Can Diversity Improve Credibility of User Review Data?

Yoshiyuki Shoji, Makoto P. Kato, and Katsumi Tanaka

Department of Social Informatics
Graduate School of Informatics, Kyoto University
Kyoto, Japan
{shoji, kato, tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract. In this paper, we propose methods to estimate the credibility of reviewers as an individual and as a group, where the credibility is defined as the ability of precisely estimating the quality of items. Our proposed methods are built on two simple assumptions: 1) a reviewer who has reviewed many and diverse items has high credibility, and 2) a group of reviewers is credible if the group consists of many and diverse reviewers. To verify the two assumptions, we conducted experiments with a movie review dataset. The experimental results showed that the diversity of reviewed items and reviewers was effective to estimate the credibility of reviewers and reviewer groups, respectively. Therefore, yes, the diversity does improve the credibility of user review data.

1 Introduction

The rapid growth of the World Wide Web and Internet shopping services has enabled users to select from a huge number of commercial products on the Internet. Thus, the importance of user review data has increased, as it provides opinions and impressions that help users choose a quality item. There are many reviews for a variety of items on the Web, some of which are authored by professionals and others that are authored by non-professionals. Since professional reviews are available only for a limited number of items, even non-professional reviews are also useful for users to help making a decision.

However, there is a problem of *credibility* in utilizing reviews of general users. Since user reviews can be posted by any kinds of users including experts, novices, and even spammers, each review and aggregation of reviews can be biased and different from what the general public feels. Even users familiar with a particular domain cannot always produce a widely acceptable review, as they can be highly accustomed and accordingly biased to the domain. For example, users who have watched many Science Fiction(SF) movies might be likely to give a lower score to a SF movie than ordinary users, since they know more high-quality SF movies and use them as the basis for evaluating the other SF movies.

In this paper, we focus particularly on the credibility of reviewers, where the credibility of reviewers is defined as the ability of precisely estimating the item quality. This ability is defined for a single reviewer, as well as a group of

reviewers where the quality of items is estimated by aggregated reviews (e.g. the mean of their review scores). Thus, two problems regarding credibility are addressed in this paper: 1) estimating the credibility of a single reviewer, and 2) estimating the credibility of a group of reviewer.

We tackle the first problem to discover *experts* based on their review experience approximated by the number of reviews, as well as diversity of reviewed items. Although the credibility of a reviewer possibly correlates to the number of reviews that he has posted, many reviews do not always guarantee high credibility of a reviewer. As we discussed earlier, users who have reviewed only a specific category of items might post highly biased reviews. Therefore, we also consider the diversity of reviewed items to accurately estimate the reviewer credibility, assuming that a reviewer who has reviewed in diverse categories has higher credibility. For example, we expect that users who reviewed a wide variety of movies have a higher ability to evaluate the quality of movies than those who reviewed only SF movies.

We tackle the second problem to precisely estimate the quality of items by aggregating reviews of a reviewer group. Even if the credibility of individuals is low, it is possible to achieve high credibility when their reviews are aggregated. This phenomenon is known as *the wisdom of crowds* [12], in which one of the key criteria to obtain quality results is diversity of opinions. Thus, our proposed method to estimate the credibility of a reviewer group stands on diversity of reviewers, with an assumption that a group of more diverse reviewers has higher credibility.

To verify the two assumptions mentioned above, we conducted experiments with a movie review dataset. The credibility of reviewers was measured by the similarity between their review score and a *true* score, which was approximated by the score given by a well-known professional reviewer. Our experimental results showed that the diversity of reviewed items and reviewers in a group was effective to estimate the credibility of a reviewer and a group of reviewers, respectively. Therefore, yes, the diversity does improve the credibility of user review data.

The rest of this paper is organized as follows. Section 2 describes the related work. In Section 3, we introduce methods of estimating the credibility of reviewers based on diversity. Section 4 describes our experiments, and Section 5 evaluates our method in light of the experimental results. We conclude this paper in Section 6.

2 Related Work

This section introduces research on finding experts and its application to recommendation in Section 2.1, and research on diversity in Section 2.2.

2.1 Expert Detection and its Application to Recommendation

Finding experts has a long history and has been recently conducted in consumer generated media(CGM) sites. One of the representative examples is expert find-

ing in community-based question and answering (CQA) sites. Liu and Koll [5] proposed a method to find experts from CQA sites by focusing on the past answers given by users. In this work, experts are defined as users who can answer a certain kind of questions. The basic assumption used in their method is that users are able to answer a question if they have answered similar questions in the past.

There is some previous work on discovering experts to improve the accuracy of recommendations. One of the assumptions in this line of work is that an item evaluated as high-quality by experts is likely to be high-quality for many other users. Amatriain et al. [2] proposed a recommendation method that utilizes only the *nearest experts*, which are defined as users who posted a sufficient number of reviews, and are the most similar to a user who receives a recommendation. The performance of the proposed method was comparable to traditional collaborative filtering algorithms, even when a small expert set was used. Their expert detection method was based solely on the number of reviews, and the method did not take into account reviewed items. In our work, we utilize the diversity of reviewed items to find experts, and propose a method to aggregate reviews to precisely estimate the quality of items.

Sha et al. [9] proposed a method of seeking two different kinds of experts from an online photo sharing community: *trend makers* and *trend spotters*, and recommending trends in the community esitimated by these experts.

McAuley and Leskovec proposed [7] a method to find domain experts by using their review experience. Users are expected to become more professional in a domain if they work on the domain for a longer time. This work pointed out two important perspectives of expertise: 1) a user becomes an expert if s/he has been engaged in a domain for a long time, and 2) the evaluations done by novices tends to be diverse, while those by experts tends to be focused.

2.2 Measuring Diversity

Our proposed method incorporates a diversity-based measure to find experts and evaluate the credibility of a group of reviewers. There have been various studies on diversity specialized for different problems.

Collective intelligence has been actively discussed, as the collaboration on Web sites became a popular activity. Surowiecki [12] presented in his book some conditions of data under which the wisdom of crowds work correctly: *diversity of opinion*, *independence*, and *decentralization*. Once the three requirements are satisfied, useful knowledge can be built from the data by means of *aggregation*.

Diversity has been extensively used in the field of information retrieval. One of the most active research topics is diversification of Web search results [1, 13, 3]. For example, maximal marginal relevance [4] was used to diversify search results by decreasing the score of the pages similar to ones ranked in higher positions.

The research areas that focus on diversity are not limited to computer sciences, but include sociology, ecology, life science, economics, etc. Many diversity measures have been proposed especially in the biology area. Stirling [11] summarized three key factors regarding categorical diversity: *variety*, *balance*,

and *disparity*. Biodiversity has recently received attention, and is measured by Shannon-Wiener index [8], which was developed based on Shannon entropy. The index highly correlates to the number of breeds and balance across different breeds. Another diversity index, Simpson's diversity index [6], is defined as the probability of breed coincidence of two randomly-selected individuals. An alternative to these diversity measures was proposed in our previous work [10].

Since the diversity is a multi-faceted concept as can be seen in the earlier discussion, the optimal design of a diversity measure highly depends on its application domain. In this paper, we use two different kinds of diversity measures for reviewer groups, namely, entropy-based and variance-based diversity measures. The former measures the variety and balance, while the latter measures the disparity of reviewers. These two measures were compared in our experiments.

3 Method

This section introduces methods to estimate the credibility of a reviewer and a reviewer group based on diversity measures. Our methods are designed to be applicable to a wide variety of user review data such as movies, hotels, books, restaurants, etc.

3.1 User Review Data

User review data can be modeled by a tripartite graph with a category hierarchy. The tripartite graph consists of reviewers, items, categories, as well as reviewer-item and item-category edges. The category hierarchy is a set of category-category edges. More specifically, user review data D is defined as follows:

$$D = (U, R, I, B, C, H), \tag{1}$$

where U is a set of reviewers, I is a set of items, C is a set of categories. A set of edges $R \subset U \times I$ represents reviews of reviewers for items, e.g. $(u, i) \in R$ indicates that reviewer u reviewed item i. A set of edges $B \subset I \times C$ represents categories of items, e.g. $(i, c) \in B$ indicates that item i belongs to category c. A set of edges $H \subset C \times C$ represents *is-a* relationships between pairs of categories, e.g. $(c_j, c_k) \in H$ indicates that category c_j is a sub-category of category c_k.

Category tree $T = (C, H)$ is a rooted tree whose root is $c_{\text{root}} \in C$. Children of c_{root}, i.e. $M = \{c \mid c \in C \wedge (c, c_{\text{root}}) \in H\}$, are called *main categories* and distinguished from the other categories.

Some variables used in our proposed methods are defined below. The number of reviews given by user u is defined as follows:

$$n_u = |\{i \mid i \in I \wedge (u, i) \in R\}|. \tag{2}$$

The number of items that belong to category c is defined as follows:

$$n_c = |\{i \mid i \in I \wedge (i, c) \in B\}|. \tag{3}$$

Finally, we define the number of items that belong to category c and have been reviewed by user u as follows:

$$n_{u,c} = |\{i \mid i \in I \wedge (u,i) \in R \wedge (i,c) \in B\}|. \tag{4}$$

3.2 Estimating the Credibility of a Reviewer

The first problem we tackle is to estimate the credibility of each reviewer. Recall that the credibility is the ability of precisely estimating the quality of items. Our method is based on the assumption that a reviewer who reviewed many and diverse items has high credibility. The reason why we came up with this assumption is explained as follows. Suppose that there are two reviewers: one reviewed 10 movies, while another reviewed 100 movies. According to the assumption, the latter reviewer is more credible, as his expertise is expected to be higher than the former reviewer. Then suppose that there are another pair of reviewers: one reviewed 100 SF movies, while another reviewed 100 a wide variety of movies. We assume that the latter is more credible since his review is expected to be unbiased compared to the former reviewer.

The following formula is derived if we follow the assumption on the credibility of individual reviewer:

$$\text{Credibility}(u) = \alpha n_u \text{Div}(u), \tag{5}$$

where α is a parameter, n_u is the number of items reviewed by user u, and $\text{Div}(u)$ is the diversity of items reviewed by user u. We then model the diversity of reviewed items based on the idea of Shannon-Wiener index [8], which measures the diversity by the entropy over species. Regarding main categories as species in our case, Shannon-Wiener index is defined as follows:

$$H(u) = -\sum_{c \in M} p_u(c) \log p_u(c), \tag{6}$$

where $p_u(c)$ is the probability that user u reviews an item that belongs to category c. This probability can be estimated by the number of items of category c reviewed by user u divided by the number of items reviewed by user u: $p_u(c) = n_{u,c}/n_u$.

One of the problems of Shannon-Wiener index is that it is agnostic about the prior category distribution. Suppose that there are 10 horror and 100 SF movies. Although the maximum entropy is achieved by reviewing 10 horror and 10 SF movies, this reviewer is considered as biased to horror movies, as he reviewed all the horror movies despite the small number of horror ones. Therefore, we slightly modify Shannon-Wiener index by taking into account the prior category distribution. More specifically, we measure the diversity by the difference of the category distribution of a reviewer from the prior category distribution, i.e. Kullback-Leibler divergence of the two distributions. Letting $p(c)$ be the prior category probability, Kullback-Leibler divergence is defined as follows:

$$\text{KL}(u) = -\sum_{c \in M} p_u(c) \log \frac{p_u(c)}{p(c)}, \tag{7}$$

where $p(c)$ is the number of items of category c divided by the number of items: $p(c) = n_c/|I|$.

Finally, we define the diversity of a reviewer as follows:

$$\mathrm{Div}(u) = \exp(-\mathrm{KL}(u)). \tag{8}$$

Note that the exponential function is not essential, but is applied to make the diversity function $\mathrm{Div}(u)$ positively correlate to the diversity. This diversity function becomes larger when the category distribution of a reviewer and prior category distribution are closer. Thus, a reviewer who has evenly reviewed items is considered as credible, as he is considered as unbiased to any category.

3.3 Estimating the Credibility of a Group of Reviewers

The second problem we address is to estimate the credibility of a group of reviewers. Even if the credibility of individual reviewers is not so high, the credibility of a group of reviewers can be high when their reviews are aggregated. For example, the average review score of a group can be close to true quality of items, even if no individual reviewer can precisely estimate the quality.

According to the previous studies on collective intelligence [12], the diversity of members in a group is an important factor to obtain a high-quality result from the group by means of aggregation. For example, there are two groups: one includes ten SF maniacs, while another includes five SF and five horror maniacs. Given an item to each group, the average review score given by the former group might be more biased than the latter, as the aggregated score may reflect only a specific preference of the homogeneous group.

Therefore, we propose methods to estimate the credibility of a reviewer group based on the diversity of the reviewers. Our assumption for this problem is that a group of many and diverse reviewers has high credibility. As the diversity can be measured by three types of aspects, namely, balance, variety, and disparity [11], we propose two diversity measures that take into account different aspects, i.e. entropy-based and variance-based diversity measures.

The entropy-based diversity measure is similar to the one we used in estimating the credibility of individual reviewers, and takes into account the balance and variety of reviewers[1]. A high entropy-based diversity measure indicates that there are more types of reviewers in a group and the distribution of reviewers is balanced across the types. On the other hand, the variance-based diversity measure reflects the disparity aspect of diversity, and becomes high if reviewers in a group are dissimilar each other.

To compute the two diversity measures briefly explained above, it is necessary to model the similarity between reviewers in some way. To this end, we opted to characterize reviewers by using their expertise estimated by their reviews, with an assumption that a reviewer who has reviewed diverse items in a category has

[1] Balance and variety are simultaneously measured since they are not divisible in many cases.

high expertise in the category. For instance, a reviewer who have watched and reviewed all of space opera, cyberpunk, and science fantasy movies is expected to have more knowledge in the SF category than one who have reviewed only space opera movies.

In a similar way to the diversity computation for a single reviewer, the expertise of user u in main category c is measured by Kullback-Leibler divergence of the sub-category distribution of a reviewer and prior sub-category distribution:

$$\mathrm{KL}_{\mathrm{sub}}(u, c) = -\sum_{s \in \mathrm{Sub}(c)} p_u(s|c) \log \frac{p_u(s|c)}{p(s|c)}, \tag{9}$$

where $\mathrm{Sub}(c)$ is a set of sub-categories of main category c (i.e. $\mathrm{Sub}(c) = \{s \mid s \in C \wedge (s, c) \in H\}$), $p_u(s|c)$ is the probability that user u reviews an item of category s conditioned by category c ($p_u(s|c) = p_u(s)/p_u(c)$), and $p(s|c)$ is the prior probability of category c conditioned by category c ($p(s|c) = p(s)/p(c)$).

As the Kullback-Leibler divergence negatively correlates to the expertise in a main category, we apply an exponential function in the same way as the diversity computation for a single reviewer, and define the expertise of user u in main category c as follows:

$$e_{u,c} = \exp(-\mathrm{KL}_{\mathrm{sub}}(u, c)). \tag{10}$$

Below, we explain the two diversity measures in the details.

Entropy-based Diversity Measure

The entropy-based diversity measure is the entropy of the expertise distribution of a group as a whole with consideration of the prior expertise distribution. We first model the expertise of group $G \subset U$ in category c by aggregating the expertise of reviewers in the group:

$$e_{G,c} = \frac{1}{|G|} \sum_{u \in G} e_{u,c}. \tag{11}$$

We then model the *prior* expertise in category c:

$$e_c = \frac{1}{|U|} \sum_{u \in U} e_{u,c}. \tag{12}$$

The prior expertise can be interpreted as the average expertise in all the reviewers. Although these expertise scores do not represent a probability, we could normalize the expertise scores to treat them as probabilities:

$$p^e_G(c) = \frac{1}{|G|} \sum_{u \in G} e_{u,c}, \tag{13}$$

$$p^e_c = \frac{1}{|U|} \sum_{u \in U} e_{u,c}. \tag{14}$$

Kullback-Leibler divergence of the expertise distribution of a reviewer group and the prior expertise distribution is defined as follows:

$$\mathrm{KL}^e(G) = -\sum_{c \in M} p^e_G(c) \log \frac{p^e_G(c)}{p^e(c)}. \tag{15}$$

This divergence represents the closeness between the expertise of a group and prior expertise, and becomes smaller if the group expertise is more evenly distributed against the prior expertise.

Entropy-based diversity measure EDiv is then defined as follows:

$$\mathrm{EDiv}(G) = \exp(-\mathrm{KL}^e(G)). \tag{16}$$

Note that the exponential function is not essential again.

The entropy-based diversity measure increases as the expertise of a group as a whole is evenly distributed in each category. Note that this measure does not take into account the diversity of each reviewer in a group, and becomes high in both of the following cases: 1) all the reviewers in the group have balanced expertise in each category, and 2) the expertise distribution of the group is close to the prior expertise distribution, even though the expertise distribution of each reviewer is far from the prior expertise distribution.

Variance-based Diversity Measure

As computing the variance-based diversity measure requires the dissimilarity between reviewers, we first map reviewers on a $|M|$-dimensional space, where each dimension represent the expertise in a main category. A vector of reviewer u is denoted by $\mathbf{v}_u$ and defined as follows:

$$\mathbf{v}_u = (e_{u,c_1}, e_{u,c_2}, \ldots, e_{u,c_{|C|}}), \tag{17}$$

where $e_{u,c}$ is the expertise of reviewer u in category c.

Variance-based diversity measure VDiv, which is the average dissimilarity between individual reviewers and the mean of the reviewers in the group, is defined as follows:

$$\mathrm{VDiv}(G) = \frac{1}{|G|} \sum_{u \in G} \|\mathbf{v}_u - \bar{\mathbf{v}}_G\|, \tag{18}$$

where $\bar{\mathbf{v}}_G$ is the mean of reviewer vectors of group G, i.e. $\bar{\mathbf{v}}_G = \frac{1}{|G|} \sum_{u \in G} \mathbf{v}_u$.

In summary, we proposed diversity measures to estimate the credibility of reviewers as an individual and as a group. A variant of Shannon-Wiener index was proposed to measure the diversity for both of the cases, and a variance-based diversity measure was used only for a reviewer group. Note that the entropy-based and variance-based diversity measures correlate to some extent, but behave differently in some cases. For example, the entropy-based diversity measure becomes high if reviewers in a group have similar expertise in a wide variety of

categories, whereas the variance-based diversity measure does not. In the next section, we demonstrate the correlation between the credibility and diversity measured by the proposed methods.

4 Experiment

To clarify the effectiveness of our diversity measures for estimating the credibility of reviewers, we conducted experiments by using movie review data taken from Yahoo! Movies. Through the experiments, we tested the validity of the two assumptions: 1) a reviewer who has reviewed many and diverse items has high credibility, and 2) a group of reviewers is credible if the group consists of many and diverse reviewers.

4.1 Dataset

The movie review data was taken from Yahoo! Movies[2], which is one of the biggest movie communities in Japan. We collected 27,516 movies and 158,385 reviewers. There are 1,124,555 reviews and 38 categories in this dataset.

Since some real review data including ours do not contain explicit hierarchy information in categories, we applied a heuristic method to construct a hierarchy. Our method first extracted existing categories as main categories (e.g. 38 categories in our data), and then generated sub-categories by combining any pair of co-occurring main categories. More precisely, letting M be a set of main categories, we define a set of sub-categories as $S = \{c_j \oplus c_k \mid i \in I \wedge (i, c_j) \in B \wedge (i, c_k) \in B\}$, where $\oplus$ is an operator to concatenate two category names. We let the resultant set of sub-categories belong to main categories from which the sub-categories were generated, e.g. edges (c, c_j) and (c, c_k) were added to H for $c = c_j \oplus c_k$. For example, "Star Wars" belongs to two main categories *SF* and *adventure*. We created a sub-category *SF* - *adventure* and let it belong to *SF* and *adventure*. Finally, a set of categories is defined as $C = M \cup S$.

Note that we created a special sub-category indicating that a movie belongs to only a main category and does not belong to any sub-category. Given a movie belonging only to main category c, we added subcategory $c' = c \oplus c$ to the entire category set, and edge (c', c) to H. This special type of sub-categories was added because movies without any sub-category are not taken into account in the expertise estimation. For instance, the movie "Blade Runner" belongs only to *SF* category. This movie was assigned to a *SF* - *SF* sub-category.

Tables 1 and 2 show the detailed statistics of reviewers and movies in our dataset, from which we can find many reviewers who posted a review only once, and movies with a few reviews.

4.2 Evaluating the Credibility of a Reviewer

The first assumption is that a reviewer who has reviewed many and diverse items has high credibility. To test this assumption, we compared the correlation

[2] http://movies.yahoo.co.jp/

# of reviewers	
Reviewed only 1 movie	140,180
Reviewed less than 10 movies	204,178
Reviewed 1,000+ movies	39
Reviewed 2,000+ movies	7
Total	158,385

# of reviews per reviewer	
Arithmetic mean	6.35
Mode	1
Median	1
Max	5,301

Table 1. Statistics of reviewers.

# of movies	
Reviewed by only 1 Reviewer	6,326
Reviewed by 10+ Reviewers	8,877
Reviewed by 1000+ Reviewers	158
Reviewed by 2000+ Reviewers	29
Total	27,514

# of reviewers per movie	
Arithmetic mean	40.82
Mode	1
Median	4
Max	6,304

Table 2. Statistics of movies.

between the credibility and following measures: quantity (n_u in Equation 2), diversity (Div(u) in Equation 10), and both diversity and quantity (Credibility(u) in Equation 5 ($\alpha = 1$)).

Before testing the first assumption, we start with illustrating the characteristics of these three measures. Figure 1 shows how well the three measure distinguish expert reviewers from the others, where the horizontal axis represents the value of each measure, and the vertical axis represents the entropy of review scores. Each point in the figures represents the value of a measure and review score entropy of a reviewer. According to McAuley and Leskovec's work, experienced reviewers have a higher review score entropy, while novice reviewers cannot take full advantage of the range of scores, and are likely to evaluate items in a narrow and biased manner. For example, novice reviewers may use only three or four even if they are asked to evaluate movies at a five-point scale. Thus, the review score entropy can be a good indicator of experts.

In the ideal case, points in the figures should converge towards the upper right corner: some novice reviewers gave a wide or a narrow range of scores, while the most expert reviewers gave a wide range of scores. It can be seen from Figure 1 that both of the quantity and diversity can distinguish experts (reviewers with high review entropy) from the others. The diversity measure shows a slightly better discriminative power as reviewers broadly spread along the horizontal axis.

To test the first assumption, it is necessary to obtain *true* quality of each item. Since it is hard to get exact true quality score, we approximated it by the score given by a well-known professional reviewer. We extensively compared reviewers who rated many and diverse movies, and carefully selected one who gives a widely acceptable score. Finally, we decided to use reviews authored by Yuichi Maeda, and manually collected his reviews from his Web site[3]. He is a Japanese professional critic and movie journalist who has written 1,832 reviews

[3] http://movie.maeda-y.com/

Fig. 1. Quantity, diversity, and their combination vs. review score entropy.

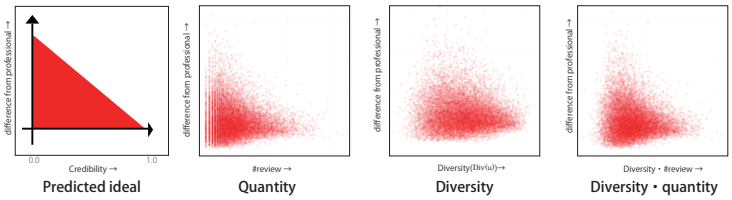

Fig. 2. Quantity, diversity, and their combination vs. RSS to professional scores.

since 2003 to 2014 on his site. We found 1,689 movies included in both of his and our review data. As the range of his review scores was different from ours, we converted them to a five-point scale and used the scores as true quality of items.

The credibility of a reviewer was estimated by the residual sum of squares (RSS) between his score and a score of reviewer u:

$$\mathrm{RSS}(u) = \frac{1}{|I_u \cap P|} \sum_{i \in I_u \cap P} (\mathrm{score}(u,i) - \mathrm{score}_{\mathrm{pro}}(i))^2, \tag{19}$$

where P is a set of movies reviewed by the professional, I_u is a set of movies reviewed by user u ($I_u = \{i \mid i \in I \wedge (u,i) \in R\}$), $\mathrm{score}(u,i)$ is a review score of u for movie i, and $\mathrm{score}_{\mathrm{pro}}(i)$ is a review score of the professional for movie i.

Figure 2 demonstrates that a reviewer becomes more similar in rating to the professional reviewer if the reviewer has reviewed more and more diverse movies.

4.3 Evaluating the Credibility of a Group of Reviewers

To test the second assumption regarding the credibility of a group of reviewers, we compared following measures: quantity ($|\{u \mid u \in U \wedge (u,i) \in R\}|$ for item i), entropy-based diversity measure ($\mathrm{EDiv}(G)$ in Equation 16), and variance-based diversity measure ($\mathrm{VDiv}(G)$ in Equation 18). The absolute error between the score of the professional and the average score of group G for item i is defined as follows:

$$\mathrm{AE}(G,i) = \left| \frac{1}{|G|} \sum_{u \in G} \mathrm{score}(u,i) - \mathrm{score}_{\mathrm{pro}}(i) \right|. \tag{20}$$

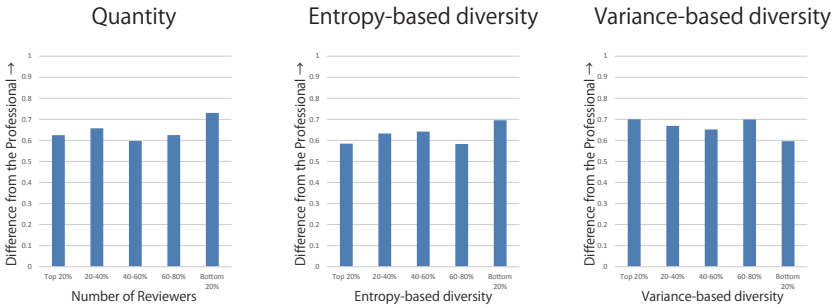

Fig. 3. Quantity, entropy-based, and variance-based diversity measure vs. RSS to professional scores (for all the groups).

If our second assumption is probable, the absolute error from large and diverse groups is smaller than that of smaller and/or more homogeneous groups.

Figure 3 shows the average absolute error of groups in each *bin*. We sorted all the groups based on one of the three measures, and categorized them into five bins based on the order of groups. For example, the leftmost bin of each figure includes groups ranked within top 20% when they are sorted in descending order of each measure. Thus, the left bins of each graph contain reviewer groups that are estimated as more credible, whereas the right bins contain reviewer groups that are estimated as less credible.

In the ideal case, the bars would slant upward to the right: the absolute error to the professional should become bigger for smaller or more homogeneous groups, while the error should be smaller for bigger or more diverse groups. of a group whose members are many or diverse is close to it. The bars of the quantity and entropy-based diversity measure show slightly similar trends to the ideal case, though they are not conclusive. The graph of the variance-based diversity measure does not show a trend similar to the ideal case. When we compare the leftmost bins, which contains the most diverse groups (top 20%), it can be seen that the entropy-based diversity measure outperforms the quantity-based measure in finding the most credible reviewers.

As we have observed from Figure 3, there is much absolute error difference between groups with different diversity. We hypothesized that the absolute error to the professional can be small enough if plenty of reviews are available for each movie, and investigated a case where a limited number of reviews are available. Figure 4 shows the average absolute error of groups with less than 100 reviews. In this case, the entropy-based diversity measure and number of reviewers can more accurately estimate the credibility of reviewer groups.

5 Discussion

Our first experiment was successful in evaluating the credibility of a reviewer, supporting our hypothesis that reviewers who see diverse movies and reviewers

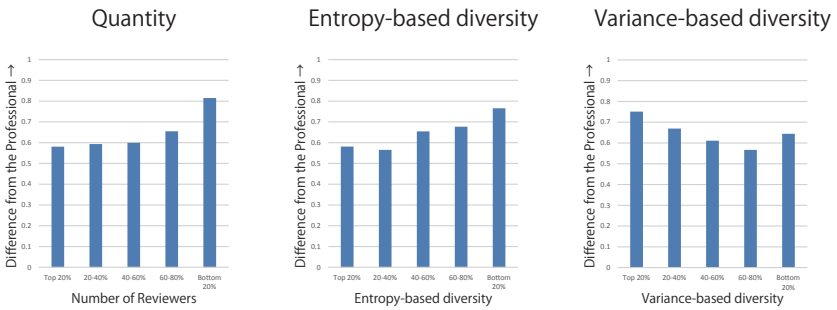

Fig. 4. Quantity, entropy-based, and variance-based diversity measure vs. RSS to professional scores (for groups with less than 100 reviewers).

who see many movies are reliable. We learned that these reviewers are characterized by a more even spread among their review scores and a amaller difference in rating with professional reviewers.

The reason why the difference of opinion of amatures and of professional does not converge to 0 is the difference of average; the professional's average rating is 3.3, and amature's is 3.6. Professionals are sometimes forced to see and to rate unfavorite movies at the job. Amateur can choose their favorite movies to review.

From the second experiment, we established that the entropy-based diversity and reviewer group size are good barometers to measure the credibility of a group. In contrast, Variance-based diversity does not work well.

The entropy-based diversity can measure the credibility of a group especially in case the number of reviewer is lower than 100. It's interesting to note that, when the number of members is small, the diversity of members is important, but when it is large, this is not the case. Generally, when the size of the a group is large enough, the most group is reliable when it is likely that the credibility of the group is saturated, we don't need to consider the size and diversity of the group. Figure 5 shows the relationship between the effect of the entropy-based diversity and the size of a group. The horizontal axis lists the groups binned by size. Each bin contains same number of groups. Groups were classified into high-diversity groups and low-diversity groups by their median entropy-based diversity. The vertical axis shows the average distance between the rating of the professional and that of the group. When the number of reviewers is less or equal to 440, a high diversity of reviewers minimized the score difference with the professional review. This fact supports our proposition. Contraly, in cases where the number of reviewers exceeds 440, the diversity of reviewers did not affect the score difference. Naturally, a larger group will be more credible because of the law of large numbers. The accuracy of the average score, however, trended down for the cluster of movies that assumed 119 to 182 reviews. This can be attributed to two possible causes. The movie reviewed by many reviewers is a popular movie, who tend to attract an audience of persons unfamiliar

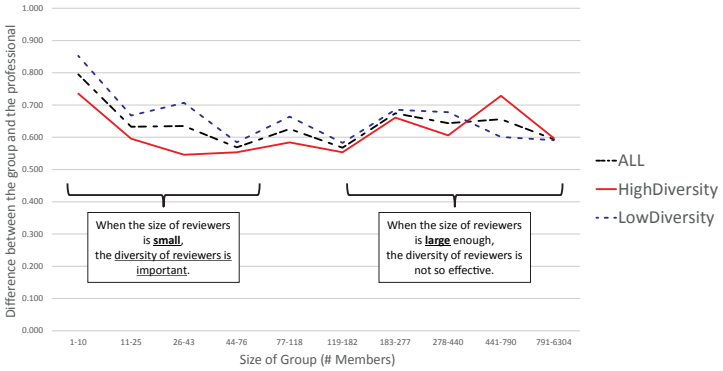

Fig. 5. Effectiveness of the Entropy-based Diversity

with movies. Their opinions are not very credible as evidenced by professional reviewers often shooting down popular movies. It refrects a characteristic of the review dataset; online user review are not implicit data, but intentional data.

The variance-based diversity does not work well, regardless of the group size. One reason could be a biased group (i.e. a community of specialists) providing a correct opinion. Another cause could be generalists. They are simillar to each other. A group that consists of non-diverse generalists can rate movies accurately.

6 Conclusion

In this paper, we proposed a method to estimate credibility of individuals and reviewer groups. We proposed two simple assumptions: a reviewer who has reviewed many and diverse items has a high credibility, and a group of reviewers is credible if the group consists of many and diverse reviewers. We modeled a general user review structure with a category tree and proposed diversity-based measurement calculations. Through experiments using a real dataset of movie reviews, the effectiveness of the assumption 1 was confirmed; a reviewer, who reviews many and diverse movies has a high credibility. The effectiveness assumption 2 was partially confirmed; when the number of members is small, the entropy-based diversity is a good indicator to measure the credibility of a group.

Acknowledgments

This work was supported in part by the following projects: Grants-in-Aid for Scientific Research (Nos. 24240013) from MEXT of Japan.

References

1. R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.
2. X. Amatriain, N. Lathia, J. M. Pujol, H. Kwak, and N. Oliver. The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 532–539, New York, NY, USA, 2009. ACM.
3. G. Capannini, F. M. Nardini, R. Perego, and F. Silvestri. Efficient diversification of web search results. *Proc. VLDB Endow.*, 4(7):451–459, Apr. 2011.
4. J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, New York, NY, USA, 1998. ACM.
5. X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pages 315–316, 2005.
6. A. E. Magurran. Measuring biological diversity. *African Journal of Aquatic Science*, 29(2):285–286, 2004.
7. J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 897–908, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.
8. K. A. Nolan and J. E. Callahan. Beachcomber biology: The shannon-weiner species diversity index. In *Proc. Workshop ABLE*, volume 27, pages 334–338, 2006.
9. X. Sha, D. Quercia, P. Michiardi, and M. Dell'Amico. Spotting trends: The wisdom of the few. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 51–58, New York, NY, USA, 2012. ACM.
10. Y. Shoji and K. Tanaka. Diversity-based hits: Web page ranking by referrer and referral diversity. In A. Jatowt, E.-P. Lim, Y. Ding, A. Miura, T. Tezuka, G. Dias, K. Tanaka, A. Flanagin, and B. Dai, editors, *Social Informatics*, volume 8238 of *Lecture Notes in Computer Science*, pages 377–390. Springer International Publishing, 2013.
11. A. Stirling. A general framework for analysing diversity in science, technology and society. *Journal of the Royal Society Interface*, 4(15):707–719, 2007.
12. J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.
13. J. Wang and J. Zhu. Portfolio theory of information retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 115–122, New York, NY, USA, 2009. ACM.