Applying Collaborative Filtering to Reputation Domain: a model for more precise reputation estimates in case of changing behavior by rated participants

Alexandre Lopes, Ana Cristina Bicharra Garcia

Instituto de Computação - Universidade Federal Fluminense (UFF)

Rua Passo da Pátria, 156 Bloco E Sala 350 - 24.210-240 – Niterói – RJ.

alexcflopes@gmail.com, bicharra@ic.uff.br

Abstract. Automated Collaborative Filtering (CF) techniques have been successfully applied on Recommendation domains. Dellarocas [1] proposes their use on reputation domains to provide more reliable and personalized reputation estimates. Despite being solved by recommendation field researches (e.g. significance weighting [2]), the problem of selecting low-trusted neighborhoods finds new roots in the reputation domain, mostly related to different behavior by the evaluated participants. It can turn evaluators with similar tastes into distant ones, contributing to poor reputation rates. A Reputation Model is proposed to minimize those problems. It uses CF techniques adjusted with the following improvements: 1) information of evaluators taste profiles is added to the user evaluation history; 2) transformations are applied on user evaluation history based on the similarities between the taste profiles of the active user and of the other evaluators to identify more reliable neighborhoods. An experiment is implemented through a simulated electronic marketplace where buyers choose sellers based on reputation estimates generated by the proposed reputation model and by a model that uses traditional CF. The goal is to compare the proposed model performance with the traditional one through comparative analysis of the data that is created. The results are explained at the end of the paper.

1 Introduction

The goal of online reputation reporting systems and models applied in e-commerce systems [1, 3] is to restrain the participation of agents who have a poor-quality service history in electronic marketplaces. These models can combine both direct and indirect [4] information sources to better estimate their participants' reputation. The direct sources contain information on past encounters between the client and the rated

supplier. The indirect one contain information indirectly acquired through other clients' witnesses [1, 3, 5] or through the analysis of their social relationship network, which is kept by the suppliers [4]. Even though reputation models based on direct sources are considered to be the most reliable way of estimating the supplier's reputation, there is a higher amount of available information through indirect sources, which should be used in case there's little probability of any two participants having a history of past encounters [4]. The problem with models based on indirect sources is the possibility that the reputation estimate will not be as reliable as desired, because it is difficult to measure precisely subjective aspects like quality of service. The reputation estimate is based on aggregates that do not reflect the differences in the client's taste or the context of the interaction [1, 3]. If it is poorly calculated, the reputation estimate can cause clients to interact with suppliers who they wouldn't choose to transact with otherwise.

1.1   Automated Collaborative Filtering applied to Reputation Models

For more reliable calculation of reputation estimates, Dellarocas [1], proposes the incorporation of the Automated Collaborative Filtering (CF) technique to the reputation model. The CF has been used in Recommendation and Information Filtering Systems [6]. It identifies similarities between an active user and other users based on the similarities in past ratings on common items, and uses this similarity to generate recommendations about items not yet rated by the active user. The neighbor users are the ones who have higher similarity factor with the active user.

   The goal of incorporating the CF to reputation domains is to estimate the reputation of a supplier in a personalized way, calculating it based on the ratings of clients that have similar tastes with those of the active client.

1.2   The "False Good Neighbor" Problem

This problem is described in the Recommendation Systems literature as being a situation in which clients calculated as having greatest similarity with the active client are, in fact, not that similar [2, 7]. It can happen because of coincident ratings and because of a low number of common ratings between the clients. The very work of Herlocker [2] already proposes solutions for this issue, however, there are other factors in the reputation domain that can contribute to situations of false neighborhood, such as changes in the supplier's behavior from one encounter to another and rating manipulation by clients with bad intentions. In this paper we intend to explore the changes in the supplier's behavior. Suppliers who change their behavior from one encounter (and rating) to another may cause deviations in the similarity detection between clients, causing clients with similar preferences not to be considered neighbors, or inversely, causing clients with different preferences to be considered neighbors. It is important to notice that these problems do not happen in conventional Recommendation Systems because the rated items are products and not rational entities (humans or computers). Products don't have behavior, their characteristics and looks are normally maintained after each rating, while a supplier has goals, which influence his actions and behavior [4]. Apart from timely aspects, if

two clients disagree on a product's rating, it is safe to say that it's because they have different tastes. In the case of supplier's rating, if they disagree on a rating, there are no guarantees that the suppliers maintain the same behavior with each client.

2   Proposed Model

This paper proposes a way of minimizing this problem adjusting the CF technique to calculate more precisely the similarity between users, considering, besides the ratings they enter, their preferences, which are presented through reputation rating issues (price, quality, etc.), that have been used as base for the rating. The model also aims to be applied on application domains as Electronic Marketplaces and partially decentralized P2P information sharing systems.

2.1   Automated Collaborative Filtering applied to Reputation Models

An example of a matrix of ratings in the reputation domain is presented in table 1.

Table 1. Matrix of ratings in a reputation system

Item rated / Rater	Client 1	Client 2
Supplier X	3.8	3.8
Supplier Y	1.1	3

To minimize the negative effects of a supplier's behavior change on the reliability of the recommendations, we propose a rating history adjustment based on taste similarity between the clients, which general scheme is presented in the following algorithm.

2.2   Proposed model's high level algorithm:

1. The active client chooses one of two operational modes [8]: the Prediction Mode – in which you estimate the reputation value of a supplier with whom to interact; or the Recommendation Mode – in which you generate a list of recommendations sorted by the highest reputation values of the estimated suppliers; in the first case, a premise is that the client can use any mediation resource available to locate suppliers and negotiate with them. In the second case, choosing the supplier is made based on the recommendation list generated.
2. The Reputation Service recovers all common rating history between the active client and other clients (neighbors), creating an Active Client Matrix of Ratings;
3. Before calculating the similarity with each neighbor, and to minimize the problems of changing behavior, the Reputation Service adjusts the Active Client Matrix of Ratings, with each cell being recalculated as described in the "Matrix of Ratings Transformation" section;

4. It is applied the traditional CF algorithm over the transformed matrix, which will generate the chosen supplier's predictive reputation value, or a recommendation list.
5. The client decides if he/she will start a transaction according to the recommendations or predictions generated.
6. At the end of the transaction, the client writes a testimonial on the supplier's reputation. Besides the general reputation, every testimonial must have the reputation values rated for each rating issue and the preferences of the client (more details in the following section).

2.3   Matrix of Ratings Transformation and reputation calculation

Extended Matrix of Ratings

In order for the transformation indicated in step 3 to take place, it is necessary to work with an Extended Matrix of Ratings (table 3) that contemplates, besides the General Reputation Ratings, the client's tastes and the reputation values given by rating issue (contract clause). The client's tastes are represented in the model as "Reputation Preferences", which are data structures inspired in the Behavioural Aspects φ and Ontological Structures defined in the ReGreT [9] model. The conceptual representation of the Extended Matrix of Ratings, as well as of the Reputation Preferences, can be seen in the class diagram, Fig.1, and described below.

Each and every Reputation Rating is associated to a Contract and to the Client's current Reputation Preference. The Reputation Preferences change as time goes by and are used in the supplier's reputation rating task. They are composed of Rating Issues. Each contract clause is related to a Rating Issue that has a weight and a rating formula. The Rating Issues indicate how deviations of the final values in relation to the agreed values influence (negatively or positively) the Behavioural Aspect. In this sense, they have similar function to the Ground Relations defined in the ReGreT model. Such influence must be calculated through a domain-dependent formula that is described in the Expression attribute of the RatingFormula class. The proposed model shares the ReGreT's premise that reputation is a complex concept (Complex Behavioural Aspects) rated through the combination of various simpler rating dimensions (Simple Behavioural Aspects), table 2. Thus, every Reputation Preference is made up of various Rating Issues, which, combined with its weight, would determine the general reputation value.

Table 2. Example of Rating Issues combined into a Reputation Preference

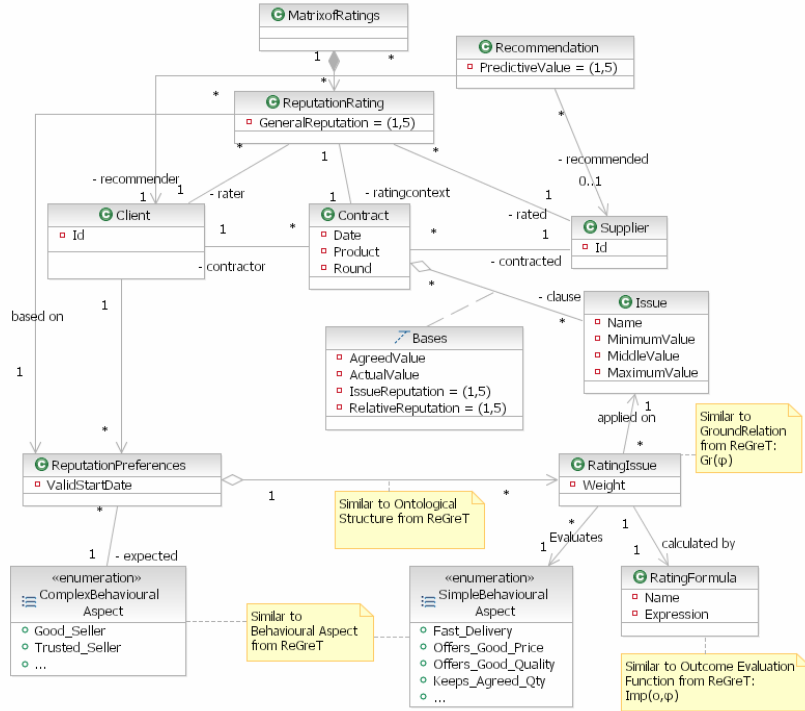| Reputation Preferences (Complex Behavioural Aspect)	Rating Issue (Simple Behavioural Aspect)	Issue	Weight	Influence
Good-seller	Offers_High_Price	Price	0,6	Negative
	Offers_Good_Quality	Quality	0,4	Positive

Fig. 1. Complete conceptual model (UML 2.0 notation)

Table 3 is an Extended Matrix of Ratings from table 1. The reputation preferences are in the "weight" line, while the reputation values given by rating issues are in the "Rating" line. The relative rating is the product of the weight of the issue and the reputation rating of the issue.

Table 3. Example of Extended Matrix of Ratings

	Rater	Client 1				Client 2			
Rated	Issue	Quality	Price	Date	General	Quality	Price	Date	General
Supplier X	Weight	0.3	0.2	0.5		0.2	0.3	0.5	
	Rating	5	4	3		4	5	3	
	Relative Rating	1.5	0.8	1.5	3.8	0.8	1.5	1.5	3.8
Supplier Y	Weight	0.3	0.2	0.5		0.3	0.2	0.5	
	Rating	3	1	0		4	4	2	
	Relative Rating	0.9	0.2	0	1.1	1.2	0.8	1	3

2.4 How to perform the adjustments over the Extended Matrix?

The equations (1) (2) and (3) implement the necessary calculations for reputation rating as well as the Matrix of Ratings adjustments shown in step 3 of the proposed model's high level algorithm.

$$(1) \qquad Ar_{neighbor,supplier} = R_{activeclient,supplier} * sf_{activeclient,neighbor,supplier}$$

$$(2) \quad R_{activeclient,supplier} = \frac{\sum_{issue} (rt_{supplier,issue} * w_{activeclient,issue})}{\sum_{issue} (w_{activeclient,issue})}$$

$$(3) \quad sf_{activeclient,neighbor,supplier} = \\ cosine (preferences_{activeclient,supplier} , preferences_{neighbor,supplier})$$

Where,

Ar$_{neighbor,supplier}$ – is the reputation value of each Matrix of Ratings cell, adjusted accordingly to the similarity factor between the Active Client and the Neighbor. The examples in tables 4 and 5 illustrate how the Matrix of Ratings transformation takes place.

R$_{activeclient,supplier}$ - is the supplier's reputation according to the perspective of the active client. It is the pondered average of the active client's reputation rating on every issue of the contract and not in only one reputation value, as it happens in other systems like eBay (www.ebay.com). The calculation formula is independent of the application domain and the quantity of issues of the negotiated contracts, and always result of growing scale values of real numbers between 1 and 5.

rt$_{supplier,issue}$ - is the supplier's reputation accordingly to a determined issue of the contract. The calculation format depends on the application domain and the rating of the Behavioural Aspect (examples described in the "Experiment" section), but it must produce values between 1 and 5 so as to not compromise the supplier's reputation calculation (eq. 2).

w $_{activeclient,issue}$ – is the weight and the importance given by the active client to the issue. The weight is a real number between 0 and 1 given by the client, which may vary as time goes by. The sum of the issue's weights must always total 1.

sf$_{activeclient,neighbor,supplier}$ - represents the similarity factor between the active client and a determined neighbor. It is determined through a cosine function, which calculates the distance between the rating issues' weight vectors, and so, identifying similarities in the reputation preferences applied by the active client and its neighbors when rating a common supplier's reputation.

2.5   Example of the Matrix of Ratings transformation

Considering Client 1 as being the active client, and applying the equations (1) (2) and (3) on a Matrix of Ratings from table 3 (and simplified table 4), we have the results in the Adjusted Matrix from table 5. This Matrix is used as input to step 4 of proposed model's high level algorithm.

Table 4. Active client 1's Matrix of Ratings

	Rater	
Rated	Client 1	Client 2
Supplier X	3.8 (R)	3.8
Supplier Y	1.1 (R)	3

Table 5. Adjusted Matrix of Ratings

	Rater	
Rated	Client 1	Client 2
Supplier X	3.8	3.7 (Ar)
Supplier Y	1.1	1.1 (Ar)

3   Experiment

The experiment's goal is to prove that the proposed model minimizes the effects on the seller's behavior change. It was assembled so it could be possible to compare the performance of the proposed model accordingly to a model based on traditional CF technique, through a system that simulates an e-commerce product marketplace, and that registers the effects of the rated seller's changing behavior. The experiment was inspired in the work conducted in [10], incorporating its organization and way of measuring the performance of the tested models. However, the system's architecture is different because the proposed reputation model is based exclusively on witnesses recovery by collaborative filtering mechanisms.

The simulation is made of 16 buyers and 64 sellers. Half of the buyers receive seller's recommendations based on the proposed CF algorithm, and the other half, based on the traditional algorithm. The transactions occur in 64 rounds, being 51 training rounds and 13 test rounds. The goal of the training rounds is to prepare the reputation database and to prevent the low dispersion of ratings from damaging the tested CF algorithm performance. During these rounds, the reputation module does not provide recommendation (sellers are randomly selected), it is only fed by the buyer's ratings. The test rounds are for monitoring buyer performance and for complementing the comparative analysis between the tested reputation models. Buyers select sellers with the highest reputation prediction value in the recommendation lists generated by the reputation module. In each round, it is possible to occur up to 16 transactions, totaling 1024 transactions per simulation. The maximum number each buyer can close per round is one transaction. Each seller can participate in one or more transactions per round.

The system has modules that simulate buyers and sellers, and a reputation service capable of generating recommendation based in the traditional CF technique and in the proposed technique. The reputation service implements two CF algorithms with user-to-user correlation developed from the algorithm originally proposed by Resnick in [6]. They share common configurations, like similarity calculation between neighbors through Pearson's coefficient, and the usage of a neighborhood selection method by a maximum amount of neighbors (best-n-neighbors) [2, 7] (configured to 30 neighbors).

When starting a transaction, both buyer and seller agree on the price and the quality of the product. The initial agreement is established based on the middle values (30,00 for price and 3 for quality). During the transaction, the seller can change the agreed values with the buyer according to his/her behavior, what will influence in the outcome of the transaction. There are three types of behavior:

Bad – During training rounds, in 60% of the transactions they increase the initially agreed price in ¼ and decrease quality in ¼. In the rest of the transactions, they present a similar behavior than the Good one. The originally defined percentage in the experiment with ReGreT was 75%, however, in this case, it makes more sense to define it as 60% so as to make the behavior changes more frequent. During test rounds, they increase price and decrease quality in 100% of the rounds. The sellers are configured like this to reproduce the false neighborhood situations in which other buyers with similar preferences are not considered neighbors, and vice-versa.

Good - Along all rounds, this type of seller increases the quality of product in ¼, and decreases its price in ¼. They are configured this way so as to benefit buyers who effectively receive the best recommendations, and then, represent a counterpoint relative to the disappointment of a buyer in case he/she receives a bad recommendation.

Neutral – There are no changes on the agreement of the contract in any of the rounds.

When closing a transaction, the buyers rate the seller's reputation, and update their cash. The seller's reputation is rated in a growing scale of real numbers which go from 1 to 5, being calculated as the ratings' pondered average of the issues of the product, as shown in the simplified equation (1):

$$(1)\ R_{b,s} = \frac{(rt_{s,qual} * w_{b,qual}) + (rt_{s,price} * w_{b,price})}{(w_{b,qual} + w_{b,price})}$$

The weight w is chosen randomly to each buyer in the beginning of the simulation, and remains unchanged during the whole simulation.

The Rating Issue formula specific in this domain are:

Quality Issue: (4) $rt_{s,qual} = fv_{s,qual}$ **Price Issue:** (5) $rt_{s,price} = 6 - (fv_{s,price}/10)$

where $\mathbf{fv_{s,price}}$ represents the final selling price, and $\mathbf{fv_{s,qual}}$ the final quality value.

Each buyer initiates the simulation with 5000,00 in cash, which are updated at the end of each transaction, as shown in equations 6 and 7:

$$(6)\ cb_t = cb_{t-1} - fv_{s,price} + rp$$

where ***cb*** represents the buyer's cash (cb_t the current round cash and cb_{t-1} the previous round cash) and ***rp*** the resale price. It was defined that, in test rounds, the buyers should resell the acquired products, with the resale price being determined by the quality of the acquired product: (7) $rp = fv_{s,qual} * 10$

The resale is lucrative every time the buyer transacts with a seller who has good behavior, and prejudicial every time he transacts with a seller who has bad behavior. With this premise, there is a performance comparison between the two tested reputation models. The buyers are separated in two groups: the ones who select sellers based on the given recommendations according to the proposed reputation model (ACF1); and the ones who select their partners based on the recommendations generated by the traditional CF technique (ACF). At the end

of each round the average cash value is collected between the buyers by group, and after the closing of the simulation, the distribution of averages by round can be analyzed so as to verify if the proposed method is better than the traditional one and if this difference is statistically significant.

3.1 Performed tests

Several test scenarios were anticipated in the simulation, representing different proportions of buyers according to their behavior. We applied statistical analysis in all of the scenarios (the "t" test, the unilateral, for independent samples, using significance level α of 5%). Each sample refers to one of the buyer's groups, and is formed by the distribution of averages of cash per test round. We tested the following hypothesis:

H_0 – The performance of the proposed reputation model (called ACF1) is like the one in the model that uses the traditional CF technique (called ACF) in scenarios where buyers will change their behavior from one rating to another, noted as:

$$H_0: \mu_{acf1} = \mu_{acf}$$

H_1 - The performance of the proposed recommendation method tends to be better than the originally proposed CF method, in scenarios where the buyers will change their behavior from one rating to another:

$$H_1: \mu_{acf1} > \mu_{acf}$$

The performance of the proposed model (ACF1) was superior to the traditional model (CF) in every test, being statistically significant in 5 of the 7 tested scenarios. The average of buyer's cash who used ACF1 was superior after the execution rounds on every test scenario executed. We present a summary of the results in table 6 and in Fig. 2.

Table 6. Summary of test results

	Summary of results						
	Scenario I	Scenario II	Scenario III	Scenario IV	Scenario V	Scenario VI	Scenario VII
Bad	40%	50%	50%	55%	60%	65%	75%
Good	50%	40%	50%	35%	30%	25%	25%
Neutral	10%	10%	0%	10%	10%	10%	0%
Best performance	ACF1	ACF1	ACF1	ACF1	ACF1	ACF1	ACF1
Significance probability	0.005	0.004	0.007	0.008	0.018	0.057	0.139

As the number of Bad buyers increase, the significance probability decreases. Even with the ACF1 model keeping better performance in relation to ACF, the difference between both is no longer statistically significant as the proportion of Bad Buyers is higher or equal to 65%.

Due to space constraints, we show only one of the comparative performance graphics of the test scenarios.
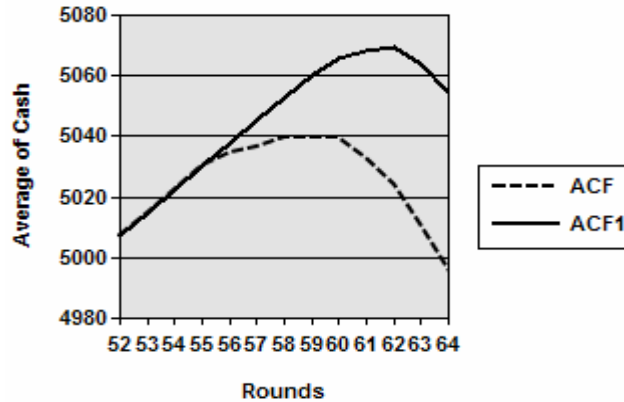
Fig. 2. 40% Bad, 50% Good and 10% Neutral Scenario

4   Conclusions

In this paper, we propose a model so that personalized reputation ratings based on CF can work adequately in case of changing behavior from the rated participants. The good results obtained in described simulations allow us to continue this work, performing field experiments so as to ratify the preliminarily results obtained.

5   References

1.   C. Dellarocas, Immunizing Online Reputation Reporting Systems Against Unfair Ratings and Discriminatory Behavior, in: Proceedings of the Conference on Electronic Commerce, (Minneapolis, Minnesota, 2000), pp. 150-157.
2.   J.L. Herlocker, J.A. Konstan, A. Borchers and J. Riedl, An algorithm framework for performing collaborative filtering, in: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, (1999).
3.   P. Resnick, R. Zeckhauser, E. Friedman and K. Kuwabara, Reputation Systems, *COMMUNICATIONS OF THE ACM*, **43**(12), 45-48 (2000).
4.   J. Sabater and C. Sierra, Reputation and Social Network Analysis in Multi-Agent Systems, in: AAMAS´02, (Bologna, Italy, 2002), pp. 475-482.
5.   T.D. Huynh, N.R. Jennings and N.R. Shadbolt, On Handling Inaccurate Witness Reports, in: 8th Int. Workshop on Trust in Agent Societies, (Utrecht, The Netherlands, 2005), pp. 63-77.
6.   P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom and J. Riedl, GroupLens: An Open Architecture for Collaborative Filtering of Netnews, in: CSCW, (1994), pp. 175-186.
7.   R. Torres, *Personalização na Internet* (Novatec Editora, São Paulo, 2004).
8.   S.T.K. Lam and J. Riedl, Shilling Recommender Systems for Fun and Profit, in: Proceedings of the 13th international conference on World Wide Web, (2004), pp. 393-402.
9.   J. Sabater. Trust and reputation for agent societies *Institut d'Investigació en Intel.ligència Artificial*, Consell Superior d'Investigacions Científiques, Catalonia, Spain. (2003).
10.   J. Sabater, Evaluating the ReGreT system, Applied Artificial Intelligence, **18**(9/10), 797-813 (2004).