IEICE TRANS. INF. & SYST., VOL.E95–D, NO.11 NOVEMBER 2012

PAPER

Estimating Reviewer Credibility Using Review Contents and Review Histories

Yuya TANAKA[†], **Nobuko NAKAMURA**[†], ***Nonmembers***, **Yoshinori HIJIKATA**[†a)], ***Member***, ***and*** **Shogo NISHIDA**[†], ***Fellow***

SUMMARY In recent years, user-supplied reviews have increased to become widely prevalent on many websites. Some reviewers (users who comment on items) provide valuable information. Others provide information many people already know. Our goal is to identify credible reviewers who provide valuable information. Two methods can be used to measure reviewer credibility: assessing reviewers based on the content of reviews that they have written in the past and assessing reviewers based on their review histories. By comparing these methods, we aim at obtaining knowledge to determine which method is most useful for identifying credible reviewers. Additionally, many features have been proposed for assessing reviews or reviewers in the previous methods, but they have not been compared. We compare these attributes and clarify what kinds of attribute are useful for identifying credible reviewers.

key words: *information credibility, product review, credible reviewer estimation, feature investigation, methodology investigation*

1. Introduction

Reviews have increased with the increasing number of reviewers (users who comment on items). However, the spread of review sites has made reviewers more diverse. Some reviewers provide valuable information, but others provide information that many people already know. We think that if we can find credible reviewers, we can identify highly helpful reviews that include useful information for making purchase decisions.

Two methods exist to measure reviewer credibility. One is to assess a reviewer based on the content of the reviews that the reviewer has given before (hereinafter, content-based method (CBM)). Another method is to assess reviewers based on their review histories (hereinafter, history-based method (HBM)). Examples of review histories include the number of items on which a reviewer has commented and the time of review.

Actually, the content of the review has been used for assessing the usefulness of the review. However, it is not used for assessing the credibility of the reviewer. We test the capability of the review content for assessing reviewers. Actually, we compare the two methods described above in experiments. By comparing these methods, we aim at obtaining knowledge to determine which method to use for identification of credible reviewers.

Additionally, in related works of CBM and HBM, some attributes have been proposed, but they have not been compared. Therefore, which attributes are highly associated with the quality of a review has not been revealed. We compare all the attributes proposed in related works.

Finally, we combine CBM and HBM by using the attributes proposed in these methods at the same time. We call this combination *Hybrid Method*. We see the improvement of the ability for assessing reviewer credibility of the *Hybrid Method*. We also compare it with the four existing methods of the related works and show the *Hybrid Method* achieves the best result.

The remainder of this paper is organized as follows. Section 2 describes the nature of trust. Section 3 introduces works related to reputation analysis and reviewer credibility assessment. Section 4 explains the design of our experiments. Section 5 presents experiment results. We conclude this study and suggest directions of future work in Sect. 6.

2. Nature of 'trust'

In this work, we examine methods for assessing reviewer credibility. However, 'credibility' and its upper notion 'trust' have wide meanings in our daily life. Therefore we need to discuss the concept of 'trust'. In this section, we reconsider the nature of trust. The nature of trust has been examined in diverse fields such as economics and sociology. We organize the nature of trust based on results of three studies from the above diverse fields.

The first of the studies was conducted by Yamagishi [23], a social psychologist, who stated that 'trust' consists of "trust as expectations for other's capability" and "trust as expectations for other's intention". Specifically, he stated that trust as expectations for other's capability includes expectations that another person can carry out a role in social relationships or social systems. He also stated that trust as expectations for other's intention is expectations for that another person carry out entrusted duties and responsibilities. That is to say, it is the trust that whether or not another person carries out actions without betraying. Additionally, trust is the nature of trustors, and that trustors judge whether or not another person is credible based on information about the person's capability and intention. He stated that a person's crediblity means the information about the person.

The second study was conducted by Falcone and Castelfranchi [4], who proposed a model of 'trust' based

Manuscript received June 29, 2011.
Manuscript revised May 25, 2012.

[†]The authors are with the Graduate School of Engineering Science, Osaka University, Toyonaka-shi, 560–8531 Japan.
a) E-mail: hijikata@sys.es.osaka-u.ac.jp

DOI: 10.1587/transinf.E95.D.2624

on a cognitive sociological analysis. They assume a situation in which an agent determines whether or not the agent entrusts unavoidable tasks to reach a goal to other agents. In the proposed model, an execution of trust includes three stages. First, an agent who entrusts others conducts "trust disposition" by measuring others' capabilities and intentions. Then, the agent conducts a "decision to trust" based on that trust disposition. Finally, the agent carries out an "act of trusting" by entrusting the tasks to other agents in this study.

The third of the studies is one conducted by Fogg et al., who are experimental psychologists [5], [6]. They define 'credibility' as a perceived quality for information receivers. Credibility is divided into two types. One is trustworthiness, which indicates whether or not information senders have goodness or morals. The other is expertise, which denotes whether or not information senders have skills or knowledge.

To summarize, we have the following findings related to 'trust'.

- Trust is the nature of a trustor.
- Trustors have some expectations for another's capabilities or intentions.
- Trustors trust another person when they determine that the person has capabilities or intentions based on the person's information.

In this work, we aim to automatically identify credible reviewers who provide valuable information. We can say that these reviewers have much knowledge or high capabilities. Therefore, trust to reviewers in this study corresponds to "trust as expectations for other's capability", as stated by Yamagishi. In this work, we use reviews as information to determine whether or not the reviewers have capabilities. Actual acquirement of reviewers' credibility is defined in Sect. 5.2.

3. Related Work

In this section, we introduce related works in the field of computer science. Our study is related to reputation analysis, review quality assessment and reviewer quality assessment. Studies of reputation analysis include sentiment classification and sentiment summarization. The former was studied by Turney [20], Pang *et al.* [16], and Dave *et al.* [3]. They classified a review according to whether the review is positive or negative using semantic orientation of words or machine-learning techniques. The latter was investigated by Hu *et al.* [10] and Hijikata *et al.* [9]. Hu *et al.* proposed a method for extracting opinion sentences and summarizing them according to product features. Hijikata *et al.* proposed a method for summarizing a review comment in an online auction using social relationships in the auction.

Some studies assess review quality based on the review content. Kim *et al.* [12] and Zhang *et al.* [24] examined some candidate attributes that might influence the review quality. Especially, Zhang *et al.* argued that good product review includes objective information and subjective evaluation, and used attributes reflecting this hypothesis. They used machine-learning techniques to train a regression model that estimates review quality, and assessed the model using correlation coefficient. Kim *et al.* adopted Support Vector Regression (SVR) [1], and Zhang *et al.* adopted SVR and Simple Linear Regression [22]. Liu *et al.* [14] pointed out amount of information, readability, and subjectiveness as influential factors of good reviews, and used attributes related to these factors. They classified reviews in datasets into those of high-quality and those of low-quality using SVM [21], and evaluated their method using the precision of classification. Liu *et al.* raised reviewer expertise, writing style and timeliness as factors influencing review helpfulness [15]. They use almost the same attributes as Chen *et al.* [2], Zhang *et al.* [24] and Riggs *et al.* [17] proposed. Tsur *et al.* [19] took an un-supervised approach for finding the most helpful book reviews. They identified a lexicon of dominant terms that constituted ideal reviews, and used it for finding helpful reviews.

Riggs *et al.* [17] and Chen *et al.* [2] assessed reviewers based on their review histories. Riggs *et al.* [17] considered that reviewers who rate items near the average ratings of the items in the early stage are credible. They used the attributes reflecting this hypothesis, and estimated reviewers' qualification. Chen *et al.* [2] considered that if a review receives a high evaluation in a category, then a reviewer who has given the review is credible in the category. They aggregated evaluation values for a reviewer's reviews in a category. Lim *et al.* [13] proposed methods for determining whether or not a reviewer is a spammer. They used reviews that reviewers have given and their review history. Their aim is to detect spammers, and they do not assess reviewer quality.

Our research aims at detecting the effective method and the effective attributes for assessing reviewer credibility. The previous studies related to sentiment analysis provide the knowledge that is a basis for credibility assessment of reputation. However, they do not assess the credibility of reputation. In the previous studies of review quality assessment and reviewer quality assessment, each study uses a different set of attributes. Therefore, the most effective attribute for reviewer credibility assessment remains elusive. Additionally, simple comparisons of methods are not possible because datasets used in the previous works differ. In this work, we compare CBM and HBM using the same dataset. We also compare attributes including the attributes used in the related works of review quality assessment and reviewer quality assessment. Although the previous studies related to review quality assessment use contents of only single review, we use contents of reviews that a reviewer has given before for assessing reviewer credibility.

4. Experiment Design

4.1 Purposes of the Experiments

The purposes of the experiments are as follows.

(1) Clarifying effective attributes for assessing reviewers' credibility

We clarify which attributes are effective for assessing reviewer credibility among the attributes used in works related to CBM and HBM.

(2) Clarifying an effective method for assessing reviewers' credibility

We clarify which method should be used among CBM and HBM when we assess reviewer credibility.

(3) Verifying the effectiveness for combining CBM and HBM

We verify whether combination of CBM and HBM improves the prformance of prediction. We use both attributes of CBM and those of HBM at the same time. We name the combination method *Hybrid Method*. We place it as our proposed method.

(4) Comparing our method and the related works

We clarify whether our method (*Hybrid Method*) outperforms the assessment methods in the related works.

In Purpose (1), we calculate a correlation coefficient between values of each attribute and the actual helpfulness of reviewers for investigating the attributes. Actual helpfulness of reviewers is evaluated by readers. In this study, we use an aggregation of users' votes as the actual helpfulness. This is also used in the experiments for Purpose (2)-(4). The calculation of the actual helpfulness is explained concretely in Sect. 5.2.

In Purpose (2), In CBM, we use the reviews that a reviewer has given in the same category because we infer that a reviewer's writing style differs according to the category. First, we learn a model that estimates reviewer evaluation values from the attribute values in each of the two methods. Then, we clarify which method is more effective between CBM and HBM by calculating a correlation coefficient and mean absolute error (MAE) between the estimated values and the actual helpfulness of reviewers. This evaluation method is also used in the experiments for Purpose (3) and (4).

4.2 Attributes Used for This Study

We describe methods for calculating values for the attributes used for this study (Table 1). The attributes of CBM used for this work include all of those used for the studies of Kim *et al.* [12], Liu *et al.* [14] and Zhang *et al.* [24]. The attributes of HBM used for this work include all of those used for the studies of Riggs *et al.* [17] and Chen *et al.* [2]. Our methods for calculating values for the attributes are the same as those of the related works. In fact, however, our calculation method is different from those of the related works to some of the attributes. We think that the calculation should be done using commonly-used tools or algorithms. Some of the related works use more advanced algorithms to calculate the attribute value. In that case, we use more general and simple method instead of them. The attribute values are normalized by dividing them by the maximum attribute value in the same attribute.

Table 1 Attributes used for this study.

Attributes of CBM			
1	no. words	15	ratio superlative adjective
2	no. product names	16	ratio wh-phrase
3	tf-idf (*uni-cate*)	17	no. subjective words
4	tf-idf (*uni-item*)	18	no. product features
5	tf-idf (*bi-cate*)	19	freq. product features
6	tf-idf (*bi-item*)	20	no. paragraphs
7	ratio proper noun	21	paragraph length
8	ratio noun	22	no. sentences
9	ratio interjection	23	sentence length
10	ratio verb	24	no. sentences with product features
11	ratio numeral	25	ratio negative sentences
12	ratio adjective	26	ratio positive sentences
13	ratio adverb	27	sim. product specification
14	ratio comparative adjective	28	stars (the rating score)
Attributes of HBM			
29	no. items reviewed	32	star differences
30	no. reviews of an item	33	num. reviews (category)
31	time of review		

4.2.1 Attributes of CBM

We describe the methods used for calculating values for the attributes of CBM. The values for the attributes of CBM are calculated using all the reviews that a reviewer has given in the same category. We calculate the values for the attributes of CBM in each of the above reviews and sum the values in each attribute. We divide the sums by the number of reviews that the reviewer has given.

The number of words that exist in a target review is "no. words". The number of the target item names used in the review is "no. item names". Here, "tf-idf" is the average of the tf-idf values of all the words in the review. We prepare document sets of two types for calculating df values. One document set consists of all reviews in a category to which the target item belongs (*cate*). The other consists of all reviews for the target item (*item*). We use unigrams (*uni*) and bigrams (*bi*) to calculate tf and df values.

To calculate the attribute values of "the ratio of part-of-speech" (attribute No. 7-16), we count each part-of-speech in the review conducting part-of-speech analysis. We divide the number of occurrences of each part-of-speech by the number of words in the review. For this study, we use Apple Pie Parser[†] to conduct part-of-speech analysis. We regard 'what', 'where', 'when', 'which', 'why', 'who', and 'whose' as wh-phrases.

The number of subjective words in the review is denoted as "no. subjective words". Subjective words are words existing in a list of subjective words. Zhang *et al.* learned this feature using the methods proposed in other four studies. Some of them use more advanced algorithm like bootstrapping algorithm. Therefore, we use General-Inquirer Dictionaries provided by Harvard University as the list of subjective words[††]. This dictionary is one of the most

[†]http://nlp.cs.nyu.edu/app/

[††]http://www.wjh.harvard.edu/ inquirer

popular dictionaries for obtaining subjective words.

The number of product features in the review is "no. product features". Product features are words existing in a list of product features. The number of times the product features occur in the review is "freq. product features". We create a list of product features in accordance with the method described by Kim *et al.* [12]. They created it automatically using pro and con keywords in Epinions.com†.

The number of break tags in the review is "no. paragraphs". The average number of words between the break tags is "paragraph length". The number of end points such as a period and a question mark in the review is "no. sentences". The average number of words that exist between the end points is "sentence length". The number of sentences that include the product features in the review is "no. sentences with product features". "ratio negative sentences" and "ratio positive sentences" respectively represent the ratios of sentences that include negative words and positive words in the above list of subjective words to all sentences in the review. "sim. product specification" is the cosine similarity between the review and the product specification of the target item written by editors of Amazon.com. "stars" is the rating score assigned by the reviewer.

4.2.2 Attributes of HBM

We describe the methods for calculating values for attributes of HBM. These methods use reviewers' review histories in the same category. "no. items reviewed" is the number of items for which the target reviewer has given a review. "no. received reviews of the item" (Num_i) is calculated using the following formula. This formula is based on the idea that reviewers who have given a review to an item that has few reviews are highly evaluated.

$$Num_i = \frac{Max\_num - num_i}{Max\_num}$$

Here, num_i is the average number of reviews received from other reviewers of the items that reviewer i has reviewed. Max_num is the maximum value among all the num_i.

The attribute value of "time of review" ($time_i$) is calculated by the following formula.

$$time_i = \frac{\sum_{j \in S_i} t_{ij}/m_j}{n_i}$$

Here S_i stands for the set of items about which reviewer i has given reviews. Furthermore, t_{ij} represents the rank of the review that reviewer i gave to item j among the reviews to item j (from other reviewers including reviewer i) in chronological order. m_j signifies the number of reviews that item j has received from other reviewers including reviewer i, and n_i stands for the number of reviews that reviewer i has given.

The following formula yields the attribute value of "star difference" (dif_star_i).

$$dif\_star_i = 5 - \frac{\sum_{j \in S_i} |a_j - r_{ij}|}{n_i}$$

Here S_i stands for the set of items for which reviewer i has given a review; a_j signifies the average of the ratings from other reviewers for item j. r_{ij} denotes the rating that reviewer i gave to item j. n_i represents the number of reviews that reviewer i has given. We used the number of stars in Amazon.com as the rating of the reviewer. The maximum number of stars is five in Amazon.com. We want to make this attribute as the reviewer credibility becomes higher when its value increases. This is why we subtract the average difference of stars from five.

The "no. reviews (category)" used in the method of Chen *et al.* [2] is the number of reviews that the reviewer has done in the same category as the target review.

4.3 Method for Learning a Model

For this study, we use machine learning techniques to learn a model that estimates reviewers' evaluation values from attribute values. We adopt SVM regression [18] to learn the model as Kim *et al.* [12] and Zhang *et al.* [24] did. We divide our reviewer data into training data and test data randomly, and learn the model from the training data using the SVM regression tool SVM^{light} ††. We estimate the reviewers' evaluation values from the test data using the model, and assess the model measuring a correlation coefficient and mean absolute error (MAE) between the estimated values and the actual helpfulness. Pearson product-moment correlation coefficient is used to measure the correlation. MAE is calculated using the following formula [7].

$$MAE = \frac{\sum_{i=1}^{N} |p_i - q_i|}{N}$$

where p_i stands for the correct value and q_i stands for the predicted value, N stands for the number of data for prediction.

5. Execution of the Experiments

We conducted the experiments explained in Sect. 4 using data of reviewers who have given reviews on an e-commerce site. This section describes our dataset and actual reviewer helpfulness used for the experiments. Then we present and discuss the experiment results.

5.1 Our Dataset

We collected data from Amazon.com during December 2009. Specifically, we collected reviews and reviewers for all items in four categories: *Mystery Movie*, *Rock Music*, *MP3 Player*, and *Digital Camera*. In our experiments, we use users' votes (clicking the Yes or No button to the question "Was this review helpful to you?") to acquire the actual

†http://www.epinions.com
††http://svmlight.joachims.org/

Table 2 Data in respective category.

	Mystery	*Rock*	*MP3*	*DCamera*	Total
# items	1,972	3,717	686	1,660	8,035
# all reviewers	52,410	67,029	41,913	60,135	221,487
# target reviewers	1,879	3,739	4,859	14,263	24,740
# average reviews	3.779	2.750	4.759	1.621	2.188

reviewer helpfulness. In order to make the helpfulness be robust, we used only reviews with at least 10 votes as Zhang *et al.* [24] did. The data of reviewers who have not given any reviews with at least 10 votes was discarded. Table 2 shows the number of items, number of all reviewers, number of target reviewers, and average number of reviews that a reviewer has given in each category.

5.2 Actual Reviewer Helpfulness

In this study, we use the votes for reviews in Amazon.com to acquire the actual reviewer helpfulness. Specifically, we used the ratio of the total of helpful votes to all votes for the reviews that a reviewer has given in the same category as the actual reviewer helpfulness as Zhang *et al.* [24] and Kim *et al.* [12] did. We restricted reviews to the same category for the same reason we used reviews for the same category in CBM. The actual helpfulness of the reviewer i is calculated using the following formula.

$$helpfulness_i = \frac{total\_helpful_i}{total\_helpful_i + total\_nothelpful_i}$$

Here, $total\_helpful_i$ represents the number of helpful votes that reviewer i received; $total\_nothelpful_i$ denotes the number of not-helpful votes that reviewer i received. The votes that we used to calculate the actual reviewer helpfulness were originally intended for Amazon users to assess the helpfulness of reviews. We take a user's helpful vote for a review as a judgment that the user considers the reviewer who has given the review has skills or knowledge. We consider that if we sum up the votes over all the reviews that a reviewer has given, we can regard the sum as the actual reviewer helpfulness.

When using this formula, the reliability of the actual reviewer helpfulness might differ among reviewers. Some reviewers might receive many votes while others might not receive so many votes. We examined the number of votes each reviewer has received. Figure 1 shows the result. X-axis shows the number of received votes per reviewer and Y-axis shows the number of reviewers. From this figure, we can see that the number of reviewers receiving n votes follows the power-law distribution. The numbers of votes received by most reviewers are less than 50. Although we agree with the imbalance of the number of received votes among reviewers, we think that most reviewers' received votes are in fixed range (10-49). Therefore we used this formula for calculating the reviewer's helpfulness.

We think that the actual reviewer helpfulness can be obtained more accurately using all of the past review that the reviewer has written than using only the latest review.

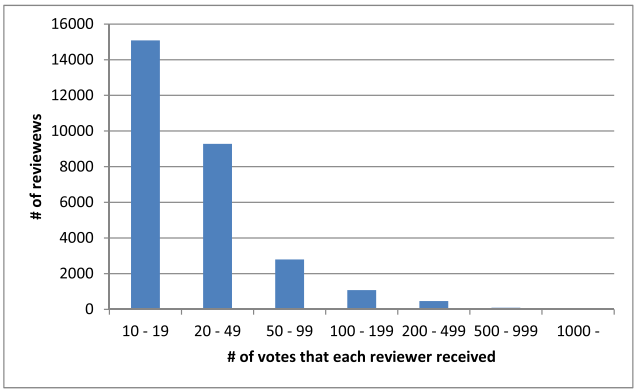


Fig. 1 The number of votes that each reviewer has received.

Table 3 The top 10 attributes of correlation coefficients r. (all categories)

Rank	No.	Attributes	r
1	28	stars	0.5421
2	32	star differences	0.4886
3	18	no. product features	0.2799
4	19	freq. product features	0.2292
5	1	no. words	0.2034
6	30	no. reviews on an item	−0.1880
7	21	paragraph length	0.1821
8	5	tf-idf (*bi-cate*)	0.1619
9	31	time of review	0.1476
10	24	no. sentences with product features	0.1330

Therefore $helpfulness_i$ is calculated from the users' votes to all of the past review that the reviewer has written.

5.3 Results

5.3.1 Effectiveness of All Attributes

In this section, we describe the results of the experiment for the effectiveness of all attributes. We clarify the effective attributes calculating correlation coefficients between values of each attribute and the actual reviewer helpfulness.

We examine effective attributes using all the reviewer data in all categories. We present the top 10 attributes of the correlation coefficients in Table 3. In Table 3, we specifically consider the top 5 attributes whose correlation coefficients are 0.2 and higher[†]. Results show that star ratings influence the credibility assessment for the reviewer because the attributes related to stars achieve high correlation coefficients. It also shows that whether a reviewer refers to product features is related to a credibility assessment for the reviewer because the attributes related to product features are correlated with the actual reviewer helpfulness. These results show that users tend to support positive opinions with an average rating, and to trust opinions that include product features. Additionally, the length of reviews is related to a

[†]Interpretation of the correlation coefficient r is categorized as follows: weak correlation ($0.2 \leq |r| \leq 0.4$), moderate correlation ($0.4 \leq |r| \leq 0.7$), and strong correlation ($0.7 \leq |r| \leq 1$) [11].

Table 4 Top 5 attributes of correlation coefficients. (in each category)

Rank	Mystery	Rock	MP3	DCamera
1	stars	stars	no. product features	stars
2	star differences	star differences	freq. product features	star differences
3	time of review	no. product features	stars	no. product features
4	no. words	freq. product features	no. words	freq. product features
5	no. reviews of an item	no. words	sim. product specification	no. words

credibility assessment for the reviewer because "no. words" are correlated with the actual reviewer helpfulness. Therefore, probably reviewers who give reviews including many contents are more credible.

We give our intuition why these attributes influence the reviewer credibility. We think that most users tend to avoid review comments whose star rating is low because some of them include the reviewer's excessive negative feelings. These comments are not helpful for making a decision of purchase. 'star differences' also shows the similar tendencies. We think that most users rely on reviewers who can give correct judgments for items. This means that reviewers do not give irrelevant ratings compared to the mass agreement (usual users' ratings). Finally, long reviews usually explain why the reviewer gave his rating to the full extent. This might make users understand the validity of the rating.

We examine whether the effective attributes differ according to the categories. We show the top five attributes in each category in Table 4. Table 4 shows that the attributes related to stars, product features and length of reviews are correlated with the actual reviewer helpfulness in every category. From Table 4, we can confirm that the effective attributes differ according to the categories. The attributes related to stars rank high in *Mystery*, *Rock*, and *DCamera*. However, the attributes related to product features rank high in *MP3*.

5.3.2 Effectiveness of Each Method for Assessing Reviewer Credibility

This section describes the results of the experiment for examining the effectiveness of the two methods for assessing reviewer credibility. First, in each of the two methods, we learn a model that estimates reviewers' evaluation values from the top five attributes of correlation coefficient described in Sect. 5.3.1 in each category because the number of the attributes of HBM is five. We also learn the model from top five attributes in all categories (*All*). Although we showed the ranking of all the attributes without any distinction among the two methods in Table 3 and Table 4, we use top five attributes in each method here. We use a correlation coefficient and MAE between the estimated values and the actual reviewer helpfulness for model assessment. Hereinafter, we describe the correlation coefficient and MAE as the estimation accuracy. For each of the two methods, we calculate the estimation accuracy. Then we compare the estimation accuracies of the two methods.

Subsequently, we clarify the effective combination of the attributes in CBM. The experiment described above only uses the top five attributes in each method. However, CBM has 28 attributes. This method might improve the estimation accuracies increasing the number of the attributes used for learning the model. Therefore, we learn the model increasing the number of the attributes in descending order according to correlation coefficients described in Sect. 5.3.1 in each category. We also learn the model in all categories (*All*). After we identify the appropriate number of the attributes, we compare the estimation accuracies of the two methods.

Table 5 Correlation coefficient. (top five attributes)

	CBM-past	CBM-latest	HBM
All	**0.5790**	0.5749	0.5224
Mystery	**0.7049***	0.6399	0.5616
Rock	**0.7881**	0.7737	0.6552
MP3	**0.5656**	0.5525	0.4241
DCamera	**0.6446**	0.6346	0.4541

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

Table 6 MAE. (top five attributes)

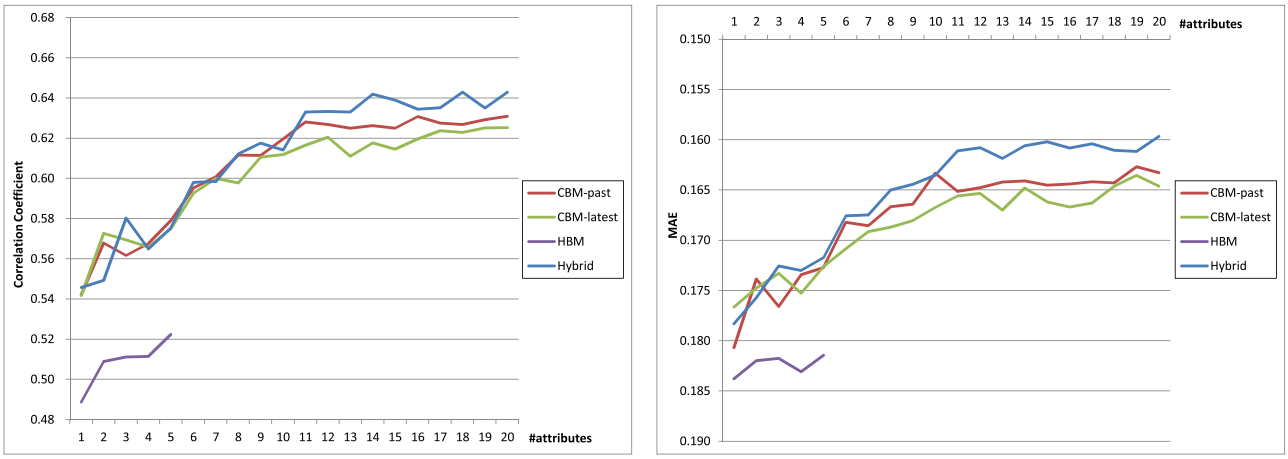
	CBM-past	CBM-latest	HBM
All	0.1727	**0.1726**	0.1814
Mystery	**0.1663****	0.1895	0.2079
Rock	**0.1361**	0.1380	0.1624
MP3	**0.1963**	0.2001	0.2119
DCamera	**0.1403**	0.1407	0.1600

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

Here, we try two different settings in CBM. One of them uses all the past reviews written by the target reviewer and the other uses only the latest review written by the target reviewer (Hereinafter, we describe these settings as CBM-past and CBM-latest respectively). Although the amount of information used in CBM-latest is lower than that used in CBM-past, the computation of CBM-latest is faster than that of CBM-past. In CBM-latest, the actual usefulness of the reviewer $helpfulness_i$ is calculated from all of the past review that the reviewer has written the same way as be calculated for CBM-past and HBM.

(1) Comparison of the two methods with top five attributes

We show the results of the experiment for comparing the two methods with top five attributes. Table 5 and Table 6 show the estimation accuracies (correlation coefficient and MAE respectively) in each of the two methods. In these tables, the value in bold font represents the best result in each category. Note that smaller values mean better prediction in



(a) Correlation Coefficient

(b) MAE

Fig. 2 Estimation accuracies vs. the number of attributes. (CBM-past, CBM-latest, HBM and Hybrid Method)

MAE. We can see that CBM (both CBM-past and CBM-latest) outperforms HBM in every category. For the *Mystery* category, we confirmed that the result of CBM-past reached statistical significance in correlation coefficient and MAE compared to results of HBM and CBM-latest ($\alpha \leq 0.05$ for correlation coefficient and $\alpha \leq 0.01$ for MAE). We used the test for equality of correlation coefficients [8] and *t*-test for MAE to test the statistical significance of the results. We can also see that the estimation accuracy of CBM-past is better than CBM-latest in each category. However, the difference of CBM-past and CBM-latest is small in MAE.

(2) Detection of the effective combination of the attributes

We show a transition of the estimation accuracies to the number of attributes in Fig. 2. Figure 2-(a) shows the transition of the correlation coefficient and Fig. 2-(b) shows the transition of MAE. In Fig. 2, we show only the transition with reviews in all categories (*All*) because the transition of each category is similar to it. The line named "Hybrid" will be explained later. We can see that the transitions of CBM-past and CBM-latest resemble each other. The estimation accuracy rises gradually until the number of the attributes reaches 11; then it changes little. Subsequently, we infer that using the top 11 attributes is appropriate in light of calculation costs. We can also see that CBM-past achieves better results than CBM-latest when the number of attributes used for learning the model is more than four.

We show the results using the top 11 attributes in CBM-past and CBM-latest, and using all five attributes in HBM. Table 7 shows the correlation coefficient calculated for each method. Table 8 shows MAE calculated for each method. In these tables, the value in bold font represents the best result in each category. Like the results using top five attributes, CBM outperforms HBM using the top 11 attributes. The correlation coefficient of CBM-past is better than that of CBM-latest (Unfortunately, the result of CBM-past did not reach statistical significance compared to that of CBM-latest in any category). However, we cannot see the clear difference in *MP3* category and *DCamera* category in MAE.

Table 7 Correlation coefficient. (top 11 attributes)

	CBM-past	CBM-latest	HBM
All	**0.6282**	0.6165	0.5224
Mystery	**0.6826**	0.6505	0.5616
Rock	**0.7920**	0.7852	0.6552
MP3	**0.5680**	0.5595	0.4541
DCamera	**0.6452**	0.6389	0.5224

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

Table 8 MAE. (top 11 attributes)

	CBM-past	CBM-latest	HBM
All	**0.1651**	0.1656	0.1814
Mystery	**0.1720**	0.1823	0.2079
Rock	**0.1336**	0.1388	0.1624
MP3	0.1983	**0.1973**	0.2119
DCamera	0.1403	**0.1394**	0.1600

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

The results of the experiment for comparing the two methods (including two different settings in CBM) are summarized as follows. CBM outperforms HBM in both cases using the top five attributes and using the top 11 attributes. CBM-past achieves better results than CBM-latest to most of the number of attributes used for learning the model in Fig. 1. In addition, the estimation accuracy of CBM-past is better than that of CBM-latest using top 11 attributes (In MAE, clear difference does not exist in some categories). These experimental results show CBM-past as the most effective method among the two methods (including two different settings of CBM) for assessing reviewer credibility.

5.3.3 Comparison between CBM and the Hybrid Method

In the previous section, we showed that CBM-past is the most effective method for assessing reviewer credibility. We

can probably obtain a better result using the effective combination of the attributes of CBM and HBM. Therefore, we clarify the effective combination using all 33 attributes (*Hybrid Method*) to determine whether the result improves. In the Hybrid Method, we learn the model increasing the number of the attributes in descending order according to correlation coefficients described in Sect. 5.3.1 in each category. We also learn the model in all categories (*All*). To assess the model, we use the estimation accuracy described in Sect. 5.3.2. We portray the transition of the correlation coefficient and MAE to the number of attributes in Fig. 2-(a) and 2-(b) respectively. In Fig. 2, we show only the transition with reviews in all categories (*All*) because the transition of each category is similar to it. "Hybrid" in Fig. 2 corresponds to the *Hybrid Method*.

The estimation accuracy of the *Hybrid Method* rises until the number of the attributes reaches 14. In light of calculation costs, we inferred that using the top 14 attributes is appropriate for the *Hybrid Method*. We compare the estimation accuracies of the *Hybrid Method* and CBM-past. We use the top 14 attributes in the *Hybrid Method* and the top 11 attributes in CBM-past. The estimation accuracies in each method are shown in Table 9 and Table 10. Table 9 shows the correlation coefficient and Table 10 shows MAE. In these tables, the value in bold font represents the best result in each category.

We can see that the *Hybrid Method* achieves better results than CBM-past in every category (Unfortunately, we cannot see the clear difference in *Rock* category in MAE).

Table 9 Correlation Coefficient. (CBM-past vs. Hybrid methods)

	CBM-past	*Hybrid Method*
All	0.6282	**0.6419**
Mystery	0.6826	**0.7216**
Rock	0.7920	**0.7926**
MP3	0.5680	**0.6138***
DCamera	0.6452	**0.6524**

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

Table 10 MAE. (CBM-past vs. Hybrid methods)

	CBM-past	*Hybrid Method*
All	0.1651	**0.1606***
Mystery	0.1720	**0.1631**
Rock	**0.1336**	0.1366
MP3	0.1983	**0.1885***
DCamera	0.1403	**0.1339**

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

We confirmed that the correlation coefficient of the *Hybrid Method* was significant compared to the correlation coefficient of CBM-past in the *MP3* category. We also confirmed that MAE of the *Hybrid Method* was significant compared to MAE of CBM-past in the *All* and *MP3* category.

These results demonstrate that the *Hybrid Method* is the most effective method when assessing reviewer credibility.

5.4 Comparison Between Our Method and the Related Works

As explained in Sect. 5.3.3, we confirmed that the *Hybrid Method* (hereinafter, the "proposed method") is the most effective when we assess reviewer credibility. Here, we confirm the effectiveness of the proposed method in comparison with combinations of the effective attributes reported in the related works.

The effective combination of attributes proposed by Kim *et al.* is "no. words", "tf-idf", "no. product features", "no. sentences", "sentence length", and "stars". That proposed by Liu *et al.* is "no. words", "no. item names", "no. product features", "freq. product features", "no. paragraphs", "paragraph length", "no. sentences", "sentence length", "no. sentences with product features", "ratio negative sentences", "ratio positive sentences", and "similarity to product specification". That proposed by Zhang *et al.* are attributes related to part-of-speech.

The effective combination of the attributes proposed by Riggs *et al.* is "no. items reviewed", "no. received reviews of the item", "time of review", and "star difference". That proposed by Chen *et al.* is "star difference" and "no. reviews (category)".

We learned the model from each effective combination of the attributes in each category, and also learned the model in all categories (*All*). We assessed the model using the estimation accuracy described in Sect. 5.3.2. The correlation coefficient and MAE of each related work and the proposed method are shown in Tables 11 and 12: the proposed method achieves the best results of all the methods. We tested statistical significances for correlation coefficient and MAE. We could confirm that the correlation coefficient of the proposed method reached statistical significance compared to those of all the other methods in "All", "Mystery" and "MP3" categories (All, Mystery: $\alpha \leq 0.01$, MP3: $\alpha \leq 0.05$). We could also confirm that MAE of the proposed method reached statistical significance compared to those of all the other meth-

Table 11 Correlation coefficient. (proposed method vs. related works)

	Kim	Liu	Zhang	Riggs	Chen	Proposed
All	0.6045	0.3762	0.2159	0.5123	0.4915	**0.6419** **
Mystery	0.6517	0.2337	0.1176	0.5673	0.5045	**0.7216** **
Rock	0.7651	0.4061	0.1558	0.6767	0.6627	**0.7874**
MP3	0.5674	0.5011	0.1486	0.4412	0.3824	**0.6138** *
DCamera	0.6349	0.4991	0.2363	0.4564	0.4444	**0.6524**

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

Table 12 MAE. (proposed method vs. related works)

	Kim	Liu	Zhang	Riggs	Chen	Proposed
All	0.1669	0.2039	0.2147	0.1829	0.1863	**0.1606****
Mystery	0.1847	0.2697	0.2722	0.2075	0.2206	**0.1631****
Rock	0.1434	0.2352	0.2451	0.1610	0.1577	**0.1366**
MP3	0.1977	0.2107	0.2486	0.2155	0.2227	**0.1885***
DCamera	0.1388	0.1629	0.1749	0.1596	0.1625	**0.1339***

" * " shows that the result was statistically significant compared to the result of others. (* : $\alpha \leq 0.05$, ** : $\alpha \leq 0.01$)

ods in "All", "Mystery", "MP3" and "DCamera" categories (All, Mystery: $\alpha \leq 0.01$, MP3, DCamera: $\alpha \leq 0.05$).

The proposed method is the most effective among the related works for assessment of reviewer credibility.

6. Conclusion

In this work, we conducted experiments to clarify the effectiveness of a content-based method (CBM) and a history-based method (HBM) for assessing reviewer credibility. In addition, we compared some attributes that characterize reviewer credibility and examined which attributes are effective. We also compared the most effective method that we clarified with the method proposed in the previous studies.

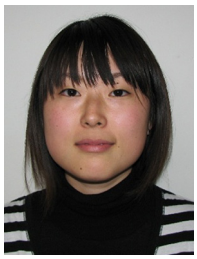
The experiment results clarified that the attributes related to star ratings and product features are effective for assessing reviewer credibility, and that CBM using the target reviewer's past review outperforms HBM for assessing reviewer credibility. We also found that we can improve the results of assessing reviewer credibility using a combination of CBM and HBM. The hybrid method is more effective than the methods proposed in the previous studies.

In future work, we think that providing an explanation of the reviewer's credibility. We show users why the system outputs the reviewer credibility values by showing the values of attribute groups in a chart.

References

[1] C-C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," http://www/csie.ntu.edu.tw/ cjlin/libsvm, 2001.
[2] M. Chen and J.P. Singh, "Computing and using reputations for internet ratings," Proc. Third ACM Conference on Electronic Commerce, pp.154–162, 2001.
[3] K. Dave, S. Lawrence, and D.M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," Proc. 12th International Conference on World Wide Web, pp.519–528, 2003.
[4] R. Falcone and C. Castelfranchi, "A belief-based model of trust," Trust in Knowledge Management and Systems in Organizations, IGI Publishing, pp.306–343, 2004.
[5] B.J. Fogg and H. Tseng, "The elements of computer credibility," Proc. SIGCHI Conference on Human Factors in Computing Systems, pp.80–87, 1997.
[6] B.J. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Ranqnekar, J. Shaon, P. Swani, and M. Treinen, "What makes Web sites credible?: A report on a large quantitative study," Proc. SIGCHI Conference on Human Factors in Computing Systems, pp.61–68, 2001.
[7] J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl, "Evaluating collaborative filtering recommender systems," ACM Trans. Information Systems, vol.22, no.1, pp.5–53, 2004.
[8] T. Hidano, M. Seya, N. Ohkawa, and K. Endo, Introduction to statistics for psychology, sociology and education, Baifukan Press, 1995.
[9] Y. Hijikata, H. Ohno, Y. Kusumura, and S. Nishida, "Social summarization of text feedback for online auctions and interactive presentation of the summary," Proc. 11th ACM International Conference on Intelligent User Interfaces, pp.242–249, 2006.
[10] M. Hu and B. Liu, "Mining and summarizing customer reviews," Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.168–177, 2004.
[11] S. Ishimura, Easily Understandable Statistics Analysis, Tokyo Tosho, 1993.
[12] S. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti, "Automatically assessing review helpfulness," Proc. Conference on Empirical Methods in Natural Language Processing, pp.423–430, 2006.
[13] E-P. Lim, V-A. Nguyen, N. Jindal, B. Liu, and H. Lauw, "Detecting product review spammers using rating behavior," Proc. 19th ACM International Conference on Information and Knowledge Management, pp.939–948, 2010.
[14] J. Liu, Y. Cao, C-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.334–342, 2007.
[15] Y. Liu, X. Huang, A. An, and X. Yu, "HelpMeter: A nonlinear model for predicting the helpfulness of online reviews," Proc. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp.793–796, 2008.
[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," Proc. Conference on Empirical Methods in Natural Language Processing, vol.10, pp.79–86, 2002.
[17] T. Riggs and R. Wilensky, "An algorithm for automated rating of reviewers," Proc. First ACM/IEEE-CS Joint Conference on Digital Libraries, pp.381–387, 2001.
[18] B. Scholkopf and A.J. Smola, "Making large scale SVM learning practical," in Advances in Kernel Methods: Support Vector Learning, pp.41–56, MIT Press, 1999.
[19] O. Tsur and A. Rappoport, "REVRANK: A fully unsupervised algorithm for selecting the most helpful book review," Proc. International AAAI Conference on Weblogs and Social Media, 2009.
[20] P.D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," Proc. 40th Annual Meeting on Association for Computational Linguistics, pp.417–424, 2001.
[21] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[22] I.H. Witten and E. Frank, Practical machine learning tools and techniques, Morgan Kaufmann, San Francisco, 2005.
[23] T. Yamagishi, The structure of trust: The evolutionary games of mind and society, Tokyo University Press, 1998.
[24] Z. Zhang and B. Varadarajan, "Utility scoring of product reviews," Proc. 15th ACM International Conference on Information and Knowledge Management, pp.51–57, 2006.

Yuya Tanaka received the B.E. and M.E. degree from Osaka University in 2010 and 2012 respectively. Currently, he works in Central Japan Railway Company. While in school, he studied information credibility and information filtering.



Nobuko Nakamura received the B.E. degree from Osaka Prefecture University in 2007. She received the M.E. degree from Osaka University in 2009. In 2009, she joined Panasonic Electric Works. While in school, she studied text mining and information credibility assessment.

Yoshinori Hijikata was born in Kobe, Japan. He received the B.E. and M.E. degrees from Osaka University in 1996 and 1998, respectively. In 1998, he joined IBM Research, Tokyo Research Laboratory. After working on Web technologies there, he received Ph.D. degree from Osaka University in 2002. Currently, he is an associate professor in Osaka University. His research interests are on Web intelligence, recommender system and text mining. He received the best paper awards from IPSJ Interaction'05 and ACM IUI'06. He is a member of the IPSJ, JSAI, HIS and DBSJ.

Shogo Nishida was born in Hyogo Prefecture in 1952. He received the B.S. M.S. and Ph.D. degrees in Electrical Engineering from the University of Tokyo, in 1974, 1976 and 1984, respectively. From 1976 to 1995, he worked for Mitsubishi Electric Corporation, Central Research Laboratory. From 1984 to 1985, he visited MIT Media Laboratory, Boston, Massachusetts, as a visiting researcher. In 1995, He moved to Osaka University as a Professor of Graduate School of Engineering Science. He was the dean of Graduate School of Engineering Science from 2004 to 2007. He is currently the Trustee and Vice President of Osaka University. His research interests include CSCW, Media Technology, Human Interfaces and Human Communication. He is a Fellow of IEEE (1998), a Fellow of IEICE in Japan (2005), a Fellow of IEE in Japan (2008) and Honorary Member of Human Interface Society of Japan (2007).