Computing
DOI 10.1007/s00607-013-0345-x

Improving the accuracy of Business-to-Business (B2B) reputation systems through rater expertise prediction

Heidi Dikow · Omar Hasan · Harald Kosch · Lionel Brunie · Renaud Sornin

Received: 2 February 2013 / Accepted: 23 August 2013

Abstract Digital ecosystems rely on reputation systems in order to build trust and to foster collaborations among users. Reputation systems are commonplace in the Customer-to-Customer and Business-to-Customer contexts, however, they have not yet found mainstream acceptance in Business-to-Business (B2B) environments. Our first contribution in this paper is to identify the particularities of feedback collection in B2B reputation systems. An issue that we identify is that the reputation target in the B2B context is a business, which requires evaluation on a large number of criteria. We observe that due to the wide variation in user expertise, feedback forms that require users to evaluate all criteria have significant negative consequences for rating accuracy. Our second contribution is to propose an expertise prediction algorithm for B2B reputation systems, which filters the criteria describing the target business such that each user rates only on those criteria that he has expertise in. Experiments based on our real dataset show that the algorithm accurately predicts the expertise of users

H. Dikow
Albert-Ludwigs-Universität Freiburg, Friedrichstr. 39, 79098 Freiburg, Germany
e-mail: heidi.dikow@venus.uni-freiburg.de

O. Hasan (✉) · L. Brunie
University of Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, 69621 Lyon-Villeurbanne, France
e-mail: omar.hasan@insa-lyon.fr

L. Brunie
e-mail: lionel.brunie@insa-lyon.fr

H. Kosch
University of Passau, Innstrasse 43, 94032 Passau, Germany
e-mail: harald.kosch@uni-passau.de

R. Sornin
ALG-Attestation Légale, 20 Bd Eugène Deruelle-CS 63753, 69432 Lyon, France
e-mail: renaud.sornin@attestationlegale.fr

Published online: 02 October 2013

in given criteria. The algorithm may also increase the motivation of users to submit feedback as well as the confidence of users in B2B reputation systems.

Keywords Digital ecosystems · Reputation systems · Business-to-Business · Collaborative filtering · Trust

Mathematics Subject Classification 68M14

1 Introduction

A digital ecosystem is an open, loosely coupled, demand-driven, domain clustered, self-organized collaborative environment where users as well as agents form coalitions for specific goals, and everyone is proactive and responsive for their own benefit or profit [5]. Digital ecosystems aim to promote collaboration instead of competition to cultivate networked and enriched communities. Due to the inherent openness and loose coupling in digital ecosystems, lack of trust between users is a common issue. In order to build trust and to foster collaborations, one of the principal technologies that digital ecosystems rely on is reputation systems.

Reputation systems are widely used on Business-to-Customer (B2C) and Customer-to-Customer (C2C) platforms such as ebay.com or amazon.com to build trust between users. However, they have not yet found similar success in mainstream Business-to-Business (B2B) environments. This paper focuses on a little explored field: the integration of reputation mechanisms into B2B platforms.

Concentrating on the feedback collection part of reputation systems, this paper elaborates the main particularities of reputation systems in the context of B2B environments. There are some basic differences in reputation targets (rates) and sources (raters) in the B2B context in contrast to other environments. The particularities entail a range of issues for the collection of feedback for reputation systems in this context.

One of these challenges arises from the fact that the reputation source consists of several raters, having different expertise concerning the reputation target. As our experimental results show, user expertise has substantial influence on rating accuracy. Therefore, it is important to ensure that each user evaluates only those aspects of a business in which he has expertise.

To address this problem we propose an expertise prediction algorithm, which predicts users expertise such that only those criteria in which a user has expertise are added to the feedback form presented to him. The algorithm adapts the idea of collaborative filtering algorithms which are used for recommender systems. This is possible since we observe that the issue of expertise prediction is similar to the issue of item recommendation in recommender systems.

The remainder of this paper is organized as follows: In Sect. 2, the current state of feedback collection for reputation systems is described in general. Section 3 identifies the particularities of feedback collection for reputation systems in the B2B context through an analysis of the two components, the source and the target, of a reputation system which differ fundamentally from those of other contexts. This allows us to identify the problems that occur due to these particularities in Sect. 4. The proposed solution is explored in Sect. 5. Section 6 describes the experiment setup for the evaluation of the solution. The results are presented and discussed in Sect. 7. We present some related work in Sect. 8. The paper is concluded in Sect. 9.

2 Reputation systems

In digital ecosystems, users often need to communicate and interact with other users whom they do not know. When dealing with a complete stranger, a user does not have any information or experience about the trustworthiness of that stranger. Online rating and reputation systems are one solution to decrease the absence of this information [29]. Reputation systems compensate for the lack of trust between unacquainted users [29,39,34,9]. They collect, aggregate and distribute feedback about the behavior of participants [33]. Reputation systems can be viewed as the digitization of word-of-mouth [6].

Reputation systems are already widely used in different contexts. The online trading platform ebay.com is an example for the use of reputation systems to increase trust between users in the C2C environment. The ebay reputation system helps buyers to identify trustworthy sellers [15]. Another example of reputation systems is the iovation.com reputation system, which protects businesses from online fraud by exposing devices such as computers, tablets and smart phones that are associated with chargeback, identity theft, and account takeover attacks. Reputation systems are also used by online programming communities such as advogato.org and stackoverflow.com to filter users who post spam.

2.1 Components of reputation systems

Online reputation systems can be divided into three main components shown in Fig. 1. Feedback collection is the process of eliciting feedback from users. Generally, a rating scale is used on which users can express their opinions about the items that they rate. The feedback collection component is an important component of a reputation system because the other two components rely on the quantity and the quality of the collected feedback. To obtain a high quantity of responses, users need to be motivated to submit ratings [7]. For a high quality of feedback, the expertise of users is an important element (cf. Sect. 7.1, [24]).

Feedback aggregation is the compilation and aggregation of the collected information. One of the simple ways of aggregation is to calculate the mean of the collected feedback. Reputation dissemination finally distributes the aggregated reputation information [29,33,13] to interested users.

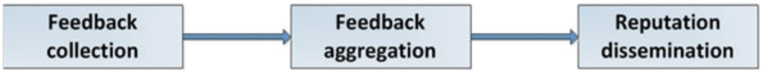


Fig. 1 Components of a reputation system

2.2 Aspects of feedback collection

In this article, we focus on the issues of eliciting feedback for reputation systems. The following subsections describe these issues.

2.2.1 User motivation for submitting feedback

To reach a high volume of rating input, the users need to be motivated to rate an item [7]. Giving a rating costs the user mental effort as well as time. The mental effort describes the cognitive load which is necessary to make a rating. The reason a user is willing to spend this mental effort and time is that he perceives to gain some benefits from the system following the economical model of Harper et al. [12]. The motivation of users to rate an item thus depends on:

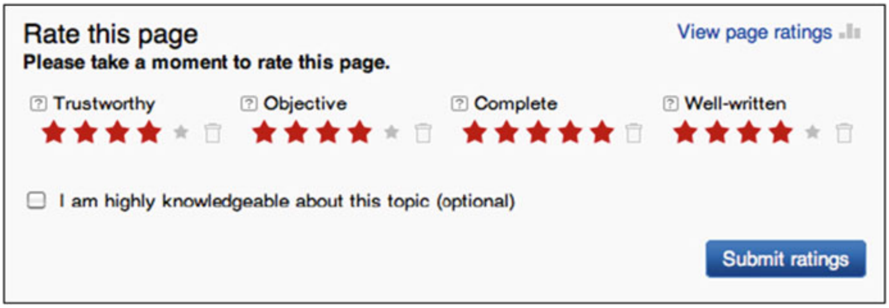
1. The mental effort required to complete the rating.
2. The time required to give a rating.
3. The perceived benefits of the rating system.

The rating interface has an impact on the mental effort that a user has to exert and the time that he needs to spend in giving a rating. Sparling and Sen [36] compared several rating scales in terms of cognitive load and rating time. They found that finer grained scales require more mental effort as well as more time.

Dellarocas [7] investigated the benefits that raters receive from rating systems. Examining the ratings system of ebay, he found that raters are mainly motivated to rate by self-interest with users tending to be reciprocal towards partners who rated them before. The feeling of belonging to the community also seems to be a component for the motivation to rate [3].

2.2.2 User expertise

Collecting reliable rating input necessitates raters with high expertise (cf. Sect. 7.1, [24]). Liu and Munro [24] differentiate between the *expertise granularity* which defines the level of expertise of the evaluator in the target item's area and the *interaction granularity* which defines whether the evaluator is in direct interaction with the target item.



Rate this page

Please take a moment to rate this page.

☐ I am highly knowledgeable about this topic (optional)

Fig. 2 Rating tool for Wikipedia articles

The article feedback tool of wikipedia.org, shown in Fig. 2, gives an additional checkbox "I am highly knowledgeable about this topic (optional)". This provides the opportunity to detect the experts of the article's subject. The project page [8] of the article feedback tool also poses the question on which criteria readers can provide a reasonable level of assessment and whether ratings meaningfully predict quality in those categories.

3 Business-to-Business (B2B) reputation systems

The current literature on reputation systems focuses primarily on reputation systems in the C2C and the B2C contexts. Research on reputation systems in the B2B context is very limited. Contrary to the C2C and B2C contexts, reputation systems are less common in the B2B environments.

The purpose of a B2B reputation system is inter-business evaluation and subsequently computing the reputation of businesses. B2B reputation systems are intended for business users to evaluate other businesses that their business has interacted with. The most common case would be that an employee of a business that consumed a service would evaluate the business that provided the service. Similarly, an employee of a provider business could evaluate the client business.

A study by Carlsson [3] examined the basic issues of reputation systems in relation to the B2B context using an online questionnaire. He confirmed that reputation systems in the B2B context are significantly less widespread: 75 % of respondents never rated any products or services in the role of a business user. According to his results, reputation systems in the B2B context also offer the possibility to increase trust. However, users seem to have less confidence in current systems and reviews. According to the online questionnaire, 89 % of respondents agreed or strongly agreed that "It is hard to know if the rater/reviewer actually has enough experience of the product or service".

Concerning the motivation of users, Carlsson found that for business users the time factor as well as complicated sign-in/identification processes are inhibitors to rate or review. Similarly, the absence of personal involvement ("I don't know what's in it for me") decreased a business user's motivation to rate.

An article [19] published by kompass.fr, a directory service for businesses, analyzed the process of choosing new suppliers. It characterized the group of actors who are involved in the decisions and identified that each of them needs a different type of information. The actual decision maker is more interested in the financial stability and reputation of a business while the final user of the product needs to know about concrete product functionality and technical information [19].

In the following subsections, we outline the particularities of reputation systems in the B2B context. We focus on the properties that the B2B context entails for the feedback collection aspects of reputation systems.

3.1 Properties of a business as the reputation target

Contrary to the reputation targets that we observed in Sect. 2 such as products, ebay users, online community members, etc., the reputation target in a B2B reputation system is a complete business. This is a more complex reputation target due to the fact that several criteria are needed to form a complete reputation. The reputation of a business depends not only on its product or service but also on other aspects of the business which are of interest for a client business. A larger set of criteria is thus needed in order to form a significant reputation which describes all the aspects of a business and its products. These criteria differ depending on the type of a business. Certain core criteria are always of interest, such as quality, price, innovation, delivery time and reliability. The target supplier business is thus composed of several criteria describing different aspects of the business.

Some C2C and B2C reputation systems also use multiple criteria to describe a target. Wikipedia and ebay both describe the target items on four criteria. On ebay, sellers are evaluated on the criteria “Item as described”, “Communication”, “Postage time” and “Postage and handling charges”, while the criteria used by Wikipedia can be seen in Fig. 2. However, a complete business is a much more complex entity and thus a B2B reputation system requires a much higher number of criteria for the evaluation of a target business.

3.1.1 Internal evaluation systems

Even though online B2B reputation systems are not yet common, supplier evaluation is already a key issue in purchasing departments of businesses and suppliers are already evaluated through internal evaluation systems. Supplier evaluation is currently used to avoid choosing a supplier whose product or service is not satisfactory and to remove hidden cost drivers. Suppliers are evaluated on the core criteria, as well as specific criteria aligned to the needs of the business and the type of supplier business. Internal evaluation systems use either evaluation forms or interviews. Most evaluation forms employ the Likert-Like rating [16] scale to evaluate the criteria [18]. However, internal evaluation systems do not have the potential of online B2B reputation systems because internal evaluation systems can only provide information about suppliers that the business has already interacted with. They cannot provide information about new suppliers.

3.2 Properties of a business as the reputation source

The reputation source (i.e., the source of feedback) of a reputation system in the B2C context (e.g. amazon.com) and the C2C context (e.g. ebay.com) is one person. On the contrary, in a B2B reputation system, the reputation source consists of a group of raters, i.e., the employees of a business. The group of raters in a B2B reputation system is composed of people with different expertise concerning the reputation target.

In a business, every employee has a specific field of activity, depending on his competences and his position in the business. Accordingly, their knowledge is limited to the areas of the business concerning their field of activity. People working in the IT department for example know which hardware and software is used in the business, while employees of the human resources have knowledge about the employees working in the business. The knowledge that an employee has in a certain area of a target business depends not only on his position but on several criteria, for example, the length of time spent in a certain position. Therefore, each rater needs a different subset of criteria for evaluating a business.

The reputation source business is thus composed of a group of employees with a certain profile. The profile has several elements including information about the position in the business, knowledge, competences, time spent in the business, etc. Section 3.1 explained how the target business consists of a set of criteria representing different aspects of a business. Each employee of the source business is an expert for a subset of this set, namely the criteria concerning their specific field.

Additionally, the user type in a B2B reputation system is different. Buying and evaluating a book on amazon.com or evaluating an ebay.com seller after a transaction are activities that take place mostly with a private purpose. Thus, in B2C and in C2C environments, the user is acting in their own free time and in their role as a private user. However, in B2B reputation systems, the user is acting in their role as an employee of a business during their working hours.

3.3 C2C versus B2B context

Table 1 gives an overview of the particularities for reputation systems in the B2B context compared to the B2C and the C2C contexts. We discussed that the reputation target business consists of several criteria. A business as the reputation source in the B2B context differs from the reputation sources in other contexts in that it consists of a group of people, namely several employees of the business. In this group each member has expertise in a different subset of the criteria describing the target business. The user type is the business user, in contrast to reputation systems such as amazon.com where people act in the role of private users.

4 Challenges for feedback collection in B2B reputation systems

The particularities of reputation systems in the B2B context discussed in Sect. 3 entail several challenges for their design. This section identifies the challenges for B2B reputation systems, concentrating on the collection of feedback.

Table 1 C2C vs. B2B context

	C2C/B2C	B2B
Reputation target	Product/user	Business
Number of criteria describing the target item	Around five	In double-digits or higher
Reputation source	Customer	Business
User type	Private user	Business user
User expertise	Expertise in the whole set of criteria describing the target item	Expertise in a subset of criteria describing the target item

4.1 Accuracy of ratings

In Sect. 3.2, we discussed that one main particularity of reputation systems in the B2B context is that the expertise of raters is limited to a subset of criteria describing the target item. We assume this fact to become an issue for the feedback collection of a reputation system concerning the feedback quality. As seen in Sect. 2.2.2, literature [24] already states that the accuracy of collected feedback depends highly on the expertise of raters. Our experimental results (Sect. 7.1) reconfirm that low user expertise has a negative influence on the rating accuracy.

4.2 Motivation of business users

As seen in Sect. 2.2.1, the motivation of users to submit a rating depends primarily on the benefits he retrieves from the reputation system and of the costs (time and mental effort) he needs to invest in it. While a private user can determine how much time he wishes to invest in a rating, the timetable of a business user is not that flexible. Additionally, an accurate business reputation consists of several criteria, which leads automatically to higher fill-out times of the reputation form. One main challenge of a B2B reputation system is thus to find an equilibrium between the number of rating criteria and the motivation of the users. The system needs therefore to be as less cost-consuming as possible. This means that the time it takes a user to submit an evaluation should be reduced as much as possible.

4.3 Limited trust in reputations

Users seem to have less confidence in current B2B reputation systems and reviews. As we noted in Sect. 3, according to an online survey [3], 89 % of respondents agreed or strongly agreed that “It is hard to know if the rater/reviewer actually has enough experience of the product or service”. Uncertainty about the expertise of the rating person thus deters business users from trusting B2B reputations.

5 Expertise prediction for B2B reputation systems

We propose an algorithm that filters the rating criteria such that the feedback form contains only those criteria that the feedback provider has expertise in. The objectives of the expertise prediction algorithm (EPA) are as follows:

1. **Increase the accuracy of the reputation.** As we discussed in Sect. 2.2.2, the accuracy of a reputation is highly correlated with the expertise of the feedback providers. We can thus assume that the accuracy of the reputation increases if each feedback provider is evaluating only those criteria that he has expertise in.
2. **Increase the motivation of business users to submit feedback.** As seen in Sect. 2.2.1, the motivation of a user to submit feedback depends highly on the time that he needs to invest in it. We can thus assume that a reputation form, which is composed of only those criteria that the user has expertise in, decreases the time and increases the motivation.

3. **Increase the trust in the reputation.** As discussed in Sect. 3, one of the reasons for the absence of trust in B2B reputation systems is the uncertainty about the expertise of the feedback providers. We can thus assume that a reputation system that ensures the distribution of criteria to those users who are competent to evaluate them increases the trust in B2B reputation systems.

The EPA is inspired by the k-nearest neighbor collaborative filtering algorithm for recommender systems. Therefore, we give a short overview of recommender systems and the collaborative filtering approach in the next section (Sect. 5.1). The similarities between the problem of item recommendation and the problem of expertise prediction are then discussed in Sect. 5.2. We present the expertise prediction algorithm in Sect. 5.3.

5.1 Building block: recommender systems

The objective of recommender systems is to suggest a personalized subset of items to users. The subset of items suggested to a user depends on the utility of the items for the user. The utility of items in recommender systems can be defined by a rating, which indicates the preference of a user towards a certain item. For example, the rating 8 out of 10 for a movie would imply a high utility of this movie for a user. The utility u of an item for a user can be described by the following function:

$$u : User \times Item \rightarrow Utility \tag{1}$$

where the set *Item* can consist of books (amazon.com), music (last.fm), friends (facebook.com), etc. and *User* is the set of users. Each user can be described by a user profile containing their preferences. The user profile can be based on implicit or explicit datasets. Implicit feedback is obtained by measuring interaction of users with different items, for example, a music listening log or clicking on web pages. Explicit feedback, on the other hand, is obtained by direct input through the user on some scale [31,23].

The utility function calculates the utility of an item for the user. In this case it calculates an estimation of the rating a user might give to an item. Having computed the utility of items for a user, those items with the highest utility are recommended to a user [1]. The main objective of recommender systems is thus to estimate the rating a user would give to a certain item. The recommender system then suggests those items with a high rating estimation to the user.

5.1.1 Collaborative filtering (CF) based recommender systems

Collaborative Filtering (CF) based recommender systems recommend items that other users with similar taste liked previously. The *memory based* approach of collaborative filtering uses the entire collection of previously rated items by the users to calculate the recommendation [1]. It compares a principal user with all the other users in order to find users who are similar to the principal user in terms of their preferences. The preferences of this subset of similar users are then combined to estimate the rating the user would give to the items. The system finally recommends the items with the highest rating prediction.

5.1.2 k-Nearest neighbor (k-NN) collaborative filtering

The k-nearest neighbor (k-NN) algorithm is one of the most used and most cited collaborative filtering algorithms [10,26]. It generates the predictions for a user in two steps.

In the first step, the first k nearest "neighbors" are calculated, i.e., the top k most similar users in terms of rating behavior. To calculate the similarity between two users x and y, the two most frequently used approaches are the correlation-based and the cosine-based approaches [11]. The correlation-based approach generally uses the Pearson correlation (Eq. 2) to measure the similarity between two users x and y.

$$Pearson(x, y) = \frac{\sum_{i=1}^{n}(r_{x,i} - \bar{r}_x)(r_{y,i} - \bar{r}_y)}{\sqrt{\sum_{i=1}^{n}(r_{x,i} - \bar{r}_x)^2 \sum_{i=1}^{n}(r_{y,i} - \bar{r}_y)^2}} \tag{2}$$

where, $r_{x,i}$ and $r_{y,i}$ are the ratings given by the users x and y respectively to the item s_i in a set of items $\mathbb{S} = \{s_1, s_2, \ldots s_n\}$. Moreover, $\bar{r}_x$ and $\bar{r}_y$ are the mean of the ratings given by the users x and y respectively. The Pearson correlation measures the level of linear dependence between two variables. The cosine-based approach (Eq. 3) treats users as vectors and defines the cosine of the angle between them in order to define their similarity [1].

$$Cosine(x, y) = \frac{\sum_{i=1}^{n} r_{x,i} r_{y,i}}{\sqrt{\sum_{i=1}^{n} r_{x,i}^2}\sqrt{\sum_{i=1}^{n} r_{y,i}^2}} \tag{3}$$

In the second step, the prediction for each item i is formed by aggregating the ratings of the k nearest neighbors. A number of different functions can be used as the aggregation function [1]. A simple and commonly used aggregation function is the standard mean function (Eq. 4).

$$\frac{1}{k} \sum_{v \in \mathbb{K}} r_{v,i} \tag{4}$$

where, $\mathbb{K} = \{v_1, v_2, \ldots, v_k\}$ is the set of the top k most similar users, and $r_{v,i}$ is the rating given by a user $v \in \mathbb{K}$ to the item s_i.

5.2 Item recommendation vs. expertise prediction

In Sect. 5.1, we identified the problem of recommender systems as the problem of calculating the utility of an item for a user (Eq. 1) in order to recommend items with a high utility to the user. The problem of expertise prediction for B2B reputation systems is very similar to the one of recommender systems. As shown in Table 2, it

Table 2 Item recommendation vs. expertise prediction, in terms of the utility function (Eq. 1) $u : User \times Item \rightarrow Utility$

	User	Item	Utility
Item recommendation	Set of users (each user is described by a profile containing their preferences/previous ratings)	Set of items, e.g., books, movies, etc.	A user's recommended rating for an item
Expertise prediction	Set of employees of a source business (each employee is described by a profile containing position, time spent, competences, etc.)	Set of criteria describing a target business	A user's predicted expertise for a criterion

can be defined similarly as the problem of defining the utility of an item for a user. The users in this context are the employees of the source business of the reputation. Each user can be described by a user profile containing information about his position in the business, the time that he spent in the position, his competences, etc. Contrary to items in recommender systems such as books or movies, the items in the context of B2B reputation systems are criteria. The utility of a criterion for a user is determined by his level of expertise for this criterion.

The main objective of expertise prediction is thus to estimate the expertise that a user has for a certain criterion. This leads to the selection of a subset of criteria that the user has high expertise in.

5.3 Description of the expertise prediction algorithm

Let $\mathbb{U}$ be a set of business users, such that $\mathbb{U} = \{u_1, u_2, \ldots, u_m\}$. Let $\mathbb{P}$ be a set of user profile entries, such that $\mathbb{P} = \{p_1, p_2, \ldots, p_\lambda\}$. For example, the user profile entries could be as follows: $p_1 =$ "Position in the company", $p_2 =$ "Time spent in the current position", etc. Let $f_{u,p}$ represent the value of a profile entry p for a user u. Let vector $\overrightarrow{P}_u = \langle f_{u,p_1}, f_{u,p_2}, \ldots, f_{u,p_\lambda} \rangle$ be the profile of a user u.

Let $\mathbb{C}$ be a set of rating criteria concerning the target business, such that $\mathbb{C} = \{c_1, c_2, \ldots, c_\gamma\}$. For example, the rating criteria might be $c_1 =$ "Quality", $c_2 =$ "Price", etc. Let $\mathbb{L} = \{0, 1, 2, \ldots, l\}$ be a set that represents the scale of a user's expertise for a given criterion. We consider that 0 represents the lowest expertise and l represents the highest expertise on this scale. The values between 0 and l represent expertise from low to high according to their magnitude. Let $e_{u,c}$ represent the expertise of a user u in criterion c.

Consider τ as a threshold for sufficient user expertise for rating a criterion c. A user who has expertise greater than or equal to τ can be considered as having sufficient expertise for rating the criterion c. For example, we could consider the value $\tau = 4$ for an interval $\mathbb{L} = \{0, 1, 2, \ldots, 5\}$. The subset of criteria that a user u has sufficient expertise in is given as: $\mathbb{C}_u = \{c \mid e_{u,c} \geq \tau\}$. The algorithm for computing $\mathbb{C}_u$ using the Pearson correlation (Eq. 2) is given below. The Pearson correlation function can be substituted by the cosine function (Eq. 3).

```
EXPERTISE PREDICTION ALGORITHM($u, \tau, k, \mathbb{C}, \mathbb{U}$)
 1  $\mathbb{C}_u \leftarrow \phi$
 2  for each user $v \in \mathbb{U}$, where $v \neq u$
 3      do calculate the similarity value $s_{uv}$ between
           the profile $\overrightarrow{P}_u$ of user $u$ and the profile $\overrightarrow{P}_v$ of user $v$
           using the Pearson correlation (Equation 2)
 4  for each criterion $c \in \mathbb{C}$
 5      do $\mathbb{K} \leftarrow \phi$
 6         for each user $v \in \mathbb{U}$, where $v \neq u$, and $e_{v,c} \neq 0$
 7             do if the value of $s_{uv}$ is one of the $k$ highest
                     similarity values (computed in Line 4)
 8                   then $\mathbb{K} \leftarrow \mathbb{K} \cup \{v\}$
 9         $e_{u,c} \leftarrow \lfloor \frac{1}{|\mathbb{K}|} \sum_{v \in \mathbb{K}} e_{v,c} + 0.5 \rfloor$
10         if $e_{u,c} \geq \tau$
11            then $\mathbb{C}_u \leftarrow \mathbb{C}_u \cup \{c\}$
```

As the first step, the EPA calculates the similarity between the new user u and each user v in the set of users $\mathbb{U}$ using the correlation-based similarity approach (cf. Sect. 5.1.2). In the next step, the k users with the highest similarity to user u are selected. The EPA then predicts the expertise of user u for each criterion c in the set of criteria $\mathbb{C}$. The predicted value $e_{u,c}$ is calculated as the mean of the expertise values for criterion c of the most similar users (the set $\mathbb{K}$). If the rounded predicted value $e_{u,c}$ is higher than the threshold τ, the expertise of the user for this criterion is considered as sufficient and the criterion is added to $\mathbb{C}_u$, the set of criteria with high expertise prediction for the user.

6 Experiment setup

The experiment was set up to answer the following questions:

- *Question 1:* Does user expertise effect rating accuracy?
- *Question 2:* Does the expertise prediction algorithm (EPA) correctly filter the criteria in a reputation form such that the users receive only those criteria that they have expertise in?

6.1 Data set

The experiment relies on the values of an online reputation survey in which we asked students to evaluate their university on a set of criteria. The survey was built using the online survey tool soscisurvey.de.

The survey ran from 8 August 2012 to 13 September 2012. Approximately 200 students from 27 different universities in 9 different countries answered. Out of those users, 130 finished the survey and generated 2,504 ratings. The majority of the students who answered were from the University of Passau, Germany (66 %).

Table 3 Expertise fields of the survey

Expertise field	Number of questions
Computer science	2
Law	3
Dining hall	4
Sports facilities	1
Practical components	2
Internationality	3
Student residences	1
General questions	4
Sum of criteria	20

6.1.1 Choice of university as the use case

We choose the use case university because we assume a high portability of the experiment results into the B2B context. We base this assumption on the following reasons:

- The character of the reputation target *University* is similar to the one of the target *Business* as we described it in Sect. 3.1. The reputation target *University* is also composed of multiple, different criteria which need all to be taken into consideration in order to form a complete reputation.
- The reputation source of our study can be compared with the reputation source in the B2B context as we described it in Sect. 3.2. The reputation source of our study is a group of students. The users of this group do have different expertise in the criteria describing the target *University*.
- Finally, the users of our study being students led to a high number of participants because we could use several ways to distribute the survey and students were a group that was readily available.

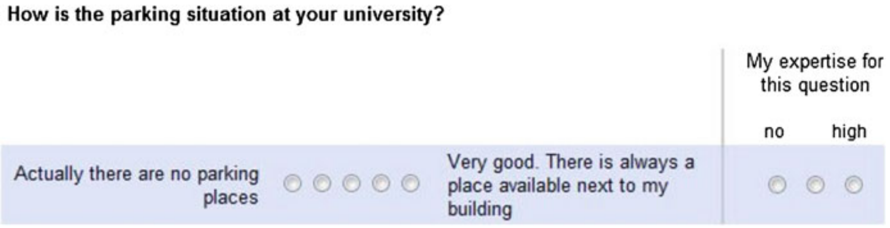
6.1.2 Survey overview

The survey consisted of two parts:

1. Profile building
2. University evaluation

The first page of the reputation asked users 9 questions about themselves, for example, which services of their university they are using and which major they are studying in order to build a user profile. Starting from the second page users got to evaluate their university on a set of 20 criteria grouped in the 8 expertise fields listed in Table 3.

The criteria were evaluated on a 5-point reference scale. The extreme values were labeled with an indication of their meaning. The 20 criteria were distributed on 6 pages, each with 3–5 questions.



How is the parking situation at your university?

			My expertise for this question no high
Actually there are no parking places	○ ○ ○ ○ ○	Very good. There is always a place available next to my building	○ ○ ○

Fig. 3 An excerpt from the University Evaluation Survey

6.1.3 Expertise measurement

A 3 point Likert scale [16] asked users for their expertise in the question next to each question (cf. Fig. 3). On the first page after the profile building site, a small text introduces the term “user expertise” to the user and gives some indication on how to estimate his own expertise. Also a tooltip (“I have some experience about the subject of this question—I can tell more or less”) explained the middle value of the 3 point scale.

6.1.4 User recruitment

Users for the survey were recruited using 4 channels: a link to the survey was shared on the social network facebook.com and posted on the facebook wall of several people. Most of the participants were recruited through the university mailing distributor of the faculty of computer science, Passau. The link was also distributed in several online student forums (studi-online.de, thestudentroom.co.uk, pruefungsgeil.de). Finally, the survey was distributed through personal emails. At the end of the survey, the users were asked to distribute the link to other friends.

6.2 Data set division

The algorithm was tested 130 times, once for each entry of the entire set of 130 user values (cf. Sect. 6.1). For each combination, we used the profile values of the current user entry as input for the algorithm and calculated the expertise predictions for these profile values. For the predictions, the algorithm relied on the values of the remaining 129 data set entries. The accuracy of a prediction was then calculated using the actual self-submitted criteria expertise entries of the current user as benchmark values. The overall accuracy of an algorithm was then calculated as an average of these 130 accuracy values.

6.3 Evaluation metrics

6.3.1 Frequency rate

The frequency rate describes how frequently one rating value of the rating scale v isused.

$$Frequency\ rate_v = \frac{V}{N} \tag{5}$$

where, V is the number of ratings with the values v and N the total number of ratings.

6.3.2 Correlation rate

The correlation rate is a widely used metric for rating accuracy [22]. The correlation determines the similarity between a rating and its benchmark value. Therefore, the Pearson correlation equation (function 2) is used. The more the value of the correlation approaches the value 1, the more the submitted rating is similar to the benchmark value and thus accurate. It can take values between -1 and 1. The value 0 means no correlation between two variables. The more the value is differing from 0, the more the variables get correlated.

6.3.3 Mean Absolute Error (MAE)

Introduced by Miller et al. [27], the MAE is one of the most commonly used accuracy metrics for CF recommender systems [21,22]. It is defined as the average difference between the predicted rating and the corresponding actual rating:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|r_i - z_i| \tag{6}$$

where, N is the number of predictions, r_i the value of the actual rating value for user i and z the predicted rating value for user i. A low value of the MAE indicates high rating accuracy.

6.3.4 Accuracy rate

The accuracy rate describes the percentage of predictions that are correct over the entire set [38] and is calculated as follows:

$$Accuracy\ rate = \frac{P}{N} \tag{7}$$

where, P is the number of correct predictions and N the number of all predictions.

7 Experiment results

7.1 Question 1: influence of user expertise

7.1.1 Frequency of rating values for different expertise

Users could rate the criteria on a scale from 1 to 5. Figure 4 shows the frequency (cf. Sect. 6.3.1) of used values for each expertise. The values of users with low expertise

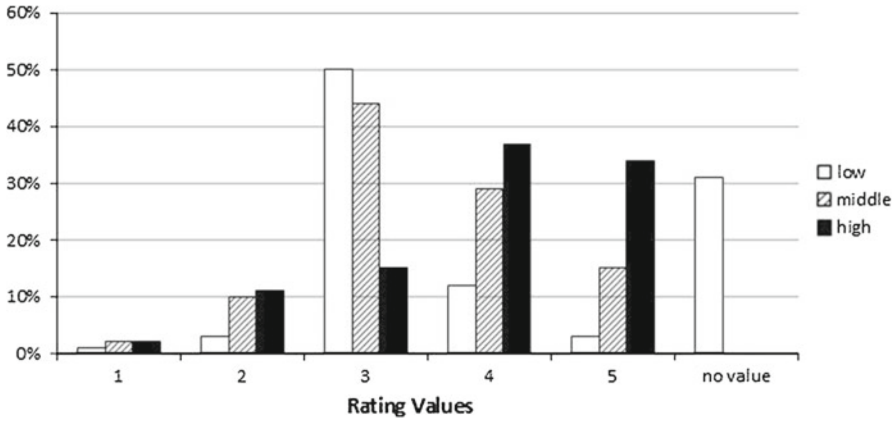


Fig. 4 Frequency of rating values for different user expertise

concentrated on the mid-point value 3 (50 %) and no value at all (31 %). Even though users with some expertise used most often (44 %) the mid-point scale 3, the rest of their values were more distributed over the complete set of available values. Users with both middle and high expertise always gave a value to a criterion. Users with high expertise did not use the mid-point value 3 very frequently (15 %). They made use of the extreme values 4 (37 %) and 5 (34 %) more often. The less used value among all the users was the value 1 which was used only up to 2 % of one expertise group.

We found that users with low expertise were less motivated to use the complete range of the rating scale. The higher the expertise the more frequently users used the extreme high values. In summary, we find that the expertise of a user has influence on the frequency of the used values. Users with low expertise tend to rate the mid-point value 3 or no value at all whereas users with high expertise make more use of the extreme values.

7.1.2 Influence of expertise on rating accuracy

Accuracy, in this study, is defined as agreement with expert raters (cf. Lampe and Garrett [22]). The benchmark scores are based on the ratings of users who indicated their expertise as high (value 3 out of 3). We calculated the benchmark score for each criteria as mean value of ratings submitted by users with high expertise. Only those criteria with a standard deviation of expert rating values lower than 1 were used. The values of users with middle and low expertise were equally calculated as mean values of ratings for each criterion. We chose the values of the user group of the University of Passau for the data set because this was the biggest group of users from one university.

For example, if the rounded mean value of all users with expertise 3 (cf. Sect. 6.1.3) for the criterion “Parking Situation” is 5 (cf. Sect. 6.1.2), we choose the value 5 as the benchmark value for this criterion.

The values of users with middle expertise and those of users with low expertise differ from the benchmark scores for each criteria. However, the values of users with low expertise show a constant bigger difference to the benchmark scores. Table 4

Table 4 Accuracy of ratings for low and middle experts

	MAE	Corr
Low expert ratings	0.77	0.33
Middle expert ratings	0.56	0.47

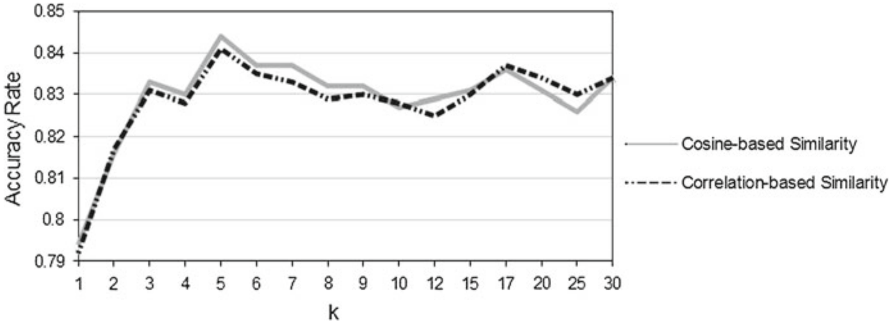


Fig. 5 Accuracy rate of the EPA depending on the neighborhood size *k*

presents the results for the two accuracy metrics MAE (cf. Sect. 6.3.3) and correlation rate (cf. Sect. 6.3.2) for middle and low expert ratings. The MAE is much higher for low experts, that is, their values differ more from the benchmark values than those of middle experts. Middle expert values are also more correlated to the benchmark values as those of low experts.

In summary, we find that the accuracy of rating values depends highly on user expertise. Values of users with low expertise are much less accurate than those of users with middle expertise.

7.2 Question 2: accuracy of expertise prediction

The accuracy of the EPA (cf. Sect. 5.3) depends on the size of the neighborhood k. In order to find the best-performing value for k, we tested the accuracy of the EPA for the set of neighbors $k = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 17, 20, 25, 30\}$. We did so for both similarity calculation approaches described in Sect. 5.1.2.

Figure 5 shows the results for the accuracy rate (cf. Sect. 6.3.4). The results for the cosine-based approach are mostly a little higher than those of the correlation-based approach till the value of $k = 15$. Starting from the accuracy rate 0.79 for $k = 1$, the accuracy increases with increasing k till the value of 5 where it reaches the value 0.84. For a k higher than 5, the accuracy is varying between 0.82 and 0.84.

Choosing the best-performing k, the EPA predicts the expertise of users for criteria up to 84 % accurately. The mean accuracy error of 0.41 shows that those values that are wrongly predicted only differ in a value of approximately 0.4 from the actual expertise value. Considering that we choose the simplest aggregation function for the feedback prediction (cf. Sect. 5.3), we can assume that the accuracy of the user expertise prediction can be further improved with other advanced aggregation functions, e.g., the weighted mean [1].

In conclusion, the EPA is able to predict user expertise for criteria with high accuracy and fulfills its design objective to accurately distribute criteria only to those users who have expertise in it.

8 Related work

8.1 B2B reputation systems

As discussed in Sect. 3, Carlsson [3] conducted a study on reputation systems in the B2B context. One of the aims of the study was to understand how and when business users use online ratings and reviews, and what are the perceived benefits and barriers particularly in comparison to consumer users. A significant disparity was observed between the adoption of online rating and reviewing systems by business users and consumer users. For instance, the ratio between consumer users and business users for having used online ratings and reviews at least 10 times was found to be approximately 4 to 1. The ratio between consumer users and business users who had never done so was observed to be almost 1–10. Moreover, Carlsson also inferred from the gathered data that business users do not yet rely on reputation systems as a measurement of trust.

Another study [19] by kompass.fr found that different business users seek reputation information about different aspects of a target based on their profile. For example, an executive officer would be interested in the financial stability and reputation of a target business, whereas the consumer of a product of that business would be more concerned about the reputation of that specific product.

There are a number of surveys that have covered the broader domain of reputation systems. Jösang et al. [17] present a survey of trust and reputation systems for online service provisioning. Possible attacks and defense mechanisms for reputation systems are identified by Hoffman et al. [14] along with a comparison of existing reputation systems in that context. Pinyol and Sabater-Mir [32] give an overall overview of computational trust and reputation models.

8.2 Recommender systems

The objective of recommender systems is identified as the distribution of an adjusted subset of items to users [1]. Recommendations can be based on items similar to those a user preferred in the past (content-based filtering) or on items that users with similar tastes and preferences liked in the past (collaborative filtering). Recommender systems are a well treated subject in the literature [2,35,37].

Several improvements have been proposed in recent recommender systems in order to increase the accuracy of the recommendations. Liu et al. [25] cluster users based on their criteria preferences and derive recommendations from ratings by other users from similar clusters. They show that the accuracy of recommendations can be improved through this multi-criteria recommendation. Ortega et al. [28] improve collaborative filtering using Pareto dominance to exclude irrelevant users from the *k*-neighbor selection. Another approach is the use of matrix factorization as filtering method for collaborative recommender systems presented by Koren et al. [20]. This approach represents items and users as a vector of factors inferred from item rating patterns. Recommendations are then based on the correspondence between those vectors. The matrix factorization approach has been found to be more accurate than the nearest neighbor approach.

8.3 Expertise prediction

The issue of directing certain items to users according to their expertise can be viewed, for example, in community question answering services. They need to direct questions to users based on their knowledge to obtain accurate answers. They include the various areas of expertise of users in their profiles to be able to distribute each question to users with a high expertise for a question. Pal and Konstan [30] present a mathematical model to distinguish experts from ordinary users in community question answering services. Zhang et al. [40] present an expertise finding mechanism for help-seeking communities which can automatically infer expertise level.

To identify a user's expertise, Chen and Singh [4] present a mechanism that computes the reputation of raters based on the quantity and quality of the ratings they submitted. Using this mechanism, the reliability and quality of ratings submitted by a certain user can be identified.

Contrary to the approaches noted above, in the expertise prediction algorithm that we propose, the input value is not a new criteria for which a high expertise user needs to be found. Our approach takes a new user as input and calculates his expertise for a given set of criteria.

9 Conclusion and future work

Although reputation systems are widely used and treated in the literature, they mostly occur in the C2C or B2C contexts. In this paper, the main particularities of reputation systems in the B2B context were identified, concentrating on the feedback collection part of reputation systems. In the case of a business as the reputation target, we observe that each business may be described by a large number of different criteria. Moreover, in the case of a business as the reputation source, we note that the business is characterized as a group of employees with the particularity that each user has only expertise in a subset of criteria describing the target business.

Based on the above mentioned particularities, we stated some issues of feedback collection for reputation systems in the B2B context. One of the most important issues is the difficulty in the collection of accurate feedback due to a wide variety of criteria in the feedback forms. Our experiment results confirmed the assumption that low user expertise has a negative influence on the accuracy of reputation. Additionally, the motivation of business users to submit ratings is a challenge in contrast to the motivation of non-business users.

As a solution, we presented the EPA, an algorithm which filters and presents only those criteria of the feedback form to a user that he has expertise in. The EPA is built on the idea of collaborative memory-based filtering algorithms used for recommender systems. It predicts the expertise of a user in the criteria describing a business. We conducted an experiment to evaluate two research questions: the impact of user expertise for rating accuracy and the accuracy of expertise prediction of the EPA. The experiment relied on the data of an online survey which was set up for this study and asked students to evaluate their university. We argued that the context of a university has similarities to the B2B context. The results of the experiments showed that the EPA predicts user expertise with an accuracy of up to 84 %.

In future work, we intend to experiment with other recommender system approaches for the expertise prediction algorithm. In particular, we would like to evaluate an implementation of the expertise prediction algorithm with the matrix factorization approach (discussed in Sect. 8.2), which has been found to be more accurate than the nearest neighbor approach. Another direction for future work is to evaluate the expertise prediction algorithm on data from a real B2B environment or on synthetic data that closely simulates such an environment.

References

1. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17(6):734–749. doi:10.1109/TKDE.2005.99
2. Burke R (2002) Hybrid recommender systems: survey and experiments. User Model User Adapt Interact. 12(4):331–370. doi:10.1023/A:1021240730564
3. Carlsson C (2008) Ratings and reviews for the business user—a presentation of results from the author's mba dissertation at henley management college, UK. https://sites.google.com/site/businessuserreviews/
4. Chen M, Singh JP (2001) Computing and using reputations for internet ratings. In: Proceedings of the 3rd ACM conference on electronic commerce, ACM, New York, NY, USA, EC '01, pp 154–162. doi:10.1145/501158.501175
5. Cheng J, Condry M, Karduck AP (2013) Call for papers—7th IEEE international conference on digital ecosystems and technologies (IEEE dest 2013). http://dest2013.digital-ecology.org/
6. Dellarocas C (2003) The digitization of word of mouth: promise and challenges of online feedback mechanisms. Manage Sci 49(10):1407–1424. doi:10.1287/mnsc.49.10.1407.17308
7. Dellarocas CN, Fan M, Wood CA (2004) Self-interest, reciprocity and participation in online reputation systems. Working Papers 4500–04, MIT Sloan. http://ssrn.com/abstract=585402
8. Fung H (2011) Rate this page? is coming to the english wikipedia. http://blog.wikimedia.org/2011/07/15/
9. Golbeck J, Hendler J (2006) Inferring binary trust relationships in web-based social networks. ACM Trans Internet Technol 6(4):497–529. doi:10.1145/1183463.1183470
10. Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: a constant time collaborative filtering algorithm. Inf Retr 4(2):133–151. doi:10.1023/A:1011419012209
11. Grcar M (2004) User profiling: collaborative filtering. In: Proceedings of the 7th international multi-conference information society IS, pp 75–78
12. Harper FM, Li X, Chen Y, Konstan JA (2005) An economic model of user rating in an online recommender system. In: Proceedings of the 10th international conference on user modeling. Springer, Berlin, Heidelberg, UM'05, pp 307–316. doi:10.1007/11527886_40
13. Hasan O, Brunie L, Bertino E (2012) Preserving privacy of feedback providers in decentralized reputation systems. Comput Secur 31(7): 816–826
14. Hoffman K, Zage D, Nita-Rotaru C (2009) A survey of attack and defense techniques for reputation systems. ACM Comput Surv 41(4)
15. Houser D, Wooders J (2001) Reputation in auctions: theory and evidence from eBay.Mimeo
16. Jamieson S (2004) Likert scales: how to (ab)use them. Med Educ 38(12):1217–1218. doi:10.1111/j.1365-2929.2004.02012.x

17. Jøsang A, Ismail R, Boyd C (2007) A survey of trust and reputation systems for online service provision. Decis Support Syst 43(2):618–644
18. JStueland V (2004) Supplier evaluation: best practices and creating or improving your own evaluation. In: 89th annual international supplier management conference
19. Kompass (2012) Le processus d'achat b2b. http://fr.kompass.com/espace-business/index.php?option=com&view
20. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37. doi:10.1109/MC.2009.263
21. Lakiotaki K, Matsatsinis NF, Tsoukias A (2011) Multicriteria user modeling in recommender systems. IEEE Intell Syst 26:64–76. doi:10.1109/MIS.2011.33
22. Lampe C, Garrett RK (2007) It's all news to me: the effect of instruments on ratings provision. In: Proceedings of the 40th annual Hawaii international conference on system sciences, IEEE Computer Society, Washington, DC, USA, HICSS '07, pp 180b. doi:10.1109/HICSS.2007.308
23. Lee S, Song Si, Kahng M, Lee D, Lee Sg (2011) Random walk based entity ranking on graph for multidimensional recommendation. In: Proceedings of the fifth ACM conference on recommender systems, ACM, New York, NY, USA, RecSys '11, pp 93–100. doi:10.1145/2043932.2043952
24. Liu L, Munro M (2012) Systematic analysis of centralized online reputation systems. Decis Support Syst 52(2):438–449. doi:10.1016/j.dss.2011.10.003
25. Liu L, Mehandjiev N, Xu DL (2011) Multi-criteria service recommendation based on user criteria preferences. In: Proceedings of the fifth ACM conference on recommender systems, ACM, New York, NY, USA, RecSys '11, pp 77–84. doi:10.1145/2043932.2043950
26. McLaughlin MR, Herlocker JL (2004) A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, ACM, New York, NY, USA, SIGIR '04, pp 329–336. doi:10.1145/1008992.1009050
27. Miller B, Riedl JT, Konstan JA (1997) Experiences with grouplens: making usenet useful again. In: Proceedings of the 1997 Usenix Winter technical conference, pp 219–231
28. Ortega F, SáNchez JL, Bobadilla J, GutiéRrez A (2013) Improving collaborative filtering-based recommender systems results using pareto dominance. Inf Sci 239:50–61. doi:10.1016/j.ins.2013.03.011
29. Ozakca M, Lim YK (2006) A study of reviews and ratings on the internet. In: CHI '06 extended abstracts on human factors in computing systems, ACM, New York, NY, USA, CHI EA '06, pp 1181–1186. doi:10.1145/1125451.1125673
30. Pal A, Konstan JA (2010) Expert identification in community question answering: exploring question selection bias. In: Proceedings of the 19th ACM international conference on information and knowledge management, ACM, New York, NY, USA, CIKM '10, pp 1505–1508. doi:10.1145/1871437.1871658
31. Parra D, Karatzoglou A, Amatriain X, Yavuz I (2011) Implicit feedback recommendation via implicit-to-explicit ordinal logistic regression mapping. In: In CARS Workshop at RecSys
32. Pinyol I, Sabater-Mir J (2011) Computational trust and reputation models for open multi-agent systems: a review. Artif Intell Rev. doi:10.1007/s10462-011-9277-z
33. Resnick P, Kuwabara K, Zeckhauser R, Friedman E (2000) Reputation systems. Commun ACM 43(12):45–48
34. Sabater J, Sierra C (2005) Review on computational trust and reputation models. Artif Intell Rev 24(1):33–60
35. Schafer JB, Konstan J, Riedi J (1999) Recommender systems in e-commerce. In: Proceedings of the 1st ACM conference on electronic commerce, ACM, New York, NY, USA, EC '99, pp 158–166
36. Sparling EI, Sen S (2011) Rating: how difficult is it? In: Proceedings of the fifth ACM conference on recommender systems, ACM, New York, NY, USA, RecSys '11, pp 149–156
37. Swearingen K, Sinha R (2002) Interaction design for recommender systems. In: Designing Interactive Systems, Vol. 6, No. 12. ACM Press, pp. 312–334
38. Yesilada Y, Brajnik G, Harper S (2009) How much does expertise matter? A barrier walkthrough study with experts and non-experts. In: Proceedings of the 11th international ACM SIGACCESS conference on computers and accessibility, ACM, New York, NY, USA, Assets '09, pp 203–210
39. Zacharia G (2000) Trust management through reputation mechanisms. Appl Artif Intell 14:881–907
40. Zhang J, Ackerman MS, Adamic L, Nam KK (2007) Qume: a mechanism to support expertise finding in online help-seeking communities. In: Proceedings of the 20th annual ACM symposium on user interface software and technology, ACM, New York, NY, USA, UIST '07, pp 111–114