | 715 | 12 |
| Sweeper3 | 106 | 1 |
| Sweeper4 | 109 | 5 |
| Sweeper5 | 665 | 4 |
| Sweeper6 | 161 | 9 |
| Trimmer | 539 | 9 |
| Watch | 220 | 9 |

Table 3. Distribution of stars in dataset

| Number of stars | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Number of reviews | 155 | 66 | 57 | 59 | 155 |

$$V_R = \bigcup_{r_i \in R} V_{r_i} \quad (2)$$

$R \backslash \{r_i\}$ is a set of the reviews for a product except for review $r_i$. In formula (1), the denominator is the summation of the importance of all the keywords that appear in the set of reviews $R$, and the numerator is the summation of the importance of all the keywords that appear in review $r_i$. When we compute the coverage of $r_i$, we do not use $r_i$ for DF. Otherwise, the coverage becomes large when $r_i$ is too long, and $r_i$ contains unimportant words such as advertising words.

## 4.2 Deciding Threshold for Estimating Helpfulness

We set a threshold for coverage. If the coverage is greater than or equal to this threshold, we regard the review as helpful. The distributions of coverage depend on sets of reviews, and they have many differences. Therefore, we use a relative threshold instead of an absolute threshold.

To find an optimal threshold, we try the mean, median, and top 25% as the candidates. To compare each threshold, we use three measures: precision for helpful ($\mathrm{precision_h}$), precision for unhelpful ($\mathrm{precision_u}$), and accuracy. They are defined as follows:

$$\mathrm{precision_h} = \frac{\text{number of helpful reviews that are classified correctly}}{\text{number of reviews that are estimated as helpful}} \quad (3)$$

$$\mathrm{precision_u} = \frac{\text{number of unhelpful reviews that are classified correctly}}{\text{number of reviews that are estimated as unhelpful}} \quad (4)$$

$$\text{accuracy} = \frac{\text{number of reviews that are classified correctly}}{\text{number of reviews in test data}} \quad (5)$$

We show the results in Table 4. The mean and median have almost the same performance and have better precision for unhelpful than the top 25%. However, the top 25% has better precision for helpful and accuracy. In practical use, users want helpful reviews rather than unhelpful ones. Therefore, we consider that precision for helpful is more important than precision for unhelpful, and we use the top 25% as the threshold of coverage.

Table 4. Results for different coverage thresholds

| Threshold | Precision for helpful | Precision for unhelpful | Accuracy |
|---|---:|---:|---:|
| Mean | 0.64 | 0.71 | 0.67 |
| Median | 0.64 | 0.71 | 0.67 |
| Top 25% | 0.76 | 0.64 | 0.68 |

## 4.3 Analyzing Effects by Number of Reviews

The coverage depends on the number of reviews. As more reviews are posted, its value changes and finally it will converge to a certain value. However, when many reviews have not been posted yet, it can drastically change, and it is not a reliable value for estimating review helpfulness. This is a cold start problem.

Therefore, we analyzed the effects of the number of reviews on estimation of review helpfulness. For the analysis, we sampled several reviews from the review set and computed the coverage using only them. We changed the sample size of the review set from 10 to 60 in increments of 10 and repeated this procedure 100 times for each sample size. We show the relationship between the sample size and the accuracy as boxplots (Figure 2). The horizontal lines in the boxes, mean the average accuracy for each sample size. The boxes mean the accuracy distribution from the top 25% to 75%. The whiskers mean the highest and lowest accuracy, and the plotted points are the outliers. It can be seen that the accuracy distribution is large and the accuracy average is low when the sample size is small.

We also show the accuracy distribution when the sample size is 10 (Figure 3). It can be seen that there are only three times when the accuracy is higher than or equal to 0.68, which is the accuracy when all reviews are used for computing the coverage. From these experiments, we conclude that the method using coverage doesn't work well when there are too few reviews. In the next section, we propose another approach, estimation using writing style, which is independent of the number of reviews.

# 5. WRITING STYLE

## 5.1 Dataset for Learning Writing Style

We classify writing style by a machine learning technique. To learn whether a writing style is formal or informal, we prepared a dataset. We prepared 400 reviews from the collected data described in Section 3. We gave the following criteria to two labelers, and they labeled them as formal, neutral, informal.

- Informal : There are many emoticons, marks and typos.
- Formal : Writing style of review is consistent, and it is not informal.

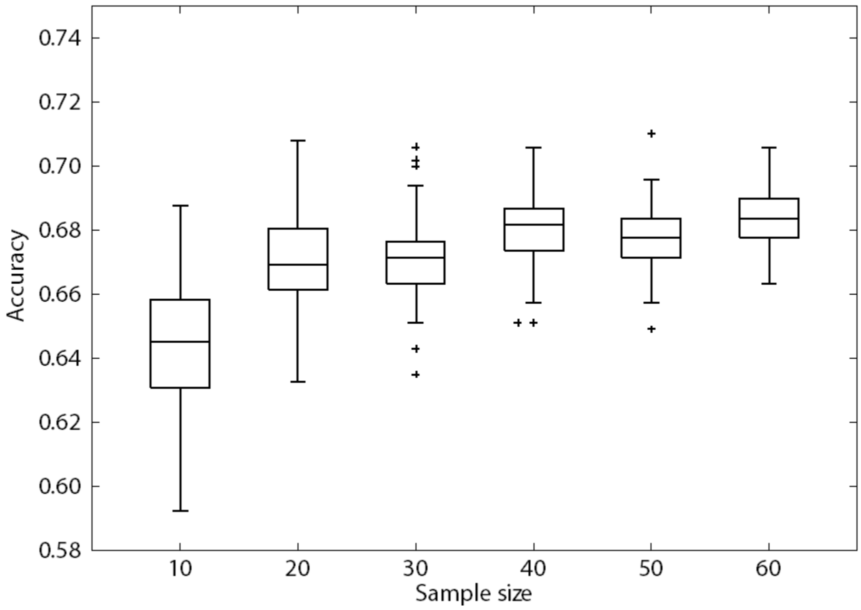


Figure 2. Relationship between sample size and accuracy

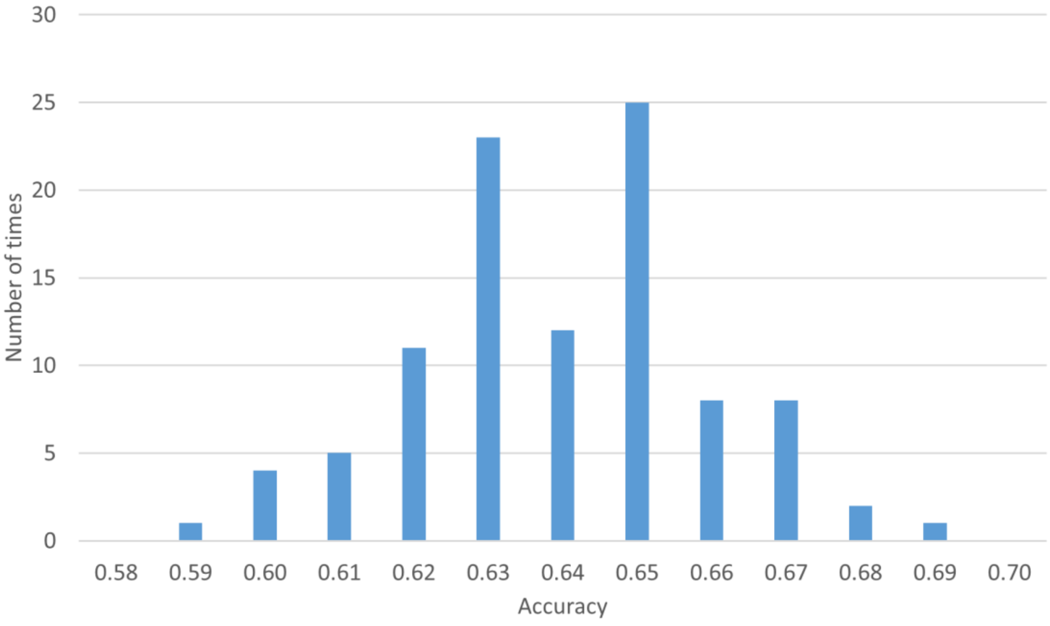


Figure 3. Distribution of accuracy by coverage (sample size is 10)

The labeled results are shown in Table 5. The $\kappa$ statistics between the two labeled data is 0.414 and shows medium association. We use 246 labeled reviews that both labelers labeled as formal or informal (the underlined data in Table 5).

## 5.2 Classification of Writing Style

We used a support vector machine (Vapnik 1998) to classify writing styles of reviews. We prepared the following features:

Table 5. Result of labelling writing style

| | | Labeler B | | | |
|---|---|---|---|---|---|
| | | Formal | Neutral | Informal | Sum |
| Labeler A | Formal | <u>155</u> | 54 | 20 | 229 |
| | Neutral | 19 | 11 | 14 | 44 |
| | Informal | 10 | 26 | <u>91</u> | 127 |
| | Sum | 184 | 91 | 125 | 400 |

- Noun
- Adjective
- Pattern of the end of sentence
- Biased keywords

The total number of features is about 3000. In Japanese sentences, expressions related to writing styles often appear at the end of sentences. Therefore, we focus on patterns of the end of sentences. To explain biased keywords, we define a bias score of a keyword $w$ as follows:

$$bias(w) = \left| \frac{DF(w, R_f)}{|R_f|} - \frac{DF(w, R_i)}{|R_i|} \right| \quad (6)$$

$R_f$ is the set of reviews that are labeled as formal, and $R_i$ is the set of reviews that are labeled as informal. The bias score becomes smaller when keyword $w$ evenly appears in formal reviews and informal reviews. It becomes larger when $w$ appears only to formal reviews or informal reviews. We compute the bias scores of all keywords in all reviews of each product. We regard the keywords whose bias score is in the top 100 as biased keywords.

## 5.3 Evaluation of Writing Style Classification

We evaluated our writing style classification method by accuracy. We used the RBF kernel for SVM and obtained optimal gamma and cost parameters by grid search. We combined several features described in 5.2 and searched for the best combination of features. We executed 5-fold cross-validations on our dataset, which is explained in 5.1. The results are shown in Table 6. END means the pattern of the end of sentence, and BIAS means biased keywords. From the results, END + BIAS is the best combination to classify writing styles. We built a classifier for writing styles using END + BIAS.

Table 6. Result of classification of writing style

| Features | Accuracy |
|---|---|
| END + noun | 0.78 |
| END + adjective | 0.66 |
| END + BIAS | 0.79 |
| END + noun + adjective + BIAS | 0.77 |

# 6. COMBINING COVERAGE AND WRITING STYLE

## 6.1 Rule-based Combination

We assume that

1. If a review has higher coverage, it tends to be helpful.
2. If a review is described in formal style, it tends to be helpful.

From these assumptions, we make classification rules for helpfulness of reviews. We consider that a review in formal style is more likely to be helpful. Therefore, we set the threshold of coverage for reviews described in formal style lower than the one for informal reviews. If a review is formal and its coverage is greater than or equal to the median, the review is helpful. If the review is informal and its coverage is greater than or equal to the top 25%, the review is helpful. Otherwise, the review is not helpful. We summarize the classification rules in Table 7.

Table 7. Rules for classifying helpfulness

| | Coverage | | |
|---|---|---|---|
| Writing style | ≥ top 25% | ≥ median | < median |
| Formal | Helpful | Helpful | Unhelpful |
| Informal | Helpful | Unhelpful | Unhelpful |

## 6.2 Machine Learning-Based Combination

Another approach combining coverage and writing style is to include coverage as one of the features of SVM. In Section 5, we build the classifier for writing styles. Here, we build a classifier for helpfulness. We use the same features (pattern of the end of sentence and biased keywords) as the best writing style classifier. We also use coverage as the feature. We use helpful or unhelpful labels as supervisory signals for SVM. We use the RBF kernel and experimentally decide other parameters for SVM.

## 6.3 Experiments

We compared our methods, which combine coverage and writing style, and the methods using coverage and writing style. In the method using coverage, we used the top 25% as a threshold. In the method using writing styles, we regarded that formal reviews are helpful. We show the results in Table 8. Rule-based combination of coverage and writing style is the best. The machine learning-based combination is worse than when each measure is used separately.

Table 8. Result of classification of helpfulness

| Method | Accuracy |
|---|---|
| Rule-based | 0.69 |
| Machine learning-based | 0.56 |
| Coverage | 0.68 |
| Writing style | 0.65 |

We show the distribution of accuracy per product by the rule-based method in Figure 4. The horizontal axis is accuracy and the vertical axis is the number of products. In $45/51 \approx 88.2\%$ of products, the accuracy values are more than 0.5. However, in two products, "Bath scales" and "Car navigation system2", their accuracy values are less than 0.1. This is caused by too few labeled data for these products. The number of labeled data of "Bath scales" and "Car navigation system2" are 1 and 2 respectively. We conducted a two-tailed t-test on the results of the rule-based method and the method using only coverage. We obtained $t(50) = 2.279$ and $p < 0.05$, and we concluded that the rule-based method performs significantly better than the method using only coverage.

Next, we examined the rule-based method when the number of reviews is limited. We sampled the reviews and computed the accuracy in the same way as in section 4.3. We show the results in Figure 5. Blue means the results using the rule-based method and red means the results using the coverage for each sample size. The plotted points mean the accuracy average, and the whiskers mean the standard deviation. The broken line means the result using coverage when all of the reviews are used, and the solid line means the result using writing style. We found the rule-based method is the best for any sample size. Furthermore, it is better than the results using coverage with all reviews except when the sample size is 10.

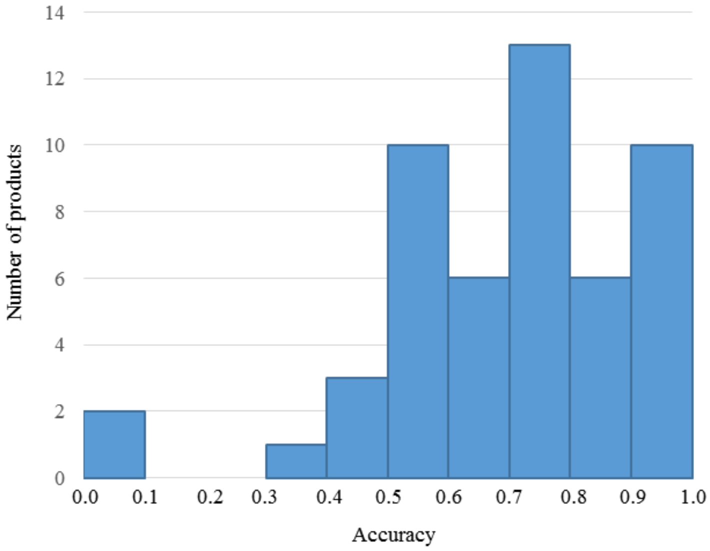


Figure 4. Distribution of accuracy per product by rule-based approach

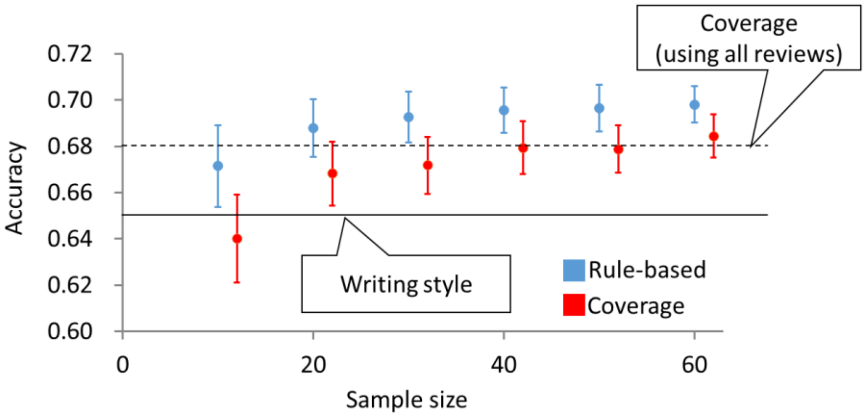


Figure 5. Comparison of accuracy

We also analyzed the accuracy distribution using the rule-based approach per product. We show the result in Figure 6. By comparing it with Figure 3, it can be seen that the distribution using the rule-based method is located in a higher accuracy area than the distribution using only the coverage. This means that the rule-based method successfully reduces the effects of the cold start problem.

## 7. CONCLUSION

We proposed methods to estimate helpfulness of customer reviews. We focused on coverage and writing style of reviews. Coverage expresses how many important words are contained in a review in comparison with other reviews for the same product. We defined coverage as a frequency-based measure. Writing style means the review text is described in formal style or informal style. We built a writing style classifier by support vector machines. We assume that a review whose coverage is high and is written in formal style is likely to be helpful. From this idea, we build a rule-based classifier for helpfulness of reviews by combining coverage and writing style. Our method achieved better accuracy than classifiers using coverage and writing style separately. We also analyzed the accuracy when the number of reviews is limited. In these cases, the accuracy using only coverage became worse. On the other hand, our rule-based method was less affected than the method using only coverage.

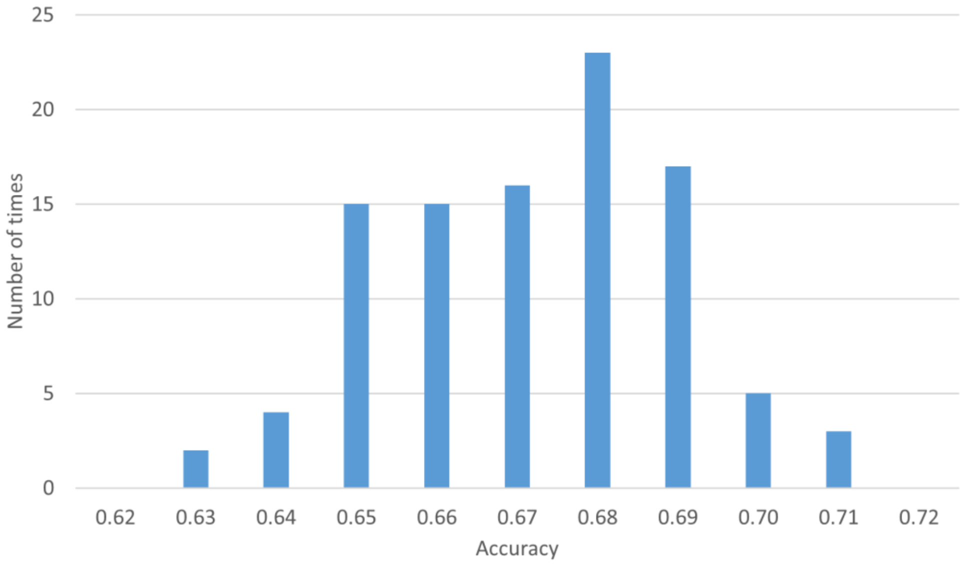


Figure 6. Accuracy distribution using rule-based approach (sample size is 10)

Future work is as follows:

- In this paper, we conducted offline evaluation using actual reviews. We plan to conduct the evaluation by users to examine the accuracy and interview them.
- We will consider using other features such as information about reviewers to improve accuracy.
- We combined the coverage and writing style in a simple way. We will try other methods to combine them.

## ACKNOWLEDGEMENT

This work was supported by Grant-in-Aid for Young Scientists (B) Grant Number 24700097 from Japan Society for the Promotion of Science.

## REFERENCES

Danescu-Niculescu-Mizil, C. et al., 2009, How opinions are received by online communities: a case study on amazon.com helpfulness votes. *Proceedings of the 18th international conference on World Wide Web*. Madrid, Spain, pp. 141-150.

Gamon, M., 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, Switzerland, Article No.841.

Hu, M. and Liu, B., 2004, Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. Seattle, USA, pp.168-177.

Jindal, N. and Liu, B., 2008, Opinion Spam and Analysis. *Proceedings of the 2008 international conference on Web search and data mining*. Stanford, USA, pp.219-230.

Kim, T. et al., 2006. Automatically Assessing Review Helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Sydney, Australia, pp. 423-430.

Kudo, T. et al., 2004. Applying Conditional Random Fields to Japanese Morphological Analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. Barcelona, Spain, pp. 230-237.

Liu, B. et al., 2005. Opinion observer: analyzing and comparing opinions on the Web. *Proceedings of the 14th international conference on World Wide Web*. Chiba, Japan, pp. 342-351.

Liu, J. et al., 2007. Low-Quality Product Review Detection in Opinion Summarization. *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing*. Prague, Czech Republic, pp. 334-342.

Lu, Y. et al., 2010. Exploiting Social Context for Review Quality Prediction. *Proceedings of the 19th international conference on World Wide Web*. Raleigh, USA, pp. 691-700.

Mudambi, S.M. and Schuff, D., 2010. What makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com. *MIS Quarterly,* Vol.34, No.1, pp.185-220.

Pang, B. and Lee, L., 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. *Proceedings of the 43th Annual Meeting on Association for Computational Linguistics*. Ann Arbor, USA, pp. 115-124.

Turkey, P.D. et al., 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Philadelphia, USA, pp. 417-424.

Vapnik, N. V., 1998. *Statistical Learning Theory*. Wiley-Interscience, New York, USA.

Zhuang, L. et al., 2006. Movie review mining and summarization. *Proceedings of the 15th ACM international conference on Information and knowledge management*. Arlington, USA, pp. 43-50.