From Ratings to Trust: An Empirical Study of Implicit Trust in Recommender Systems

Guibing Guo[1], Jie Zhang[1], Daniel Thalmann[1], Anirban Basu[2], Neil Yorke-Smith[3]
[1]Nanyang Technological University, Singapore
[2]KDDI R&D Laboratories, Inc., Japan
[3]American University of Beirut, Lebanon; and University of Cambridge, UK
{gguo1, zhangj}@ntu.edu.sg, basu@kddilabs.jp, nysmith@aub.edu.lb

ABSTRACT

Trust has been extensively studied and its effectiveness demonstrated in recommender systems. Due to the lack of explicit trust information in most systems, many *trust metric* approaches have been proposed to infer implicit trust from user ratings. However, previous works have not compared these different approaches, and oftentimes focus only on the performance of predictive item ratings. In this paper, we first analyse five kinds of trust metrics in light of the properties of trust. We conduct an empirical study to explore the ability of trust metrics to distinguish explicit trust from implicit trust and to generate accurate predictions. Experimental results on two real-world data sets show that existing trust metrics cannot provide satisfying performance, and indicate that future metrics should be designed more carefully.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Experimentation

Keywords

Recommender systems, trust metrics, ratings, similarity

1. INTRODUCTION

Trust-based recommender systems [11] is an emerging field to provide users personalized item recommendations (e.g., books, movies, etc.) based on the historical ratings given by users (i.e., user ratings) and the trust relationships among users (e.g., social friends). The intuitions are that users tend to adopt items recommended by trusted friends rather than strangers, and that trust is positively and strongly correlated with user preference. It has been reported that trust-based recommender systems can alleviate many issues besetting traditional systems, such as *data sparsity* and *cold start* [5].

Trust in recommender systems can be broadly classified into two types: explicit and implicit trust. Explicit trust refers to the trust information explicitly specified by users in the systems. For example, users in FilmTrust [3] can directly add others as trusted neighbours. Many explicit trust-based recommender systems have been proposed [3, 11, 5] and their effectiveness has been empirically demonstrated. However, several issues have been observed. First, although specific trust values are possible in real systems, publicly available datasets such as FilmTrust [6] only contain trust links without real values due to the concern of privacy. The indifferent and binary trust will prevent achieving better performance. Second, trust could be noisy in terms of user preference. For example, trusted friends (due to offline relationships, e.g., colleagues) may have different tastes towards movies. Lastly, the amount of trust information is relatively little compared to the number of ratings. Although trust propagation can alleviate this issue to some extent, it is risky to raise new noise *per se* [5].

On the contrary, implicit trust is generally inferred from user behaviors, such as user ratings in our case, rather than specified by users. By analyzing the value that is conveyed via ratings given by users, it is possible to identify the valuable users who are trustworthy and whose ratings are useful for item recommendation. A number of studies have been conducted to interpret such a perspective, such as [13, 9, 16]. As numerical real values, implicit trust can be simply compared and well distinguished with each other. Further, implicit trust is richer than explicit trust since there are a greater number of ratings. However, most existing trust inference approaches (i.e., *trust metrics*) merely focus on the performance of predicting item ratings, and ignore the comparison with explicit trust. In contrast, we claim that it is critical for a trust metric to recover explicit trust as accurate and many as possible using inferred trust values.

In this paper, we first give a definition of trust in recommender systems, and review the four distinct properties of trust. Five representative trust metrics are then introduced and analyzed from those properties. However, we find that none of these metrics can satisfy all the trust properties. In addition, an empirical study of implicit trust is conducted to explore the ability of trust metrics to recover and distinguish from explicit trust, and to generate accurate predictions. Experimental results on two real-world data sets show that the existent trust metrics cannot provide an effective trust list where explicit trust should be ranked higher than

SAC'14 March 24-28, 2014, Gyeongju, Korea.

implicit trust, and that the inferred implicit trust cannot achieve consistent results and improvements across different data sets. Better metrics should be designed more carefully.

2. TRUST IN RECOMMENDER SYSTEMS

In this section, we will first introduce the definition of trust in recommender systems, and review the four aspects of trust. Then, five trust metrics are summarized and discussed to infer implicit trust from user ratings.

2.1 Trust Definition and Properties

Although trust is generally known as a complex and ambiguous concept, in recommender systems it is mostly defined as correlated with similar preferences towards the items commonly rated by two users [15, 13, 12, 9, 16]. For example, users in Citeulike.com can connect to others who have bibliography of interest by adding them to the 'watch list', while users in Epinions.com can specify and put others whose reviews or ratings are consistently valuable into the 'web of trust'. Formally, we adopt the definition given by Guo [4]: "Trust is defined as one's belief towards the ability of others in providing valuable ratings" since our focus is to infer trust from user ratings in this paper.

Trust theory [2] contends that a number of distinct properties can be attributed to trust[1] described as follows.

- **Asymmetry.** Trust is personal and subjective. Even towards a same user, different people may hold various opinions according to their understanding or experience with the target user. Hence, for two users u and v involved in a trust relationship, user u trusting user v cannot guarantee that user v will trust user u to the same extent. It is possible that user v does not trust user u at all. Hence, trust is directed and asymmetric.

- **Transitivity.** An important property of trust that is heavily used in trust-based recommender systems is transitivity. It says if users u trusts v, and v trusts p, it can be inferred that users u trusts p to some extent. It is consistent with real life in which people tend to trust the friend of a friend rather than a stranger. By propagating trust in social networks, we may identify more trusted friends and hence improve the predictive performance of recommender systems.

- **Dynamicity.** In general, trust is built in a continuous way, that is, gradually established and changed over time as more evidences or experience arrive. Trust can be increased with positive evidences and decreased with negative evidences. Another commonplace is that trust is hard to establish but easy to crash. That is, more evidences are needed to form a high trust but few evidence may sufficiently and greatly decrease trust.

- **Context Dependence.** Trust is context-specific in that, for example, a user who is trustworthy in movies may not be trustable in IT technology. One reason is that the accumulated evidences for trust building are contextual. The context in recommender systems refers to the context in which ratings are issued, such as time and location, or the profiles of users and items.

2.2 Trust Metrics

There are many *trust metrics* proposed to calculate implicit trust from user ratings, mainly based on the intuition that the users whose ratings are close to or similar with each other tend to be trustworthy [1]. To facilitate discussion, we introduce a number of notation. Denote U, I and R as the set of all the users, items and ratings, respectively. For simplicity, we keep symbols u, v for users and i, j for items, hence $r_{u,i}$ represents a rating given by user u on item i. Let I_u be the set of items rated by user u, and $t_{u,v}$ be the trustworthiness of user v towards user u. Five trust metrics (denoted by TM1-TM5) are elaborated as follows due to their representativeness and popularity [10].

TM1 Lathia et al. [9] stress the value of providing ratings by other users. In other words, a user who provides even opposite ratings is more trustworthy than the one who is not willing to share opinions. Trust is defined as the average of provided values over all the rated items.

$$t_{u,v} = \frac{1}{|I_{u,v}|} \sum_{i \in I_{u,v}} \Big(1 - \frac{|r_{u,i} - r_{v,i}|}{r_{\max}}\Big), \qquad (1)$$

where $I_{u,v} = I_u \cap I_v$ is the set of items commonly rated by users u and v, and $r_{\max}$ is the maximum rating scale predefined by a recommender system. According to Equation 1, it is trivial to find that trust is symmetric, i.e., $t_{u,v} = t_{v,u}$, and there is no consideration of time and contextual information.

TM2 Papagelis et al. [13] define trust through user similarity computed by Pearson correlation coefficient (PCC):

$$s_{u,v} = \frac{\sum_i (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_i (r_{u,i} - \bar{r}_u)^2}\sqrt{\sum_i (r_{v,i} - \bar{r}_v)^2}}, \qquad (2)$$

where $s_{u,v}$ is the similarity between users u and v, and trust is assigned as similarity, i.e., $t_{u,v} = s_{u,v}$. More general, researchers tend to use a threshold to determine if a user with a certain similarity is trustworthy. For example, Yuan et al. [18] threshold similarity to form binary trust values. Without lack of generality, we propose below formalization:

$$t_{u,v} = \begin{cases} s_{u,v}, & \text{if } s_{u,v} > \theta_s, |I_{u,v}| > \theta_I; \\ 0, & \text{otherwise;} \end{cases} \qquad (3)$$

where θ_s, θ_I are the threshold of the user similarity and the number of co-rated items, respectively. One characteristic of similarity measures is symmetry, i.e., $s_{u,v} = s_{v,u}$. Hence, the trust based on PCC is also symmetric. In addition, Sotos et al. [17] posit that PCC is not transitive unless under strict conditions, that is, $s_{u,v} > 0.707$ when users are highly correlated. Thus, to enable the transitivity of trust, it is necessary to set similarity threshold $\theta_s = 0.707$. Further, Guo et al. [6] report that PCC is not reliable when the length of rating vectors is short, hence the threshold θ_I is to ensure that the computed PCC value is more reliable.

TM3 Hwang and Chen [7] compute a predicted rating using a simple version of Resnick's prediction formula only based on a single user:

$$p_{u,i} = \bar{r}_u + (r_{v,i} - \bar{r}_v), \qquad (4)$$

[1] As a distinct concept, distrust is not covered in this paper.

where $\bar{r}_u$ and $\bar{r}_v$ refer to the mean ratings of users u and v, respectively. The trust score is then derived by averaging the prediction error on co-rated items:

$$t_{u,v} = \frac{1}{|I_{u,v}|} \sum_{i \in I_{u,v}} \Big(1 - \frac{|p_{u,i} - r_{u,i}|}{r_{\max}}\Big) \qquad (5)$$

Shambour and Lu [16] adopt the same strategy, but compute trust based on mean squared distance (MSD):

$$t_{u,v} = \frac{|I_{u,v}|}{|I_u \cup I_v|} \Big(1 - \frac{1}{|I_{u,v}|} \sum_{i \in I_{u,v}} \big(\frac{p_{u,i} - r_{u,i}}{r_{\max}}\big)^2\Big) \qquad (6)$$

The users whose trust value is greater than a threshold λ, i.e., $t_{u,v} > \lambda$ are regarded as trusted neighbours. We note that both Equations 5 and 6 result in symmetric trust. However, both equations do not take into account the dynamic and context property of trust.

TM4 O'Donovan and Smyth [12] regard a rating provided by others as *correct* if the absolute difference between the predicted rating $p_{u,i}$ and the ground truth $r_{u,i}$ is smaller than a threshold ϵ:

$$\text{correct}(r_{u,i}, r_{v,i}) \iff |p_{u,i} - r_{u,i}| \leq \epsilon, \qquad (7)$$

where $p_{u,i}$ is given by Equation 4. Then two kinds of trust are defined using the notion of correctness: profile-level and item-level trust. The former trust is defined as the ratio of correct ratings over all the ratings provided to generate predictions:

$$t_{u,v} = \frac{|CorrectSet(v)|}{|RecSet(v)|}, \qquad (8)$$

where $CorrectSet(v)$ represents the set of correct ratings provided by user v, and $RecSet(v)$ denotes the set of recommendations that user v has involved. The item-level trust is a finer-grain trust of a user for a certain item. To be consistent with other metrics, we only consider the trust at the level of users, i.e., profile-level trust rather than item-level trust. Since the absolute value is adopted in Equation 7, the proposed trust metric is also symmetric. Note that setting a low value of ϵ is not in favor of trust formation whereas setting a high value will tend to treat other users equally trusted.

TM5 Pitsilis and Marshall [15] adopt the *subjective logic* [8] to define trust. In particular, the *uncertainty* is redefined as the inability to make accurate predictions:

$$u_v = \frac{1}{|I_{u,v}|} \sum_{i \in I_{u,v}} \frac{|p_{u,i} - r_{u,i}|}{r_{\max}}, \qquad (9)$$

where u_v is the uncertainty towards user v, and $p_{u,i}$ is derived by Equation 4. Then, they define the *belief* and *disbelief* as:

$$\begin{aligned} b_v &= \tfrac{1}{2}(1 - u_v)(1 + s_{u,v}); \\ d_v &= \tfrac{1}{2}(1 - u_v)(1 - s_{u,v}); \end{aligned} \qquad (10)$$

where $s_{u,v}$ is the similarity between users u and v computed by Equation 2. Hence, it satisfies the requirement for subjective logic: $b_v + d_v + u_v = 1$. The belief b_v is used as the direct trust that user u has on user v, i.e., $t_{u,v} = b_v$. The advantage of this metric is that other than belief, disbelief is also considered. Both disbelief and uncertainty can be involved in trust propagation which is left as a part of our future work.

Table 1: A comparison of different trust metrics in terms of trust properties

Method	Asymm.	Transitive	Dynamic	Context
TM1 [9]	No	Yes	No	No
TM2 [13, 18]	No	Yes, iff $s_{u,v} > \theta_s$	No	No
TM3a [7], TM3b [16]	No	Yes	No	No
TM4 [12]	No	Yes	No	No
TM5 [15]	No	Yes, iff $s_{u,v} > \theta_s$	No	No

2.3 Discussion

A comparison of these five trust metrics is provided in Table 1 from the perspective of trust properties. The table shows that these existing trust metrics are not asymmetric since they are based on similarity or error measures which are symmetric in general. Therefore, all five can be regarded as similarity-based trust metrics; they only differ in the manner of computing rating distances. For the transitivity, the metrics based on the PCC (TM2, TM5) are not transitive unless the PCC value is greater than 0.707. Further, none of the trust metrics explicitly consider the dynamics and context-dependency of trust, though they may be re-defined at a specific time point or in a specific domain. In contrast, recent works on explicit trust such as Peng and Chou [14] have deeply explored context reliance. To capture the dynamics and context-dependency, it is not sufficient to infer trust only based on users' ratings: one must incorporate the contextual information around rating-giving, and users' interactions pertaining to the items. As a conclusion, the current approaches can be recognized as only partially valid in terms of trust properties. New trust metrics are needed to better satisfy the semantics of trust.

More specifically, to achieve the asymmetry of trust, one must involve one-side rating information (e.g, the number of user u's ratings). For the overlapping ratings, alternative similarity measures such as the Bayesian similarity [6] should be used instead of the PCC to ensure the transitivity. The time that a rating is issued may be useful to track the dynamics of user preferences as well as user trust. Fortunately, rating time is usually available in most data sets and real applications. Lastly, ratings can be further classified according to the domains or categories of items, or the attributes of users such that the context information could be involved in trust computation. Further, the potential noise of ratings should be taken into consideration [10]. Current trust metrics are based on a common assumption: the provided ratings are accurate and reflecting users' real preferences. However, in many cases this assumption is not valid.

To sum up, richer information of user ratings should be considered to design a better trust metric. In this paper, we focus on the performance of existing trust metrics rather than the design of a new metric, which will be targeted in our future work.

3. EVALUATION

We conduct a series of experiments to evaluate the effects of different trust metrics in predicting item ratings and ranking trusted users. Our aim is to give an intuitive understanding regarding the different trust metrics and their performance in recommender systems.

3.1 Experimental Settings

Two real-world data sets are used in our experiments, namely FilmTrust and Epinions. They are chosen because both user ratings and explicit trust are available. FilmTrust data set[2] is provided by Guo et al. [6], containing 35,497 movie ratings given by 1508 users ranging from 0.5 to 4.0 with step 0.5. Users can share their ratings and specify other users as trustworthy. In total, there are 1853 trust statements. The Epinions data set[3] is publicly available, including 664,824 ratings issued by 40,163 users on 139,738 items. The ratings are integer from 1 to 5. The items could be electronics, books, sports, etc. Similarly, users can add other users to their 'web of trust'. In particular, there are 487,183 trust statements issued by users. The detailed statistics of data sets is illustrated in Table 2.

Table 2: Summary statistics of the data sets

Data set	users	items	ratings	trust	density
FilmTrust	1508	2071	35497	1853	1.14%
Epinions	40,163	139,738	664,824	487,183	0.05%

In our experiments, five-fold cross validation is used. That is, we split the rating set into five folds and for each experiment, four of them will be used as training test and the left one as test set. Five rounds of executions are conducted such that each fold has been used to test. Then the average performance will be adopted as the final results. We adopt the most general or suggested settings for the parameters in different trust metrics. Specifically, we set $\theta_s = 0.707$ and $\theta_I = 2$ for TM1, $\lambda = 0.15$ for TM3b as suggested by [16], and ϵ for TM4 is set to 0.8 and 1.5 for FilmTrust and Epinions respectively through which the best performance is achieved.

3.2 Evaluation Metrics

In trust-based recommender systems, the effectiveness of trust is often evaluated via the performance of recommendation using inferred trust. However, in this paper, we propose that trust should be useful not only to generate item predictions, but also to suggest reliable users, i.e., to distinguish explicitly defined trust from implicit trust. To the authors' knowledge, our work is the first step to evaluate implicit trust with respect to the explicit trust.

3.2.1 *Metrics for Trust Ranking.*

We contend that the inferred trust should recover the explicit trust as accuracy and fully as possible. The basic idea is that, for a list of users ranked by inferred implicit trust (i.e., *trust list*), the users who are explicitly specified as 'trustworthy' should be ranked higher than those who are not. Therefore, we adopt the widely used Normalized Discounted Cumulative Gain (NDCG) to measure the quality of a ranked trust list. It is computed by:

$$\text{NDCG} = \frac{1}{\text{IDCG}} \times \sum_{i=1}^{n} \frac{2^{rel_i} - 1}{log_2(i+1)}, \tag{11}$$

where rel_i is 1 if the user at position i is relevant (i.e., an explicitly trusted user), and 0 otherwise; n is the length of the trust list; IDCG is the ideal DCG which ensures that the perfect ranking has a NDCG value 1. The second metric is the *Recall*, to measure the average of percentages of explicit trust that can be identified via implicit trust:

$$\text{Recall} = \frac{1}{N} \sum_{u} \frac{|TL_u \cap TN_u|}{|TN_u|}, \tag{12}$$

where TN_u is the set of users explicitly trusted by user u, TL_u is the trust list generated for user u and N is the number of test users who have trust statements. Thus, higher recall indicates more explicitly trusted users recovered.

3.2.2 *Metrics for Rating Prediction.*

On the other hand, since the inferred trust is used to generate recommendations for the active users, the effectiveness of computed trust values can be reflected by the predictive performance in terms of both accuracy and coverage. To have a fair and consistent comparison, we use trust instead of similarity to weigh user ratings when generating predictions for unknown items as used in [11]:

$$\hat{r}_{u,j} = \bar{r}_u + \frac{\sum_{v \in N_u} t_{u,v}(r_{v,j} - \bar{r}_v)}{\sum_{v \in N_u} t_{u,v}}, \tag{13}$$

where $\hat{r}_{u,j}$ is the predicted value for user u on item j, $\bar{r}_u, \bar{r}_v$ are the average of ratings given by users u and v respectively, and N_u is the set of nearest neighbours for user u. We adopt the generally-used mean absolute error (MAE) to measure the accuracy of predictions.

$$\text{MAE} = \frac{\sum_j |\hat{r}_{u,j} - r_{u,j}|}{\kappa}, \tag{14}$$

where κ is the number of predicted ratings. Thus, smaller MAE value means better predictive accuracy. In addition, we use the rating coverage (RC) to measure to what extent the ratings of test items can be predicted:

$$\text{RC} = \frac{P}{M}, \tag{15}$$

where P and M refer to the number of predictable and all the test ratings, respectively. Higher RC value indicates that more items can be recommended to users.

3.3 Results and Analysis

The performance of ranking a user list by inferred trust to show its effectiveness in identifying explicitly defined trust, and that of predicting item ratings using implicit trust to show its effectiveness in making accurate predictions, will be explored in this section.

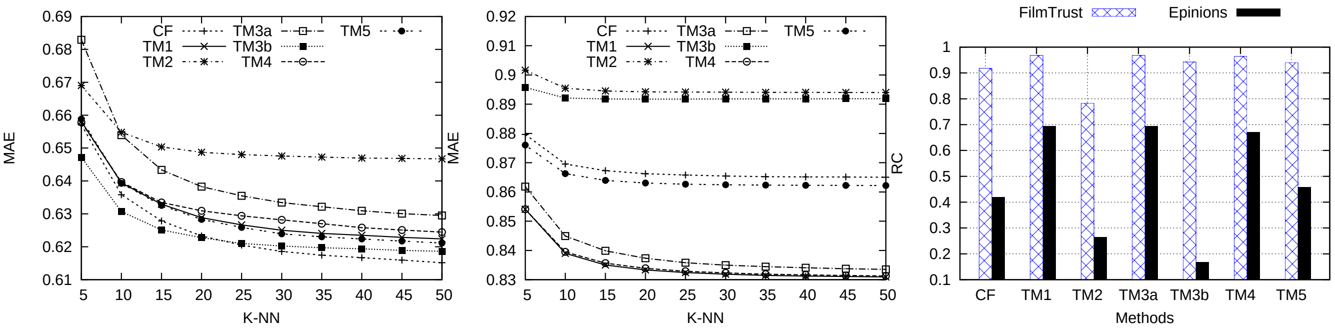
3.3.1 *Performance of Trust Ranking.*

For each method considered, we first compute implicit trust based on their respective formulation, and then rank users in descending order according to the inferred trust. After that, the quality of ranked list is measured using Equations 11 and 12, according to the set of explicitly defined trust. The average of results using the five-fold cross validation is adopted and illustrated in Figure 1.

The results show that the NDCG on both data sets is small, indicating that current methods cannot well distinguish explicit trust from implicit trust and hence there is large room to be improved in the future. When ratings are dense (i.e., on FilmTrust), it is likely more implicit trust can be inferred from user ratings, resulting in a relatively high recall. In contrast, if ratings are sparse (i.e., on Epinions), smaller amount of implicit trust can be computed from user

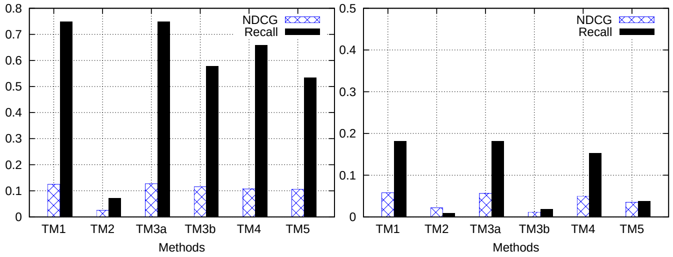
[2] trust.sce.ntu.edu.sg/~gguo1/funcs/filmtrust.zip

[3] www.trustlet.org/wiki/Epinions_datasets



(a) FilmTrust - Accuracy (b) Epinions - Accuracy (c) Rating Coverage

Figure 2: Performance of predicting item ratings



(a) FilmTrust (b) Epinions

Figure 1: The performance of ranking trusted users on FilmTrust and Epinions data sets

ratings, and thus the recall is relatively low. More specifically, across the two data sets, methods TM1, TM3a, TM4 achieve comparative performance in both NDCG and recall due to less constraints (e.g., thresholding), whereas the others are constrained to some extent. TM5 depends on the effectiveness of similarity measure used. TM3b requires a proper setting for parameter λ as illustrated by the great discrepancy in terms of recall on the two data sets. TM2 performs the worst due to the strong constraint of similarity threshold. In conclusion, the current trust metrics are not satisfactory to produce distinguishable trust lists, and may be further limited by the used similarity measures or required thresholds.

3.3.2 Performance of Rating Prediction.

For the purpose of comparison, we adopt the classic user-based collaborative filtering (denoted by CF) [5] as the baseline, where the ratings of the top-K most similar users are aggregated to make a prediction for the active user on a certain unknown item. We vary the values of K from 5 to 50 with step 5. Similarly, for each step, the five-fold cross validation is used and the average performance is recorded. The results are shown in Figure 2.

In addition, we also conduct multiple paired t-tests to investigate the significance of the MAE differences between the various methods and CF. The results are presented in Tables 3 and 4 for the FilmTrust and Epinions data sets, respectively. Note that the column 'Mean Diff.' refers to the average difference of MAEs between two tested methods. For example, 0.0055 in Table 3 is the average MAE difference between TM1 and CF across different values of K-NN. Hence, positive values mean the first method (e.g., TM1) works worse than the second (CF) (and vice versa), and the p-value indicates whether it is statistically significant.

Table 3: Significance tests of the MAE differences relative to the CF method in FilmTrust

Methods	Mean Diff.	df	t	p-value
TM1 – CF	0.0055	9	9.075	7.978e-6
TM2 – CF	0.0258	9	12.926	4.077e-7
TM3a – CF	0.0162	9	14.254	1.756e-7
TM3b – CF	-0.0005	9	-0.357	0.729
TM4 – CF	0.0073	9	7.719	2.943e-5
TM5 – CF	0.0049	9	11.691	9.614e-7

Table 4: Significance tests of the MAE differences relative to the CF method in Epinions

Methods	Mean Diff.	df	t	p-value
TM1 – CF	-0.0324	9	-38.992	2.386e-11
TM2 – CF	0.0276	9	39.747	2.009e-11
TM3a – CF	-0.0284	9	-20.978	5.957e-9
TM3b – CF	0.0248	9	23.531	2.155e-9
TM4 – CF	-0.0320	9	-40.137	1.841e-11
TM5 – CF	-0.0031	9	-37.747	3.190e-11

Based on the empirical results obtained, we see that CF performs significantly better on a dense data set (FilmTrust, see Table 3) than on a sparse data set (Epinions, see Table 4) in comparison with other trust-based approaches. One explanation is the ineffectiveness of traditional similarity measures in cold conditions [6], i.e., computed similarity tends to be misleading and unreliable when the length of rating vectors is short. In contrast, methods TM1, TM3a and TM4 significantly outperform CF on Epinions in terms of both accuracy and coverage, but not on FilmTrust. In other words, trust-based approaches may be more effective if ratings are sparse than the case if ratings are dense. In addition, different approaches have distinct predictive performance. TM2 consistently functions the worst since the largest mean differences are obtained across the two data sets. Although high thresholds ensure more relevant users, the amount of such users could be few. What is worse, the fact that the computed similarity may be error-prone leads to further unreliable predictions and thus bad performance. Interesting to note, methods TM1, TM3a and TM4 are outperformed by methods TM3b and TM5 on FilmTrust in both accuracy and coverage whereas the former methods exceed the latter on Epinions in accuracy. To sum up, there is no single trust metric that is superior to the others across both data sets. For some possibly future trust metric, its effectiveness is necessary to be validated on multiple data sets.

3.3.3 Summary and Discussion.

From the obtained performance of ranking trusted users and predicting item ratings, we find that the evaluation metrics proposed in Section 3.2.1 do not conflict with the traditional metrics. For example, the method giving the worst trust lists also shows the worst performance in rating prediction, while the methods that perform better in terms of NDCG and recall also tend to be more useful in predicting item ratings on some data sets. However, such consistency may not be valid on all other data sets. Using both kinds of metrics give more comprehensive understanding about the quality of inferred trust.

We find that the trust metrics are not yet able to give convincing trust lists where explicit trust ranks higher than implicit trust. More powerful trust metrics are expected to improve the quality of computed implicit trust. To achieve this, explicit trust should be taken into consideration to learn a proper trust metric, especially when there are some parameters that needs to be well tuned. In addition, the investigated trust metrics do not provide consistent utility in predicting item ratings across over different data sets and may not outperform the traditional collaborative filtering technique. Directly depending on traditional similarity measures may prevent trust metrics from obtaining expected properties such as asymmetry and transitivity. Further, the drawbacks of similarity measures could be inherited by and further hinder the effectiveness of trust metrics.

4. CONCLUSIONS AND FUTURE WORK

This paper proposed an empirical study of implicit trust in recommender systems. Five kinds of typical trust metrics were studied in several aspects. First, they are analyzed according to the properties of the concept of trust. We found that none of them can satisfy all the properties. Second, the quality of inferred trust was empirically investigated. Different from the traditional works, we claimed that at least two cases should be considered: to rank trusted users and to predict ratings of unknown items. We proposed two evaluation metrics for trust ranking. The results on two real-world data sets showed that the current trust metrics cannot well distinguish explicit trust from implicit trust. In addition, the predictive performance further demonstrated their inconsistency and ineffectiveness to generate item ratings. Overall, the achievements of current trust metrics are still small, and it is expected to have better trust metric in the future.

We consider four directions for future work. First, besides memory-based approaches, investigation of trust-aware model-based approaches, such as matrix factorization. Second, comparison between explicit and implicit trust in predicting item ratings. Third, investigation of the performance when trust propagation is adopted. Finally, our goal is to develop new trust metrics by considering more fine-grained rating information such as time as well as contextual and interactional information, to better suit trust properties, give more satisfying ranked trust, and have a higher impact on rating predictions.

Acknowledgments

GG thanks the Institute for Media Innovation for a Ph.D. grant at NTU, Singapore. This work is also partially supported by the MoE AcRF Tier 2 Grant M4020110.020. NYS thanks the Operations group at the Judge Business School and the fellowship at St Edmund's College, Cambridge.

5. REFERENCES

[1] A. Abdul-Rahman and S. Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS)*, 2000.

[2] C. Castelfranchi and R. Falcone. *Trust Theory: A socio-cognitive and computational model.* Wiley, 2010.

[3] J. Golbeck and J. Hendler. Filmtrust: Movie recommendations using trust in web-based social networks. In *Proceedings of the IEEE Consumer Communications and Networking Conference (CCNC)*, 2006.

[4] G. Guo. Integrating trust and similarity to ameliorate the data sparsity and cold start for recommender systems. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys)*, 2013.

[5] G. Guo, J. Zhang, and D. Thalmann. A simple but effective method to incorporate trusted neighbors in recommender systems. In *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2012.

[6] G. Guo, J. Zhang, and N. Yorke-Smith. A novel bayesian similarity measure for recommender systems. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

[7] C.-S. Hwang and Y.-P. Chen. Using trust in collaborative filtering recommendation. In *New Trends in Applied Artificial Intelligence.* 2007.

[8] A. Jøsang. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9(03):279–311, 2001.

[9] N. Lathia, S. Hailes, and L. Capra. Trust-based collaborative filtering. In *Trust Management II*, 2008.

[10] N. Lathia, S. Hailes, and L. Capra. The role of trust in collaborative filtering. `http://www0.cs.ucl.ac.uk/staff/l.capra/publications/lathia_recsys_handbook09.pdf`, 2009. online; accessed at Sep 3, 2013.

[11] P. Massa and P. Avesani. Trust-aware recommender systems. In *Proceedings of the 2007 ACM Conference on Recommender Systems (RecSys)*, 2007.

[12] J. O'Donovan and B. Smyth. Trust in recommender systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI)*, 2005.

[13] M. Papagelis, D. Plexousakis, and T. Kutsuras. Alleviating the sparsity problem of collaborative filtering using trust inferences. In *Trust management.* 2005.

[14] T. Peng and S. Chou. iTrustU: a blog recommender system based on multi-faceted trust and collaborative filtering. In *Proceedings of the 24th Annual ACM Symposium on Applied Computing (SAC)*, 2009.

[15] G. Pitsilis and L. Marshall. *A model of trust derivation from evidence for use in recommendation systems.* Technical Report. University of Newcastle upon Tyne, 2004.

[16] Q. Shambour and J. Lu. A trust-semantic fusion-based recommendation approach for e-business applications. *Decision Support Systems*, 54:768–780, 2012.

[17] A. Sotos, S. Vanhoof, W. Van Den Noortgate, and P. Onghena. The transitivity misconception of pearson's correlation coefficient. *Statistics Education Research Journal*, 8(2):33–55, 2009.

[18] W. Yuan, L. Shu, H. Chao, D. Guan, Y. Lee, and S. Lee. itars: trust-aware recommender system using implicit trust networks. *Communications, IET*, 4(14):1709–1721, 2010.