Journal of Information Processing Vol. 19 332–344 (July 2011)

Regular Paper

Exploring the Use of Explicit Trust Links for Filtering Recommenders: A Study on Epinions.com

Pern Hui Chia[†1] and Georgios Pitsilis[†2]

The majority of recommender systems predict user preferences by relating users with similar attributes or taste. Prior research has shown that trust networks improve the accuracy 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 are used to select the best neighbors or predictors, to generate accurate recommendations. We conducted a series of evaluations using data from *Epinions.com*, a popular collaborative reviewing system. We find that, for highly active users, using trusted sources as predictors does not give more accurate recommendations compared to the classic similarity-based collaborative filtering scheme, except in improving the precision to recommend items that are of users' liking. This cautions against the intuition that inputs from trusted sources would always be more accurate or helpful. The use of explicit trust links, however, provides a slight gain in prediction accuracy when it comes to the less active users. These findings highlight the need and potential to adapt the use of trust information for different groups of users, besides to better understand trust when employing it in the recommender systems. Parallel to the trust criterion, we also investigated the effects of requiring the candidate predictors to have an equal or higher experience level.

1. Introduction

The use of recommender systems (e.g., on Amazon.com, Youtube.com, NetFlix) has been popular to assist users to choose from multiple products or options available online. Different from reputation systems that give global ratings, recommender systems provide personalized recommendations to individual users based on attributes such as users' past purchases, search keywords, likings and ratings. Collaborative Filtering (CF) is employed to consider only inputs from relevant users, that are referred to as 'neighbors' or 'predictors', in order to provide accurate recommendations. A widely adopted CF technique is the similarity-based k-Nearest Neighborhood (kNN) scheme which identifies the k most similar neighbors, measured using Pearson's correlation coefficient, so to use their inputs for predicting recommendations on individual user-items.

Other than similarity in user attributes, trust information is also harnessed by some recommender systems. A trust-based recommender system incorporates the network of trust links into its recommendation prediction algorithm. A simple case is to make recommendations using only the inputs from those that have been explicitly indicated as 'trusted' by individual users. Trust relationship can also be propagated across the user network or inferred when not explicitly indicated. For example, if Alice trusts Bob and Bob trusts Carol, it is likely that Carol's inputs would be considered trustworthy by Alice. Also, if there is a consistent trend that Dan could provide Eva with good advices, even though Eva has not explicitly indicated that she trusts Dan, it is likely that Dan's inputs would be useful for Eva and could be implicitly trusted. The use of propagated or inferred trust values can be particularly helpful when trust information is scarce. Several schemes have been proposed to compute such implicit trust values, including MoleTrust [9] and Subjective Logic [7].

While it seems intuitive that considering inputs from trusted sources mimics the way that people get advices from trusted friends or experts in real life, there is no obvious support to that trusted inputs will *always* result in accurate recommendations for individual users. In this work, we are motivated to re-examine the links between trust and accurate recommendation. We investigate whether it is beneficial in terms of accuracy to employ inputs from explicitly trusted sources.

In addition to trust, we also study if the experience level of potential predictors can be utilized to help improving prediction accuracy. While experience is necessary for expertise, the link between user experience and providing accurate recommendation is not obvious. An intuition is that users are more likely to seek and accept opinions from users that are equally or more experienced in practice. Inputs from inexperienced users are also generally thought to be error-prone and inconsistent.

†1 Norwegian University of Science and Technology (NTNU), Norway
†2 University of Luxembourg (UNILU), Luxembourg

1.1 Motivation & Contribution

The usefulness of applying trust onto the existing CF has been explored in a number of prior works [2),10)], predominantly using complex mechanisms (e.g., trust propagation) devised by individual researchers. For example, Massa, et al. [10)] proposed using the trust values computed using their MoleTrust [9)] algorithm to replace the Pearson's correlation coefficient in the classic Resnick's prediction formula, to improve prediction accuracy. With that in mind, we considered not worthy to further investigate in the direction of inferring trust for better performance. Instead, we focus on the neighborhood selection mechanisms (i.e., in selecting the best predictors for kNN CF), which in our opinion, have not shown their full potential. Our work here is also strongly motivated to better understand the properties of trust and the link with user preferences. It is widely accepted by now that when one assigns his trust on another user, it is likely that the particular user has a similar rating profile to his. However, it is important to note that trust captures much more than just overall similarity. Realizing that the nuanced relationship of trust and profile similarity is a relatively unexplored area, Golbeck [3)] conducted a series of surveys and found that, in addition to overall similarity, trust is also correlated to a number of other properties such as the largest single difference in ratings, and the agreement on extreme ratings. Her work serves to better understand the link between trust and preferences. Here, we look at another important facet: how it could be different, and thus should be adapted, when predicting the preferences of different groups of users.

We find that, for highly active users, using explicit trust to filter for predictors does not give more accurate recommendations, other than improving the precision to predict items of users' liking, compared to the classic similarity-based kNN approach. Inputs from the more experienced users also do not help in improving predictive accuracy. These caution against the intuition that inputs from trusted and/or more experienced sources would always be more accurate or helpful.

Nevertheless, our evaluation does show a trend that the trusted (and more experienced) sources can give accurate recommendations for the less active users, albeit there are challenges in eliciting sufficient trust links from the less active users in the first place. Use of implicit trust values for the less active users can thus be helpful. Put together, our work highlights the need to adapt the use of trust information for different groups of users and to better understand the properties of trust in a collaborative system.

In the following, we first describe the related work in details in Section 2. We elaborate on how we incorporate trust and experience criteria into different predictor selection schemes in Section 3, followed by the evaluation settings and metrics in Section 4. We describe our findings in Section 5, and discuss the limitations of our evaluation and related concerns in Section 6.

2. Related Work

As aforementioned, use of trust information for improving the performance of recommender systems has been a popular research area. Massa, et al. [9)] reported interesting findings on "controversial" users, who are simultaneously trusted and distrusted by many, in the *Epinions.com* dataset. They argued that personalized trust metrics are needed given the fact that the controversial users take up to a large fraction of the population (20% in Epinions.com). They proposed MoleTrust – a personalized metric that is designed to propagate trust links explicitly indicated by users to a controllable distance. The same authors in another work [10)] modified the classic Resnick's prediction formula to make use of propagated trust values, computed using MoleTrust [9)], in place of Pearson's similarity weights. They found that this approach outperforms the classic CF in terms of prediction accuracy, especially for the "cold start" users, i.e., those who have only contributed a few ratings [⋆1]. Our evaluation strategies are different from theirs in the way that we incorporate trust information into neighborhood selection schemes rather than modifying the classic rating prediction formula. Nevertheless, our finding that trust is more helpful in predicting the preferences of the less active users than the highly active ones, is interestingly similar to theirs.

Several other works have devised ways to infer the implicit trust values between users and use the computed trust values to improve the performance of recommender systems. In the absence of explicit trust information in the MovieLens dataset, Lathia, et al. [8)] proposed an algorithm to determine who and how

⋆1 The advantage of better accuracy for the "cold start" users, however, reduces as trust is propagated through a longer social distance.

much users should trust one another based on individual users' past ratings. Meanwhile, O'Donovan and Smyth [13] proposed to compute the trustworthiness of individual users on the item-level, by comparing the ratings of individual users on a specific item, and to use the trustworthiness values, in conjunction with similarity values between users, for recommendation prediction.

The aforementioned works render support for the use of trust to improve the accuracy of recommender systems; however, we note that the findings are dependent on the individually devised algorithms and the conditions of the specific dataset used. It is both interesting and important to re-visit the links between trust and user preferences in the absence of complex inference rules. Golbeck further investigated the relationship between trust and profile similarity in Ref. 3), and found that *trust captures more than just the overall user similarity*. By grouping the users into three groups of different activity levels, we attempt in this work to investigate the link between trust and user preferences from another perspective. Specifically, we are interested in how useful trust information could be in predicting the preferences of different groups of users.

Besides the kNN collaborative filtering scheme, another technique used in recommender systems is the clustering scheme. Several existing works have applied user clustering to improve the prediction accuracy. We mention the work by Truong, et al. [16] as an important work in the field, in which the authors propose a method to group users into clusters of common interests. We have not applied clustering technique in our evaluation. A potential future work is to find out under which context or for which item category users have indicated their explicit trust on the others and to exploit the information. A better understanding towards the topological properties and how they are related to real user interaction or trust in social networks can also be helpful. We mention the work of Wilson, et al. [18] as a recent and complete analysis.

3. Predictor Selection Schemes

To evaluate the potential of explicit trust, we constructed several variations of the classic kNN Collaborative Filtering (CF) that make use of the trust links explicitly indicated by the users. We also investigated the use of experience level for selecting the most suitable predictors in our evaluation. Before elaborating on the details of the various schemes, we first describe the basic mechanisms used in conventional collaborative filtering systems.

Rating prediction. Classic CF systems typically use the Resnick's formula for recommendation prediction. As shown in Eq. (1), Resnick's formula computes the predicted rating $\hat{r}_{a,i}$ of item i for user a using existing ratings $r_{p,i}$ given to this item i by a set of predictors p.

$$\hat{r}_{a,i} = \bar{r}_a + \frac{\sum[w_{a,p}(r_{p,i} - \bar{r}_p)]}{\sum |w_{a,p}|} \tag{1}$$

Similarity computation. Resnick's formula requires that the similarity $w_{a,b}$ between two users a and b is known. The best known formula for computing the similarity between the users is Pearson's correlation coefficient, which measures the correlation of the preferences of two users. As shown in Eq. (2), Pearson's similarity $w_{a,b}$ between two users a and b is computed based on their ratings on the set of commonly-rated items. $\bar{r_a}$ and $\bar{r_b}$ denote the average of all ratings of user a and b respectively, while $r_{a,j}$ and $r_{b,j}$ indicate a's and b's rating on the item j. The outcome ranges from 1 to -1, with 1 denoting a prefect match and -1 the direct opposite in preference.

$$w_{a,b} = \frac{\sum (r_{a,j} - \bar{r}_a)(r_{b,j} - \bar{r}_b)}{\sqrt{\sum (r_{a,j} - \bar{r}_a)^2 \sum (r_{b,j} - \bar{r}_b)^2}} \tag{2}$$

Two-step process of predictor selection. Resnick's formula is sensitive to the number of predictors used and does not provide accurate prediction in sparse datasets [4]. These issues have also been highlighted in prior research e.g., in Ref. 15) which demonstrates the impact of number of predictors used on the prediction accuracy. The general consensus is that, in order to achieve a good performance, a small set of the most suitable predictors should be identified and used. Thus, a predictor selection scheme can be thought of as a two-step process. First, some *filtering* criteria are applied to obtain a set of candidate predictors. These candidates are then ranked following some *ordering* criteria to identify the k most suitable predictors.

3.1 Classic Similarity-based kNN (baseline scheme)

Having explained the formulas used in rating prediction and similarity compu-

tation, we now describe the classic similarity-based kNN CF which is used as the baseline scheme in our study.

Let U be the set of all users in the system and I be the set of all rated items. We write the set of ratings given by user a as:

$$I_a = \{\forall i \in I : r_{a,i} \neq \perp\} \subseteq I \quad (3)$$

where $r_{a,i} \neq \perp$ denotes that user a has given a rating for the item i.

The first requirement that a candidate predictor should fulfill is to have rated the item of interest i, i.e., $r_{b,i} \neq \perp$, shorthanded as $R_{b,i}$. A candidate predictor should also have rated at least q common items with the querying user (i.e., the user whom a recommendation is being predicted for). This is necessary so that Pearson's correlation coefficient would be computable or meaningful. This requirement can be described formally as:

$$\varsigma_{a,b} : |I_b \cap I_a| \geq q \quad (4)$$

Thus, the set of candidate predictors (or neighborhood) to be considered in the prediction of recommendation on the item of interest i for user a, in the similarity-based kNN scheme, is given by: $\{\forall b \in U : \varsigma_{a,b} \wedge R_{b,i}\}$. Note that the symbol $\wedge$ is used to denote *logical conjunction*. Next, the candidate predictors are ranked according to their similarity with user a in order to select the top-k predictors. One can see that the set of k most suitable predictors for user a is not static as it depends on the item of interest i.

3.2 Explicit Trust and Experience Based *k*NNs

We develop two additional *filtering* criteria to identify candidate predictors that have been explicitly trusted and that have a higher experience level than the querying user. First, we denote the requirement that a candidate predictor b must be explicitly trusted by the querying user a as:

$$\tau_{a,b} : trust(a, b) = 1 \quad (5)$$

Note that trust relationship is unidirectional. Similarly, the requirement for a candidate predictor b to have equal or more experience than the user a can be expressed as follows:

$$\epsilon_{a,b} : |I_a| \leq |I_b| \quad (6)$$

Requiring a candidate to have contributed equal or more ratings than the querying user depicts the intuition that users are more likely to seek advices from those who are more or equally experienced in real life. Combining both the trust and experience filtering criteria, we can thus investigate if those who are equally or more experienced and, at the same time, explicitly trusted, would be indeed better candidate predictors to help improving the accuracy of recommender systems.

To select the k most suitable predictors from the candidate set, we have tested with different *ordering* strategies, based on *similarity*, *experience* and *trustworthiness* levels as well as the Jaccard similarity index of the explicit trust links, shorthanded as *Jaccard distance*.

In the first case, Pearson's similarity value between a particular candidate and the querying user is used for ranking and selecting the k most similar predictors. Similarly, when using the *experience* ordering criterion, a candidate who has contributed more ratings gets a higher probability of being selected as a predictor. This is done with the intuition that users who have contributed more ratings could be more reliable and knowledgeable. Meanwhile, we measure the *trustworthiness* of a candidate predictor by the number of in-degree trust links that the candidate has. This depicts the case whereby a candidate that is trusted by many others should have a higher probability to be employed as a predictor. Lastly, we define *Jaccard distance* to measure how similar are the sets of explicit trust links of two users, based on the Jaccard similarity index [6] that is typically used in the field of data mining to measure the similarity of sample sets. It captures the intuition that the more people commonly trusted by two users, the more similar the two users are in placing their trust on the others. The symmetric *Jaccard distance* value between users a and b, is computed as follows:

$$J_{a,b} = \frac{\Gamma_a \cap \Gamma_b}{\Gamma_a \cup \Gamma_b} \quad (7)$$

with Γ_a and Γ_b denoting the set of users trusted by a and b respectively.

Table 1 summarizes the *filtering* criteria, set of candidate predictors and *ordering* criteria of all kNN schemes that we have evaluated. The abbreviation of each predictor selection scheme is coded such that the capital letter describes the filtering criteria, while the subscript letter denotes the (highest priority) ordering criterion used. It matters which criterion is applied first in the *ordering* process. The ordering criteria used in each kNN scheme have a decreasing priority levels from the left to right in the last column of Table 1. Specifically, *similarity* value has always the lowest priority, while *experience* takes the second lowest priority

Table 1 The filtering criteria, set of candidate predictors and ordering criteria of all predictor selection schemes. The classic S_s kNN is used as the baseline for evaluating the trust- and/or experience-based kNNs.

Abbr.	Filtering criteria	Set of candidate predictors	Ordering criteria
S_s	Similarity	$\{\forall b \in U : \varsigma_{a,b} \wedge R_{b,i}\}$	Similarity
T_s	Trust	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge R_{b,i}\}$	Similarity
T_t	Trust	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge R_{b,i}\}$	Trustworthiness, Similarity
T_j	Trust	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge R_{b,i}\}$	Jaccard distance, Similarity
E_e	Experience	$\{\forall b \in U : \varsigma_{a,b} \wedge \epsilon_{a,b} \wedge R_{b,i}\}$	Experience, Similarity
TE_s	Trust & Experience	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge \epsilon_{a,b} \wedge R_{b,i}\}$	Similarity
TE_t	Trust & Experience	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge \epsilon_{a,b} \wedge R_{b,i}\}$	Trustworthiness, Experience, Similarity
TE_j	Trust & Experience	$\{\forall b \in U : \varsigma_{a,b} \wedge \tau_{a,b} \wedge \epsilon_{a,b} \wedge R_{b,i}\}$	Jaccard distance, Experience, Similarity

in the presence of *trustworthiness* or *Jaccard distance* criterion. Note that also the requirement in Eq. (4) has been applied to all schemes so to ensure that it is possible to compute Pearson's similarity and thus to predict recommendations using Resnick's formula. We include the graphical illustration of the work flow of various kNN schemes as well as the filtering and ordering criteria in **Fig. 1**.

4. Evaluation Settings

To evaluate the central question of whether trust helps in selecting better predictors, we performed a series of evaluations using data from a popular online system, *Epinions.com*. The choice of this particular system was made because it contains both ratings and explicit trust links that are needed for our study.

4.1 Dataset

Epinions.com is a collaborative reputation system on which members can write textual reviews about items they have experienced with and rate them in a five-star scale. The system covers a wide range of product categories, including cars, books, movies, electronics, sports equipments and travel destinations. The members of *Epinions.com* can also rate each other based on the ratings and reviews individual users have provided on some items. Specifically, users can express their explicit trust (or distrust) to those whom them consider reliable (or find their reviews offensive or inaccurate), forming a web of trusted (distrusted) reviewers. Yet, in its current form, *Epinions.com* does not make use of the trust and distrust information to compute personalized recommendations. It simply highlights the reviews and ratings from sources that have been explicitly marked as trusted to individual users, such that the trusted reviews and ratings are displayed before the inputs from general reviewers. Such arrangement allows users to observe the signals from trusted sources easily.

The dataset we used was collected by Paolo Massa [17] who crawled the public pages on *Epinions.com* in November-December 2003. The dataset has around 664 thousand ratings contributed by about 49 thousand users on about 139 thousand items. Also included are about 487 thousand trust links that have been explicitly indicated by users. The dataset, however, does not contain any distrust information. **Figure 2** depicts the distributions of rating contribution and out-degree trust links, per user, in a log-log scale. One can see that both distributions are characterized by a heavy tail where the bulk of rating contribution or out-degree trust links comes from a small number of users in the system. This suggests that the ratings on common products and the trust towards same users are sparsely distributed.

4.2 Evaluation Setup

To mitigate the sensitivity of Pearson's correlation coefficient, we used in our study a subset of Paolo Massa's dataset consisting of 1,500 top active users, selected on the basis of rating contribution, no matter how many explicit trust links they have indicated or received from the others. This ensures that Pearson's similarity value is computable with an adequate number of commonly rated items. To study the effect of user activity on the performance of recommender systems, we further divided the 1,500 users into 3 separate groups of 500 users, also on the basis of rating contribution. We refer to the 3 groups as the "most", "medium" and "less" active communities, hereafter.

In each community, we evaluated the performance of the similarity-based base-

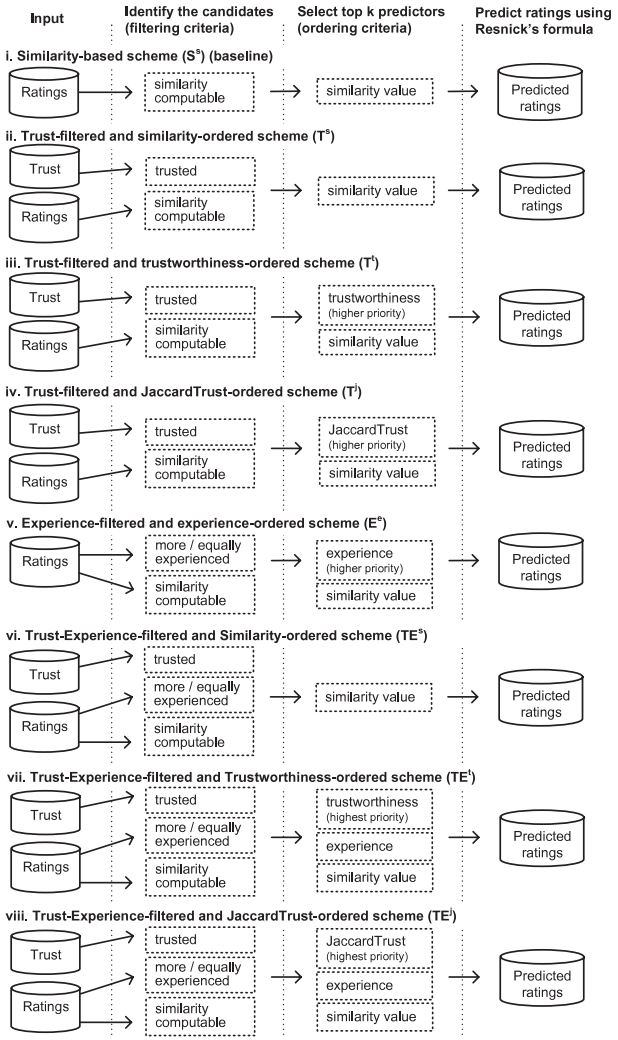

Fig. 1 The workflow and filtering/ordering criteria of the predictor selection schemes.

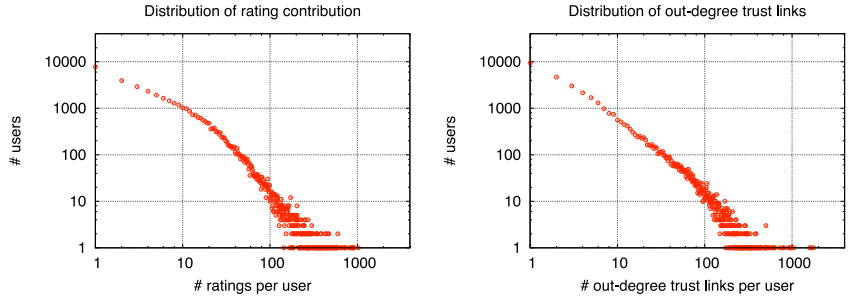

Fig. 2 Distribution of rating contribution (left) & out-degree trust links (right) per user.

foreach *predictor-selection scheme* $\{T_s, T_t, T_j, E_e, TE_s, TE_t, TE_j\}$ **do**
  foreach *community* {*most, medium, less active*} **do**
    foreach $k \in [3, 16]$ **do**
      foreach *five-fold cross validation test set* **do**
        foreach *user a in the test-set* **do**
          foreach *item i that user a has rated* **do**
            choose k best predictors
            compute predicted recommendation $\hat{r}_{a,i}$
        compute the average performance of five-fold cross validation
        compare with the average performance of baseline S_s scheme

Fig. 3 Detailed evaluation flow.

line and the seven different trust- and/or experience-based schemes, as shown in Fig. 1. As we were also interested in studying the impact of number of predictors k on the performance, we tested for different values of k ranging from 3 to 16. To be fair when comparing the predictive accuracy of a particular scheme to the classic S_s baseline, we compared only cases in which the exact number of predictors k could be found in both the baseline and the scheme under evaluation. The detailed flow of our evaluation is summarized in **Fig. 3**.

We checked the validity of our evaluation setup by investigating the relationship between trust and rating similarity in the dataset. Specifically, we computed the average similarity of individual users with the explicitly trusted users, and with all users in general (inclusive of the non-trusted users), as shown in **Table 2**. We

Table 2 Average similarity values with explicitly trusted users, and with all users.

	Avg. similarity with trusted users only	Avg. similarity with all users
All users in complete dataset	0.370	0.302
The top 1,500 active users	0.317	0.292

found the average similarity values[⋆1] are not high, ranging from $0.29 - 0.37$. A higher similarity is observed with the trusted users than with all users in general. But surprisingly the increment in similarity is very minimal, especially for the top 1,500 active users that we used in our evaluations, indicating a weak link between rating similarity and trust relationship in this dataset. In other words, users do not decide on whether to trust a particular user, solely based on how similar are their ratings. This largely agrees with Golbeck's findings that trust captures more than just overall similarity [3]; there is much to learn about the different properties of trust.

4.3 Evaluation Metrics

Trust and experience filtering criteria were expected to have implications on the number of predictable recommendations. We used *Coverage* to measure the extent to which a recommender system can provide recommendations for all items that are of interest of all users in the system. The coverage of a predictor selection scheme for a particular user a is defined as the ratio of items that are of interest to a and that the selected k predictors can recommend, divided by the total number of items that a is interested in, as shown in the following:

$$C_a = \frac{\mid I_a \cap \{\cup_{b \in K} I_b\} \mid}{\mid I_a \mid} \tag{8}$$

where K denotes the set of top-k predictors for each user-item, I_a denotes the set of items a has rated, and I_b denotes the set of items that a particular predictor b from K has rated.

For evaluating the accuracy on different selection scheme, we considered both *Predictive* and *Classification accuracy* as equally important. The former measures how accurate are the personalized recommendations made for different items of interest of individual users. It uses two typical metrics, namely *Mean Absolute Error* (MAE) and *Root Mean Squared Error* (RMSE). RMSE is useful for quantifying undesirably large errors. Both metrics compare the real ratings (given by users) with the predicted recommendations.

Typical in the field of recommender systems, it is not easy to simulate a real environment of users having received some recommendations, experienced and rated the items sequentially. Therefore, to be able to measure the difference between a predicted recommendation and the real rating, we used the technique, called 'five-fold cross validation' in our evaluation. Specifically, we divided each community of particular activity level ("less", "medium" and "most" active) into five fifths, from which one fifth of users was used as a testing set while the remaining four fifths as training sets. That was repeated 5 times using a different fifth as a test set and the results were finally averaged.

Equally important to prediction accuracy is the ability of a recommender system to provide a list of recommendations to items that the users actually like (i.e., items that the users eventually give a rating of 4 or 5 stars). This ability can be measured in terms of *Classification accuracy* which has been widely used in the field of information retrieval, e.g., in Refs. 5), 11). The metrics used for classification accuracy are *Precision* (P), *Recall* (R) and *F-Score* (F).

In the context of this paper, Precision (P) indicates the relative success that a recommendation provided by the recommender system matches a user's real liking. Recall (R) measures the relative success in retrieving all items that are liked by individual users. Meanwhile, F-Score (F) measures the trade-off of P and R by taking the harmonic mean of both values. P, R and F are computed as follows:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}, \quad F = \frac{2PR}{P + R} \tag{9}$$

where TP (True Positive) denotes the case that a product is liked by a user and the recommender system has predicted such (i.e., generating a recommendation of 4 or 5 stars). Likewise, FP (False Positive) denotes the case where an item has been wrongly predicted to be of the liking of a user, while FN (False Negative) is the case when an item has been wrongly predicted as not being of the user's liking. The TP, FP and FN cases are shown in the confusion matrix in **Table 3**.

⋆1 Note that we have omitted in our computation cases where the similarity value between two users is not computable due to a lack of commonly rated items.

Table 3 Confusion matrix for classification test.

	Predicted Value ≥ 4	Predicted Value < 4
Actual rating ≥ 4	TP	FN
Actual rating < 4	FP	TN

Five-fold cross validation was also used when evaluating classification accuracy.

5. Results

We report the most interesting results from our study. First, for the "most active" community, we observed that the predictive accuracy of the baseline similarity-based scheme (S_s) improves when the number of predictors used k increases from 3 to 9 and stabilizes until 16, as shown in **Fig. 4**. This is similar to the findings by Herlocker, et al. in Ref. 4) which reported that, using the *MovieLens* dataset (different from our *Epinions.com* dataset), an increasing number of predictors improves the predictive accuracy of the similarity-based collaborative filtering only until a certain threshold (about 15 in their findings, after which the performance starts to deteriorate). Due to the sparse distribution of commonly rated products and trust links in the *Epinions.com* dataset, we did not investigate the cases with more than 16 predictors.

Not all item ratings could be predicted when applying the trust and/or experience criteria, given the limited number of explicit trust links. Thus, when evaluating the predictive accuracy of a particular trust- and/or experience-based scheme, we considered only item ratings that can be predicted both using the baseline S_s and the scheme under evaluation. We present the relative improvement in MAE of a particular trust- and/or experience-based scheme over the similarity-based baseline scheme in **Fig. 5**. We do not include the RMSE measurements here as they follow a similar pattern to MAE.

The negative values, for all cases of different number of predictors in the trust-based schemes (T_s, T_t and T_j) in Fig. 5, show that explicit trust does not help in choosing better predictors to improve predictive accuracy among the highly active users. Using experience criterion (in E_e) or the combination of trust and experience criteria (in TE_s, TE_t and TE_j) also does not help to improve the predictive accuracy. *This can be due to that the highly active users have strong personal opinions and they may not rely on or be influenced easily by even those whom they trust and/or who are more experienced.*

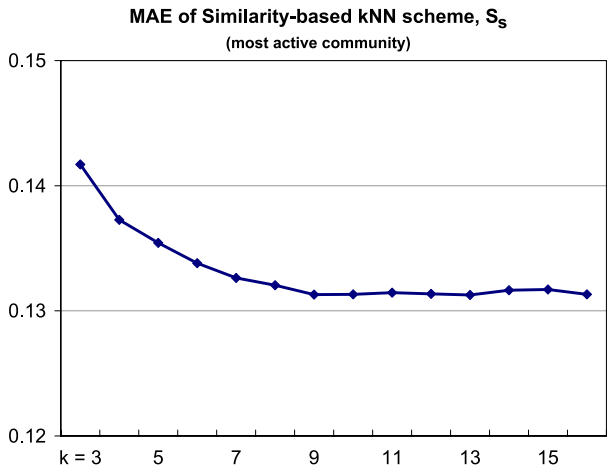

Fig. 4 MAE of the baseline S_s scheme against an increasing number of predictors used.

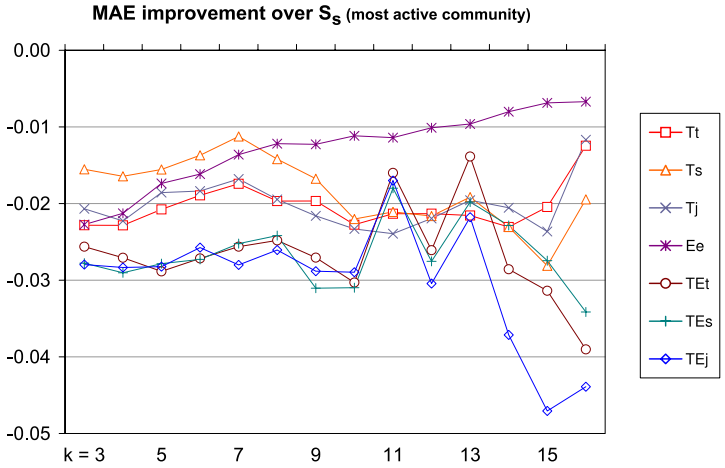

Fig. 5 MAE improvement of various schemes compared to the baseline S_s. Note that the accuracy of all the Trust- and/or Experience-based kNN schemes is lower than S_s.

We observed that the accuracy of the trust-based schemes does not differ significantly when using different ordering criteria. Other than a slightly better

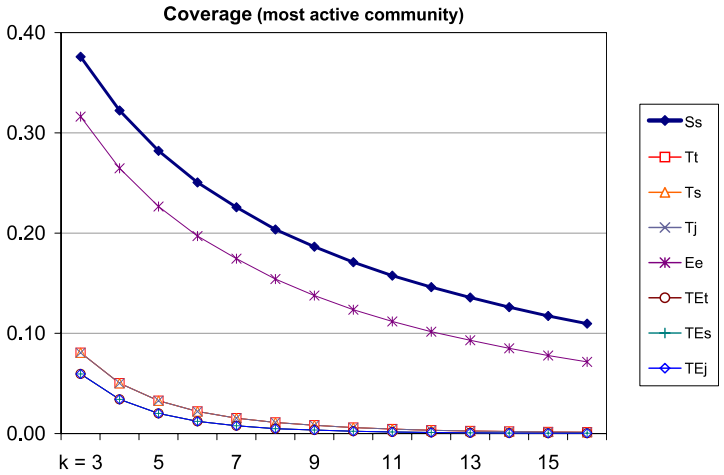

Fig. 6 Coverage of different predictor selection schemes decreases as the number of predictors required k increases.

accuracy in the T_s scheme when the number of predictors is small ($k = 3$ to 9), the performance of the T schemes seem to follow a similar pattern. A stronger pattern can be recognized among the TE schemes. The fluctuation in the performance of the TE schemes can be likely due to the limited number of candidate predictors that meet both the requirements of being explicitly trusted and more experienced.

Indeed, we found that the coverage of the trust and experience-based schemes drops significantly, especially in cases requiring a large number of predictors, as shown in **Fig. 6**. Coverage is especially affected by the limited and sparse trust information in the dataset. Inferring the implicit trust values among the users from the explicit trust links or ratings might be helpful, at least for overcoming the coverage problem. Massa and Avesani [10] found that the use of propagative trust can be useful for improving the predictive accuracy of recommender systems. We note that our work is different from theirs in that we examined the suitability of using explicit trust and experience criteria to select better predictors as an attempt to better understand how trust works in collaborative filtering systems. Inferring implicit trust values is out of the scope of this paper. Exploring whether implicit trust between users can help to select better predictors for collaborative filtering can, however, be an interesting future work.

Next, we measured the classification accuracy of the various schemes in recommending items that are of users' liking. We observed that Precision (P) improves in the T, TE and E schemes following an increasing number of predictors k. As shown in **Fig. 7** (left), the precision of these schemes is in fact better than the baseline similarity-based system when a sufficiently large number of trusted and/or more experienced predictors are used. We interpret this as: *user intuition in indicating explicit trust, especially when being assisted by the system to take only inputs from the more experienced users, can help the system to give recommendations that match users' liking.*

However, as shown in the Recall (R) values (Fig. 7, middle), the smaller set of explicitly trusted and/or more experienced candidate predictors performs poorly in recognizing all items that are of user's liking. If we weigh Precision and Recall measures equally, the overall classification accuracy of the various schemes is captured in F-Score (S), the harmonic mean of Precision and Recall. Depicted in Fig. 7 (right), we see that the F-Score values of the T, TE and E schemes are all lower than that of the classic similarity-based scheme (S_s). Hence, *with the only exception of Precision, the use of explicit trust to select predictors has not been found helpful, for the highly active users.*

Next, we evaluated the performance of the trust- and/or experience-based schemes in the "medium active" and "less active" communities. "Cold start" users, which was coined to describe those who have not provided a sufficiently large number of ratings, were found to be often receiving poor recommendations using the classic similarity-based scheme [9].

Figure 8 shows the average MAE improvement of different predictor selection schemes over the baseline similarity-based approach for different communities. The average values were taken from cases in which the number of predictors used, k was from 3 to 6 only, as the trust links within the "medium active" and "less active" communities are particularly scarce, causing it infeasible to investigate further. As shown in the figure, the MAE improvement in various T and TE schemes over the baseline S_s method follows an increasing trend from the "most active" to "medium active" and "less active" communities.

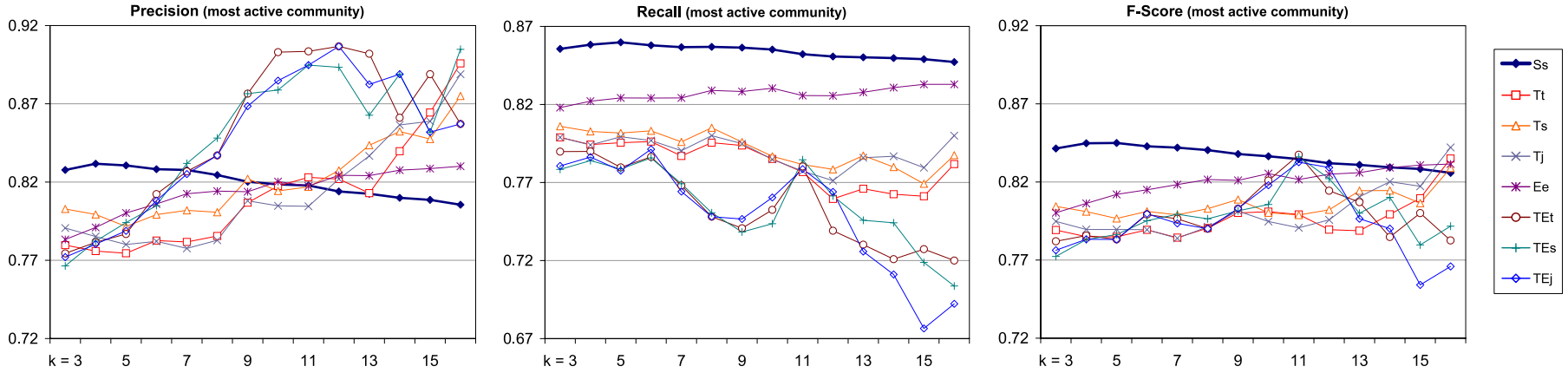

Fig. 7 Classification accuracy of all evaluated schemes.

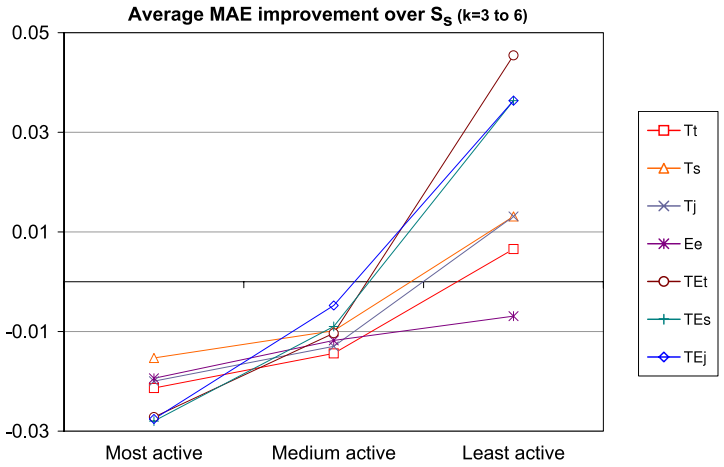

Fig. 8 Average MAE improvement, considering only cases where the number of predictors $k = 3$–6, over the baseline S_s scheme in different communities.

While the predictive accuracy of various trust and/or experience predictor selection schemes is still lower compared to the baseline similarity-based approach in the "medium active" community, the improvement becomes positive in the "less active" community (other than the experience-only E_e scheme which also follows an increasing trend). This shows that explicit trust can help to select better predictors to improve the predictive accuracy of the recommendations for the less active users.

As also shown in the TE schemes in Fig. 8, combining trust and experience filtering criteria produces a even higher predictive accuracy for the less active users. These findings point to the conclusion that *opinions from explicitly trusted sources and/or more experienced users can be helpful for the less experienced users*. More importantly, this is *opposite to the finding for the highly active users*. This highlights the need to adapt the use of trust information for different groups of users and to better understand trust when employing it in the recommender systems in the hope of improving predictive accuracy.

Nevertheless, for the less active users to benefit from trust information, it is required that they have indicated adequate trust links on those whom they think they could rely on to predict their preferences. In reality, this is not always the case. It is not counter intuitive to reason that the less active users (i.e., those who have contributed fewer ratings according to our definition) will have less energy to indicate their trust on certain reviewers. A recommender system could consider inferring implicit trust values based on various available attributes to

assist the less active (or new) users (who have not indicated sufficient trust links). The system should also make the process of indicating explicit trust links easier. It could be helpful also to harness the potential of social networks, for example, by having a default configuration in an online social network service to easily mark friends with common interests as trusted.

6. Discussion

A extensively studied subject in the field of recommender systems is data sparsity, which has been identified as a major problem that affects prediction quality. It has been investigated from several different directions. The use of mathematical models such as Latent factor analysis, generally known as Dimensionality Reduction, has shown promising results [14]. A majority of work, however, takes the empirical approach and proposes variations of the classic similarity-based kNN scheme. An important work is by Melville, et al. [12] which propose a hybrid collaborative filtering system to overcome the problem of rating sparsity. We have not treated the problem of the sparse ratings and trust links in the *Epinions.com* dataset in this work. Instead, we simply mitigated the effects of data sparsity by evaluating only on the 1,500 top active users. This may limit the generalizability of our findings [⋆1].

Our work has mainly focused on the prediction accuracy of recommender systems when applying explicit trust links and experience level as additional criteria for filtering potential recommenders. Yet, we note that accuracy is not everything. Factors such as *serendipity* (i.e., the ability to recommend items that can pleasantly surprise the users), *user experience* and *understanding of user expectation* are equally important to the success of a recommender system [11].

While trust information may not help to predict the preference of the highly active users, use of inputs from explicit trusted sources can have several advantages. For example, inputs from trusted sources can be aggregated to generate *trustworthy* outcomes as a measure to mitigate the problems of exploitation (e.g., profile-injection attack) [10]. Chia, et al. also showed that inputs from trusted and known sources are considered more *relevant* and *salient* (of stronger impact) than inputs from unknown community members [1]. Relevance and salience are important properties that can be helpful in mitigating the click-through habituation of users ignoring security warnings.

⋆1 Evaluating on some random 1,500 users will not work as it is highly likely that the similarity values of randomly selected users will not be computable.

7. Conclusions

Trust has been the subject of investigation by many researchers in the past as a solution for improving the performance and security of recommender systems. Nevertheless, little effort has been put to examine the intuition that inputs from trusted sources will always result in a higher level of recommendation accuracy. The use of experience as a criterion in predictor selection has also not been explored adequately thus far.

We performed a series of evaluations to explore the use of explicit trust and/or experience criteria in selecting better predictors to improve the accuracy of a recommender system. We did not attempt to propose a new predictor selection scheme here; our work is motivated by the purpose to better understand trust and experience in a collaborative system, and to get new insights of how they could be better incorporated in the design of recommender systems.

Our evaluation results show that trust, when used as the filtering criteria of a kNN predictor selection scheme, can only help to improve the accuracy of a recommender system in limited ways. For the highly active users, it helps to improve the precision of a recommender system to recommend items that are of users' liking, but does not help to improve the predictive accuracy and recall in comparison to the classic similarity-based scheme. Requiring the predictors to have equal or a higher experience level gives similar results. Trust and experience criteria are, however, found to be helpful to the less experienced users. There is a trend of better predictive accuracy, going from the "most active" to the "less active" communities. The opposite finding, for the highly active and the less active users, highlights the need and potential to adapt the use of trust and experience criteria for different groups of users.

We are not against the idea of using trust information in recommender systems, as we believe there are instances where trust information can contribute to improve accuracy. As discussed earlier, there are also multiple advantages

for using trusted inputs such as to mitigate the problems of manipulation (e.g., profile-injection attack) and click-through habituation (as trusted inputs increase the salience and impact of risk signaling). Yet, we note that it is important to better understand the properties of trust before applying it in complex systems. It would be also interesting to study the role of "distrust" in affecting user preferences.

Using only explicit trust, without inference of implicit trust values, as we have done in our evaluation setup, incurs a heavy loss in terms of coverage. The better performance found for the less active users is also restricted, given that we can hardly expect the less active users to provide adequate trust information. For these reasons, we render our support to the ongoing research in the computation of implicit trust values for building sophisticated trust-aware recommender systems. However, we note that the implicit trust values should be used with care so that the advantages of trusted inputs (e.g., in producing trustworthy, relevant and impactful outcomes) do not diminish. There are much to learn from other disciplines including psychology and behavioral science.

Acknowledgments We thank our colleagues and the anonymous reviewers for their helpful comments. The second author carried out this work during the tenure of an ERCIM "Alain Bensoussan" Fellowship program.

References

1) Chia, P.H., Heiner, A.P. and Asokan, N.: Use of Ratings from Personalized Communities for Trustworthy Application Installation, *NordSec '10: 15th Nordic Conference in Secure IT Systems* (2010).
2) Golbeck, J.: Computing and Applying Trust in Web-based Social Networks, PhD Thesis, University of Maryland (2005).
3) Golbeck, J.: Trust and nuanced profile similarity in online social networks, *ACM Trans. Web*, Vol.3, No.4, pp.1–33 (2009).
4) Herlocker, J.L., Konstan, J.A., Borchers, A. and Riedl, J.: An algorithmic framework for performing collaborative filtering, *SIGIR '99: Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.230–237, ACM (1999).
5) Herlocker, J.L., Konstan, J.A., Terveen, L.G. and Riedl, J.T.: Evaluating collaborative filtering recommender systems, *ACM Trans. Inf. Syst.*, Vol.22, No.1, pp.5–53, ACM (2004).
6) Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura, *Bulletin del la Société Vaudoise des Sciences Naturelles*, Vol.37, pp.547–579 (1901).
7) Jøsang, A.: A logic for uncertain probabilities, *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Vol.9, No.3, pp.279–311 (2001).
8) Lathia, N., Hailes, S. and Capra, L.: Trust-Based Collaborative Filtering, *Trust Management II*, Karabulut, Y., Mitchell, J., Herrmann, P. and Jensen, C. (Eds.), *IFIP International Federation for Information Processing*, Vol.263, pp.119–134, Springer Boston (2008).
9) Massa, P. and Avesani, P.: Controversial users demand local trust metrics: An experimental study on Epinions.com community, *AAAI'05: Proc. 20th National Conference on Artificial Intelligence*, AAAI Press, pp.121–126 (2005).
10) Massa, P. and Avesani, P.: Trust-aware recommender systems, *RecSys '07: Proc. 2007 ACM Conference on Recommender Systems*, pp.17–24, ACM (2007).
11) McNee, S.M., Riedl, J. and Konstan, J.A.: Being accurate is not enough: How accuracy metrics have hurt recommender systems, *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, pp.1097–1101, ACM (2006).
12) Melville, P., Mooney, R.L. and Nagarajan, R.: Content-Boosted Collaborative Filtering for Improved Recommendations, *Proc. 18th National Conference Artificial Intelligence*, pp.187–192 (2002).
13) O'Donovan, J. and Smyth, B.: Trust no one: Evaluating trust-based filtering for recommenders, *IJCAI'05: Proc. 19th International Joint Conference on Artificial Intelligence*, pp.1663–1665, Morgan Kaufmann Publishers Inc. (2005).
14) Sarwar, B.M., Karypis, G., Konstan, J.A. and Riedl, J.T.: Application of dimensionality reduction in recommender systems – A case study, *ACM WebKDD Workshop* (2000).
15) Shardanand, U. and Maes, P.: Social information filtering: Algorithms for automating "word of mouth", *CHI '95: Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp.210–217, ACM Press/Addison-Wesley Publishing Co. (1995).
16) Truong, K., Ishikawa, F. and Honiden, S.: Improving Accuracy of Recommender System by Item Clustering, *IEICE Trans. Inf. Syst.*, Vol.E90-D, No.9, pp.1363–1373 (2007).
17) TrustLet: A cooperative environment for the scientific research of trust metrics on social networks. http://www.trustlet.org/
18) Wilson, C., Boe, B., Sala, A., Puttaswamy, K.P. and Zhao, B.Y.: User interactions in social networks and their implications, *EuroSys '09: Proc. 4th ACM European Conference on Computer Systems*, pp.205–218, ACM (2009).

(Received November 1, 2010)
(Accepted April 8, 2011)
(Released July 6, 2011)

Pern Hui Chia was born in 1980. He received his B.Comp. (Computer Engineering, Honors) from National University of Singapore in 2005, and MSc. degree (Mobile Computing and Security) from Helsinki University of Technology, Finland and Royal Institute of Technology, Sweden in 2008. He is currently a Ph.D. student at the Centre for Quantifiable Quality of Service in Communication Systems (Q2S), Norwegian University of Science and Technology (NTNU), Norway. His Ph.D. research focuses on the economic, social and behavioral aspects of information security.

Georgios Pitsilis was born in 1967. He received his B.Eng. in Engineering and B.Eng. in Informatics from the Technological Educational Institution of Athens, Greece in 1991 and 1999 respectively, MSc. degree from Oxford Brookes University, UK in 2000, and Ph.D. degree from Newcastle University, UK in 2007. He is currently a postdoc researcher at the University of Luxembourg. Prior to this position, he has worked as a research scientist at Newcastle University, UK and as a postdoc fellow at NTNU, Norway. His research interests are in the areas of trust management, recommender systems and distributed computing.