Expert Systems with Applications 36 (2009) 3751–3760

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Q-rater: A collaborative reputation system based on source credibility theory

Jinhyung Cho [a,*], Kwiseok Kwon [b,*], Yongtae Park [c]

[a] School of Computing and Information, Dongyang Technical College, 62-160 Gochuk-Dong, Guro-Gu, Seoul 152-714, Republic of Korea
[b] Interdisciplinary Graduate Program of Technology and Management, Seoul National University, Seoul, Republic of Korea
[c] Department of Industrial Engineering, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Reputation system
E-commerce
Source credibility theory
Collaborative filtering
Rating aggregation

ABSTRACT

The consumer is an important information source and the after-transaction-feedback is used as the quality indicator for trust building in e-commerce. However, the possibility of incorrect information from unreliable users and declining trust in the electronic market looms large. Hence, there is a need for a fairer and more objective reputation mechanism. Most existing reputation systems focus on the reputation of users than items and they rely on explicit feedback information. As a solution to these problems, we propose a reputation system ("*Q-rater*") suitable for B2C e-commerce. We adopt the source credibility theory of consumer psychology and the basic mechanism of collaborative filtering methods for the overall process of the proposed reputation system. We also evaluate performance of the *Q-rater* experimentally by comparing it with the other benchmark systems and observing the performance changes with variations in the size of rater group and item rating aggregation mechanism.

1. Introduction

E-commerce involves a high level of "uncertainty" regarding both participant reliability and the quality of the items (products or services) being traded; this is because traditional authentication mechanisms based on physical inspection are not feasible online (McKnight, Choudhury, & Kacmar, 2002). Due to this uncertain nature of e-commerce, after-transaction-feedback information—such as comments, reviews, and ratings regarding the providers or the products—is becoming increasingly important in building trust in the electronic market. Such forms of information are referred to as "online word-of-mouth (WOM)." For example, a growing number of e-commerce sites such as Amazon.com display evaluations by consumers in order to offer other consumers reliable products or services on their sites and to maintain their competitiveness. Therefore, in e-commerce, the consumer is an important "information source" and the evaluation feedback of the consumers is used as the quality indicator. In other words, the consumer plays an important role as a rater.

However, unlike consumer-to-consumer (C2C) e-commerce sites, most business-to-consumer (B2C) sites do not provide users with explicit information on the reputation of a user as assessed by other users. This suggests the absence of a feedback tool for user reputation; a user is under no obligation to express an opinion regarding other users. Therefore, the possibility of incorrect information from unreliable users and declining trust in the electronic market looms large. For example, a malicious user can assign a poor rating to and/or spark off unpleasant rumors regarding a product from a specific vendor. Hence, there is a need for a fairer and more objective rating mechanism for a rater (a user who rates); this mechanism should involve measuring the credibility of the user. Further, since the small number of ratings cause difficulties in generating overall item reputation, there is a need for a more reliable aggregation mechanism for item evaluations.

As a solution to these problems, we propose an online reputation system suitable for B2C e-commerce sites where both explicit user evaluation ratings and social relationships among users are inadequate. The conceptual framework of our proposed mechanism is based on the source credibility theory for WOM (Word-Of-Mouth) communications model in consumer psychology. We adopt the basic mechanism of collaborative filtering methods for the overall process of the proposed reputation system. With the proposed approach, we implement a pilot reputation rating system, "*Q-rater*." We also evaluate the performance of the proposed system experimentally by comparing it with the other benchmark systems and observing the performance changes with variations in the size of rater group and item rating aggregation mechanism.

2. Reputation systems and e-commerce

2.1. Reputation systems, their mechanisms, and collaborative filtering

Reputation systems aggregate users' feedback after the completion of a transaction and compute the "reputation" of products,

[*] Corresponding authors. Address: School of Computing and Information, Dongyang Technical College, 62-160 Gochuk-Dong, Guro-Gu, Seoul 152-714, Republic of Korea. Tel.: +82 2 2610 1864; fax: +82 2 2610 1859.
E-mail addresses: solver3@gmail.com (J. Cho), kwiseok@gmail.com (K. Kwon).

0957-4174/$ - see front matter © 2008 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2008.02.034

services, or providers, which can assist other users in decision-making in the future (Jøsang, Ismail, & Boyd, 2007; Resnick, Zeckhauser, Friedman, & Kuwabara, 2000). Most large-scale e-commerce sites currently have reputation systems that provide guidance for purchase decisions. These systems present ranks and/or ratings of items and reputation information on users on the basis of feedback collected from users who have participated in e-commerce transactions. Many reputation systems have been proposed recently.

A reputation system can have several essential, separate mechanisms that are responsible for information gathering, scoring and ranking, and response management (Marti & Garcia-Molina, 2006). However, since there exists no systematic research on the mechanism of a reputation system, we define and split a reputation system into two mechanisms in terms of technical aspects: "user reputation generating mechanism" and "item rating aggregation mechanism".

The user reputation generating mechanism involves generating the reputations of users. Such mechanisms in e-commerce can be broadly classified into two different types depending on the e-business model. The first type is a bidirectional rating mechanism for enterprises with the C2C e-business model, such as online auction sites and peer-to-peer (P2P) services (e.g., eBay and Napster), where users are rated by other users. In this case, both the raters and the rated objects are users participating in transactions, and the reputation of a user can be extracted from explicit ratings assigned to him/her. The second type is a unidirectional rating mechanism for enterprises with the B2C e-business model, such as online shopping mall sites (e.g., Amazon.com) or online evaluation service sites (e.g., Epinion.com), where raters, products, or services are rated explicitly by general users or selected evaluators.

User reputation generating mechanisms can also be classified on the basis of source-of-reputation information. Since the conventional reputation system uses explicit feedback information such as user reputation as evaluated by other users, we term the conventional system "explicit reputation system." In most explicit reputation systems, consumers submit feedback voluntarily. Thus, in the absence of concrete incentives, online users may refrain from providing evaluation feedback, or they may submit feedback that is intentionally or unintentionally untruthful (Dellarocas, 2003).

As an alternative method, the "implicit reputation system" is a promising solution to overcome the problems of inadequate and untruthful feedback accompanying the explicit reputation system. In addition, the implicit reputation mechanism might be a viable substitute of or complement to the voluntary feedback mechanism of the bidirectional reputation system in C2C e-commerce (Dellarocas, 2003). There exist a few methods that use implicit information; in these methods, implicit reputation information is derived by analyzing the position of each user within the corresponding social network (Hogg & Adamic, 2004; Pujol, Sanguesa, & Delgado, 2002).

Therefore, user reputation generating mechanisms can be categorized into four main groups using two criteria, business model and reputation information source, as shown in Table 1.

However, in the case of B2C e-commerce sites with a wide range of product categories and innumerable anonymous users, the explicit feedback information is inadequate for all users, and the implicit information of social networks is unavailable. Therefore, it would be more practical to extract users' reputations automatically and implicitly from their past rating data for transaction items. Most existing reputation systems focus on generating ratings only for user reputation; they fail to consider the reputations of products or services, referred to as (defined as) "item reputation" or "the rating of objects" (Chen & Singh, 2001). However, it is essential for B2C e-commerce sites to have a reliable reputation rating mechanism for items since they offer guidance for decision-making by presenting the ranks or ratings of items. The item rating aggregation mechanism involves generating the reputation of the rated items. In this paper, we also term it "item reputation generating mechanism." Item ratings (also known as feedback, recommendations, reviews, and referrals) can serve as a very powerful tool to find sought-after resources and information in online communities. In this study, we evaluated several item rating aggregation mechanisms in the context of a source-credibility-theory-based reputation system.

Collaborative filtering methods compute recommendations by computing the similarities between a target user's preferences and those of other people. Collaborative filtering can be used to improve a reputation system (Jøsang et al., 2007; Zacharia, Moukas, & Maes, 2000), and we apply this method to computing reputations. In other words, we can weight the ratings assigned by a rater on the basis of the rater's own reputation. Trust-based collaborative filtering methods (Massa & Avesani, 2004; O'Donovan & Smyth, 2005) incorporate this concept. In our study, we calculate the reputation of a given user by using a few schemes of collaborative filtering methods and predict the overall ratings for item reputation by adopting the basic mechanism of collaborative filtering methods. Unlike the pure collaborative filtering methods, our Q-Rater system does not create personalized recommendations for individual users. Instead, it uses raters' ratings to estimate the raters' underlying credibility (implicit user reputation) and to predict the general users' opinion for items' reputation rating.

2.2. Reputation and the source credibility theory

"Reputation" can be defined as a collective measure of "trust" based on the ratings assigned by members in a community (Jøsang et al., 2007). In the real world, people conceptualize the reputation of a person subjectively by using multi-dimensional qualification criteria such as expertise, intelligence, reliability, honesty, attractiveness, familiarity, and so on. However, most existing trust and reputation systems have quantified the reputation value of users by using a uni-dimensional measure. Therefore, it is necessary to define and measure the concept of reputation using multidimensional criteria that are more systematic. Moreover, in the previous trust and reputation systems, trust was variously defined. We propose the source credibility theory for WOM communications model in consumer psychology for a more systematic definition and classification of trust-related concepts. According to the source credibility theory, the credibility of an information source comprises expertise (competency), trustworthiness, co-orientation (similarity), and attraction (Robertson, Zielinski, & Ward,

Table 1
Classification of the user reputation generating mechanisms

Business model	Information source	
	Explicit	Implicit
C2C (Bidirectional)	Online auction site (e.g., eBay) P2P service (e.g., Napster)	Social network analysis (e.g., Web of Trust)
B2C (Unidirectional)	Online shopping (e.g., Amazon.com) Online evaluation site (e.g., Epinion.com)	The proposed system (Q-rater)

Table 2
Four key factors of source credibility

Factors	Description
Expertise	The extent to which a source is perceived as being capable of providing correct information
Trustworthiness	The degree to which a source is perceived as providing information that reflects the source's actual feelings or opinions
Co-orientation	The degree to which a source is similar to the target audience members, or is depicted as having similar problems or other characteristics relating to use of a particular product or brand
Attraction	The extent to which a source elicits positive feelings from audience members, such as a desire to emulate the source in some way

1984; Hawkins, Best, & Coney, 2004; Schweitzer, 1969). Each factor is defined as Table 2.

Among the source credibility factors, attraction is not appropriate for the online reputation system because it is generated in the environment where an information source is revealed to an information receiver. Thereby, among the source credibility factors, we adopted three key factors for reputation measurement—expertise, trustworthiness, and co-orientation excepting for attraction. Here, expertise is defined as the extent to which an information source is perceived as being capable of providing correct information, trustworthiness implies the degree to which an information source is perceived as providing information that reflects the source's actual feelings or opinions, and co-orientation implies the degree to which an information source is similar to the target users.

Ekström, Björnsson, and Nass (2005) proposed a reputation mechanism for B2B e-commerce based on the source credibility theory. However, in order to enable the calculation of users' credibility values, users had to assign explicit ratings directly to other users. They focused only on generating the reputation of the raters. However, in our proposed reputation system, the credibility of users is implicitly extracted from their past rating data in order to conduct a more objective calculation of their reputations. Our mechanism differs from that of Ekström et al. (2005) in that we focus on generating both user and item reputations implicitly.

3. Proposed approach

3.1. Overview

Our proposed mechanism focuses on the B2C e-business model, which uses unidirectional ratings. Moreover, it generates both user reputation and item reputation implicitly from users' past rating data for transaction items, without applying direct evaluations by users.

The overall process of our proposed reputation system (*Q-Rater*) is divided into three phases as shown in Fig. 1; the "user credibility extraction phase," "user reputation generation phase" and the "item reputation generation phase." In the first phase, the source credibility factor of each user is measured implicitly from the user's explicit ratings for the items. In the second phase, the reputation of each user is generated by a combination of three source credibility factors and the qualified rater group is formed by the user reputation. The item reputation generation phase is executed by the rating tendency calculation and the item rating aggregation to predict general users' evaluation for each item. Through forming the qualified rater group, the item's reputation will be generated based on the reputation-weighted rating aggregation mechanism.

3.2. Phase I: User credibility extraction

3.2.1. Expertise measurement

We define the expertise factor on the basis of the source credibility theory as the degree of a user's competency to provide an accurate prediction and exhibit a high activity in the item category. Based on this definition, we devised a measure of expertise reflecting activity and prediction competency at the category level. The expertise of a user u for category c, $\omega_{\mathrm{E}}(u,c)$, is defined in

$$\omega_{\mathrm{E}}(u,c) = \beta(u,c)\left(1 - \frac{\sum_{j\in C(i)}\sum_{a\in U(j)}|R_{u,j} - R_{a,j}|}{N_{C(i)}}\right) \qquad (1)$$

where

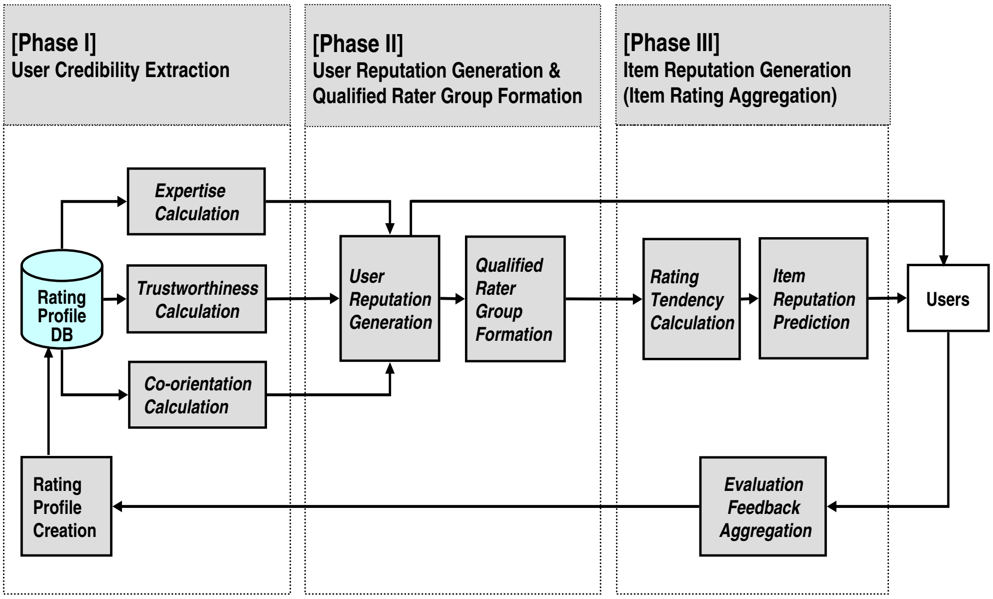

[Phase I]
User Credibility Extraction
[Phase II]
User Reputation Generation &
Qualified Rater Group Formation
[Phase III]
Item Reputation Generation
(Item Rating Aggregation)
Expertise Calculation
Trustworthiness Calculation
Co-orientation Calculation
Rating Profile DB
User Reputation Generation
Qualified Rater Group Formation
Rating Tendency Calculation
Item Reputation Prediction
Users
Evaluation Feedback Aggregation
Rating Profile Creation

Fig. 1. The overall process of the proposed reputation system (*Q-rater*).

$R_{u,j}$ is the rating of user (rater) u for item j
$U(j)$ are the users who assigned rating for item j, except for user u
$C(i)$ is the item sets with ratings in the category of target item i
$N_{C(i)}$ is the cardinality of $C(i)$
$\beta(u,c)$ is the activity weighting

The activity weighting, $\beta(u,c)$, is defined as $1 - 1/n$ (n: the number of ratings within the category) in order to obtain a higher value of expertise with more rating activities for more items within a particular category.

3.2.2. Trustworthiness measurement

We define the trustworthiness factor on the basis of the source credibility theory as the degree to which a user is perceived as providing information that reflects his actual feelings or opinions. If a user's ratings are similar to the average ratings of the general users in the community to which he belongs, we assume that he offers his true opinions. Therefore, a user's trustworthiness is measured by the similarity between his/her rating and the mean of the ratings of general users. Pearson's correlation coefficient, the most widely used coefficient in collaborative filtering methods, is employed. We define the trustworthiness of a user u, $\omega_T(u)$ in

$$\omega_T(u) = \alpha \times \frac{\sum_{i=1}^{m}(R_{u,i} - \overline{R}_u)(R_{a,i} - \overline{R}_a)}{\sqrt{\sum_{i=1}^{m}(R_{u,i} - \overline{R}_u)^2 \sum_{i=1}^{m}(R_{a,i} - \overline{R}_a)^2}} \tag{2}$$

where

$R_{u,i}$ is the rating of user (rater) u for item i
$R_{a,i}$ is the average rating of the general users for item i
α is the significance weighting
m is the number of items

The significance weighting, α, is 1 if the number of item ratings of a user is over 50; otherwise, it is $n/50$. This confers a high trustworthiness value on a user who has many ratings for the items.

3.2.3. Co-orientation measurement

We define the co-orientation factor on the basis of the source credibility theory as the degree to which a user is similar to the other users in the community that he belongs to. The co-orientation of a user u, $\omega_C(u)$, is defined as in Eq. (3). In other words, a user's co-orientation is the average of the similarities between the other users in the community

$$\omega_C(u)\frac{\sum_{a \in U(G)} \alpha \times \mathrm{Sim}(u,a)}{N_{U(G)}} \tag{3}$$

where

$U(G)$ is the general users in the community that a user u belongs to
$\mathrm{Sim}(u,a)$ is the similarity between a user u and a user a
$N_{U(G)}$ is the cardinality of the $U(G)$
α is the significance weighting

The significance weighting, α, which is 1 if the number of co-rated items between user u and user a is over 50 and $n/50$ otherwise, is included as shown in Eq. (3). This also assigns a high co-orientation to a user who has many co-ratings with the other users.

3.3. Phase II: User reputation generation & qualified rater group formation

3.3.1. User reputation generation

In this step, the user reputation, $\omega_R(u,c)$, is generated based on the measured values of the source credibility factors. There could be several mechanisms to obtain user reputation using the three credibility factors. The reputation can be obtained by using only one factor or all three factors. Further, the mechanisms that take all the factors into account can include the arithmetic mean, the harmonic mean, or the multiplication (equal to the square of the geometric mean) of the credibility factors, and the filtering by one or two factors. The filtering mechanism suggests that one or two factors is/are used as threshold criterion/criteria and is/are not included in the reputation. Based on the definition of each factor, we believe that trustworthiness and co-orientation are not the principal but supplementary factors in measuring user reputation. The higher the expertise of a user, the higher is the valuableness of the information he/she would be expected to provide. However, we also believe that high trustworthiness and co-orientation do not guarantee the valuableness of the information; they only guarantee the reliability and similarity of the information. Therefore, we propose a user reputation generating mechanism filtered by the trustworthiness and co-orientation factors. In the experiment, we will compare various user reputation generating mechanisms. It should be noted that prior to generating the user reputation, we standardized the measured values of each credibility factor to z-scores.

3.3.2. Qualified rater group formation

In order to predict the overall ratings for each item, we form a group comprising the most reputable/credible raters on the basis of the user reputation. We select as the qualified raters who had the top N reputation value in a certain user reputation generating mechanism; we term them the "best N raters."

3.4. Phase III: Item reputation generation (Item rating aggregation)

3.4.1. Rating tendency calculation

We propose "rating tendency" concept. If a rater has a tendency of rating higher than others, it is reasonable that the rating from the rater is decreased by some degree, and vice versa. Therefore, we calculate the rating tendency, E_u, of a rater by the average difference between the past ratings from a rater and the average ratings of the general users for the same items as shown in Eq. (4)

$$E_u = \overline{R_u} - \overline{R_a}(u) \tag{4}$$

where

$\overline{R_u}$ is the average ratings from a rater u
$\overline{R_a}(u)$ is the average ratings of the general users for the items that the rater u rated

3.4.2. Item reputation prediction

Our proposed reputation system can present the estimated ratings or a ranked list of items, prepared by the qualified rater group. In order to estimate item reputation, all the ratings from the qualified rater group in the previous step should be aggregated. We compare three aggregating mechanisms: non-weighted aggregation, reputation-weighted average, and adjusted reputation-weighted average. In non-weighted aggregation, all the qualified raters are considered as having the same weight. The average rating weighted by the user reputation of the evaluator group is calculated by Eq. (5) and is termed the "reputation-weighted average" of the qualified rater group.

$$R(i) = \frac{\sum_{u=1}^{n}\omega_R(u,c) \times R_{u,i}}{\sum_{u=1}^{n}\omega_R(u,c)} \tag{5}$$

where

ω_R is the reputation of a rater u
$R_{u,i}$ is the rating for item i given by the rater$u$

Further, we devise another rating aggregation scheme by modifying weighted average.

We calibrate the rating from the rater by deducting rating tendency of the rater, E_u and the estimated item reputation is calculated by the Eq. (6). This is called "adjusted reputation-weighted average"

$$R(i) = \frac{\sum_{u=1}^{n}(R_{u,i} - E_u)\omega_R(u,c)}{\sum_{u=1}^{n}\omega_R(u,c)} \tag{6}$$

where

E_u is the rating tendency of the rater u
$R_{u,i}$ is the rating for item i given by the rater u
ω_R is the reputation of the rater u

4. Performance evaluation

4.1. Assumptions for the performance evaluation

We performed experiments to verify the feasibility and advantages of the proposed reputation system and compared our system with benchmark systems.

Since there is no systematic evaluation method for a reputation system, we made the following assumptions in order to compare the different user reputation generating mechanisms and item rating aggregation mechanisms.

Assumption 1. A user is more credible (reputable) if his/her prediction is closer to the opinion of the majority of the general users in the community.

Assumption 2. If the overall item rating calculated by an "A" user reputation generating mechanism with an item rating aggregation mechanism is able to explain the real opinions of the general users better than that calculated by a "B" user reputation generating mechanism with the same item rating aggregation mechanism, then the "A" user reputation generating mechanism is superior to the "B" user reputation generating mechanism.

Assumption 3. If the overall item rating calculated by a "C" item rating aggregation mechanism with a user reputation generating mechanism is able to explain the real opinions of the general users better than that calculated by a "D" item rating aggregation mechanism with the same user reputation generating mechanism, then the "C" item rating aggregation mechanism is superior to the "D" item rating aggregation mechanism.

On the basis of the abovementioned assumptions, we will verify the performance of the user reputation generating and item rating aggregation mechanisms in view of the predictive ability of the item reputations.

4.2. Benchmark reputation systems

There are several ways to obtain reputation using the credibility factors, as mentioned in Section 3.2.2. Based on the discussion contained in that section, we constructed benchmark systems to be used in the experiments, as shown in Table 3. We set "the baseline model[BASE]" as the non-weighted average of randomly selected N raters. In calculating the average of randomly selected N users' ratings of the [BASE] mechanism, we conducted multiple random sampling ten times and obtained averages of each sample's average ratings in order to avoid sampling bias.

4.3. Major questions

The following are the major questions regarding and expectations of the performance evaluation.

Question 1. To what extent is the predictive performance of the proposed reputation system better than that of the non-weighted baseline model?

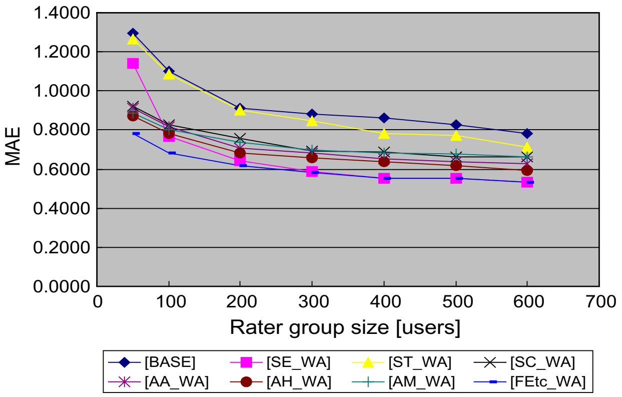
We expect that the overall item ratings predicted by the proposed user reputation generating mechanisms are closer to the average ratings of general users than those predicted by the non-weighted mechanism. In other words, we expect the [S], [A], and [F] mechanisms to be superior to [BASE].

Question 2. In what manner does each combination scheme of three credibility factors affect the performance of the proposed reputation system?

We expect that reputation measured using three factors will show better results than reputation measured using a single factor, i.e., the [A] and [F] mechanisms will be superior to [S]. This implies that the source credibility theory is applicable to measuring reputation. We also expect that it is more reasonable to use trustworthiness and co-orientation as ancillary criteria in predicting item reputations, i.e., the [F] mechanism will show the best results.

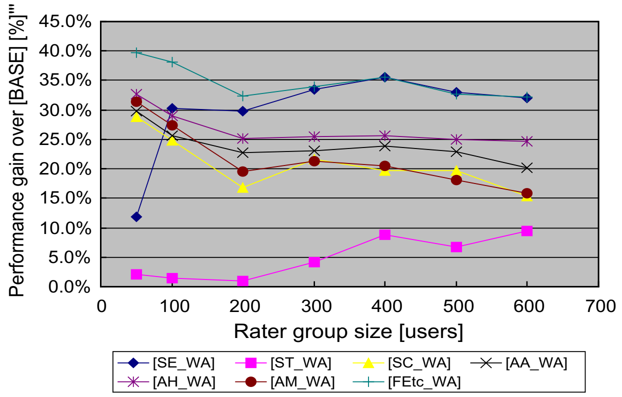
Table 3
Benchmark systems

Benchmark systems				
User reputation generating mechanism			Item rating aggregation mechanism	Acronym
Randomly selected N users			Non-weighted average	[BASE] (baseline)
Reputation-weighted mechanisms	Measured using a single factor [S]	Only expertise [E]	Weighted average [_WA] Or Weighted average reflecting rating tendency [_WT]	[SE_WA] [SE_WT]
		Only trustworthiness [T]		[ST_WA] [ST_WT]
		Only co-orientation [C]		[SC_WA] [SC_WT]
	Measured using all factors [A]	Arithmetic mean [A]		[AA_WA] [AA_WT]
		Harmonic mean [H]		[AH_WA] [AH_WT]
		Multiplication [M]		[AM_WA] [AM_WT]
	Measured using filtering [F]	Expertise with a threshold of trustworthiness and co-orientation [Etc]		[FEtc_WA] [FEtc_WT]



Systems	**Rater group size [users]**							Average
	50	100	200	300	400	500	600	
[BASE]	1.2939	1.1011	0.9135	0.8836	0.8595	0.8253	0.7844	0.8603
[SE_WA]	1.1403	0.7678	0.6417	0.5883	0.5545	**0.5531**	0.5333	0.5846
[ST_WA]	1.2669	1.0858	0.9040	0.8469	0.7841	0.7698	0.7101	0.8053
[SC_WA]	0.9208	0.8284	0.7594	0.6928	0.6897	0.6628	0.6635	0.6949
[AA_WA]	0.9095	0.8193	0.7058	0.6801	0.6543	0.6364	0.6256	0.6641
[AH_WA]	0.8709	0.7825	0.6836	0.6588	0.6388	0.6186	0.5908	0.6397
[AM_WA]	0.8883	0.8000	0.7353	0.6957	0.6833	0.6761	0.6605	0.6920
[FEtc_ WA]	**0.7810**	**0.6809**	**0.6186**	**0.5841**	**0.5544**	0.5554	**0.5318**	**0.5696**

Fig. 2a. Comparison of predictive accuracy of [_WA] mechanisms: MAE of each [_WA] mechanism with variation in the size of raters.



Systems	**Rater group size [users]**							Average
	50	100	200	300	400	500	600	
[SE_WA]	11.9%	30.3%	29.8%	33.4%	35.5%	**33.0%**	32.0%	32.3 %
[ST_WA]	2.1%	1.4%	1.0%	4.2%	8.8%	6.7%	9.5%	6.6%
[SC_WA]	28.8%	24.8%	16.9%	21.6%	19.7%	19.7%	15.4%	19.0 %
[AA_WA]	29.7%	25.6%	22.7%	23.0%	23.9%	22.9%	20.2%	22.6 %
[AH_ WA]	32.7%	28.9%	25.2%	25.4%	25.7%	25.0%	24.7%	25.5 %
[AM_WA]	31.3%	27.3%	19.5%	21.3%	20.5%	18.1%	15.8%	19.2 %
[FEtc_WA]	**39.6%**	**38.2%**	**32.3%**	**33.9%**	**35.5%**	32.7%	**32.2%**	**33.6%**

Fig. 2b. Comparison of predictive accuracy of [_WA] mechanisms: performance gain over [BASE] of each [_WA] mechanism (([BASE]-[*Systems*])/[BASE]).

Question 3. In what manner does rater group size affect the performance of the proposed reputation system as compared to that of the baseline model?

We expect that the smaller the rater group size, the better is the relative performance of the proposed reputation system over the baseline mechanism.

Question 4. In what manner does each item rating aggregation mechanism affect the performance of the proposed reputation system?

We expect the reputation-weighted item rating aggregation mechanism to be better than the non-weighted model, and the weighted average reflecting rating tendency mechanism to be better than the simple weighted average mechanism. In other words, the performances of the item rating aggregation mechanisms are in the descending order of [_WT], [_WA], and [BASE].

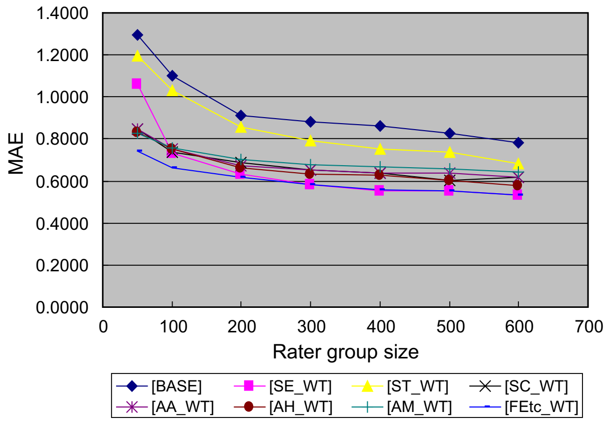
4.4. Evaluation metrics

For the performance evaluation, we adopt the predictive accuracy metric and classification accuracy metric widely used in researches related to the recommender system (Herlocker, Konstan, Terveen, & Riedl, 2004). We use the mean absolute error (MAE) measure to compare the predictive accuracy of each mechanism. Here, MAE is the absolute difference between general users' rating for an item and the predicted rating of the rater group for the same item.

For our classification accuracy measure, we use receiver operating characteristic (ROC) sensitivity. ROC sensitivity is a measure of the diagnostic power of a filtering system. Operationally, it is area under the ROC curve—a curve that plots the sensitivity and 1—specificity of the system. Sensitivity refers to the probability of a randomly selected good item being accepted by the filter, while specificity is the probability of a randomly selected bad item being rejected by the filter. ROC sensitivity ranges from 0 to 1, where 1 is perfect and 0.5 is random. Here, we assume that an item whose rating is above average is a good item, while one whose rating is below average is a bad item.

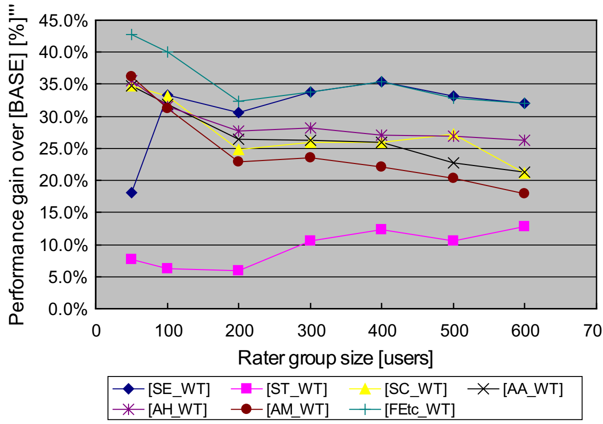
4.5. Data set

We have evaluated the feasibility and advantages of our proposed reputation system with an actual rating data collected from a commercial film rating website. Naver.com (http://www.naver.com) is the Web portal site operated by the NHN Corporation in Korea. It is the number one portal site in Korea and has approximately 24 million subscribers and 0.23 million simultaneous visitors. Naver.com provides the film review pages showing the reviews and ratings that range from 1 to 10 from numerous users for the Korean and foreign movies as the "*Netizen Ratings*." We gathered the ratings for the movies which have more than 100 ratings from the individual users who have more than 20 ratings. We, finally, could get the data set which comprises the ratings of 8124 users for 1724 movies. For the evaluation, we separated this data set into two parts—a calibration set comprising the ratings for 1000 items and a validation set comprising the ratings for the remaining 724 items. In other words, we calculated the reputation of the users in the calibration set, and we verified the performance of the benchmark systems in the validation set.



Systems	Rater group size [users]							Average
	50	100	200	300	400	500	600	
[BASE]	1.2939	1.1011	0.9135	0.8836	0.8595	0.8253	0.7844	0.8603
[SE_WT]	1.0590	0.7335	0.6343	0.5854	**0.5551**	**0.5521**	0.5338	0.5800
[ST_WT]	1.1936	1.0330	0.8589	0.7903	0.7528	0.7377	0.6839	0.7684
[SC_WT]	0.8449	0.7358	0.6869	0.6537	0.6366	0.6012	0.6188	0.6399
[AA_WT]	0.8447	0.7510	0.6718	0.6517	0.6359	0.6377	0.6168	0.6467
[AH_WT]	0.8343	0.7539	0.6603	0.6344	0.6269	0.6034	0.5780	0.6227
[AM_WT]	0.8252	0.7577	0.7038	0.6752	0.6696	0.6569	0.6439	0.6711
[FEtc_ WT]	**0.7408**	**0.6607**	**0.6180**	**0.5847**	0.5556	0.5544	**0.5334**	**0.5682**

Fig. 3a. Comparison of predictive accuracy of [_WT] mechanisms: MAE of each [_WT] mechanism with variation in the size of raters.



Systems	Rater group size [users]							Average
	50	100	200	300	400	500	600	
[SE_WT]	18.2%	33.4%	30.6%	33.8 %	**35.4%**	**33.1%**	32.0%	32.7%
[ST_WT]	7.8%	6.2%	6.0%	10.6 %	12.4%	10.6%	12.8%	10.9%
[SC_WT]	34.7%	33.2%	24.8%	26.0 %	25.9%	27.2%	21.1%	25.3%
[AA_WT]	34.7%	31.8%	26.5%	26.3 %	26.0%	22.7%	21.4%	24.5%
[AH_ WT]	35.5%	31.5%	27.7%	28.2 %	27.1%	26.9%	26.3%	27.4%
[AM_WT]	36.2%	31.2%	23.0%	23.6 %	22.1%	20.4%	17.9%	21.6%
[FEtc_WT]	**42.7%**	**40.0%**	**32.3%**	**33.8%**	35.4%	32.8%	**32.0%**	**33.7%**

Fig. 3b. Comparison of predictive accuracy of [_WT] mechanisms: performance gain over [BASE] of each [_WT] mechanism (([BASE]-[*Systems*])/[BASE]).

4.6. Experimental results

4.6.1. Answers to question 1 and 2: Combination scheme of credibility factors and predictive performances

The results of predictive accuracy of each mechanism using each credibility factor combination in Table 3 are shown in Figs. 2a and 2b—([_WA] mechanisms) and Figs. 3a and 3b—([_WT] mechanisms). As expected, all the reputation-weighted mechanisms outperformed the non-weighted baseline model, [BASE], with respect to MAE. Among the benchmark systems, the [FEtc] mechanism shows the best results. The mechanisms employing all the factors as weights ([AA], [AH], and [AM]) do not outperform the filtering mechanism, [FEtc]. The [FEtc] mechanism gives almost 34% performance gain averagely over [BASE] mechanism. Moreover, according to Figs. 2a, 2b and 3a and 3b, reputation measured by the three credibility factors is generally better than that measured by the single credibility factor. We, however, can find that the second performer is [SE] mechanism that employs reputation measured by only one factor, expertise. We give high expertise to a user when his ratings have low absolute error with the ratings of other users. That is why [SE] mechanism shows good performance with respect to MAE. [ST] mechanism shows the worst results and [SC] mechanism is the next. These results tell us that trustworthiness and co-orientation should not be used as a single criterion for best N users, but be incorporated with other factors. This implication coincides with the superiority of [FEtc] mechanism that employs trustworthiness and co-orientation as cut-off criteria. It also provides us the justification of adopting the source credibility theory to the reputation generating mechanism.

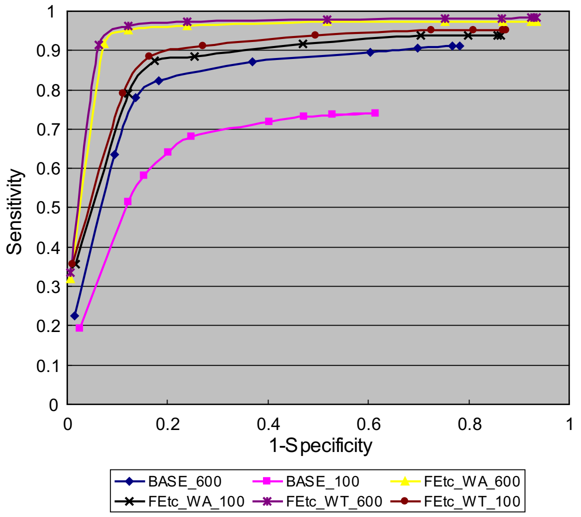
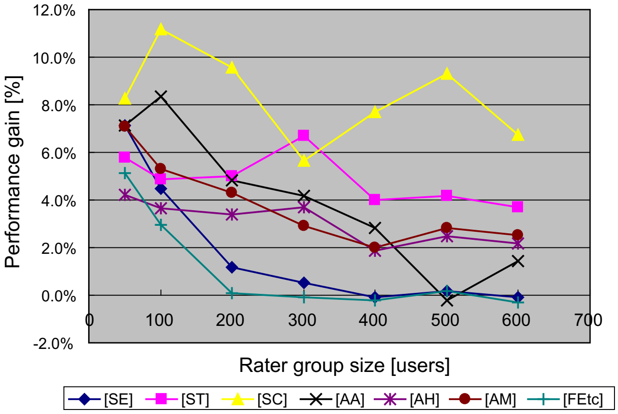


Fig. 4. Comparison of classification accuracy (ROC).

Fig. 4 shows the ROC-curves of [BASE], [FEtc_WA] and [FEtc_WT] mechanisms with the 100 and 600 raters. According to the ROC-curves, [FEtc] with 100 raters shows the better discriminating power than [BASE] with 600 raters. These results imply that multi-dimensional reputation value measured using expertise, trustworthiness and co-orientation factors is more appropriate, which rationalizes our proposed user reputation generating mechanism that considering three factors of the source credibility theory.

4.6.2. Answers to question 3: Rater group size effect

We have also checked the performance changes of the systems with respect to the variations of rater group size in Figs. 2a, 2b, 3a



Systems	Rater group size [users]							Average
	50	100	200	300	400	500	600	
[SE]	7.1%	4.5%	1.2%	0.5%	-0.1%	0.2%	-0.1%	0.55%
[ST]	5.8%	4.9%	5.0%	6.7%	4.0%	4.2%	3.7%	4.50%
[SC]	8.2%	11.2%	9.6%	5.6%	7.7%	9.3%	6.8%	7.86%
[AA]	7.1%	8.3%	4.8%	4.2%	2.8%	-0.2%	1.4%	2.46%
[AH]	4.2%	3.7%	3.4%	3.7%	1.9%	2.5%	2.2%	2.62%
[AM]	7.1%	5.3%	4.3%	2.9%	2.0%	2.8%	2.5%	2.96%
[FEtc]	5.1%	3.0%	0.1%	-0.1%	-0.2%	0.2%	-0.3%	0.18%

Fig. 5. Performance gain of [_WT] mechanisms over [_WA] mechanisms.

and 3b. We tested the systems with best 50, 100, 200, 300, 400, 500 and 600 raters. Those are corresponding to about 0.62%, 1.23%, 2.46%, 3.69%, 4.92%, 6.15% and 7.39% of entire users, respectively. According to Figs. 2a and 3a, as the number of users gets smaller, all the mechanisms are showing the increased MAE. However, the relative performance of each mechanisms over [BASE] is increasing, though. In case of [FEtc_WT], the performance gain over [BASE] is 32.0% when the rater group size is 600 users, and it is 42.7% when the rater group size is 50 users. It is noteworthy that [ST] mechanism is poor in Fig. 2b. It means only high trustworthiness does not guarantee high predictive power. Hence, it is reasonable that trustworthiness is used as an ancillary cut-off criterion. This result also agrees with the superiority of [FEtc] mechanism. The effect of the rater group size on the item rating aggregation mechanism is described in next section.

4.6.3. *Answer to question 4: The effect of the item rating aggregation mechanisms*

We expected that the performances of the item rating aggregation mechanisms are in the descending order of [_WT], [_WA], and [BASE]. As shown previously, [_WT] and [_WA] mechanisms are better than the baseline model. In Figs. 2b and 3b, we can identify the improvement of predictive accuracy by adopting [_WT] mechanism rather than [_WA] mechanism. The improvement of averages between Figs. 2b and 3b ranges from 0.1% to 6.3% comparing to the baseline model. According to the ROC-curves in Fig. 4, we can also identify that the [FEtc_WT] mechanisms show the better classification accuracy than the [FEtc_WA] mechanisms.

Fig. 5 shows the performance gain of [_WT] mechanism over [_WA] mechanism with respect to the rater group size. According to Fig. 5, on the whole, [_WT] mechanism shows better results than the [_WA] mechanism in most cases. However, the performance gain of [_WT] mechanism over [_WA] gets reducing with the larger rater group size. We found that the smaller the rater group size, the better is the relative performance of the proposed [_WT] mechanism over the traditional weighted average mechanism, [_WA].

5. Conclusion

Currently, the number of B2C e-commerce sites that provide consumers' evaluations of products or services is growing in order to help other consumers make purchase decisions. However, the non face-to-face communication and anonymity characteristics of e-commerce have led to side effects such as incorrect information from unreliable users, Internet frauds, and so on. In order to solve these problems, the need for online reputation systems that build trust among users has been highlighted. However, unlike C2C e-commerce sites, where between-peer evaluations occur frequently, most B2C e-commerce sites lack explicit information on the reputation of the users acting as raters.

We have pointed out the limitations of the existing explicit reputation systems and proposed a systematic method to develop implicit reputation systems for both products and raters that uses multidimensional credibility factors. Therefore, this study makes the following important contributions. First, we proposed an implicit user reputation generating mechanism by using the past item ratings of users, which is suitable for B2C e-commerce. Second, we devised quantitative measures for multidimensional reputation factors based on the source credibility theory in order to extract user credibility implicitly and automatically. Third, we also devised an item rating aggregation mechanism considering the rating tendencies of users and developed a prototype of the reputation system, *Q-rater*, for the experimental evaluation.

Although many B2C e-commerce sites do not have sufficient explicit evaluations by customers on products or services, *Q-rater* could improve the predictability of the item preference of general users by about 34% as compared with the non-weighted model. Our study offers a very meaningful, practical implication to new product/service developers and/or marketers. Almost all the e-commerce sites wish to know whether or not the products or services under consideration are successful. However, in most cases, the sites lack the human, financial, and time resources necessary to determine this by conducting a large-scale survey. Our proposed *Q-rater* will make it possible for these sites to improve the reliability of their surveys with only a small number of raters.

Future work in this area should be along the following directions. First, the measure of each credibility factor should be improved by employing other factors such as users' temporal behaviors. Second, the proposed user reputation generating mechanism should incorporate implicit preference information such as Web usage behaviors or recency, frequency, and monetary (RFM) data. If it is proved that the feasibility of the implicit Web usage behavior and purchase (or rejection) information is an indication of positive (or negative) preference, our proposed framework for a reputation system can be applied to most B2C e-commerce sites. Although this study has focused on reputation systems for B2C e-commerce, the proposed reputation system is generally applicable to any rating system such as recommender systems, online survey systems, electronic voting systems, and so on.

References

Chen, M., & Singh, J. (2001). Computing and using reputations for internet ratings. In *Proceedings of the 3rd ACM conference on electronic commerce (EC 01)*, Tampa, Florida, USA (pp. 246–247).

Dellarocas, C. (2003). The digitization of word-of-mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1407–1424.

Ekström, M., Björnsson, H., & Nass, C. (2005). A reputation mechanism for business-to-business electronic commerce that accounts for rater credibility. *Journal of Organizational Computing and Electronic Commerce, 15*(1), 1–18.

Hawkins, D., Best, R., & Coney, K. (2004). *Consumer behavior: Building marketing strategy*. Boston, USA: McGraw-Hill.

Herlocker, J., Konstan, J., Terveen, L., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53.

Hogg, T., & Adamic, L. (2004). Enhancing reputation mechanisms via online social networks. In *Proceedings of the 5th ACM conference on electronic commerce (EC 04)*, New York, USA (pp. 236–237).

Jøsang, A., Ismail, R., & Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision Support Systems, 43*(2), 618–644.

Marti, S., & Garcia-Molina, H. (2006). Taxonomy of trust: Categorizing P2P reputation systems. *Computer Networks, 50*(4), 472–484.

Massa, P., & Avesani, P. (2004). Trust-aware collaborative filtering for recommender systems. In *Proceedings of the 2nd international conference of cooperative information systems (CoopIS 04)* (pp. 492–508).

McKnight, D., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334–359.

O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference of intelligent user interfaces (IUI 05)* (pp. 167–174).

Pujol, J., Sanguesa, R., & Delgado, J. (2002). Extracting reputation in multi agent systems by means of social network topology. In *Proceedings of the 1st international joint conference on autonomous agents and multi agent systems (AAMAS 02)*, Bologna, Italy (pp. 467–474).

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the ACM, 43*(12), 45–48.

Robertson, T., Zielinski, J., & Ward, S. (1984). *Consumer behavior*. Scott, Foresman and Company.

Schweitzer, D. (1969). A note on Whitehead's factors of source credibility. *Quarterly Journal of Speech, 55*, 308–310.

Zacharia, G., Moukas, P., & Maes, P. (2000). Collaborative reputation mechanisms in electronic marketplaces. *Decision Support Systems, 29*(4), 371–388.