Multi-Layer Cognitive Filtering by Behavioral Modeling

Zeinab Noorian
University of New Brunswick
Fredericton,Canada
z.noorian@unb.ca

Stephen Marsh
Communications Research
Centre, Canada
stephen.marsh@crc.gc.ca

Michael Fleming
University of New Brunswick
Fredericton,Canada
mwf@unb.ca

ABSTRACT

In the absence of legal enforcement procedures for the participants of an open e-marketplace, trust and reputation systems are central for resisting against threats from malicious agents. Such systems provide mechanisms for identifying the participants who disseminate unfair ratings. However, it is possible that some of the honest participants are also victimized as a consequence of the poor judgement of these systems. In this paper, we propose a two-layer filtering algorithm that cognitively elicits the behavioral characteristics of the participating agents in an e-marketplace. We argue that the notion of *unfairness* does not exclusively refer to deception but can also imply differences in dispositions. The proposed filtering approach aims to go beyond the inflexible judgements on the quality of participants and instead allows the human dispositions that we call optimism, pessimism and realism to be incorporated into our trustworthiness evaluations. Our proposed filtering algorithm consists of two layers. In the first layer, a consumer agent measures the competency of its neighbors for being a potentially helpful adviser. Thus, it automatically disqualifies the deceptive agents and/or the newcomers that lack the required experience. Afterwards, the second layer measures the credibility of the surviving agents of the previous layer on the basis of their behavioral models. This tangible view of trustworthiness evaluation boosts the confidence of human users in using a web-based agent-oriented e-commerce application.

Categories and Subject Descriptors

[**distributed artificial intelligence**]: multi-agent systems

General Terms

Human Factors, Design, Measurement

Keywords

Trust, Reputation, Cognitive filtering, Behavioral modeling

1. INTRODUCTION

The inherent uncertainties in an open e-marketplace inhibit participants from reaching a mutual understanding and confidence about each other's intentions [3]. This matter affects the formation of agent-based e-commerce applications handled by human users since their personal dispositions are not allowed to be reflected in their decisions or, if so, other participants are not able to identify them[8]. As such, despite the intrinsic honesty of their recommendations, they may not be considered trustworthy. This is undoubtedly justifiable with the existence of malicious participants. In particular, in order to diminish the risk of being misled by unfair advisers, a consumer agent restricts itself to seek advice from the participants with the most similar ratings[5, 14].

In this paper we intend to amend this common view of trustworthiness [11, 16] by introducing a new definition for *unfairness*. We discuss that the intuition of unfairness could be examined across two categories: 1) *intentional*, a) participants consistently act malevolently and b) participants occasionally engage in deceitful activities. And 2) *unintentional*, as a result of a) lack of personal experience and b) various behavioral characteristics resulting in different rating attitudes.

We propose a two-layered filtering algorithm that combines cognitive and probabilistic views of trust [3] to mainly target the intentional group of unfair advisers. We show that modeling the trustworthiness of advisers based on a strict judgement of the quality of their recommendations is not complete unless it is accompanied by the analysis of their dispositions. Thus, through the comprehension of their rating attitudes, a consumer agent could take appropriate steps to evaluate them.

The main contributions of this paper are twofold: First, we propose competency evaluation methods to detect newcomers with a lack of experience and thereafter disqualify them from the role of advisers. Second, we introduce a classification schema to identify the behavioral characteristics of participants and design credibility assessment measures for each of them.

Our experimental results show the utility of our approach in terms of recognizing dispositions of various participants and, specifically, how consumers with personalized thought-frames evaluate the same adviser differently. Our filtering model can therefore be seen as an effective approach in modeling the reputation of advisers in a multi-agent system.

2. TWO-LAYERED COGNITIVE FILTERING ALGORITHM

To formalize the proposed cognitive filtering algorithm, we consider the scenario where, in an electronic marketplace, consumer agents with distinctive behavioral patterns want to bootstrap relationships with new neighbors. We assume

Cite as: Multi-Layer Cognitive Filtering by Behavioural Modeling, Zeinab Noorian, Stephen Marsh and Michael Fleming, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 871-878.

that the consumer agents have some record of interactions with transaction partners, i.e, providers. We also assume that participating agents are cooperative and willing to communicate with each other.

To analyze the neighbors' trustworthiness, each consumer agent C needs two types of information. The first type of information, which helps to build the first layer of our filtering algorithm, is used to identify malicious participants with a complementary model of deception. It also detects newly-joined agents with an insufficient number of personal experiences.

In the second layer of the filtering algorithm, the second type of information helps C to recognize the behavioral characteristics of the neighbors. As such, it will be able to evaluate their degree of trustworthiness. Note that, in this layer, C takes an analytical approach in order to detect deceitful participants with volatile dispositions who cheat opportunistically. By hiding their true intentions, this group of deceitful participants imposes greater risk and insecurity to the system compared with those with a frequently deceptive attitude[1, 6, 7].

The detailed explanation of this multi-dimensional filtering technique is provided in the following sub-sections.

2.1 First Layer: Evaluating the Competency Degree of Neighbors

The consumer agent C sends a query to a circle of its neighbors $N = \{N_1, N_2, ..., N_i\}$ requesting numbers of successful and unsuccessful outcomes experienced with providers $P = \{P_1, P_2, ..., P_r\} \subseteq \{P_1, P_2, ..., P_m\}, r \leq m$, occurring before a certain time T. Such a time threshold diminishes the risk of changeability in a provider's behavior. It is also accompanied by the Quality of Service (QoS) threshold Ω to imply C's belief about an acceptable minimum level of trust. For example, for a consumer with a *risk-averse* pattern, Ω could be 0.7 whereas for the *risk-taking* consumer this amount might be reduced to 0.5.

The neighbor N_k responds by providing a rating vector $R_{(N_k,P_j)}$ for each provider. It contains a tuple of $\langle r, s\rangle$ which indicates the number of successful (r) and unsuccessful (s) interaction results with provider P_j respectively. Note that, in the first layer of the filtering algorithm, neighbors are asked to provide merely a *binary rating* ("1" or "0" for example), in which "1" means that P_j is reputable and "0" means not reputable. Thus, considering a consumer's QoS threshold, they will send reputation reports as a collection of positive and negative interaction outcomes.

Once the evidence is received, for each $R_{(N_k,P_j)}$, C calculates the expected value of the probability of the positive outcome for a provider P_j [9] as:

$$E(pr_r, P_j) = \frac{r+1}{r+s+2} \qquad (1)$$

To generally present this formula to include all participants in an e-marketplace, we update the presentation of $E(pr_r, P_j)$ to $E(pr_r, P_j)_{Par}$, where $Par \in \{C\} \cup N$ implies participants of the community. Clearly, $0 < E(pr_r, P_j)_{Par} \leq 1$ and as it approaches 0 or 1, it indicates *unanimity* in the body of evidence[4]. That is, particularly large values of s or r provide better intuition about an overall tendency and service quality of providers. In contrast, $E(pr_r, P_j)_{Par} = 0.5$ (i.e, $r = s$) signifies the maximal conflict in gathered evidence, resulting in increasing the uncertainty in determining the service quality of providers. Based on these intuitions, we are able to calculate the degree of reliability and certainty of ratings provided by neighbors.

Let x represent the probability of a successful outcome for a certain provider. Based on the Definitions(2) and (3) in [12], the *Reliability degree* of each $R_{(N_k,P_j)}$ is defined as:

$$c(r,s) = \frac{1}{2}\int_0^1 \left| \frac{x^r(1-x)^s}{\int_0^1 x^r(1-x)^s\, dx} - 1 \right| dx \qquad (2)$$

Similar to $E(pr_r, P_j)_{Par}$, we update the presentation of $c(r,s)$ to $c(r,s)_{Par}$.

Theoretical analysis [12] demonstrates that, for a fixed ratio of positive and negative observations, the reliability increases as evidence increases. On the contrary, given a fixed amount of evidence, as the extent of conflict increases, the reliability of the provided ratings decreases proportionately. That is, reliability is at its minimum value when $E(pr_r, P_j)_{Par} = 0.5$. As such, the less conflict in their ratings, the more reliable the neighbors would be.

However, in the proposed filtering algorithm, C would not strictly judge the neighbors with rather low reliability in their $R_{(N_k,P_j)}$ as deceptive participants since this factor could signify both dishonesty of neighbors and the dynamicity and fraudulent behavior of providers. That is, some malicious providers may adopt a strategy of providing satisfactory quality of service in most situations when there is not much at stake and acting conversely in occasions associated with a large gain. As such, even though they retain a certain level of trustworthiness, their associated reliability degree is low. To address this ambiguity, C computes the $E(pr_r, P_j)_C$ and $c(r,s)_C$ of its personal experiences; $R_{(C,P_j)}$, for a common set of providers. Through the comparison of neighbors' metrics with its own, it would select those with a similar rating pattern and a satisfactory level of honesty as its *advisers*.

To formalize this, it measures an average level of dishonesty of N_k by differentiating their $E(pr_r, P_j)_{Par}$ as:

$$\overline{d}_{(N_k)} = \frac{\sum_{j=1}^{|P|} \mid E(pr_r, P_j)_C - E(pr_r, P_j)_{N_k} \mid}{|P|} \qquad (3)$$

As pointed out, increasing the amount of evidence leads to an increase in the reliability degree. The problem arises when malicious neighbors disseminate a large number of spurious ratings so as to promote their reliability. Besides, it may happen that a truthful neighbor lacks in number of experiences. Thus, despite its inherent honesty, its reliability degree is low and it is not qualified to play the role of adviser. To clarify these issues, we define an uncertainty function $\overline{U}_{(N_k)}$ to capture the intuition of information imbalance between C and N_k as follows:

$$\overline{U}_{(N_k)} = \frac{\sum_{j=1}^{|P|} \mid (c(r,s)_C - c(r,s)_{N_k})_{P_j} \mid}{|P|} \qquad (4)$$

In light of the uncertainty function, the opinions of deceptive neighbors who attempt to mislead consumer agents by supplying a large number of ratings are discounted. Similarly, it hinders short-term observations of newly-joined agents from having influence on a consumer agent's decision making process.

Given the formulae (3) and (4), the *competency degree* of N_k is calculated by reducing its honesty based on its certainty degree. Thus, it could be determined as:

$$Comp_{(N_k)} = (1 - \overline{d}_{(N_k)}) * (1 - \overline{U}_{(N_k)}) \tag{5}$$

By comparing their competency degree with a pre-defined incompetency tolerance threshold μ, C evaluates the qualification and eligibility of the neighbors to play the role of *adviser*. As such, It chooses the neighbors with $(1 - Comp_{(N_k)}) \leq \mu$ as its potential advisers and filters out the rest. It is worthwhile to note that, since in this layer we target the participants with a significant lying pattern, detecting fraudulent agents with oscillating rating attitudes is left for the next layer.

2.2 Second Layer: Calculating a Credibility Degree of Advisers

In the first phase of the filtering algorithm, neighbors are asked to send their subjective opinions of providers. By aggregating their opinions and computing their degree of reliability, a consumer agent has obtained a rough estimation of the honesty level of neighbors and selects a subset of them as its advisers. However, this method cannot thoroughly address the inherent complications of an open environment. To explain, the nature of the open marketplace allows various kinds of participants with distinctive behavioral characteristics [2] to engage in the system.

Besides, the basis of the employed multi-dimensional rating system provides tools for a consumer agent to objectively evaluate the performance of service providers across several criteria with different degrees of preference. Evidently, the measured QoS is mainly dependent on how much the criteria with a high preference degree are fulfilled[7]. Owing to the different purchasing behavior of the agents, it is expected that preference degrees vary from one participant to another, resulting in dissimilar assessment of the quality of the *same* service. As such, computing the credibility of advisers regardless of their behavioral characteristics and rating attitudes, and merely based on their subjective opinions, would not sufficiently ensure high quality judgements of their trustworthiness.

To tackle these problems, in a second layer of the filtering algorithm, consumer agent C steps forward and analytically gives credits to advisers to the extent that their evaluation of each criterion of a negotiated context is similar to its own experiences. For this purpose, it asks advisers about mutually agreed criteria on which they have bargained with *highly-reliable* providers[1] whose reputation values have been recently released in the form of binary ratings. They also are requested to include the most recent interaction time with such information so as to give a higher weight to more recent feedback. That is, feedback gradually loses its importance as time progresses. This improves the correctness and accuracy in predicting the credibility of advisers through alleviating the risk of changeability in a provider's behavior.

To formulate this, we adopt the concept of forgetting factor presented in [9, 16]:

$$z = \lambda^{T_A - T_C} \tag{6}$$

We customize it for our model and define a recency factor as:

$$T_{(C,A_k)_{P_j}} = \frac{1}{z} \tag{7}$$

Here, T_A and T_C indicate the adviser's and consumer's time windows when they had an experience with a provider P_j. Also, the λ represents the forgetting parameter and $0 < \lambda \leq 1$. When $\lambda = 1$, there is no forgetting and all the ratings are treated as though they happened in the same time period. In contrast, $\lambda \approx 0$ specifies that ratings from different time windows will not be significantly taken into account. Similarly to [16], in this filtering algorithm, the recency factor is characterized with a discrete integer value where 1 is the most recent time period and 2 is the time period just prior. Also, it is presumed that the adviser's ratings are prior to those a consumer agent supplies so that $T_A \geq T_C$.

Adviser A_k will respond, providing an interaction context $IC_{(A_k,P_j,T_A)}$ that contains a tuple of *weight* and *value*: $\{W_i.V_i | i = 1..n\}$ and the latest interaction time T_A for each provider.[2]

Given A_k's interaction context, a consumer agent would estimate the possible interaction outcomes of an adviser based on its own perspective. That is, C will examine its $IC_{(C,P_j,T_C)}$, which contains pairs of weight and value: $\{Y_i.R_i | i = 1..n\}$. It will then modify the interaction context of A_k by replacing A_k's preferences W_i with its own personal preference degrees Y_i. Based on this, the interaction context of A_k is updated to: $IC'_{(A_k,P_j,T_A)} = \{Y_i.V_i | i = 1..n\}$.

To formalize a similarity of A_k's rating approach with C, we compute a ratio of the consumer's interaction context $IC_{(C,P_j,T_C)}$ with the updated version of the adviser's interaction context as:

$$Sim_{(C,A_k)_{P_j}} = \frac{\sum_{i=1}^{n} Y_i \times R_i}{\sum_{i=1}^{n} Y_i \times V_i} \tag{8}$$

and then

$$Diff_{(C,A_k)_{P_j}} = 1 - Sim_{(C,A_k)_{P_j}}$$

represents the difference of C and A_k in assessing P_j.

Based on Equations (7) and (8), C would calculate the *average* differences between the transaction result of A_k and its own experiences with a same set of providers as:

$$\overline{Diff}_{(C,A_k)} = \frac{\sum_{j=1}^{|P|} \mid Diff_{(C,A_k)_{P_j}} \mid * T_{(C,A_k)_{P_j}}}{|P|} \tag{9}$$

Existing trust models [5, 9, 11, 14, 16] evaluate the trustworthiness of advisers mainly based on their average deviation from a consumer's opinion and exploit the *same* credibility measures for all types of advisers. Moreover, they define a threshold value[3] to separate the honest advisers from dishonest ones. However, adjusting a threshold to an efficient value has always been a controversial issue. The quality of advisers is compromised when a threshold is set to a high value. In this situation, deceitful participants who maintain a minimum level of trustworthiness remain undetected and could actively contribute to a consumer's decision making process. On the other hand, a lower threshold

[1] Obviously, a consumer only inquires about the providers with high reliability and ignores those that are possibly deceptive.

[2] Note that in this model we assume that each provider can only provide one particular service. Dealing with providers offering multiple services is left for future work.

[3] A threshold can be explicitly determined as in [5] and [14] or implicitly as in[16].

leads to the contribution of a smaller number of advisers. Clearly, adjusting a threshold value is a trade-off between the number of credible advisers and the risk of being misled by deceptive peers.

Furthermore, in a real-life e-commerce application, the differences in a consumer's behavioral patterns lead to divergent evaluations of the credibility degree of advisers[8]. For instance, the opinion of one particular adviser may seem highly credible for a risk-taking consumer while it is not so for a risk-averse one. We note that the credibility degree of advisers *not only* depends on their evaluator's dispositions but it is also related to their own individual behavioral patterns. That is, advisers' recommendations could be affected by endogenous factors[3]. As such, it may happen that two honest advisers with different attitudes have conflicting evaluations of the same provider. Characterizing the disposition of advisers helps a consumer agent to take a proportionate strategy in assessing their future recommendations. For instance, a risk-averse consumer would underestimate the ratings provided by optimistic advisers whilst overrating those provided by pessimistic advisers. This mechanism shows its practicality in a community where credible advisers are scarce and the majority of participants behave malevolently. In this state, modeling a behavior of advisers helps a consumer to get the most benefits from their opinions in such a way that the scarcity of credible advisers would not have a serious effect on the quality of predictions.

For all these reasons, in this model we take a further step and embrace the diversity in participants as an influential factor in our credibility measures. We believe that quantity should not necessarily be sacrificed for quality or vice versa. Instead, by employing a suitable mechanism, consumer agents are able to have a large number of advisers with high-quality judgements. As such, C captures the overall tendency of A_k in evaluating the providers' QoS as:

$$Tendency_{(C,A_k)} = \frac{\sum_{j=1}^{|P|} Diff_{(C,A_k)_{P_j}}}{|P|} \tag{10}$$

As the name suggests, the consumer agent could exploit a tendency metric to get an intuition about the general trends of advisers in rating a common set of providers. That is, a positive value of $Tendency_{(C,A_k)}$ indicates that an adviser has the attitude of overrating providers while a negative value declares that an adviser has a tendency to underrate providers.

Following that, to identify a behavioral pattern of advisers, we determine a pre-defined boundary β such that if A_k's $IC'_{(A_k,P_j,T_A)}$ is compatible with those experienced by C ($\overline{Diff}_{(C,A_k)} \leqslant \beta$), they will be counted as *credible* advisers. However, in this model, C would not thoroughly exclude the advisers who rate otherwise. Instead, it narrowly analyzes the $\overline{Diff}_{(C,A_k)}$ in such a way that if it is marginally greater than β with a negative $Tendency_{(C,A_k)}$, the corresponding adviser's attitude is identified as *pessimistic*.

Similarly, in case their differences marginally exceed β with a positive $Tendency_{(C,A_k)}$, the respective adviser's attitude is recognized as *optimistic*. We define such a marginal error ϵ as a ratio of the credibility threshold β and it is subjectively determined by a consumer agent. Evidently, if A_k's $IC'_{(A_k,P_j,T_A)}$ significantly deviates from the consumer agent's direct experiences, they will be detected as *malicious* advisers with *deceitful* behavioral models. We believe that the filtered advisers have a deceitful behavioral pattern; otherwise, they would have been expelled in the first layer.

Note that the thresholds are used to identify different kinds of unfair participants. These thresholds should be set with the goals of each particular layer in mind. In the first layer, the value of μ should be high, to ensure that dishonest participants are expelled. In the second layer, when analyzing participants' behavioral characteristics, a low value of β is desirable. Thus, we can conclude that $\beta \leq \mu$.

The classification mechanism of the behavioral pattern of A_k based on C's interaction context is formally presented as follows:

$$BP_{(C,A_k)} = \begin{cases} Realistic/Credible: \\ \quad \overline{Diff}_{(C,A_k)} \leqslant \beta \\ Optimistic: \\ \quad \beta < \overline{Diff}_{(C,A_k)} \leqslant \beta + \epsilon \ \& \ Tendency_{(C,A_k)} > 0 \\ Pessimistic: \\ \quad \beta < \overline{Diff}_{(C,A_k)} \leqslant \beta + \epsilon \ \& \ Tendency_{(C,A_k)} < 0 \\ Deceitful: \\ \quad \overline{Diff}_{(C,A_k)} > \beta + \epsilon \end{cases} \tag{11}$$

Given the $BP_{(C,A_k)}$, the credibility measure $CR_{(C,A_k)}$ is formulated as:

$$CR_{(C,A_k)} = \begin{cases} 1 - \overline{Diff}_{(C,A_k)}: \\ \qquad BP_{(A_k)} = Credible \\ (1 - \overline{Diff}_{(C,A_k)}) \times e^{-\theta * \overline{Diff}_{(C,A_k)}}: \\ \qquad BP_{(A_k)} = Optimistic \\ (1 - \overline{Diff}_{(C,A_k)}) \times e^{-\sigma * \overline{Diff}_{(C,A_k)}}: \\ \qquad BP_{(A_k)} = Pessimistic \\ 0: \\ \qquad BP_{(A_k)} = Deceitful \end{cases} \tag{12}$$

Here, θ and σ represent the optimistic and pessimistic coefficients respectively. A consumer agent takes a personalized adaptive approach to calculate them. Depending on its behavioral characteristics, such coefficients are initialized differently. For instance, recommendations of pessimistic advisers may seem more credible in the perspective of a risk-averse consumer and they are considered to be better peers to cooperate with than optimistic advisers [2]. Hence the risk-averse consumer promotes the credibility of a pessimistic adviser by adjusting the pessimistic coefficient to a lower value than the optimistic coefficient$\langle 0 \leq \sigma < \theta \rangle$. On the contrary, the disposition of a risk-taking buying agent compels it to consider the reputation information provided by optimistic advisers as more important. Therefore, it assigns a great deal of influence to their ratings by properly setting up the optimistic coefficient to a lower value than the pessimistic coefficent $\langle 0 \leq \theta < \sigma \rangle$.

As such, the coefficients are adaptively defined for each adviser. For initializing θ, a risk-averse agent considers the maximum difference of the adviser's ratings with a consumer's opinions upon evaluating the same providers. For a risk-taking agent, this process is reversed. That is, the optimistic coefficient is defined as the minimum deviation of the adviser's recommendations with the consumer's opinions across a common set of providers. Thus, coefficients θ and σ are formalized as:

$$\theta = \begin{cases} max\{| \ Diff_{(C,A_k)_{P_i}} \ | \ |i = 1...m\} & Risk\text{-}Averse\ consumer \\ min\{| \ Diff_{(C,A_k)_{P_i}} \ | \ |i = 1...m\} & Risk\text{-}Taking\ consumer \end{cases} \tag{13}$$

$$\sigma = \begin{cases} min\{\mid Diff_{(C,A_k)_{P_i}} \mid \mid i = 1...m\} & \textit{Risk-Averse consumer} \\ max\{\mid Diff_{(C,A_k)_{P_i}} \mid \mid i = 1...m\} & \textit{Risk-Taking consumer} \end{cases} \tag{14}$$

Through these principles, A_k's recommendations are discounted such that its influence on C's prediction depends on its honesty in *each* of its interaction contexts. The coefficient parameters ensure that the recommendation of advisers with volatile behavior who have a high variability in their opinions is heavily discounted.

3. EXAMPLES

In an electronic marketplace, a consumer C_1 needs to make a decision on whether to interact with a provider P_1. This depends on how much C_1 trusts P_1. To model the trustworthiness of P_1, when the consumer does not have an adequate number of experiences with P_1, it ought to seek advice from its neighbors. However, it first needs to acquire enough information about their credibility value in order to assign a proper credibility to their provided ratings.

In the first phase of the filtering algorithm, the risk-averse C_1 asks its surrounding neighbors $\{N_1, N_2, .., N_6\}$ about the overall performance of the providers $\{P_1, P_2, .., P_8\}$ before time T, given the QoS threshold $\Omega = 0.7$.

Consider the case where the neighbors $\{N_1, .., N_5\}$ have rated only the five providers $\{P_1, P_2, P_3, P_4, P_5\}$. Using Equations (1) and (2), C_1 would calculate the expected value of the probability of positive ratings along with a degree of reliability of their evidence. Table 1 lists a number of successful/unsuccessful ratings provided by $N_i (i \in \{1, .., 5\})$ and C_1 for the five providers along with their $E(pr_r, P_j)_{N_i}$ and $c(r, s)_{N_i}$.

Table 1: Ratings provided by the neighbors and C_1 along with their corresponding metrics

Participants	P_i	$\langle r, s\rangle$	$E(pr_r, P_j)_{Par}$	$c(r, s)_{Par}$
C_1	P_1	$\langle 16, 1\rangle$	0.89	0.71
	P_2	$\langle 7, 4\rangle$	0.61	0.47
	P_3	$\langle 2, 10\rangle$	0.21	0.57
	P_4	$\langle 15, 0\rangle$	0.94	0.77
	P_5	$\langle 2, 4\rangle$	0.37	0.38
N_1	P_1	$\langle 25, 0\rangle$	0.96	0.84
	P_2	$\langle 8, 3\rangle$	0.69	0.5
	P_3	$\langle 2, 5\rangle$	0.33	0.42
	P_4	$\langle 8, 0\rangle$	0.9	0.67
	P_5	$\langle 3, 2\rangle$	0.57	0.33
N_2	P_1	$\langle 8, 5\rangle$	0.6	0.54
	P_2	$\langle 9, 5\rangle$	0.62	0.51
	P_3	$\langle 5, 5\rangle$	0.50	0.44
	P_4	$\langle 11, 6\rangle$	0.63	0.55
	P_5	$\langle 3, 4\rangle$	0.44	0.38
N_3	P_1	$\langle 13, 2\rangle$	0.82	0.62
	P_2	$\langle 2, 6\rangle$	0.3	0.46
	P_3	$\langle 4, 7\rangle$	0.38	0.47
	P_4	$\langle 20, 5\rangle$	0.77	0.65
	P_5	$\langle 1, 11\rangle$	0.15	0.64
N_4	P_1	$\langle 4, 11\rangle$	0.29	0.55
	P_2	$\langle 6, 4\rangle$	0.58	0.45
	P_3	$\langle 13, 5\rangle$	0.7	0.57
	P_4	$\langle 5, 9\rangle$	0.37	0.51
	P_5	$\langle 10, 6\rangle$	0.61	0.53
N_5	P_1	$\langle 2, 0\rangle$	0.75	0.38
	P_2	$\langle 1, 0\rangle$	0.66	0.25
	P_3	$\langle 1, 2\rangle$	0.4	0.27
	P_4	$\langle 1, 0\rangle$	0.66	0.25
	P_5	$\langle 0, 1\rangle$	0.33	0.25

To calculate the competency degree of neighbors, C_1 would analyze their average dishonesty. Through $\overline{U}_{(N_k)}$, it also examines the adequacy of their ratings. Afterwards, using Equation (5), it is able to calculate their $Comp_{(N_k)}$, resulting in detection of particular neighbors with consistent deceptive attitudes and those with few experiences. Here, a risk-averse C_1 selects the neighbors $\{N_1, N_2, N_3\}$ whose competency values $Comp_{(N_k)}$ surpass $\mu = 0.65$ and filters out the rest (Table 2). Next, in the second layer, C_1 requests detailed descriptions of their negotiated criteria with the selected set of providers so as to identify the behavioral characteristics of advisers. Table 3 articulates personal ratings of each participant through the $\langle weight, value\rangle$ pair related to each criterion regarding the selected providers.

Table 2: Calculating the competency level of the neighbors

N_i	$\overline{d}_{(N_k)}$	$\overline{U}_{(N_k)}$	$Comp_{(N_k)}$
N_1	0.1	0.08	0.81
N_2	0.19	0.12	0.70
N_3	0.19	0.11	0.71
N_4	0.38	0.12	0.54
N_5	0.13	0.3	0.59

As can be perceived, in the first layer, the disposition of consumers has not been reflected in the evaluation of the competency degree of the neighbors. To observe the influence of this factor in the second layer, we introduce a risk-taking C_2 in addition to C_1 and examine their approaches in evaluating the same advisers. Finally, as indicated in Tables 4 and 5, the behavioral patterns of participants could serve as determinant factors in evaluating their trustworthiness. We notice that, consumer agents with similar deviation ($\overline{Diff}_{(C_1,N_1)} = \overline{Diff}_{(C_2,N_1)}$) and the same $\beta = 0.15, \epsilon = 0.07$ and $\lambda = 0.8$ could predict different credibility values for the same adviser. Note that, in Table 4, we use the notation $Tend_{(C_1,N_k)}$ for $Tendency_{(C_1,N_k)}$.

Table 3: The negotiated criteria of participants with selected providers

Participants	P_i	$Criteria\langle w, v\rangle$				T
		Cri_1	Cri_2	Cri_3	Cri_4	
C_1	P_1	$\langle 6, 7\rangle$	$\langle 10, 9\rangle$	$\langle 10, 8\rangle$	$\langle 5, 10\rangle$	2
	P_3	$\langle 10, 4\rangle$	$\langle 7, 2\rangle$	$\langle 5, 5\rangle$	$\langle 10, 3\rangle$	2
	P_4	$\langle 6, 5\rangle$	$\langle 3, 10\rangle$	$\langle 10, 10\rangle$	$\langle 8, 6\rangle$	1
C_2	P_1	$\langle 3, 8\rangle$	$\langle 10, 9\rangle$	$\langle 9, 7\rangle$	$\langle 8, 9\rangle$	3
	P_3	$\langle 9, 3\rangle$	$\langle 8, 4\rangle$	$\langle 10, 5\rangle$	$\langle 7, 4\rangle$	1
	P_4	$\langle 2, 8\rangle$	$\langle 10, 10\rangle$	$\langle 8, 10\rangle$	$\langle 6, 6\rangle$	1
N_1	P_1	$\langle 4, 9\rangle$	$\langle 10, 10\rangle$	$\langle 10, 8\rangle$	$\langle 6, 10\rangle$	3
	P_3	$\langle 10, 5\rangle$	$\langle 6, 2\rangle$	$\langle 4, 6\rangle$	$\langle 10, 5\rangle$	3
	P_4	$\langle 3, 8\rangle$	$\langle 10, 10\rangle$	$\langle 10, 10\rangle$	$\langle 6, 7\rangle$	7
N_2	P_1	$\langle 7, 5\rangle$	$\langle 10, 7\rangle$	$\langle 10, 4\rangle$	$\langle 4, 6\rangle$	4
	P_3	$\langle 10, 6\rangle$	$\langle 5, 6\rangle$	$\langle 2, 7\rangle$	$\langle 10, 6\rangle$	3
	P_4	$\langle 10, 3\rangle$	$\langle 5, 7\rangle$	$\langle 10, 6\rangle$	$\langle 7, 4\rangle$	2
N_3	P_1	$\langle 10, 6\rangle$	$\langle 10, 8\rangle$	$\langle 10, 7\rangle$	$\langle 5, 9\rangle$	4
	P_3	$\langle 10, 3\rangle$	$\langle 7, 2\rangle$	$\langle 5, 4\rangle$	$\langle 10, 2\rangle$	3
	P_4	$\langle 8, 5\rangle$	$\langle 3, 10\rangle$	$\langle 10, 9\rangle$	$\langle 7, 5\rangle$	1

Table 4: Calculating tendency of neighbors and their deviation degree based on C_1 and C_2's experiences

N_k	$\overline{Diff}_{(C_1,N_k)}$	$Tend_{(C_1,N_k)}$	$\overline{Diff}_{(C_2,N_k)}$	$Tend_{(C_2,N_k)}$
N_1	0.18	0.13	0.18	0.14
N_2	0.60	-0.22	0.53	-0.22
N_3	0.17	-0.14	0.18	-0.16

Table 5: Behavioral pattern and credibility degree of neighbors determined by C_1 and C_2

Consumer	N_k	$BP_{(C,N_k)}$	$CR_{(C,N_k)}$	θ	σ
C_1	N_1	Optimistic	0.79	0.18	N/A
	N_2	Deceitful	0	N/A	N/A
	N_3	Pessimistic	0.82	N/A	1.08
C_2	N_1	Optimistic	0.82	0.02	N/A
	N_2	Deceitful	0	N/A	N/A
	N_3	Pessimistic	0.76	N/A	0.34

4. EXPERIMENTAL RESULTS

Our approach models the trustworthiness of advisers, not only based on their honesty degree but also by examining their competency level. That is, an honest adviser but with insufficient experiences is not qualified to provide advice. Furthermore, we claim that having a good comprehension of the adviser's disposition leads to a more adaptive credibility assessment. For this purpose, we have conducted two classes of experiments. The first class is designed to indicate the effectiveness of the proposed model in detecting malicious neighbors as well as newcomers with insufficient experiences. In the second class of experiments we put the second layer to the test and observe how the *same* advisers could have *different* credibility values according to different consumers. We also estimate the accuracy of our prediction by comparing it with the actual trustworthiness value of advisers, obtained by averaging over multiple experiences.

The first series of experiments evaluates the competency level of an intrinsically *honest* neighbor having different numbers of experiences. It involves one consumer C asking a neighbor N_k about its common experiences with 2 and 50 providers. N_k provides percentages (ranging from 0% to 100%) indicating the level of difference between the number of experiences for C and the number of its own experiences. The results indicate that the competency of even an honest neighbor degrades as its number of experiences decreases (Figure 1). We also observe that C can effectively evaluate the competency level of advisers even with a limited set of providers. Figure 2 illustrates the experiment in which N_k provides different percentages (0% to 100%) of unfair ratings. Given similar conditions as in Figure 1, we observe that as the number of unfair ratings increases, the competency level of N_k decreases. It also indicates that the competency level of N_k drops more significantly if it provides unfair ratings (Figure 2) in comparison with the situation where it has insufficient ratings(Figure 1). Note that, in both experiments, it is noticeable that C can effectively evaluate the competency level of N_k with a few providers - e.g., 2.

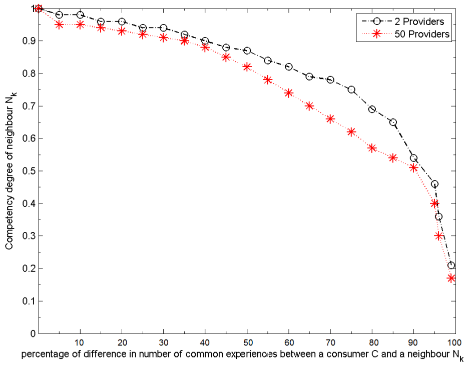

Figure 1: The competency degradation of N_k having different percentages of common experiences

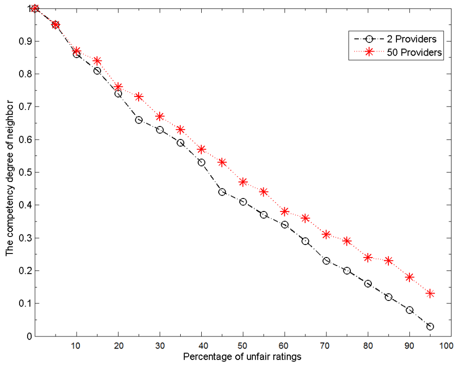

Figure 2: The competency degradation of N_k having different percentages of unfair ratings

The next class of experiment targets the second layer of the filtering algorithm. It involves 80 providers, 4 advisers and 2 consumers. The consumers and the advisers rate 50 randomly selected providers. We assume that the advisers have passed the first layer and are qualified to play the role of advisers. We model the credibility ratings the consumers have of participating advisers and compare them with the actual credibility value of advisers. More explicitly, we examine how the consumers C_1 and C_2 with different dispositions (risk-averse and risk-taking, respectively) evaluate the set of advisers A_1 , A_2 , A_3 and A_4. Note that A_1 and A_3 have a tendency to overrate the providers while A_2 and A_4 have a tendency to underrate the providers. These advisers have different credibility values from 0.0 to 1.0. Also, in order to examine the effect of the recency factor in prediction of the trust value, we assume that A_1 and A_2 provide ratings in the same window with consumers ($T_A - T_C = 0$) while the other advisers provide ratings in different time windows, differing by at most 3 time intervals ($T_A - T_C \leq 3$). Figures 3 and 4 illustrate the trustworthiness of advisers predicted by C_1 and C_2, respectively. Adjusting the threshold values and the forgetting parameter to $\beta = 0.25$, $\epsilon = 0.75$ and $\lambda = 0.9$, we can observe how C_1 and C_2 evaluate the credibility of advisers differently.

As shown in Figure 3, C_1 identifies the behavioral model of advisers and evaluates their credibility adaptively. Results indicate that C_1 assigns higher credibility to the pessimistic adviser A_4 (with $T_{A_4} - T_{C_1} = 3$) when compared with the optimistic adviser A_1 (with $T_{A_1} - T_{C_1} = 0$). Similarly, C_2 considers the old opinion of the optimistic Adviser A_3 more valuable than a recent opinion of pessimistic adviser A_2 (Figure 4).

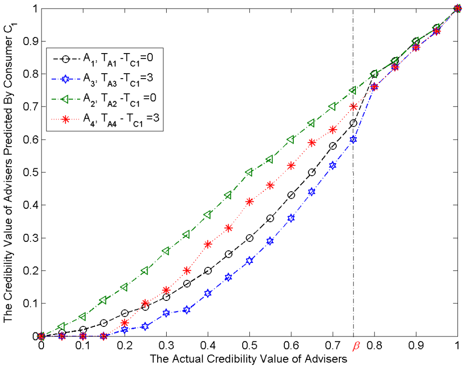

Figure 3: The predicted credibility of advisers by C_1 in comparison with their actual credibility

Table 6 measures the deviation (i.e., Mean-Square-Error and Mean-Absolute Percentage error) between advisers' actual and predicted credibility values determined by C_1 and C_2 across different values of $T_A - T_C$.

To examine how C_1 and C_2 adaptively calculate the coefficients θ and σ, Table 7 depicts the values of these coefficients across various percentages of advisers' dishonesty. That is, the advisers provide different percentages (0% to 100%) of unfair ratings. We observe that consumer agents with different characteristics take different approaches in computing such coefficients, resulting in different evaluations of the credibility degrees of the same advisers.

The final experiment examines the effect of the recency factor $T_{(C,A_k)_{P_j}}$ in evaluating the credibility of advisers.

Table 6: Calculating the error parameters for C_1 and C_2 having various time difference

Agent	Adviser's Pattern	Erorr	$T_A - T_C = 0$	$T_A - T_C = 1$	$T_A - T_C = 2$	$T_A - T_C = 3$	$T_A - T_C = 4$	$T_A - T_C = 5$
C_1	*Optimistic Adviser*	**MSE**	0.06	0.068	0.07	0.079	0.086	0.092
		MAPE	1.43%	1.56%	1.72%	1.58%	2.00%	2.15%
	Pessimistic Adviser	**MSE**	0.048	0.049	0.052	0.057	0.06	0.068
		MAPE	0.58%	0.80%	0.99%	1.22%	1.39%	1.59%
C_2	*Optimistic Adviser*	**MSE**	0.048	0.049	0.052	0.059	0.061	0.067
		MAPE	0.58%	0.8%	0.99%	1.19%	1.39%	1.59%
	Pessimistic Adviser	**MSE**	0.063	0.068	0.073	0.079	0.085	0.092
		MAPE	1.42%	1.57%	1.72%	1.86%	1.99%	2.14%

Table 7: The coefficients parameters calculated by consumers C_1 and C_2

$Buyer'sDisposition$	Coefficient	Percentage of Unfair Ratings										
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Risk-Averse Consumer C_1	θ	0	0.36	0.46	0.56	0.66	0.76	0.86	0.96	1.06	1.16	1.26
	σ	0	0.001	0.002	0.04	0.14	0.24	0.34	0.44	0.54	0.64	0.74
Risk-Taking Consumer C_2	θ	0	0.001	0.002	0.04	0.14	0.24	0.34	0.44	0.54	0.64	0.74
	σ	0	0.36	0.46	0.56	0.66	0.76	0.86	0.96	1.06	1.16	1.26

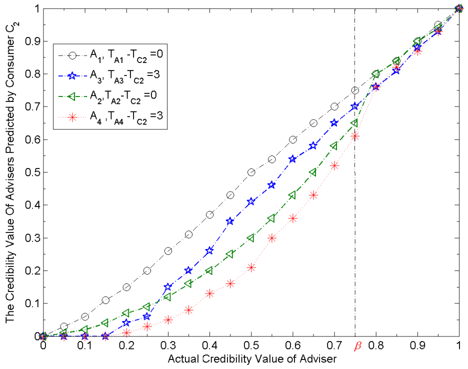

Figure 4: The predicted credibility of advisers by C_2 in comparison with their actual credibility

That is, we define a consumer C and adviser A regardless of the behavioral patterns. We also assume that A has successfully passed the first layer. Adjusting $\beta = 1$ and $\epsilon = 0$, we observe that A with $CR_{(A)} = 0.95$ loses its credibility as the differences between their time window $(T_A - T_C)$ increases. Figure 5 illustrates this by initializing the forgetting factor λ from 0.0 to 1.0.

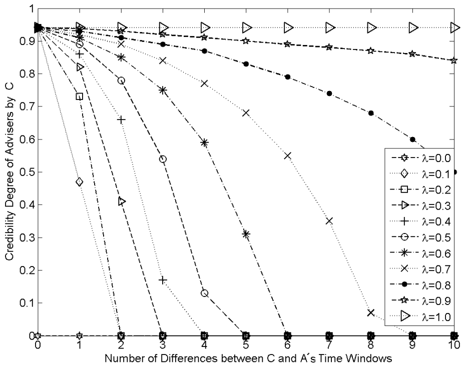

Figure 5: Aging the credibility value of A as time passes

5. RELATED WORK

Several reputation systems and mechanisms have been proposed for modeling the trustworthiness of advisers and coping with the problem of unfair ratings in multi-agent online environments.

In the beta reputation system (BRS) proposed by Jøsang and Ismail [9], which is based on a beta distribution, the agents can only provide binary ratings for each other. He further extends the proposed BRS to adopt a multinomial rating model that computes reputation scores by statistically updating the Dirichlet Probability Density Function (PDF) [10, 7]. In this context, participating agents are allowed to rate each other within any level from a set of predefined rating levels. To handle unfair feedback provided by adviser agents, Whitby et al.[13] use the endogenous discounting method to exclude advisers whose probability distributions of ratings significantly deviate from the overall reputation scores of the target agent. That is, it dynamically determines upper and lower bound thresholds in order to adjust the iterated filtering algorithm's sensitivity tailored to different environmental circumstances. For instance, if the majority of participants act deceitfully in the environment, the lower bound would be set to a higher value so as to increase the sensitivity of the BRS, which can lead to the exclusion of more unfair raters.

Teacy et al. [11] proposed TRAVOS, which is a probabilistic trust and reputation system for agent-based virtual organizations. To derive a measure of trust, this model relies heavily on its direct experiences and refuses to combine others' opinions unless it is not confident about the adequacy of its personal experiences. In such conditions, advisers share the history of their interactions in a tuple that contains the frequency of successful and unsuccessful interaction results. To evaluate the credibility of advisers, it uses a beta distribution and calculates the probability that a particular adviser provides accurate reports given its past opinions and proportionately adjusts the influence of its current observation afterwards.

PeerTrust [14] is a coherent dynamic trust model for peer-to-peer e-commerce communities. To evaluate the quality of the feedback provider, it proposes a *personalized similarity measures* mechanism to compute a feedback similarity rate between the evaluating peer and advising peer over a common set of peers with whom they have had previous interactions. Particularly, this model calculates the root-mean-error or standard deviation of the two feedback vectors to compute the feedback similarity. Through this principle, the evaluating peer discounts the future feedback released by feedback providers.

Yu and Singh[15] have proposed a decentralized reputation management model to locate the rightful advisers in multi-agent systems. In fact, one of the major concerns of this model is detecting malicious agents who deliberately disseminate misinformation through a network. The proposed model considers three types of deceptions: *complementary*, *exaggerative positive* and *exaggerative negative*. It defines an exaggeration coefficient to differentiate between exaggerative and complementary deceptive agents. This model uses the same credibility measure to calculate the trustworthiness of different kinds of advisers by considering how much their ratings deviate from the actual value experienced by a consumer agent. Note that, in this model, all the advisers have an initial credibility of 1 and as a consumer agent interacts with more provider agents, its credibility will be updated.

Zhang and Cohen [16] proposed a personalized approach for handling unfair ratings in centralized reputation systems. It provides a public and private reputation approach to evaluate the trustworthiness of advisers. In this model, advisers share their subjective opinions over a common set of providers. To estimate the credibility of advisers, it exploits a probabilistic approach and calculates the expected value of advisers' trustworthiness based on their provided ratings.

Our work differs in a number of ways. Unlike other models, which mainly evaluate the credibility of advisers based on the percentage of unfair ratings they provided, this model takes the steps to aggregate several parameters in deriving the trustworthiness of advisers. That is, in addition to the similarity degree of advisers' opinions, we aggregate their behavioral characteristics and evaluate the adequacy of their reputation information in our credibility measure. In this model, every consumer with different behavioral characteristics is able to objectively evaluate the similarity degree of advisers through a multi-criterion rating approach. Also, consumer agents could adaptively predict the trustworthiness of advisers using different credibility measures well-suited for various kinds of advisers.

6. CONCLUSION AND FUTURE WORK

In this paper, we propose a two-layered filtering algorithm that cognitively elicits the behavioral characteristics of the participating agents in an e-marketplace. The principles of the two-layer filtering algorithm mainly target malicious agents with complementary rating patterns, agents with insufficient experiences and fraudulent participants who retain a minimum level of trust to cheat opportunistically.

In the first layer, consumer agents take a probabilistic approach and narrow a circle of neighbors by expelling those with significant deceptive patterns, as well as those with an inadequate number of experiences. The basis of the second layer provides mechanisms to cognitively derive the actual intentions of the surviving agents of the previous layer. Here, consumer agents conduct additional evaluations and objectively estimate the similarity degree of advisers through a multi-criterion rating model. Thereafter, they classify their behavioral characteristics based upon their own attitudes. Our model articulates that consumers could have different credibility degrees for the same advisers. Also, it enables consumer agents to include more participants as advisers through a variety of credibility assessment measures. This matter is mostly practical in an environment where the majority of participants are unfair. In order to articulate the effectiveness of our approach in dealing with a community where a majority of participants are unfair, in future work, we will conduct extensive experiments to compare our model with others in identifying honest participants in such situations. Another avenue for future work is to propose a mechanism to dynamically adjust the presented thresholds of the layers based on the environmental conditions and the quality of participants.

7. REFERENCES

[1] K. Barber, Karen Fullam, and Joonoo Kim. Challenges for trust, fraud and deception research in multi-agent systems. In *Trust, Reputation, and Security: Theories and Practice*, volume 2631, pages 167–174. 2003.

[2] Cristiano Castelfranchi, Rino Falcone, and Michele Piunti. Agents with anticipatory behaviors: To be cautious in a risky environment. In *ECAI*, 2006.

[3] Rino Falcone and Cristiano Castelfranchi. Generalizing trust: Inferencing trustworthiness from categories. In *AAMAS-TRUST*, pages 65–80, 2008.

[4] Chung-Wei Hang, Yonghong Wang, and Munindar P. Singh. An adaptive probabilistic trust model and its evaluation. In *AAMAS (3)*, pages 1485–1488, 2008.

[5] T. D Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[6] Reid Kerr and Robin Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *AAMAS (2)*, pages 993–1000, 2009.

[7] Zeinab Noorian and Mihaela Ulieru. The state of the art in trust and reputation systems:a framework for comparison. *J. Theor. Appl. Electron. Commer.*, 2010.

[8] Marsh S. Optimism and pessimism in trust. *Proceedings of the Ibero-American Conference on Artificial Inteligence, McGraw-Hill*, 1994.

[9] Audun Jøsang and Roslan Ismail. The Beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[10] Audun Jøsang and Walter Quattrociocchi. Advanced features in bayesian reputation systems. In *TrustBus*, pages 105–114, 2009.

[11] W. T. L Teacy, J Patel, N. R. Jennings, and M. Luck. TRAVOS: Trust and reputation in the context of inaccurate information sources. *Journal of Autonomous Agents and Multi-Agent Systems*, 12(2), 2006.

[12] Yonghong Wang and Munindar P. Singh. Formal trust model for multiagent systems. In *IJCAI*, 2007.

[13] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of 7th International Workshop on Trust in Agent Societies*, 2004.

[14] Li Xiong and Ling Liu. PeerTrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):843–857, 2004.

[15] Bin Yu and Munindar P. Singh. Detecting deception in reputation management. In *AAMAS '03: Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 73–80, New York, NY, USA, 2003. ACM.

[16] Jie Zhang and Robin Cohen. Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. *Electronic Commerce Research and Applications*, 7(3), 2008.