Does Trust Matter for User Preferences? A Study on Epinions Ratings

Georgios Pitsilis, Pern Hui Chia

Does Trust Matter for User Preferences? A study on Epinions ratings

Georgios Pitsilis[†], Pern Hui Chia

Q2S[*] NTNU
O. S. Bragstads plass 2E
Trondheim 7491 Norway

{pitsilis, chia}@q2s.ntnu.no

Abstract. Recommender systems have evolved during the last few years into useful online tools for assisting the daily e-commerce activities. The majority of recommender systems predict user preferences relating users with similar taste. Prior research has shown that trust networks improve the performance of recommender systems, predominantly using algorithms devised by individual researchers. In this work, omitting any specific trust inference algorithm, we investigate how useful it might be if explicit trust relationships (expressed by users for others) are used to select the best neighbours (or predictors), for the provision of accurate recommendations. We conducted our experiments using data from *Epinions.com*[1], a popular recommender system. Our analysis indicates that trust information can be helpful to provide a slight performance gain in a few cases especially when it comes to the less active users.

Keywords: Trust, Epinions, Recommender system

1 Introduction

Reputation systems compute global scores about products (people, companies, etc) based on opinions that users hold about them and assist prospective users in deciding whether to buy these products. Different from reputation systems that provide global scores, *recommender systems* provide personalized (local) recommendations based on correlations of ratings (browsing history, search keywords or other actions) made by likeminded users. Recommendations are generated automatically to assist users to choose from multiple options available on the Internet. Amazon.com and Youtube.com, for example, correlate the users' browsing history to determine the

[†] The first author carried out this work during the tenure of an ERCIM "Alain Bensoussan" Fellowship program.

[*] Centre of Quantifiable Quality of Service in Communication Systems (Q2S), Centre of Excellence, appointed by The Research Council of Norway, is funded by the Research Council, Norwegian University of Science and Technology (NTNU) and UNINETT. http://www.q2s.ntnu.no

[1] Epinions.com – an online consumer review site: http://www.epinions.com

similarity between users. In order to learn from similar users, recommender systems employ filtering techniques such as *Collaborative filtering* to identify influential users with similar behaviour and label them as "neighbours" or "predictors".

Although involving more predictors (users whose ratings are taken into account in predictions) may help in improving the prediction accuracy, the number of predictors should be kept low to avoid expensive computation. Herlocker et.al. reported that having too many predictors will conversely reduce the accuracy of predicted recommendation [3]. Rather than trying to guess the magical number of predictors, we believe that much focus could be placed in determining (a small set of) the most suitable predictors.

In classic recommender systems the criteria used for selecting suitable neighbours regard only users' past behaviour and liking (ratings) towards products that are common with other users to compute a similarity value for each user-pair. There are then several strategies to pick the most suitable predictors. These include clustering, correlation thresholding and best-k-neighbours. Clustering organizes the whole user population into groups of similar tastes but in general it is done statically in such a way that predictions are always made using the same set of predictors (or neighbourhood). Correlation thresholding considers only neighbours that are correlated with a particular user over a certain threshold. Meanwhile, the best-k-neighbours technique selects some k best (most similar) neighbours to be considered in the prediction algorithm. The best-k-neighbours technique is also widely known as the *k*-Nearest Neighbourhood (*k*NN) approach, and it has been found that *k*NN outperforms the correlation thresholding approach (in both accuracy and coverage) with reasonable *k* values [3]. Various *k*NN-based algorithms have been proposed for selecting the best predictors [7][8].

In real life, users consult opinions of people whom they trust for forming their own decisions. Such trust relationship is being collected by advanced recommender services. The network of explicitly formed trust relationships is known as web-of-trust [22]. A trust-based recommender system incorporates the web-of-trust into its recommendation algorithm, mimicking the way that people get good advices from trusted sources in real life. A trivial case is to consider only reviews and ratings from sources that have been explicitly indicated as 'trusted' by individual users. Explicit-trust is hence binary. On the other hand, sophisticated systems propagate trust relationship across the user network in order to infer (non-binary) trust values. Several approaches have been proposed to compute implicit-trust values, including Advogato [15], Mole-trust [21] and Subjective logic [14]. Implicit-trust values are particularly helpful when explicit trust information is scarce. For example, if there is a consistent trend that user *a* could provide user *b* with good advices, even though user *a* has not explicitly indicated that she trusts user *b*, it is likely that user *b*'s suggestion would be useful in the future (or considered with a higher weight) and should be implicitly trusted.

Our intuition is that explicit trust information can be exploited to improve the traditional recommender systems in the selection of the most suitable predictors needed in the collaborative filtering. In this work, we investigate whether it is beneficial in the form of improved prediction accuracy when (a small set of) the most suitable predictors are selected with and without explicit trust. If trust helps to select better predictors, combining the classical collaborative filtering with users' personal

assessments (i.e. trust) towards the usefulness of others' recommendations might be of much benefit to improve prediction quality. To enhance comparison, we also present a trust-experience-selection strategy with the intuition that users will be more likely to take into account opinions from trusted users that are more experienced, in real life.

The research contribution and purpose of this work is two fold. First, we investigate the potential benefits of using trust in selecting better predictors with the objective to improve the collaborative filtering scheme in classical recommender system. Second, we investigate the possibility of using usage experience along with trust as a criterion to help selecting suitable predictors.

The rest of the paper is organized as follows. In section 2 we discuss the related work in the field of recommender systems, social networks and trust. Next in section 3, we elaborate on how explicit trust and usage experience could be incorporated to potentially improve the selection of suitable predictors. We describe our experimental setting in section 4 and present the evaluation results in section 5. Finally, in section 6 we discuss our findings before concluding.

2 Related Work

Research related to the *Epinions* dataset includes that of conducted by Massa et. al. [21] that reports interesting findings on "controversial" users who are simultaneously trusted and distrusted by many. They argue that personalized trust metrics are needed given the fact that the controversial users take up to a fraction of 20%. The same authors in [2] address the problem of information overload by exploiting the trust information that users provide explicitly. Even though their concept of making use of the trust graph is quite similar with ours, in their work they used different mechanism/formulae for working out predictions for user items. In our approach we tested various strategies for deciding the best-k neighbours while keeping the similarity-based methodology.

Liu et.al. [6] propose a classification approach for predicting the trust between users from reputation in the absence of first-hand knowledge. Their solution addresses the problem of sparse webs-of-trust by using pre-trained classifiers, but some minimum information, such as user attributes, is required to exist. There is also much research done concerning the topological properties of social networks. We mention that of Wilson et.al. [4] as one of the most recent and complete piece of work.

As for approaches that cluster users into groups of similar tastes, there is a wealth of literature that essentially focused on overcoming the poor prediction quality (e.g. in [9]). Geng et.al. [19] explores the idea of clustering users by imitating the way that people of common interests can be grouped together. Other clustering-based proposals include that by Truong et.al. [17] which uses the common knowledge that exists about the rating behaviour of people for allocating them into clusters of interests. Kwon [7] proposes a technique for selecting the best neighbours as improvement to the *k*-nearest neighbour approach. As opposed to our approach he uses the variance of predictions in a user-specific metric that describes the deviation of the examined user. Other work related to finding best neighbours is that of Lathia

et.al. [16] in which a policy based on trustworthiness is proposed. Contrary to our approach they use implicit trust for finding the best k-trusted neighbours to forming groups of collaborative users. In their idea the trust for some user is derived from the knowledge about her particular ratings. The work in [20] introduced a hybrid Collaborative Filtering System which differs from the standard similarity-based approaches by using weighted similarities computed from the number of common experiences with the predictor.

Sparsity is also recognized as a problem that affects the quality of predictions, and in the past it has been investigated from two different directions. Latent factor analysis, known as *Dimensionality Reduction*, has shown very promising results [18]. However, the simplicity and intuitiveness of *Neighbourhood-based* methods have made themselves more applicable and suitable for social-networking-based models. Trust has also been the subject of investigation by many researchers in the past as a solution for alleviating issues concerning sparsity and security of recommender systems. O'Donovan and Smyth [23] proposed to use implicit trust derived from the reliability of partners as another factor to influence predictions in conjunction with similarity.

To our knowledge, the concepts to use both users' usage experience and explicit trust for building dynamic clusters of suitable predictors have not been explored adequately so far. In the existing solutions no emphasis has been given on the phenomenon of social connectivity in online communities neither on how this could benefit the provision of electronic recommendation services.

3 Neighbourhood Selection Schemes

3.1 Conventional Similarity-based Neighbourhood (S)

Central to most recommender systems that employ collaborative filtering is the computation of similarity between users. Pearson's similarity is the best known formula for user-based recommender systems. It measures the proximity between two users and is computed along the rows (or columns) in the Users by Items matrix. Formula (1) computes the Pearson's similarity $w_{a,b}$ between users *a* and *b* using the set of common items between the two users. The outcome is in the range of [-1,1].

$$w_{a,b} = \frac{\sum_k (r_{a,k} - \bar{r}_a)(r_{b,k} - \bar{r}_b)}{\sqrt{\sum_k (r_{a,k} - \bar{r}_a)^2 \sum_k (r_{b,k} - \bar{r}_b)^2}} \quad \textbf{(1)}$$

$\bar{r}_a$ and $\bar{r}_b$ are the average of all ratings by user *a* and *b* while $r_{a,k}$ and $r_{b,k.}$ are the ratings given by users *a* and *b* respectively for item *k*.

Pearson's similarity is then used in conjunction with Resnick's formula to work out the predicted recommendation [13]. Formula (2) computes the predicted rating $\hat{p}_{a,i}$ of item *i* for user *a* using the set of existing ratings $r_{j,i}$ given to this item *i* by predictor *j*.

$$\hat{p}_{a,i} = \bar{r}_a + \frac{\sum_j \{w_{a,j} \cdot (r_{j,i} - \bar{r}_j)\}}{\sum_j |w_{a,j}|} \quad \textbf{(2)}$$

Resnick's formula (as it is highly sensitive to the number of predictors) does not provide accurate prediction in sparse datasets [3]. The selection of the most suitable predictors is hence of major significance when it comes to the performance of collaborative filtering [3]. Previous research [12] highlighted the importance of selecting the most suitable predictors (for achieving good prediction accuracy) and suggested that only those who are most similar in terms of product ratings should be chosen. We refer this selection scheme as Similarity-based neighbourhood (S), as shown in Figure 1a.

We provide a formal description of the Similarity-based neighbourhood. Let *U* be the set of all users and *I* is the set of all items that have been rated in the system. Let $r_a(i) \neq \perp$ denote that user *a* has given a rating for an item *i* (i.e. not null), the set of ratings $I_a \subset I$ given by a user *a* can then be written as:

$$I_a = \{\forall i \in I : r_a(i) \neq \perp\} \quad \textbf{(3)}$$

We require that only those who have similar rating behaviour with some user *a* and have experienced the item *i* (that user *a* is interested in) be considered. The set of these similar neighbours can be expressed as:

$$S_{a,i} = \{\forall b \in U : b \neq a \wedge | I_b \cap I_a | \geq q \wedge r_b(i) \neq \perp\} \quad \textbf{(4)}$$

with *q* denoting the minimum number of common items that have been rated by both user *a* and her neighbour. Setting such as minimum count is necessary as correlation coefficient (Pearson's similarity value) would not be computable (or not meaningful) unless both users have rated at least some common set of items. From equation (4), one can realize that the top-*k* neighbourhood is not static, depending also on the item of interest.

In addition, to require that Pearson's similarity to be computable, we introduce two neighbourhood formation schemes with additional criteria based on explicit trust and rating experience. We elaborate on these schemes in subsections 3.2 and 3.3.

3.2 Trust-based Neighbourhood (T)

In our first extension to the neighbourhood selection scheme, we consider only neighbours who have been explicitly indicated as trusted by a user, on top of the

requirement that Pearson's similarity value is computable. Note that we only consider binary explicit trust; we do not infer or compute implicit trust values between users. Let $t(a,b)=1$ denotes the existence of an explicit directional trust relationship from user *a* to user *b*, the set of neighbours fulfilling the criterion of being trusted and have also rated item *i* (that user *a* is interested in) can be expressed as:

$$T_{a,i} = \{\forall b \in U : b \neq a \wedge |I_b \cap I_a| \geq q \wedge t(a,b) = 1 \wedge r_b(i) \neq \perp\} \quad \textbf{(5)}$$

These neighbours are then ordered by their respective trustworthiness index, which is simply the count of in-degree trust links. Finally, the top-*k* neighbours (predictors) are selected to predict recommendations for this user using the Resnick's formula. We refer this selection scheme as Trust-based neighbourhood (T), as shown in Figure 1b.

3.3 Trust-Experience-based Neighbourhood (T-E)

We further explore if user experience could be taken into account to help selecting better predictors. The intuition is that users are more likely to seek advice from people who are more (or equally) experienced than those who are less experienced. For that reason our objective is to examine whether users who have given more product ratings should be considered better candidates. Considering this as an opinion flow, we impose that the direction is from the more experienced to the less experienced ones.

The experience level of a user (can also be thought as the amount of knowledge) is quantified according to number of recommendations that have been submitted by the user. We order users according to rating count such that:

$$\forall a,b \in U \text{ if } |I_a| \geq |I_b| \text{ then } e_a \geq e_b$$

where e_a denotes the experience level of user *a* and $|I_a|$ is the number of items that have been rated by user *a*. In this work, we categorize all users *U* into five experience levels, each consisting of $\frac{1}{5}|U|$ users. In this way, the set of neighbours that have rated some item *i* (that user *a* is interested in) and have a higher (or equal) experience level than user *a*, can be expressed as:

$$E_{a,i} = \{\forall b \in U : b \neq a \wedge e_b \geq e_a \wedge r_b(i) \neq \perp\} \quad \textbf{(6)}$$

Combining the computable similarity, trust and experience criteria in (4), (5) and (6), the set of neighbours that can potentially become the top-*k* predictors can be expressed as:

$$TE_{a,i} = \{\forall b \in U : b \neq a \wedge |I_b \cap I_a| \geq q \wedge t(a,b) = 1 \wedge e_b \geq e_a \wedge r_b(i) \neq \perp\} \quad \textbf{(7)}$$

We refer to this selection scheme as Trust-Experience-based neighbourhood (T-E), as depicted in Figure 1c. When selecting the top-k neighbours, the trustworthiness index has a higher priority over the experience.

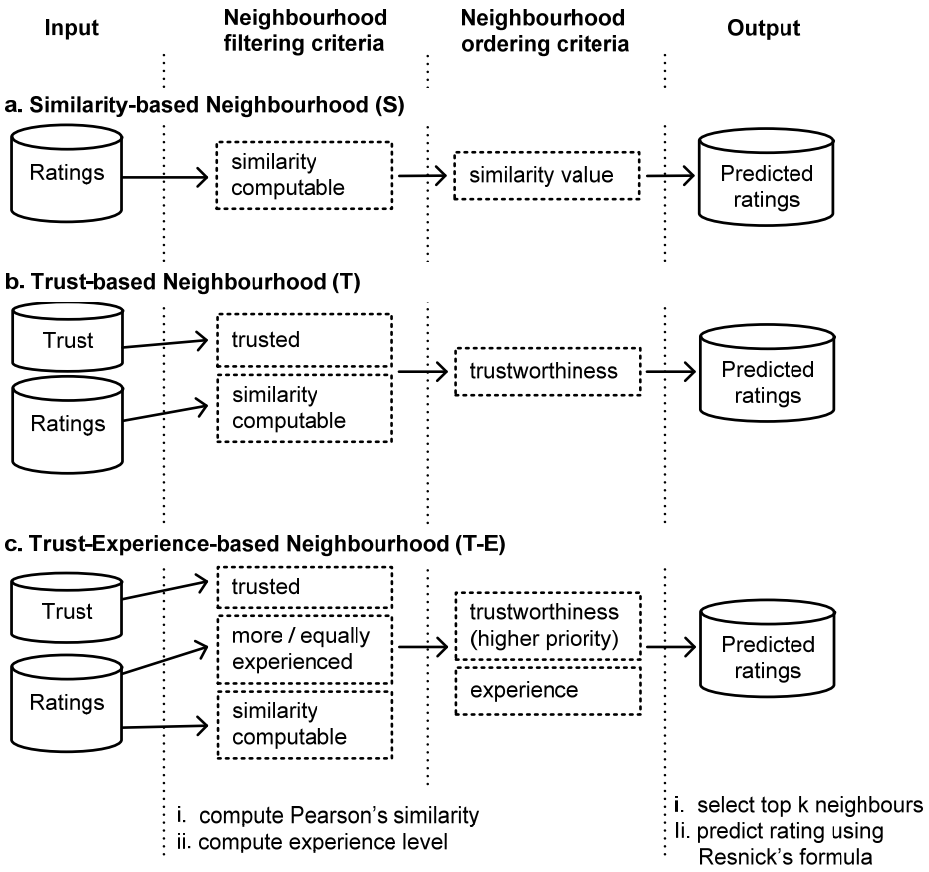

Fig. 1. Neighbourhood selection schemes

4 Experimental Setting

To evaluate our central question of whether trust helps to select better predictors we performed a series of simulations and compared the performance of the neighbourhood selection schemes as described in section 3.

4.1 Data

We used data from a popular recommender system, *Epinions.com* for the reason that it contains both product ratings and trust information we needed for our experiments. *Epinions.com* allows member users to write reviews about products consisting of a text and a quantitative rating from 1 to 5 stars. *Epinions.com* allows also users to build their web-of-trust by indicating other users whom they find have given consistently valuable reviews as trusted. In the current form of the system assistance is limited to the provision of textual and rating information from trusted users about the products of interest; input from trusted sources have to be digested by users manually.

The dataset was collected by Paolo Massa [11] by crawling the *Epinions.com* website during Nov-Dec 2003. In total, there are over 664K ratings given by 49K users on 139K products. Also included are 487K outward trust statements from users. Shown in Figure 2, both the distribution of rates and trust relationship in *Epinions.com* seem to follow power-law distribution, which is a feature of most social networks [4]. The figures are plotted in log scale showing the number of outward trust links per user (Fig. 2a) and the number of ratings given by individual users (Fig. 2b). The number of outward trust links and products rating count per user decrease sharply going from the very active users to the non-actives ones. This further suggests that the ratings on common products and the trust towards same users are sparsely distributed. As such, we used only a subset of Paolo Massa's dataset consisting of 1500 active users selected on the basis of number of ratings given by each user (no matter how many inward or outward trust links they have). This was also done to ensure that the Pearson's similarity value between users is computable with an adequate number of commonly rated products.

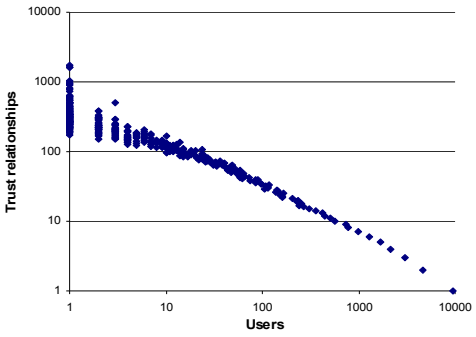

Fig. 2a. Outward trust links vs. trustors. There are 487K outward trust links in total.

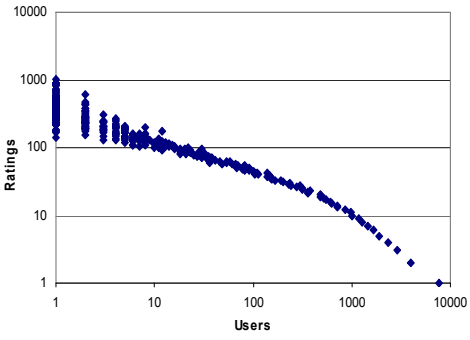

Fig. 2b. Individual rating count vs. raters. There are 664K item ratings in total.

The 1500 users were then divided into three communities referred to as "most active", "medium active" and "least active" of 500 users each (again based on the number of ratings that have been given by each user). Table 1 shows the average outward trust links and average number of ratings of the different communities.

Table 1. Trust links and average rating count of different communities

Community	Average outward trust links	Average number of ratings
Most active	40.33	260.91
Medium active	9.02	114.09
Least active	4.32	83.02

4.2 Evaluation Metrics

We considered both Predictive and Classification accuracy as being equally important to be measured. The former is demonstrative of the efficiency of the system in making accurate predictions for users. The latter as suggested by many researchers in *Information Retrieval* [1][10] is useful for measuring the frequency at which the system decides correctly or incorrectly about if an item is potentially liking for a particular user. Its usefulness in recommender systems is found in the creation of lists of products that are of high interest to users.

Predictive accuracy means the ability of the algorithm in producing accurate predictions for individual products. To demonstrate this we used the metrics *Mean Average Error (MAE)* and *Root Mean Square Error (RMSE)*. The latter is especially useful for identifying undesirably large errors. MAE and RMSE are both computed by comparing the real ratings (given by users) and the predicted ones. In our experiment the predicted values were rounded to the closest integer. We also applied correction on any predicted values that were out of range. Specifically, predicted values lower than 1 or higher than 5 were corrected to 1 and 5 respectively.

For evaluating the Classification accuracy we used the rating of 4 as the threshold for indicating a product that is of user's interest, meaning that a predicted value of 4 or 5 would be considered successful. *Recall* (*R*) is a metric to express the relative success in retrieving items of interest (either highly rated or lowly rated) in relation to the number of all items claimed to be of interest. *Precision* (*P*) is the relative success in retrieving items that are of user's interest. Both metrics can then be combined to express the effectiveness of retrieval with respect to the cost of retrieval to give the *F-Score*. *F-Score* is also known as *Harmonic Mean* [10] and it describes the trade-off between true positive (TP) and false positive (FP). *Precision (P)* and *Recall (R)* and *F-score* can be computed as follows:

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN} \qquad F = \frac{2PR}{P + R}$$

We used *Precision* (*P*) to measure the improvement in the relative success that the T and T-E schemes can possibly provide for identifying products that are of user's interest. We called a *hit* (or True Positive) for the case where some product that is of user's interest (i.e. has been rated as 4 or 5 by the user) and at the same time, using the algorithm, a high rating (4 or 5) has been predicted for it. We did not measure *Recall (R)* as is often impractical to do so in a recommender system [10]. The true positive (TP), false positive (FP) and false negative (FN) instances for our analysis are as defined in the confusion matrix in Table 2.

Table 2. The confusion matrix for classification test

	Predicted Value ≥ 4	**Predicted Value < 4**
Actual Rating ≥ 4	TP	FN
Actual Rating < 4	FP	TN

Since the application of filters (selection criteria) has implications to the number of items that can be actually predicted it is necessary that the level of this is also captured as well for every testing scenario. *Coverage* is a suitable metric for capturing this implication. It is defined as the ratio of recommendations for products that are of interest to the querying user and which the selected top-k neighbours can recommend, divided by the total number of items (that the querying user is interested in). The coverage of some particular user a can be computed using formula (8) where K denotes the set of top-k neighbours of user a, I being the set of all items in the system and I_a is the set of items that have been rated by user a .

$$C_a = \frac{1}{|I_a|} \cdot \left| I_a \cap \left\{ \bigcup_{b \in K} I_b \right\} \right| \quad \textbf{(8)}$$

4.3 Test Scenarios

The Similarity-based (S), Trust-based (T) and Trust-Experience-based (T-E) neighbourhood selection schemes formed the three main testing scenarios in our experiments. For each of the neighbourhood schemes, we studied the impact of the number of predictors on the performance by repeating *k*NN computation for different k values ranging from 3 to 13. When comparing the performance of neighbourhood schemes (e.g. T against S), we considered only predicted recommendations where the exact number of predictors (k) could be found.

As the ratings of common products and trust relationship are sparsely distributed, our experiments involved only a subset of 1500 active users from Paolo Massa's dataset. The 1500 users were divided into three communities referred to as "most active", "medium active" and "least active" (each with 500 users). We ran our experiments starting with the "most active" community.

When running a test scenario on a community of particular activity-level (i.e. most active, medium active and least active), we used the five-fold cross-validation method to further divide the community (500 users) into five fifths from where one fifth would be regarded as test set while the other four were used as training sets. This was repeated five times with a different fifth being used as the test set each time and the results were finally averaged.

5 Results - Discussion

We report the most interesting results from our experiments. First, in Table 3, we present the effects of trust criteria (both T and T-E schemes) on prediction accuracy

for the "most-active" community with *k* denoting the number of predictors used in each experiment. Due to the use of sparse dataset not all predictions can be made with the T and T-E neighbourhood schemes. When comparing the predictive accuracy (MAE, RMSE), we considered only user items that could be both predicted using the S neighbourhood scheme and the alternative scheme (T or T-E), in a pairwise manner. Thus, Table 3 and Table 4 show, for each of the experiment (using T or T-E scheme), the corresponding MAE and RMSE values measured using the S neighbourhood.

Table 3. Predictive accuracy and Coverage for the "most active" community

	MAE (%)				**RMSE**				**Coverage (%)**		
k	**S**	**T**	**S**	**T-E**	**S**	**T**	**S**	**T-E**	**S**	**T**	**T-E**
3	12.40	14.68	12.28	14.84	0.91	1.05	0.90	1.06	37.58	8.07	5.94
4	12.01	14.29	11.86	14.56	0.88	1.02	0.87	1.03	32.22	5.02	3.42
5	12.11	14.19	11.78	14.67	0.88	0.99	0.86	1.03	28.20	3.28	2.00
6	12.04	13.93	11.61	14.32	0.87	0.98	0.86	1.01	25.04	2.20	1.22
7	12.22	13.97	11.58	14.14	0.88	0.98	0.86	1.00	22.57	1.53	0.78
8	11.93	13.90	11.81	14.29	0.85	0.97	0.87	1.02	20.36	1.11	0.49
9	11.79	13.75	11.44	14.14	0.84	0.96	0.87	1.03	18.65	0.82	0.35
10	11.55	13.82	10.30	13.33	0.84	0.97	0.82	0.97	17.10	0.59	0.23
11	11.75	13.89	10.70	12.30	0.85	0.97	0.81	0.94	15.76	0.44	0.16
12	11.82	13.95	10.29	12.90	0.86	0.98	0.79	0.99	14.61	0.33	0.10
13	11.50	13.65	10.89	12.27	0.84	0.96	0.81	0.93	13.58	0.25	0.07

From the results it can be seen that trust does not help in choosing better neighbours to improve prediction accuracy as both the MAE and RMSE are lower for similarity-based selection scheme (compared to T and T-E schemes) for all *k* number of predictors. This suggests using explicit trust for selecting better predictors does not help to improve predictive accuracy for the "most active" community. Experienced users may be characterized by having stronger personal opinions; they may not rely on or be influenced easily by even those whom they trust.

Using the MovieLens dataset, Herlocker et.al. [3] show that an increasing neighbourhood size (using the S scheme) will improve the predictive accuracy until a certain threshold (about 15) where performance starts to deteriorate with more neighbours. Our results show a similar trend using the Epinions dataset; predictive accuracy improves following an increasing number of predictors, for all neighbourhood schemes (S, T, and T-E). However, due to the sparse distribution of commonly rated products and trust links with the Epinions dataset, we did not investigate further on larger neighbourhood; we stopped with 13 predictors.

As for classification accuracy, the performance of *Precision (P)* for the "most active" community improves in the T-E scheme, when a sufficiently large number of trusted and more experienced predictors are employed. This is shown in Figure 3 and can be interpreted as: user intuition, in choosing neighbours (by indicating explicit trust, when assisted by the system to filter out the less experienced ones), can work better than a system-determined similarity-based neighbourhood. Another observation is the increasing trend of *P* in both the T and T-E schemes, which starts low but catches up with (or overtakes) the S scheme with an increasing number of predictors . This suggests that user intuitions on trusting others to give good recommendations may not be individually reliable but can be helpful when aggregated.

Finally we observed serious implications on coverage value for all test cases when involving large numbers of predictors. Trust-based filters (T and T-E) affect the coverage even worse. The results are shown in Table 3.

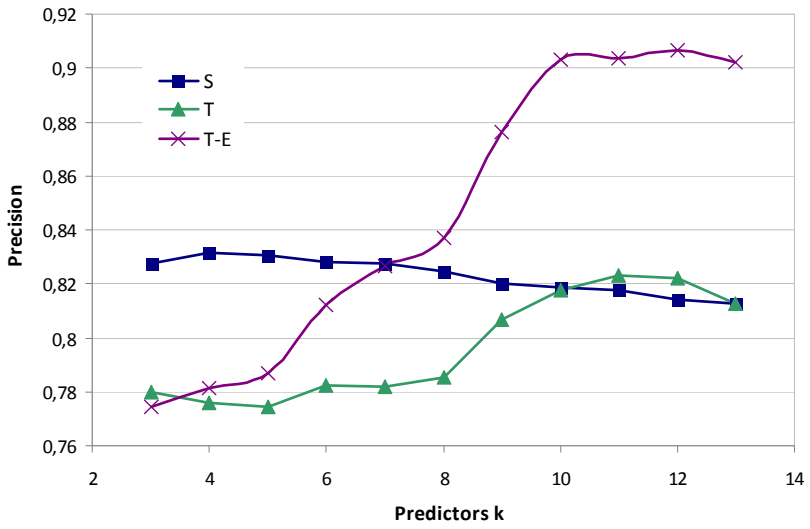

Fig. 3. Classification accuracy *(Precision)* for the "most active" community with different neighbourhood selection schemes.

So far, with the only exception of *Precision*, the use of explicit trust to select predictors has not been found very helpful. Coverage is strongly affected due to the limited trust information. Deriving trust from existing relationships (e.g. inferred from propagated trust) might be helpful for, at least, overcoming the coverage problem.

We should note that in other research works the use of propagative trust has been found quite successful for improving the predictive accuracy in user communities (e.g. in [2]). However, the distinctive difference here is that we examine the suitability of using explicit trust to select better predictors. Trying propagation trust frameworks (e.g. in [14]) to infer implicit trust values is outside the scope of this paper; we have deferred this to future investigation.

For the reason that trust is known to improve the predictive accuracy for cold-start users [21] we continued our experiments with the "medium active" and "least active" communities. Cold-start users are those who have not provided a sufficiently large number of ratings; as a consequence they often receive poor recommendations. The MAE and RMSE values for these communities are shown in Table 4a and 4b. *Diff* denotes the improvement of T or T-E scheme over the baseline S scheme. We show only results for some small k predictors as the trust links within the "medium active" and "least active" communities are scarce, causing it infeasible to investigate further.

In Table 4b, it can be seen that with k=3,4 predictors, contrary to the "most active" community, predictive accuracy is better with both the T and T-E criteria compared to the baseline S scheme. This suggests that in the "least active" community, as users are less experienced, opinions from explicitly trusted sources can be very useful.

Table 4a. Predictive accuracy for the "medium active" community.

	MAE (%)						RMSE					
k	**S**	**T**	**Diff**	**S**	**T-E**	**Diff**	**S**	**T**	**Diff**	**S**	**T-E**	**Diff**
3	12.26	14.54	-2.29	13.44	15.87	-2.43	0.94	1.09	-0.15	1.00	1.17	-0.17
4	13.10	14.71	-1.61	15.34	15.34	0.00	0.97	1.07	-0.10	1.05	1.12	-0.07
5	13.72	15.58	-1.86	15.68	18.38	-2.70	0.96	1.08	-0.12	1.05	1.29	-0.24
avg			**-1.92**			**-1.72**			**-0.12**			**-0.16**

Table 4b. Predictive accuracy for the "least active" community.

	MAE (%)						RMSE					
k	**S**	**T**	**Diff**	**S**	**T-E**	**Diff**	**S**	**T**	**Diff**	**S**	**T-E**	**Diff**
3	13.44	12.79	0.66	17.27	12.72	4.54	1.00	0.97	0.03	1.11	0.90	0.20
4	16.92	07.69	9.23	-	-	-	1.14	0.62	0.52	1.00	1.00	0.00
avg			**4.61**			**4.54**			**0.19**			**0.10**

(- denotes absence of data due to scarcity of trust links)

The average MAE *Diff* values (i.e. average improvement over the baseline S scheme) for the "most active" community are -2.22 and -2.72, for T and T-E schemes respectively. The corresponding value pairs for the "medium active" community and "least active" community are (-1.92, -1.72) and (4.61, 4.54) respectively. These average *Diff* values follow an increasing trend going from "most active" to "medium active" and to the "least active" communities. In other words, using trust to select better predictors can be more helpful to the less experienced users than the more experienced ones. Similar result, but more generalized as far as the number of predictors *k*, has also been found by Massa et.al. in [21].

We could not read much into the trend of predictive accuracy for the "medium active" and "least active" communities as the number of predictors available using the T and T-E schemes is very limited.

Table 5 presents the classification accuracy for the "medium active" community. *Precision* performs better in the T and T-E schemes for the "medium active" community even with small neighbourhood size. This conforms to the result on better predictive accuracy (as shown in Table 4a) as trust information is helpful for less experienced users. However, the combined trust-experience (T-E) criterion does not help as much as than the trust (T) criterion alone, different from the case with the "most active" community. On average, the *Precision* value has an improvement of 2.6% for T and just over 0.22% for T-E, compared to S.

Table 5. Classification accuracy for the "medium active" community.

	Precision			F-score		
k	**S**	**T**	**T-E**	**S**	**T**	**T-E**
3	0.880	0.901	0.867	0.863	0.864	0.835
4	0.884	0.919	0.894	0.864	0.850	0.849
5	0.887	0.898	0.895	0.867	0.822	0.756
avg	**0.883**	**0.906**	**0.885**	**0.865**	**0.845**	**0.813**

In short, trust is more helpful for the less active users. Nevertheless that requires that users have provided adequate trust inferences for people they can rely on, which is not always the case. When new users have little incentive or have not indicated

their trusted counterparts, a recommender system could consider inferring the implicit trust values based on product ratings.

6 Further Analysis - Discussion

We further investigated the performance based on individual ratings. As we have found that trust is more useful for less active users, we excluded the "most active" community from the analysis. The accuracy gain against the Similarity-based neighbourhood selection (S) for the "medium active" community is shown in Figure 4a. As can be seen, the application of trust-based criteria has helped to improve the predictive accuracy for items that users have given a real rating of 3 or 4. Note that also the combination of trust and experience (T-E) gives, for both cases when users have rated with 3 or 4, better predictive accuracy over using trust criterion alone.

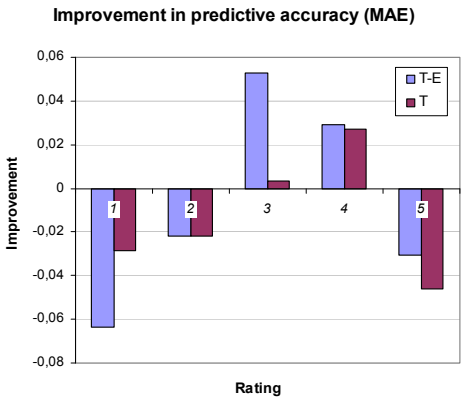

Fig. 4a. Improvement in prediction accuracy for individual ratings.

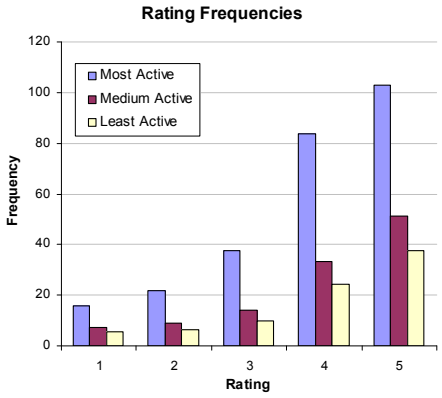

Fig. 4b. The Ratings distribution for different communities.

We further investigated whether the better predictive accuracy in rating 3 and 4 is related to rating distribution itself. Figure 4b showed that the (uneven) rating distribution is characterized with an increasing frequency going from rating 1 to 5. However, unlike those with a rating of 3 or 4, trust criteria do not help for items that have been rated with 5. This allows us to believe that there are other factors in effect (not due to rating distribution).

An attempt to explain the observation is that it is more likely that (Epinions) users would believe in non-extreme ratings (compared to ratings of 1 and 5) and therefore indicate their trust on these reviewers. When there are more non-extreme reviewers being trusted, it is likely that the predictive accuracy for non-extreme ratings will work out better. It would be interesting to further explore this matter from the perspectives of behavioral and cognitive sciences in the future.

7 Conclusions

We have performed a series of experiments in the context of recommender systems with the purpose to investigate our central question of whether explicit trust information can be useful in predicting user preferences. We presented two neighbourhood selection schemes involving trust criteria (Trust-based and Trust-Experience-based schemes) and compared their performances relative to the conventional Similarity-based *k*NN approach.

Our results show that trust criteria can help to improve the performance of recommender systems in a few cases. Specifically, trust information helps to improve the *Precision* in our classification test to provide good recommendations on items that are of users' interest.

Trust criteria are shown to be more helpful to the less experienced users judging from the increasing trend of better predictive accuracy (compared to the similarity-based scheme) going from the "most active" to the "least active" communities. Although trust-based schemes do not seem to help for active users in this work, we believe there might be other prediction algorithms where trust information can contribute. An interesting future work would be to explore if 'distrust' can be helpful for these active users.

Meanwhile, other than the *Precision* value for the "most active" community, the combined trust and experience criteria does not perform better than trust criterion alone. Although it is intuitive to filter out neighbours that are less experienced, strict selection criteria proves not to be very helpful in the Epinions dataset, where trust links and commonly rated products are scarce.

Using only explicit trust (without inference of implicit trust values) as what we have done for the purpose of our experimental setup incurs a heavy loss in terms of coverage. The performance is also restricted to the lack of trust information especially when it comes to the less active users. For these reasons, we render our support to the ongoing research in the computation of implicit trust values and building more sophisticated trust-aware recommender systems. There are also much to learn from other related disciplines including psychology and behavioural science.

References

1. McNee, M.S., Riedl, J., and Konstan, J.A., Accurate is not always good. How Accuracy Metrics have hurt. Recommender Systems, In Proc. ACM, CHI '06 April 22-27, Montreal, Canada, (2006)
2. Massa, P., and Avesani, P., Trust-aware recommender systems, In Proc. RecSys 2007: pp.17-24, (2007)
3. Herlocker, J. L., Konstan, J. A., Borchers, A., and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering, In Proc. 22nd ACM SIGIR'99 Conference on Research and Development in Information Retrieval, Aug.15–19, Berkeley, United States, pp.230-237, (1999)
4. Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y., User interactions in Social Networks and their Implications, In Proc. of 4th ACM European Conference on

Computer Systems, EuroSys '09, Nuremberg, Germany, Apr.01-03, 2009, pp.205-218, (2009)
5. Pitsilis, G., and Marshall, L.F., Modeling Trust for Recommender Systems Using Similarity Metrics, In Proc. IFIPTM, Trondheim, Norway, Springer, Vol. 263, pp.103-118, (2008)
6. Liu, H., Lim, E., Lauw, H. W., Le, M., Sun, A., Srivastava, J., and Kim, Y., Predicting Trusts among Users of Online Communities: An epinions case study, In Proc. of 9th ACM Conference on Electronic Commerce, Chicago, Il, USA, July 08 - 12, 2008, pp.310-319, (2008)
7 Kwon, Y. O., Improving Top-N Recommendation Techniques Using Rating Variance, In Proc. RecSys '08, Oct.23-25, Lausanne, Switzerland, ACM, (2008)
8. Kim, T-H., Yang, S-B., Using Attributes to Improve Prediction Quality in Collaborative Filtering, In Proc. EC-Web 2004, LNCS 3182, pp.1-10, (2004)
9. Truong, K., Ishukawa, F., and Hodinen, S., Improving the Accuracy of Recommender System by Item Clustering, In IECE Trans. Inf. & Syst. Vol.E90-D.NO9, (2007)
10. Herlocker, J. L., Evaluating Collaborative Filtering Recommender Systems, In ACM Transactions on Information Systems,Vol.22(1), pp.5-53, Jan 04, (2004)
11. A cooperative environment for the scientific research of trust metrics on social networks, http://www.trustlet.org/ (last accessed on 1st Oct 2009)
12. Shardanand, U., and Maes, P., Social Information Filtering: Algorithms for automating "word of mouth". In Proc. ACM CHI'95 Conference on Human Factors in Computing Systems, pp.210-217, (1995)
13. Melville, P., Mooney, R.L., and Nagarajan R., Content-Boosted Collaborative Filtering for Improved Recommendations. In Proc. of 18th National conf. of Artificial Intelligence, pp.187-192, (2002)
14. Jøsang, A., A Logic for Uncertain probabilities. International Journal of Uncertainty, Fuzziness & Knowledge Based Systems, Vol.9(3), (2001)
15. Levien, R., Advogato's trust metric, White Paper, (2009) http://www.advogato.org/trust-metric.html
16. Lathia, N., Hailes, S., and Capra, L., Trust-Based Collaborative Filtering, In IFIP Volume 263, Trust Management II, Karabulut Y., Mitchell J., Herman P., Jensen C.D., Trondheim, Norway, pp.87-102, (2008)
17. Truong, K., Ishukawa, F., and Hodinen, S., Improving the Accuracy of Recommender System by Item Clustering, In IECE Trans. Inf. & Syst. Vol.E(90)-D.N(O9), (2007)
18. Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. T., Application of Dimensionality Reduction in Recommender Systems–a case study. In ACM WebKDD Workshop, Boston, USA, (2000)
19. Geng, H., Deng, X., and Ali, H., A new Clustering Algorithm using Message Passing and its applications in Analyzing Microarray Data, In Proc. of 4th ICMLA, Dec.15-17, IEEE, Washington, USA, pp.145-150, (2005)
20. Melville, P., Mooney, R.J., and Nagarajan, R., Content-Boosted Collaborative Filtering for Improved Recommendations, In Proc. of 18th National ACM Conf. of Artificial Intelligence, Alberta, Canada, July 28-Aug.01, pp. 187-192, (2002)
21. Massa, P., and Avesani, P., Controversial Users demand Local Trust Metrics: an Experimental Study on Epinions.com Community, In Proc. of 20th National conf. AAAI'05, pp.121-126, (2005)
22. Boyd, D., and Ellison, N.B., Social network sites: Definition, History, and Scholarship, Journal of Computer-Mediated Communication, Vol. 13(1), (2007)
23. O'Donovan, J., and Smyth, B., Trust No One: Evaluating Trust-based Filtering for Recommenders, In Proc. of the 19th International Joint Conference on Artificial Intelligence IJCAI., Edinburgh, Scotland. Morgan Kaufmann Publishers, pp.1663-1665, (2005)