A Personalized Approach for Recommending Useful Product Reviews Based on Information Gain

Joon Yeon Choeh[1], Hong Joo Lee[2] and Sung Joo Park[3]
[1] Department of Digital Contents, Sejong University
209 Neungdongro, Gwangjin, Seoul – Republic of Korea
[e-mail: zoon@sejong.ac.kr]
[2] Department of Business Administration, The Catholic University of Korea
43 Jibongro, Wonmi, Bucheon, Gyeonggi – Republic of Korea
[e-mail: hongjoo@catholic.ac.kr]
[3]KAIST Business School
85 Hoegiro, Dongdaemun, Seoul – Republic of Korea
[e-mail: sjpark@business.kaist.ac.kr]
*Corresponding author: Sung Joo Park

Received November 12, 2014; revised February 19, 2015; accepted April 19, 2015; published May 31, 2015

Abstract

Customer product reviews have become great influencers of purchase decision making. To assist potential customers, online stores provide various ways to sort customer reviews. Different methods have been developed to identify and recommend useful reviews to customers, primarily using feedback provided by customers about the helpfulness of reviews. Most of the methods consider the preferences of all users to determine whether reviews are helpful, and all users receive the same recommendations.

In this paper, we assessed methods for generating personalized recommendations based on information gain. The information gain approach was extended to consider each individual's preference together with votes of other users. A total of 172 respondents rated 48 reviews selected from Amazon.com using a 7-point Likert scale.

The performance of the devised methods was measured by varying the ratio of training sets and number of recommendations for the data collected. The personalized methods outperformed the existing information gain method, which takes into account the votes from all users. The greatest precision was achieved by the personalized method and a method employing selective use of predictions from the personalized method combined with the existing method based on all users' reviews. However, the personalized method, which classified helpful reviews based on each user's threshold value, showed statistically better performance.

Keywords: Product review, Personalization, Text mining, Information gain

This work was supported by a Korean Research Foundation Grant, funded by the Korean Government (MEST)(KRF-2008-220-B00007).

http://dx.doi.org/10.3837/tiis.2015.05.008 **ISSN : 1976-7277**

1. Introduction

Thanks to the proliferation of e-commerce and the great influence of customer reviews on purchase decisions, many products are now being sold and purchased online [1]. Customers believe that reviews written by others who have already had an experience with the product offer more objective and reliable information than that provided by sellers [2]. As a result, an increase in the average review rating leads to a growth in product sales [3], which, in turn, can strengthen the product's price competitiveness [4]. However, if there are too many products and reviews, the advantage of e-commerce can be overshadowed by increasing search costs. Reading all of the reviews to find out the advantages and disadvantages of a particular product can be tedious and exhausting [5,6]. To help users find the most useful information about products without much difficulty, e-commerce companies try to provide various ways for customers to write and rate product reviews. Amazon.com asks customers whether a review on a certain product is helpful, and it places the most helpful favorable and the most helpful critical review at the top of the list of product reviews. Some companies also predict the usefulness of a review based on certain attributes including length, author(s), and the words used, publishing only reviews that are likely to be useful [7, 8].

The methods typically used by e-commerce companies begin from the same assumption, namely that all users share the same concept of helpfulness [7, 9, 10, 11, 12, 13, 14]. In contrast, we assume that every user has his or her own concept of helpfulness. To this end, the present study aimed to develop a model that recognizes individual preferences and to test the models used to predict usefulness and make recommendations that consider individual differences. To do this, we extended the information gain approach to consider the votes of all users as well as each individual's preference. To compare various approaches, the study compared methods that use voting in rating the usefulness of reviews.

For this study, we collected data from 172 people who assessed the usefulness of product reviews through online surveys on a website. Using these data, we identified various types of algorithms and compared the results of personalized product review recommendation methods.

2. Related Studies

2.1 Provisions of Product Reviews

E-commerce companies offer platforms for product reviews in order to provide product information to consumers. Because product reviews play a significant role in making purchase decisions, it is important to single *out* those reviews that provide useful information. Also, due to the large number of product reviews, retailers have provided ways for customers to sort them. **Table 1** shows a summary of information about product reviews for shoppting sites among the top 100 sites on the Web as defined by Alexa[1].

[1] http://www.alexa.com/topsites accessed on August 20, 2014. We inserted only one representative company if there are many local branches; thus, Amazon.com represents both Amazon.co.de and Amazon.co.jp. Some portals may have shopping pages, but we selected sites that are more focused on shopping itself.

Table 1. Provisions and Evaluations of Customer Reviews

Online stores	Information provided about customer reviews	Evaluation of customer reviews
tmall.com (Taobao.com)	Average Ratings, Total Number of Ratings, Distributions of Ratings, Most Recent Customer Reviews, Tags	--
Amazon.com	Average Ratings, Total Number of Ratings, Distributions of Ratings, Most Helpful Customer Reviews, Most Recent Customer Reviews, Excerpts from Customer Reviews	Vote (Yes or No) Helpfulness (Total Number of Helpful Votes / Total Number of Votes)
Yahoo.co.jp	Average Ratings, Total Number of Ratings, Most Helpful Customer Reviews, Most Recent Customer Reviews, Highly Rated Customer Reviews	Vote (Yes) Helpfulness (Total Number of Helpful Votes)
Apple.com	Average Ratings, Total Number of Ratings, Distributions of Ratings, Most Helpful Customer Reviews, Most Recent Customer Reviews	Vote (Yes or No) Helpfulness (Total Number of Helpful Votes / Total Number of Votes)
Aliexpress.com	Average Ratings, Total Number of Ratings, Distributions of Ratings, Most Recent Customer Reviews	Vote (Yes or No)
Booking.com	Average Ratings, Total Number of Ratings	Vote (Yes)
Rakuten.co.jp	Average Ratings, Total Number of Ratings, Distributions of Ratings, Most Helpful Customer Reviews, Most Recent Customer Reviews, Highly Rated Customer Reviews	Vote (Yes or No) Helpfulness (Total Number of Helpful Votes)

Most shopping websites provide customer reviews and offer the following information: the average preference of a product, the number of customers who have participated in preference voting, and preference distribution. The reviews are sorted in categories such as “Helpful Reviews,” “Recent Reviews,” and “Preference Score.” Most information on the helpfulness of product reviews is collected through a voting system. The voting system can be divided into two types; the first type asks whether a review is “helpful” or “not helpful,” and the second type asks whether it is “useful.” The helpfulness of a product review is expressed as the total number of users that rated a review as helpful or as the ratio of helpful votes to the total number of votes.

2.2 Recommending Useful Product Reviews

A number of studies have been conducted regarding ways to recommend useful reviews or likeable products to customers using product reviews. Cao et al. (2011)[7] demonstrated that the prediction of a review's helpfulness is more accurate if usefulness is judged by a combination of the review's basic information, style, and semantic information than when the evaluation is based on a single factor. The basic information includes whether a review describes advantages and disadvantages and the length of time since it was posted; style information includes the word count and the number of sentences; and the semantic information takes into account the meanings of words used in the review. Kim et al. (2006)[11] offered a method to predict the helpfulness of reviews through support vector machine (SVM) regression using structural, semantic, and morphological information about a review. Liu et al. (2008)[12] suggested a model that can predict the helpfulness of a review by considering the reviewers' experience and writing styles, and the date when the review was written. Ghose and Ipeirotis (2011)[9] predicted the helpfulness of reviews by considering product attributes, review attributes, reviewer attributes, reviewers' experiences, reviews' readability, and reviews' subjectivity. The predictions of usefulness produced by these methods are based on indicators of reviews' actual helpfulness, and the results are verified by comparing them with the review rankings.

Some studies have beeen based on the assumption that the prediction of review helpfulness is a matter of distinguishing useful reviews from non-useful reviews. Zhang and Tran (2011)[13] measured the contribution of of particular words to the frequency with which a review was rated as "helpful" or "not helpful." The helpfulness of a review was predicted, and specific reviews were recommended, based on the sum of the contributions of words contained in the review.

Some studies have assessed the meanings of words and sentences within a review to determine whether the review is positive or negative in tone [15, 16, 17, 18]. **Table 2** summarizes the methods of estimating the helpfulness of a review.

Table 2. Methods of estimating helpfulness of a product review

Factors	Features considered	Reference
Reviewer	Experience	[9, 11]
Rating	Star rating	[8, 11]
Basic information of review	Posting date	[7, 11]
Structural information of review	Number of words and sentences, writing style, percentage of nouns or verbs	[7, 8, 11, 12]
Semantic information of review	Meanings or moods of words, appearance of product features, review readability and subjectivity	[7, 9, 11,15, 16, 17]
Information gain	Words' contributions to identifying review as helpful or not helpful review	[13]

Studies predicting the helpfulness of reviews by employing attributes of the review and of the reviewers as well as values reflecting helpfulness voting tend to be used in systems that recommend the most useful reviews for all users instead of making personalized recommendations.

3. Information Gain-based Review Recommendation Algorithm

For a given product $P = \{p_1, p_2, p_3, \dots, p_w\}$, customer $C = \{c_1, c_2, c_3, \dots, c_m\}$, and review $R = \{r_1, r_2, r_3, \dots, r_p\}$, we define the vote V as a matrix of the customers' votes on product reviews v_{c_k, r_i}, which can be expressed by the following formula:

$$V = \begin{pmatrix} v_{c_1,r_1} & v_{c_1,r_2} & \cdots & v_{c_1,r_p} \\ v_{c_2,r_1} & v_{c_2,r_2} & \cdots & v_{c_2,r_p} \\ \cdot & \cdot & \cdot & \cdot \\ v_{c_m,r_1} & v_{c_m,r_2} & \cdots & v_{c_m,r_p} \end{pmatrix} \tag{1}$$

Here, v_{c_k, r_i} includes helpful (if c_k voted r_i as helpful), not helpful (if c_k voted r_i as not helpful), and null (if c_k has not voted for r_i). Let the set of all the "helpful" votes about review r_i be denoted as h_i, and the set of all "not helpful" votes about review r_i be denoted as $\overline{h_\imath}$. We define the helpfulness value of review r_i using the following equation:

$$\frac{|h_i|}{|\overline{h_\imath}| + |h_i|} \tag{2}$$

The average information gain required to classify a review into a group can be defined using entropy as follows:

$$H(S) = -\sum_{i=1}^{q} Pr(s_i) log Pr(s_i) \tag{3}$$

Therefore, the average amount of information contributed by a word or term *t* in a class s_i can be calculated as follows:

$$H(S|t) = -\sum_{i=1}^{q} Pr(s_i|t) log Pr(s_i|t) \tag{4}$$

In the area of text mining and text classification, information gain is the amount of information provided by a word or term. The information gain of the term *t* can be calculated as follows [13]:

$$G(t) = -\sum_{i=1}^{q} Pr(s_i) log Pr(s_i) + Pr(t) \sum_{i=1}^{q} \Pr(s_i|t) log Pr(s_i|t) + Pr(\bar{t}) \sum_{i=1}^{q} Pr(s_i|\bar{t}) log Pr(s_i|\bar{t}) \tag{5}$$

In the above equations:

– $Pr(s_i)$ is the probability of documents' being included in the category s_i among all documents;

– $Pr(t)$ is the probability of documents containing *t* among all documents;

– $Pr(s_i|t)$ is the probability of documents with *t* being included in category s_i among all documents containing *t*; and

– $Pr(s_i|\bar{t})$ is the probability of documents containing *t* belonging to category s_i among all documents that do not contain *t*.

3.1 Information gain—Total method

The above algorithm (Equation 5) was suggested by Zhang and Tran (2011) [13]. In this approach, if the total user usefulness rating is greater than 0.6, it is classified in the helpful review group (s_1); otherwise, it is classified in the not helpful review group (s_2). In the vote matrix V above, let a "helpful vote" be assigned the value 1, and a "not helpful vote" be assigned the value 0. If the helpfulness value calculated from Equation (2) is greater than 0.6, review r_j is classified into the helpful review group.

Based on $G(t)$ from Equation (5), the final $Gain(t)$ can be calculated as follows:

$$Gain(t) = \begin{cases} G(t) & if\ Pr(s_1|t) < Pr(s_2|t), \\ -G(t) & otherwise. \end{cases} \tag{6}$$

By using the information gain ($Gain(t)$) of *t* as calculated above, the predicted helpfulness score of a new review (r_i) for all customers can be calculated as:

$$Score(r_i) = \sum_{j=1}^{M} Gain(t_j) * f(r_i, t_j), \tag{7}$$

where M is the total number of stemmed words in review r_i, and t_j is the j th stemmed word. If (t_j) is included in (r_i), then $f(r_i, t_j)$ is 1; if (t_j) is not included in (r_i), then $f(r_i, t_j)$ is 0. Among the reviews that were not evaluated by a particular customer, that customer will receive recommendations for reviews with high predicted usefulness scores as calculated using Equation 7. Because this method reflects the opinions of all users, it recommends the same reviews to all users, provided only that they did not yet vote on the reviews.

3.2 Information gain – Personalized method

Based on the concept of using information gain to classify helpful reviews, in this paper, we devised personalized recommendation algorithms. The first personalized algorithm assessed the helpfulness scores of reviews based on each individual's review preference and the information gain of *t*.

In contrast to the total method described in Section 3.1, which considered the votes of all users, this personalzed method classified a review into helpful or not helpful review groups based on each individual's vote. If the value of an element (v_{c_k,r_i}) is 1 in matrix V, it is classified as representing the helpful review group (s_1); if the value of (v_{c_k,r_i}) is 0, it is classified as representing the not helpful group (s_2). Based on this,the information gain of *t* for customer c_k can be calculated using Equation (5-1) based on the review classification described above.

$$G_{c_k}(t) = -\sum_{i=1}^{2} Pr_{c_k}(s_i) log Pr_{c_k}(s_i) + Pr_{c_k}(t) \sum_{i=1}^{2} Pr_{c_k}(s_i|t) log Pr_{c_k}(s_i|t) + Pr_{c_k}(\bar{t}) \sum_{i=1}^{2} Pr_{c_k}(s_i|\bar{t}) log Pr_{c_k}(s_i|\bar{t}) \quad (5\text{-}1).$$

In the above equation:

– $Pr_{c_k}(s_i)$ is the probability of reviews' being included in category s_i among all reviews rated by c_k;

– $Pr_{c_k}(t)$ is the probability of reviews containing *t* among all reviews rated by c_k;

– $Pr_{c_k}(s_i|t)$ is the probability of reviews with *t* being included in category s_i among all reviews rated by c_k containing *t*; and

– $Pr_{c_k}(s_i|\bar{t})$ is the probability of reviews containing *t* belonging to the category s_i among all reviews rated by c_k that do not contain *t*.

Finally, $G_{c_k}(t)$ can be calculated as follows.

$$Gain_{c_k}(t) = \begin{cases} G_{c_k}(t) & if\ Pr_{c_k}(s_1|t) < Pr_{c_k}(s_2|t), \\ -G_{c_k}(t) & otherwise. \end{cases} \quad (6\text{-}1)$$

By using the information gain ($Gain_{c_k}(t)$) of the word t as calculated above, the predicted helpfulness of a new review (r_i) for customer c_k can be calculated as follows;

$$Score_{c_k}(r_i) = \sum_{j=1}^{M} Gain_{c_k}(t_j) * f(r_i, t_j)\ . \tag{7-1}$$

Customer c_k will get recommendations of reviews with high predicted helpfulness scores. Though the concept of using the information gain of a word is similar to the total method described in Section 3.1, this personalized method uses the concept to estimate helpfulness scores based on each individual's preferences. By classifying a review's helpfulness, calculating the information gain of a word and estimating the helpfulness score, this personalized algorithm only takes into account the individual's preference in selecting reviews.

3.3 Information gain—Weighted personalized method

In Section 3.1, when estimating the helpfulness scores of reviews, we included votes from all users in calculating the information gain of *t*, and in Section 3.2, information gain was calculated using only a single user's vote. However, when people decide or evaluate something, they are influenced by others as well as by their own experience and subjective evaluation.

Thus, when calculating the final helpfulness predictions, we considered estimates made using the votes of all users as well as the preferences of the individual. To do this, we devised a weighted personalized method by averaging predicted helpfulness scores from Equation (7) and Equation (7-1) as follows:

$$Score_{c_k}(r_i) = {\left\{\sum_{j=1}^{M} Gain_{c_k}(t_j) * f(r_i, t_j) + \sum_{j=1}^{M} Gain(t_j) * f(r_i, t_j)\right\}}\Big/_{2}. \tag{8}$$

By taking the average of the predicted helpfulness scores, we considered the opinions of all users as well as each individual's subjective evaluations.

3.4 Information gain—Selective personalized method

Using the weighted personalized algorithm in Section 3.3, we can selectively base helpfulness predictions on the opinions of all users or on each individual's preferences. To this end, we devised Equation (9), which uses predictions selectively based on the values of the prediction scores:

$$Score_{c_k}(r_i) = \begin{cases} Score_{c_k}(r_i) & if\ Score(r_i) < Score_{c_k}(r_i), \\ Score(r_i) & otherwise. \end{cases} \tag{9}$$

where $Score(r_i)$ is the predicted helpfulness score considering the ratings of all users, taken from Equation (7), and $Score_{c_k}(r_i)$ is the predicted helpfulness score based on a user's preference, introduced in Section 3.2.

If the prediction score from the personalized method is greater than that from the total method, then the final prediction score is the score from the personalized method. Otherwise, the predicted scores from the total method are used for the final predictions.

Using this algorithm, we usually followed the opinions of other users because raters generally showed broad agreement in evaluating helpful reviews. However, we selectively used prediction scores from the personalized method if a review was considered more helpful to a certain user.

3.5 Information gain—Personalized method with average threshold

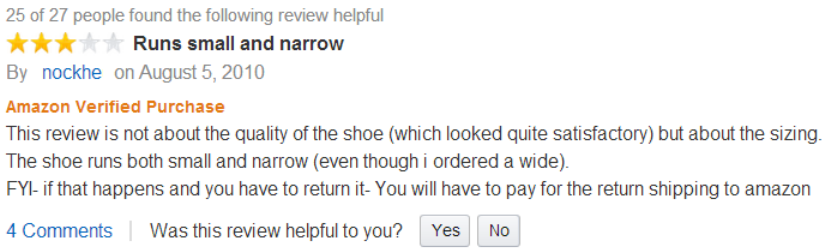
The methods introduced in Section 3.1 and 3.2 employed the vote matrix V in which entries were scored as 1 if helpful or 0 if not helpful. As shown in **Table 1**, it is common to rate a review by indicating that it is helfpul or not helpful. As discussed in Section 3.1, if the ratio of helpful to not helpful votes for a review is over 0.6, the review is considered helpful. In the personalized algorithm, a review is classified by each user's vote regardles of whether the review is regarded as helpful.

Another approach is to rate a review using a Likert scale-like star rating system. We will explain our data later in Section 4; briefly, we obtained helpfulness ratings of reviews using a 7-point Likert scale. Because we employed a 0.6 ratio for classifying helpful reviews in Section 3.1, here, we classified the ratings as helpful (or 1) if the Likert rating was 4 or above, and as not helpful otherwise. In Section 4, however, the threshold for classifying helpful reviews differs among users. Thus, when we applied personalized methods, we used each user's average ratings as the threshold to classify helpful reviews. The personalzed method described in Section 3.2 used a score of 4 or above on the Likert scale to designate a review as helpful. However, the personalized method with an average threshold used a different classification scheme for each user based on that user's average rating scores. If one user's average rating score was 4.8, we classified reviews with ratings of rated 5 and above by that user as helpful, and all other reviews were classified as not helpful, even if they had a rating of 4.

The remaining procedures used to estimate helpfulness scores are the same as those in the personalized method described in Section 3.2.

4. Data and Experiment

Product reviews used for this study were collected from Amazon.com. **Fig. 1** shows an example of a product review posted on the site. At the top of the review, the total number of people who have rated the product and the number of people who considered the review helpful are indicated. The website allows customers to leave a textual review and to give a star rating for the products. In the present study, perfumes and books were selected as "experience goods," and shoes and cameras were selected as "search goods." [19, 20] As shown in **Table 3**, two product items per product group and six reviews per item were selected. The selected reviews were composed of two entries with the greater number of votes, another two with medium-level number of votes, and two entries with the lower number of votes.



25 of 27 people found the following review helpful

★★★☆☆ **Runs small and narrow**

By nockhe on August 5, 2010

Amazon Verified Purchase

This review is not about the quality of the shoe (which looked quite satisfactory) but about the sizing.
The shoe runs both small and narrow (even though i ordered a wide).
FYI- if that happens and you have to return it- You will have to pay for the return shipping to amazon

4 Comments | Was this review helpful to you?

Fig. 1. A screen shot of a product review on amazon.com

Photos and information regarding the selected products were shown to study participants, and a website was built to allow them to rate the helpfulness of each product's set of reviews. A total of 172 people participated in the experiment. Each participant was instructed to read six reviews and rate their helpfulness on a scale of 1 to 7, where 1 meant "Not helpful at all" and 7 indicated "Very helpful." To exclude the influence of the sequence of reviews on the evaluation, the order of reviews was randomly selected.

Table 3. Products and review data used in the experiment

Product type	Category	Product
Search goods	Shoe	Adidas Men's Kanadia Trail Running Shoe
		Hi-Tec Men's Altitude IV Hiking Boot
	Camera	Nikon D80 10.2 MP Digital SLR Camera Kit
		Sony Cyber-shot DSC-HX5V 10.2 MP CMOS
Experience goods	Perfume	Acqua Di Gio By Giorgio Armani For Men
		Lovely by Sarah Jessica Parker Eau de Parfum
	Book	Outliers: The Story of Success
		The 8th Habit: From Effectiveness to Greatness

The average rating given by the participants to the forty-eight reviews was 4.862 (SD = 0.83). The distribution of scores was as follows: a rating of 1, 205 reviews; a rating of 2, 323 reviews; of 3, 695 reviews; of 4, 2,153 reviews; of 5, 1,888 reviews; of 6, 1,790 reviews; and a rating of 7, 202 reviews. The most frequently selected rating was 4 points, and the least frequently selected was 1, followed by 2 and 3. As the ratings increased from 5 to 6 and then to 7, the frequency tended to decrease.

After calculating the frequency of the words included in the 48 reviews, we created a matrix of words and documents. The tm package of R [21] was used to extract the words. Stemming was done using the tm_map function in the tm package after we removed numbers, stopwords, and symbols. Then, we extracted a total of 2,029 words. **Table 4** shows the 10 most frequently used words in our dataset.

Table 4. Top 10 most frequently used words

Word	Frequency
camera	63
book	47
nikon	22
people	21
mode	20
shoes	19
boots	18
gladwell	17
success	16
time	15

The average frequency of use for all words was 2.0001; the highest number was for words used only one time (1,289 words). To apply the above data to an information gain-based review recommendation algorithm, reviews with ratings of 4 and above were assigned a helpfulness value of 1, and the remaining reviews were assigned value of 0. To recommend reviews based on the information gain found by using the ratings of all users, those reviews that were rated as helpful by more than 60% of all users (104/172 users) were classified in the helpful review group; others were classified in the not helpful review group. This conversion was applied for all of the methods introduced in Section 3, except the personalized method with an average threshold, described in Section 3.5, which used each user's average rating as the threshold value for classifying reviews as helpful or not.

5. Experimental Results

The experiment was conducted using 30%, 50%, and 70% of the entire dataset as training data. Splitting for each case was conducted randomly and was repeated 30 times. Then, the average performance measures were obtained using the 30 randomly split datasets. Additionally, recommendations for reviews were repeated using reviews with the top three, top five, and top seven helpfulness values.

The ratio of helpful reviews included in the recommendation was taken as a measure of the precision of the recommendation, and recall was calculated as the ratio of helpful reviews in the recommendation to the total number of helpful reviews in the test dataset. Given that precision decreasaed as the number of recommended reviews increased, whereas recall showed a tendency to increase, and given its wide use as a performance indicator, an F measure indicator was used as an indicator of performance:

$$F\ measure = \frac{2*Precision*Recall}{Precision+Recall}. \tag{10}$$

Table 5 shows a summary of the recommendation performance of the methods suggested in Section 3. The performance of a method that randomly recommends reviews for users was used as a baseline for comparison. To compare the performance of the various methods, the F test was conducted for F scores, and a Tukey test was performed as a post hoc test to compare the F scores for the methods.

The F test revealed that the mean values for all three measures differed significantly among methods (**Table 5**). The random method showed the lowest F scores in all situations, as shown in the reuslts of the Tukey test. The F scores of the total method were higher than those of the random method, but lower than those of the personalized methods. It was hard to distinguish among the performances of the personalized methods. Generally, the personalized with average threshold and the personalized method had greater F scores than did the weighted personalized and the selective personalized methods. However, in some cases no signifant differences were found among the personalized methods.

When the size of the training set decreased from 70% to 50% and then to 30%, the F values for the top three, five, and seven helpfulness scores showed a tendency to decrease. Though the larger training sets showed higher performance in terms of recall, the precision values were between 0.6 to 0.7. This can be attributed to the fact that recall values decreased because more helpful reviews were included in the test sets. If we consider only precision, we believe that the influence of the volume of the training set on the information gain methods was minimal.

Table 5. Experiment results

		Top 3			Top 5			Top 7		
		Prec.	Recall	F	Prec.	Recall	F	Prec.	Recall	F
Training set ratio: 0.7										
A	Random	0.587	0.208	0.291	0.583	0.342	0.406	0.571	0.471	0.485
B	Total	0.660	0.241	0.332	0.624	0.369	0.436	0.610	0.501	0.518
C	Personalized	0.687	0.270	0.365	0.653	0.417	0.479	0.623	0.550	0.550
D	Weighted personalized	0.678	0.264	0.360	0.643	0.413	0.475	0.615	0.544	0.544
E	Selective personalized	0.687	0.262	0.355	0.657	0.409	0.471	0.629	0.535	0.542
F	Personalized with average threshold	0.633	0.280	0.375	0.599	0.437	0.487	0.565	0.568	0.547
p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Tukey test				F, C, D, E > B > A			F, C, D, E > B > A			C, F, D, E > B > A
Training set ratio: 0.5										
A	Random	0.612	0.127	0.202	0.585	0.202	0.283	0.596	0.291	0.367
B	Total	0.670	0.143	0.224	0.664	0.234	0.327	0.648	0.315	0.400
C	Personalized	0.696	0.163	0.250	0.672	0.257	0.350	0.650	0.344	0.422
D	Weighted personalized	0.673	0.156	0.241	0.648	0.247	0.336	0.627	0.330	0.407
E	Selective personalized	0.695	0.154	0.239	0.684	0.250	0.345	0.669	0.338	0.422
F	Personalized with average threshold	0.650	0.169	0.261	0.629	0.269	0.366	0.602	0.357	0.436

p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Tukey test				F, C = E, D > B > A			F, C = E, D > B > A			F, C, E, D > B > A
Training set ratio: 0.3										
A	Random	0.583	0.085	0.142	0.602	0.147	0.225	0.589	0.203	0.285
B	Total	0.630	0.093	0.157	0.632	0.156	0.239	0.626	0.215	0.304
C	Personalized	0.684	0.112	0.185	0.663	0.181	0.270	0.645	0.244	0.335
D	Weighted personalized	0.639	0.103	0.170	0.625	0.165	0.249	0.611	0.226	0.313
E	Selective personalized	0.677	0.103	0.172	0.665	0.168	0.255	0.657	0.232	0.323
F	Personalized with average threshold	0.633	0.114	0.190	0.616	0.186	0.279	0.604	0.255	0.349
p-value		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Tukey test				F, C > E, D > B > A			F, C > E, D > B > A			F, C = E, D = B > A

6. Conclusion

The personalized recommendation methods suggested in this study performed better than the total method, which recommends the same reviews to all users. This result was consistent in the cases of the top three, five, and seven helpfulness scores for various sizes of training sets (70%, 50%, and 30% of the total data). Among the personalized methods, the personalized with average threshold and the personalized method had higher F scores than did the weighted personalized and the selective personalized methods. However, their performances did not always differ significantly. Thus, caution should be used when selecting an appropriate method for personalized review recommendations. Data should be tested empirically to find the best algorithm in individual cases.

When we recommended the top three reviews using 70% of the data and the personalized methods, the F score was around 0.3; F scores with 50% and 30% of the data were around 0.2 and less than 0.2 respectively. This was also true for the top five and top seven recommendations. This implies that more training data resulted in better performance.

Though the personalized methods outperformed the total method, the total method has the advantage of being able to recommend reviews for a user even if that reviewer had not voted on a review at all, whereas the personalized methods require that users have voted on other reviews because the methods need to understand the user’s preferences.

An information gain approach, whether total or personalized, can recommend a review that has not been rated yet because the method only investigates the contents of the review. Therefore, it is useful for deciding whether a relatively new review is helpful when that review has not received enough votes for other algorithms to be applied.

In this study, we collected the preferences of users in only four product categories: perfume, books, cameras, and shoes. Thus, caution should be used when applying the suggested methods to other product categories. For personalization, we may use other information such as user preferences for products. Thus, opportunities remain for further studies that combine votes on reviews with other information so as to improve recommendations.

References

[1] C. Dellarocas, G. Gao, and R. Narayan, "Are consumers more likely to contribute online reviews for hit or niche products?," *Journal of Management Information Systems*, vol. 27, no. 2, pp. 127-157, 2010. Article(CrossRef Link)

[2] C. Dellarocas, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, vol. 49, no. 10, pp. 1407-1424, 2003. Article(CrossRef Link)

[3] F. Zhu and M. Zhang, "The influence of online consumer reviews on the demand for experience goods: the case of video games," in *Proc. of the Twenty-Seventh International Conference on Information Systems*, Paper No. 25, Available at http://aisel.aisnet.org/icis2006/25, 2006.

[4] Xinxin Li, Lorin Hitt, John Z. Zhang, "Product Reviews and Competition in Markets for Repeat Purchase Products," *Journal of Management Information Systems*, vol. 27, no. 4, pp. 9-42, 2011. Article(CrossRef Link)

[5] S. David and T. Pinch, "Six Degrees of Reputation: The Use and Abuse of Online Review and Recommendation Systems," *First Monday*, vol. 11, no. 3, 2006. Article(CrossRef Link)

[6] Y. Liu, "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing*, vol. 70, no. 3, pp. 74-89, 2006. Article(CrossRef Link)

[7] Q. Cao, W. Duan, and Q. Gan, "Exploring determinants of voting for the 'helpfulness' online user reviews: A text mining approach," *Decision Support Systems*, vol. 50, no. 2, pp. 511-521, 2011. Article(CrossRef Link)

[8] S.M. Mudambi and D. Schuff, "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly*, vol. 34, no. 1, pp. 185-200, 2010. Available at http://aisel.aisnet.org/misq/vol34/iss1/11/

[9] A. Ghose and P. G. Ipeirotis, "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 10, pp. 1498-1512, 2011. Article(CrossRef Link)

[10] H. Hsiao, C. Wei, Y. Ku, and A. Luisa, "Predicting The Helpfulness of Online Product Reviews: A Data Mining Approach," in *Proc. of PACIS 2012*, Paper No. 134, Available at http://aisel.aisnet.org/pacis2012/134/, 2012.

[11] S.M. Kim, P. Pantel, and T. Chklovski, "Automatically assessing review helpfulness," in *Proc. of the Conference on Empirical Methods in Natural Language Processing*, pp. 423-430, 2006. Article(CrossRef Link)

[12] Y. Liu, X. Huang, A. An, and X. Yu, "Modeling and Predicting the Helpfulness of Online Reviews," in *Proc. of the Eighth IEEE International Conference on Data Mining*, pp. 443-452, 2008. Article(CrossRef Link)

[13] R. Zhang, and T. Tran, "An Information gain-based approach for recommending useful product reviews," *Knowledge and Information Systems*, vol. 26, no. 3, pp. 419-434, 2011. Article(CrossRef Link)

[14] R. Zhang, T. Tran, and Y. Mao, "Real-time helpfulness prediction based on voter opinions," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 17, pp. 2167-2178, 2012. Article(CrossRef Link)

[15] Hatzivassiloglou, V., and K., McKeown, "Predicting the semantic orientation of adjectives," in *Proc. of 8th conference on European chapter of the association for computational linguistics*, pp.174-181, 1997. Article(CrossRef Link)

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proc. of the ACL-02 conference on empirical methods in natural language processing*, pp. 79-86, 2002. Article(CrossRef Link)

[17] P. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *Proc. of the 40th annual meeting of the association for computational linguistics*, pp. 417-424, 2002. Article(CrossRef Link)

[18] H. Yu, and V. Hatzivassiloglou, "Toward answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences," in *Proc. of the 2003 conference on*

empirical methods in natural language processing, pp. 129-136, 2003. Article(CrossRef Link)

[19] P. Nelson, "Information and Consumer Behavior," *Journal of Political Economy*, vol. 78, no. 2, pp. 311-329, 1970. Article(CrossRef Link)

[20] P. Nelson, "Advertising as Information," *Journal of Political Economy*, vol. 82, no. 4, pp. 729-754, 1974. Article(CrossRef LInk)

[21] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R," *Journal of Statistical Software*, vol. 25, no. 5, pp. 1-54, 2008. Article(CrossRef LInk)

Joon Yeon Choeh is an assistant professor at the Department of Digital Contents, Sejong University, Korea. He received M.S. and Ph.D. degree in management engineering from KAIST, respectively. His main research interests include recommender system, context-aware system and social media mining. He has published papers which have appeared in Expert Systems with Applications, AI Communications, International Journal of Computational Intelligence Systems, Expert Systems etc.

Hong Joo Lee is an associate professor of business administration, the Catholic University of Korea. He has a Ph.D. from the KAIST Business School (2006) and was with the MIT Center for Collective Intelligence as a postdoctoral fellow in 2006 and a visiting scholar in 2011. His research areas are utilizing intelligent techniques and harnessing collective intelligence in business settings, and analyzing effects of intelligent aids in e-commerce and m-commerce. His papers have been published in International Journal of Electronic Commerce, Decision Support Systems, Expert Systems with Applications, Information Systems Frontiers, and other journals.

Sung Joo Park is a Professor of Information Systems at the KAIST Graduate School of Management in Seoul, Korea. He holds a B.S. degree in Industrial Engineering from the Seoul National University, an M.S. in Industrial Engineering from the Korea Advanced Institute of Science, and a Ph.D. in Systems Science from the Michigan State University. He has been a senior researcher at the Software Develop- ment Center, KIST, and a professor at the KAIST since 1980. His areas of research interests include intelligent information systems and the application of agent technology to management decision-making.