Conference contribution :
Wu, Q., Liu, S., Miao, C., Liu, Y. & Leung, C. (2016). *A Social Curiosity Inspired Recommendation Model to Improve Precision, Coverage and Diversity.* 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), (pp. 240-247). Omaha, NE, USA:
http://dx.doi.org/10.1109/WI.2016.0042

A Social Curiosity Inspired Recommendation Model to Improve Precision, Coverage and Diversity

Qiong Wu*, Siyuan Liu*, Chunyan Miao*, Yuan Liu*† and Cyril Leung*‡
*Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly,
Nanyang Technological University, Singapore
†Software College, Northeastern University, China
‡Department of Electrical and Computer Engineering,
The University of British Columbia, Canada
Email: {wu.qiong, syliu, ascymiao}@ntu.edu.sg,liuyuan@swc.neu.edu.cn,cleung@ece.ubc.ca

***Abstract*—With the prevalence of social networks, social recommendation is rapidly gaining popularity. Currently, social information has mainly been utilized for enhancing rating prediction accuracy, which may not be enough to satisfy user needs. Items with high prediction accuracy tend to be the ones that users are familiar with and may not interest them to explore. In this paper, we take a psychologically inspired view to recommend items that will interest users based on the theory of social curiosity and study its impact on important dimensions of recommender systems. We propose a social curiosity inspired recommendation model which combines both user preferences and user curiosity. The proposed recommendation model is evaluated using large-scale real world datasets and the experimental results demonstrate that the inclusion of social curiosity significantly improves recommendation precision, coverage and diversity.**

I. Introduction

Recommender systems are gaining tremendous popularity in e-commerce sites, such as Amazon and Netflix, due to their effectiveness and efficiency in helping users filter through enormous numbers of items and in helping enterprisers increase their sales. However, traditional recommendation systems only consider the user-item rating information for making recommendations, and omits the abundant social information about users. With the prevalence of social network augmented sites, such as Epinions and Douban, more and more attention is being paid to social recommendations.

In general, existing social recommendation approaches focus on improving rating prediction accuracy by evaluating the similarity among users through social trust [19] or by adding social regularization terms to the objective functions of matrix factorization [18]. However, the quality of recommendations should be evaluated along a number of dimensions and accuracy alone may not be sufficient to meet user satisfaction [20]. For example, the importance of diverse recommendations has been emphasized in several studies [21]. Moveover, it has been shown that high accuracy may be secured by recommending users with the most popular items to users [2]. However, it is very likely that users will have learnt about such items from multiple sources, including advertisements, news, or friends. Recommending these items may not interest users as they already knew them. Therefore, it is important for a recommender system to be able to discover items that can truly elicit users' interests to explore.

In this work, we take a psychologically inspired view to recommend items that interest users and explore the potential impact on important dimensions of recommender systems. In human psychology, a person's feeling of interestingness is closely related to curiosity, a quality related to inquisitive thinking such as exploration, investigation, and learning [23]. Curiosity is generally not centered around a person's preferences; it is more focused on the unexpectedness in the environment. In real life, a person often gets curious about the surprising behaviors of his/her friends. For example, as illustrated in Figure 1, if Alice knows that her friend Bob hates horror movies, the incidence of Bob giving a high rating to a horror movie (e.g., House of wax) will likely catch Alice's attention. In order to find out why Bob gave this surprising rating, Alice may be driven by curiosity to watch this horror movie. This phenomenon is generally known as *social curiosity* [22][25], which is the desire to acquire new information about how other people behave, think, and feel.

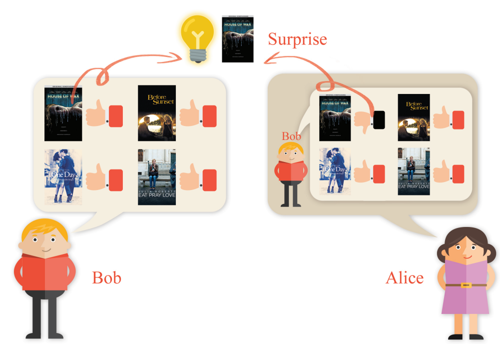


Figure 1: A real world example for social curiosity

Motivated by the above, we propose a social curiosity inspired recommendation model to recommend items that users may be interested in. More specifically, we propose a model for measuring user curiosity in the social recommendation context. This model takes into consideration the different responses given by a user to different friends' surprising ratings. Three strategies are proposed to evaluate users' curiosity when multiple friends give surprising ratings for the same item. After that, users' interests are evaluated by combining their preferences and curiosity. Finally, items are recommended based on a descending order of interest scores. We conduct experiments based on two large-scale real world datasets. The experimental results show that the incorporation of social curiosity information significantly improves the precision, coverage, and diversity of recommender systems.

II. Related Works

A. Traditional vs. Social Recommendation

Traditional recommender systems make use of the user-item rating information for recommendation. One popular idea is Collaborative Filtering (CF), which recommends items based on the similarities between users or items. CF approaches can be classified into two main categories: heuristics-based approaches and model-based approaches. Heuristics-based approaches utilize similar users or items to generate predictions, which can be further categorized into user-based [7] and item-based [15] approaches. In contrast, model-based approaches use the observed ratings to train a predictive model, typically through statistical or machine-learning methods, which is then used to predict ratings. One of the model-based approaches, i.e., Matrix Factorization, has recently gained popularity in recommender system applications due to their high recommendation accuracy and their efficiency in dealing with large-scale user-item rating matrices.

Recent research in social recommendation utilizes the social relationships among users to improve the recommendation accuracy [11], [18], [19], [24]. Social recommendation approaches can also be classified into heuristics-based approaches and model-based approaches. Heuristics-based approaches usually measure the similarity between two users using the degree of social trust [19]. It is shown to increase the number of predictable ratings without decreasing the overall accuracy. Model-based approaches usually make use of social information to constrain the matrix factorization objective function. In [18], Ma et al. propose social regularization terms to enhance the performance of the matrix factorization algorithm. Existing social recommendation methods mainly consider two types of social information: trust and friend preference similarity. In this work, we introduce another dimension of social information, social curiosity, and study its impact on recommendations.

B. Accuracy vs. Measures Beyond Accuracy

A typical recommender system attempts to estimate ratings of items accurately based on users' rating history. Accordingly, many researchers focus on improving the recommender systems' rating prediction accuracy which can be measured by Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) [16], and accuracy for recommendation lists which can be measured by precision and recall [4]. However, accuracy alone may not be enough to meet users' satisfaction and in recent years, researchers have shown a growing interest in studying other aspects beyond accuracy.

Some researchers have pointed out that recommender systems should provide users with highly idiosyncratic and personalized items [2]. With this goal in mind, a lot of works have been proposed to increase the diversity of recommendation lists, often measured by the dissimilarity between all pairs of recommended items [21][3]. On the other hand, algorithms striving for high accuracy often provide recommendations with high quality, but only for a small number of items [8]. To address this issue, some researchers propose to improve recommendation coverage, which refers to the percentage of items for which a recommender system is able to generate recommendations [6][14][1].

Recent research works have also studied factors such as novelty and serendipity. Novel recommendations are defined as recommendations of items that are interesting but unknown to the users. In [9], the system explicitly asks users what items they know to derive novel recommendations in a CF framework. However, novelty only emphasizes the fact that an item is unknown; it does not consider items that are known but unexpected. Serendipity introduces the concept of unexpectedness and measures how surprising the recommendations are. In [10], serendipity is enhanced in a content-based recommender system by recommending items whose description is semantically far from users' profiles. In [12], the surprise of each user is estimated when presented with recommendations by predicting their purchasing trend based on the purchase history of users with similar preferences. However, the concept of serendipity only considers the unexpectedness with respect to a person's own historical preferences. It does not consider the unexpectedness with respect to other people's historical preferences in a social context.

III. The Proposed Recommendation Model

An overview of the proposed recommendation model is shown in Figure 2. The model takes all users' historical ratings and their friend relationships as input. Then, it models the target user's preferences and curiosity, the combination of which forms user interests. Based on user interests, the top ranked items are output as the recommendation list. Next, we will discuss each component of the model in detail.

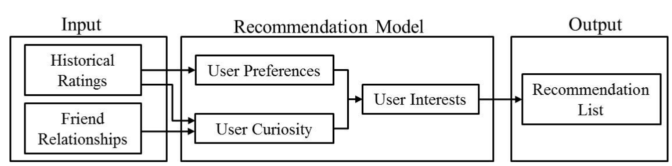


Figure 2: Overview of the proposed recommendation model

A. User Preferences: Predicted Rating Score

Various recommendation methods have been proposed to predict user preferences for items that are not experienced, e.g., neighborhood based collaborative filtering [7] and matrix factorization [13], [17]. In this work, we adopt the Matrix Factorization (MF) method for measuring user preferences, due to its high recommendation accuracy and its efficiency when dealing with large-scale user-item rating matrices. It should be noted that our model could use a different prediction method, if such a method would yield a more accurate result. In other words, our model is not dependent on MF.

The basic MF model maps both users and items to a joint latent factor space of dimensionality d, such that user-item interactions are modeled as inner products in that space. Accordingly, each user u is associated with a vector $p_u \in \mathbb{R}^d$, where the elements in p_u measure the user's preferences with respect to d latent factors. Each item i is associated with a vector $q_i \in \mathbb{R}^d$, where the elements in q_i measure the item's importance weights for the d latent factors.

The preference of target user u towards a not experienced item i is measured by a *predicted rating score*, denoted by $\hat{R}(u,i)$, which is obtained by:

$$\hat{R}(u,i) = p_u^T q_i. \tag{1}$$

The values in p_u and q_i are initially assigned arbitrarily and then iteratively updated by a simple gradient descent technique. For each observed rating $R(u,i)$, the latent variable vectors p_u and q_i are updated as follows:

$$p_u \leftarrow p_u + \gamma(\Delta_{ui} \cdot q_i - \lambda \cdot p_u), \tag{2}$$

$$q_i \leftarrow q_i + \gamma(\Delta_{ui} \cdot p_u - \lambda \cdot q_i), \tag{3}$$

where

$$\Delta_{ui} = R(u,i) - p_u^T q_i. \tag{4}$$

Here, γ is the learning rate and λ is a regularization parameter to minimize overfitting. The algorithm iterates until an accuracy threshold is reached.

B. User Curiosity: Curiosity Score

According to the theory of curiosity in human psychology [25], surprise is one of the key factors that stimulate curiosity and is the result of unexpectedness. This can be readily applied to the social context: a friend's unexpected behaviors create a feeling of surprise which will then lead to curiosity. Based on this theory, a user u will be surprised if a friend v's rating for an item i significantly differs from u's expectation of v's preference towards i.

The target user's expectation of a friend's preference towards an item can be estimated by a predictive model, e.g., MF, that is trained without the friend's observed rating for this item. To differentiate this idea from the traditional recommender systems in which predictions are usually made for items that have no observed ratings, we define the *pseudo-predicted rating* of a user v for an item i, if an observed rating $R(v,i)$ exists:

$$\check{R}(v,i) = \check{p}_v^T \check{q}_i, \tag{5}$$

where $\check{p}_v$ and $\check{q}_i$ are pseudo-latent vectors trained from all the available user-item ratings except $R(v,i)$. The pseudo-predicted rating $\check{R}(v,i)$ is an estimation of the target user u's expectation of his/her friend v's preference towards item i.

Surprise occurs in two cases: when $R(v,i)$ is much larger than $\check{R}(v,i)$, or when $R(v,i)$ is much smaller than $\check{R}(v,i)$. The former case corresponds to a friend giving a high rating to something he/she generally does not prefer, i.e., positive surprise, whereas the latter case corresponds to a friend giving a low rating to something he/she generally likes, i.e., negative surprise. In real life, people will not give importance to the negative surprise during a recommendation task because a person's taste in a field he/she likes is often trusted by his/her friends. For example, if Bob loves comedy movies and gives a low rating to a particular comedy movie, Alice will likely ignore this movie because she trusts Bob's judgement on comedies.

Hence, we focus on positive surprise and define the surprise caused by a friend v's rating for an item i, denoted by $S(v,i)$, as follows:

$$S(v,i) = \begin{cases} R(v,i) - \check{R}(v,i), & \text{if } R(v,i) - \check{R}(v,i) > T_E \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

and

$$T_E = \frac{\sum_{u,i}(R(u,i) - \hat{R}(u,i))^+}{N^+} \tag{7}$$

where T_E is an error threshold calculated by the mean of the positive predicted rating errors. It should be noted that the difference between the observed rating and the pseudo-predicted rating forms surprise, whereas the difference between the observed rating and the predicted rating is error. $(R(u,i) - \hat{R}(u,i))^+$ represents ratings that satisfy $R(u,i) > \hat{R}(u,i)$ and N^+ denotes the number of such ratings. In this way, we rule out the surprises caused by the innate errors of the MF algorithm.

As MF has been shown to yield fairly accurate predictions, it is expected that an item's pseudo-predicted rating $\check{R}(v,i)$ should be similar to the observed rating $R(v,i)$. In other words, $S(v,i)$ should usually be smaller than the error threshold. If $S(v,i)$ is much larger than the error threshold, it is most likely due to an unexpected rating behavior that deviates from v's usual rating patterns. For example, for Bob, who always gives low ratings to horror movies, MF tends to predict low ratings for such movies. If one day Bob gives a high rating to a horror movie, then this rating cannot be accurately predicted by the previously trained MF model, and results in a high surprise score for this particular movie.

In this work, we focus on the surprises from directly linked friends. In social networks, most users have many friends and it is possible that multiple friends give surprising ratings for the same item. Hence, to model the target user's curiosity, two issues should be addressed: (1) how to model a user's responses to the surprising ratings from different friends and (2) how to evaluate curiosity when multiple friends give surprising ratings for a given item.

To address the first issue, we propose a *surprise correlation* between two users who are mutual friends, denoted by $SC(u,v)$, as follows:

$$SC(u,v) = 1 - \frac{\sum_{i \in M(u,v)}(\frac{abs(S(u,i) - S(v,i))}{R_m})}{|M(u,v)|}, \tag{8}$$

where $M(u,v)$ denotes the set of items that both u and v have given surprising ratings to, $|\cdot|$ is the cardinality operator, abs is the absolute value operator, and R_m is the maximum rating scale difference for normalization (e.g., if the rating scale is from 1 to 5, then $R_m = 4$). Note that $0 \leq SC(u,v) \leq 1$. A high value for $SC(u,v)$ indicates that u's historical surprising ratings tend to be similar to v's. Hence, the surprise correlation can be used to predict whether the target user will also be surprised at the items that surprise his/her friend.

To address the second issue, we propose three strategies for evaluating the target user's curiosity towards an item i, denoted by the *curiosity score* $C(u,i)$, as follows:

$$C_1(u,i) = \min_{v \in F_s(u)} (SC(u,v) \cdot S(v,i)), \tag{9}$$

$$C_2(u,i) = \operatorname*{ave}_{v \in F_s(u)} (SC(u,v) \cdot S(v,i)), \tag{10}$$

$$C_3(u,i) = \max_{v \in F_s(u)} (SC(u,v) \cdot S(v,i)), \tag{11}$$

where $F_s(u)$ is the set of u's friends who give surprising ratings for item i. Equation (9) represents a *conservative* strategy: it takes the minimum curiosity response aroused by the friends' surprising ratings. Equation (10) represents an *average* strategy: it takes the average curiosity response aroused by the friends' surprising ratings. Equation (11) represents a *bold* strategy: it takes the maximum curiosity response aroused by the friends' surprising ratings.

C. User Interests: Interest Score

A curiosity-stimulating item may not catch a user's attention unless the user is in favor of that item to some extent. For example, the fact that Bob gives a high rating to a horror movie may not interest Alice if she also hates horror movies. Hence, we propose an interest score to balance the predicted rating score and the curiosity score for the recommended items.

The *interest score* of a user u for an item i is modeled by a weighted sum of the predicted rating score and the curiosity score:

$$I(u,i) = (1-\omega) \cdot \hat{R}(u,i) + \omega \cdot C(u,i), \omega \in [0,1] \tag{12}$$

where ω is the weight for balancing between the predicted-rating score and the curiosity score. A higher value of ω indicates more consideration of curiosity score for recommendation and hence the recommendation result will incorporate more effects brought by the social curiosity information. On the other hand, a lower value of ω will result in a recommendation list more similar to the one based purely on predicted ratings (user preferences). It should be noted that our method do not affect the value of predicted ratings but combines them with the social curiosity information strategically to make recommendations. Finally, items are recommended based on a descending order of interest scores.

IV. Metrics

We now discuss the metrics for evaluating the proposed recommendation model. As this model makes use of users' preferences without affecting its accuracy for predicted ratings, it does not impact the commonly used accuracy metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Therefore, we study the accuracy for recommendation lists through precision. We also explore two metrics beyond accuracy, including coverage and diversity.

Precision: Precision is usually defined as the percentage of items in the recommended list that have been rated by the target user in the testing set [4], given by

$$Precision(L(u)) = \frac{|\{i | i \in L(u) \wedge R(u,i) \in R_T(u)\}|}{|L(u)|} \times 100\% \tag{13}$$

where $L(u)$ is the set of items recommended to a user u and $R_T(u)$ is the set of ratings given by u in the testing set. The system-level precision is calculated as the average precision of the recommendation lists for all users. Here, a historical rating for an item can be treated as an evidence showing that the user got interested and indeed explored the item. A higher precision indicates a higher chance that the user will explore the recommended items and finally give ratings for them. Therefore, precision can reflect the extent of users' interest on the recommended items.

Coverage: The coverage of a recommender system is a measure of the domain of items in the system over which the system can make recommendations and evaluates a recommender system's ability to recommend long-tail items [8]. The coverage metric is given by:

$$Coverage = \frac{|\bigcup_{u \in U} L(u)|}{N} \times 100\% \tag{14}$$

where $\bigcup$ is the union operator and N is the total number of items in the recommender system.

Diversity: Diversity evaluates a recommender system's ability to provide users with highly idiosyncratic or personalized items and is usually defined as the average dissimilarity between all pairs of items [5]. The dissimilarity between two items is determined by the dissimilarity between ratings given to them by the set of users who have rated both. Let $U(i,j)$ represent the set of users who have rated both items i and j. The dissimilarity between i and j is given by:

$$dis(i,j) = \frac{\sum_{u \in U(i,j)} (\frac{abs(R(u,i)-R(u,j))}{R_m})}{|U(i,j)|}, \tag{15}$$

where R_m is the maximum rating scale difference.

The diversity of a recommendation list is given by:

$$diversity\,(L(u)) = \frac{\sum_{i \in L(u)} \sum_{j \in L(u), i \neq j} (dis(i,j))}{\frac{|L(u)|}{2}(|L(u)|-1)}. \tag{16}$$

The system-level diversity is calculated as the average diversity of the recommendation lists for all users.

V. EXPERIMENTS

Two publicly available datasets, Douban and Flixster, were used in our experiments to study the performance of the proposed recommendation model with respect to precision, coverage, and diversity. Both datasets include user-item ratings as well as the social network connecting different users. The Douban dataset [18] contains 129,490 distinct users, 58,541 distinct movies and 16,830,839 ratings. The social network contains 1,692,952 undirected friend links between users. Each user gives an average of 129.98 ratings and each item receives an average of 287.51 ratings. The average number of friends per user in the social network is 13.07. The Flixster dataset [11] contains 1,049,508 distinct users, 66,726 distinct movies and 8,196,077 ratings. The social network contains 7,058,819 undirected friend links between users. We preprocessed the Flixster dataset by removing the large portion of users who have social relations but no expressed ratings because our approach is interested in friends' rating behaviors. After preprocessing, the average number of ratings given per user is 55.52, and the average number of ratings received per item is 122.83. The average number of friends per user in the social network is 17.20.

A. Methods and Parameter Settings

We compare our method with the baseline MF method and various ranking methods:

1) **MF:** the baseline MF method [13].
2) **PopR**: the item popularity based ranking, which ranks items whose ratings are predicted above the ranking threshold T_H based on their popularity [2].
3) **AbsLikeR**: the item absolute likeability based ranking, which ranks items whose ratings are predicted above the ranking threshold T_H based on how many users liked them (i.e., rated the item above T_H) [2].
4) **RelLikeR**: the item relative likeability based ranking, which ranks items whose ratings are predicted above the ranking threshold T_H based on the percentage of the users who liked them (i.e., rated the item above T_H) [2].
5) **SC_Min**: the proposed recommendation model with the $conservative$ strategy for curiosity evaluation.
6) **SC_Ave**: the proposed recommendation model with the $average$ strategy for curiosity evaluation.
7) **SC_Max**: the proposed recommendation model with the $bold$ strategy for curiosity evaluation.

For the MF parameters, we set the learning rate γ to 0.001, the regularization parameter λ to 0.02, and the latent factor dimension d to 10. For the ranking methods, the ranking threshold T_H is set to 4.5, which ensures that the recommended items are generally preferred by the users. The selection strategy for the interest score weight ω will be discussed in the following subsection. All the results are obtained based on top 10 recommendations.

B. Impact of Interest Score Weight: ω

In this experiment, we empirically study the impact of the interest score weight ω. This weight provides a balance between the predicted-rating score and the curiosity score when making recommendation decisions. A higher value of ω indicates that the proposed recommender system gives more importance to the curiosity score. On the other hand, a lower value of ω leads to a more similar list to the baseline method that uses only the predicted rating score for ranking.

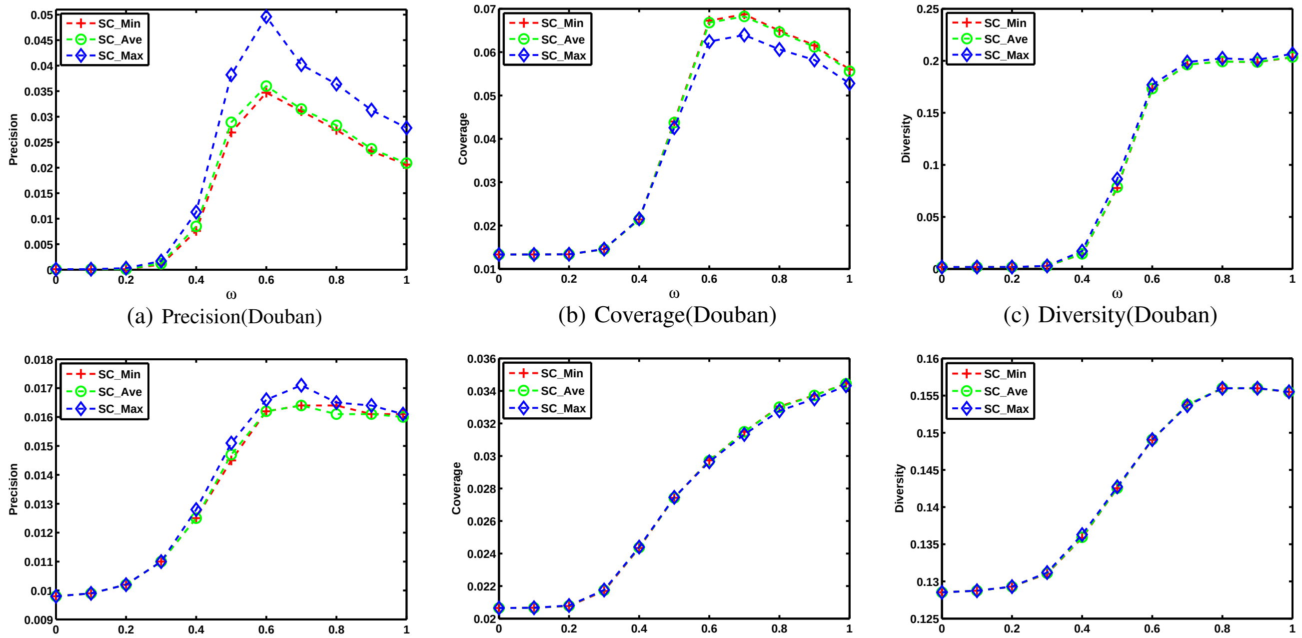
In order to find a good value for ω, we conducted a grid search for both Douban and Flixster datasets from 0 to 1, with an interval of 0.1. For each dataset, we use different training data settings (40%, 60%, 80%) for grid search. A training data setting of 40%, for example, means that we randomly selected 40% of the ratings from the user-item rating matrix as the training data to predict the remaining 60% of ratings. Similar experimental settings have been used in previous works for social recommendation [18]. Due to space limitation, we present here the result for a training data setting of 60%. Similar trends were observed with the other two training settings (40% and 80%). The experimental results are shown in Figure 3.

From Figure 3, it can be observed that the best value for ω differs for different metrics. Taking the Douban dataset for example (Figure 3a, 3b, 3c), the best performing ω for precision is 0.6, for coverage is 0.7, and for diversity is 1. This means that the impacts of social curiosity on the three metrics are not equally strong and the positive influence of social curiosity on diversity is highest when compared with the other two metrics. When ω equals to 0, the results are obtained based on the predicted rating score alone. When ω equals to 1, the results are obtained based on the curiosity score alone. It can be observed from Figure 3 that the values for all the metrics at ω=1 are always higher than those at ω=0. In other words, *the results obtained based on the curiosity score alone are always better than those obtained based on the predicted rating score alone*. Nevertheless, in most of the cases, a combination of the two scores achieves the best results. Taking the precision metric for the Douban dataset as an example (Figure 3a), the best precision value achieved based on the curiosity score alone is 0.0279 (at $\omega = 1$) and the best precision value achieved based on the predicted rating score alone is 2×10^{-4} (at $\omega = 0$). The best precision value is 0.0496, which is obtained based on the combination of both the curiosity score and the predicted rating score with the weight $\omega = 0.6$.

Based on the above observations, in the following experiments, for each training setting in each dataset, we choose ω based on the best performing precision value for all three methods. For example, for Douban dataset with 60% data for training, we set ω to 0.6, according to Figure 3a.

C. Performance Comparison

For each dataset, we use three training data settings (40%, 60%, 80%) to study the performance of the proposed approach. The experimental results are shown in Table I. The best performance for each setting is highlighted in bold. From Table I, we can observe that the baseline method (**MF**) and the ranking methods (**PopR**, **AbsLikeR** and **RelLikeR**) all



(a) Precision(Douban) (b) Coverage(Douban) (c) Diversity(Douban)

(d) Precision(Flixster) (e) Coverage(Flixster) (f) Diversity(Flixster)

Figure 3: Impact of ω on Algorithm Performance

show comparatively poorer performances for all three metrics (precision, coverage and diversity) than the proposed social curiosity methods. Taking the Douban dataset with 40% data for training as example, the minimum precision value achieved by the social curiosity methods is 0.0545, which is two orders of magnitude higher than the baseline method as well as the ranking methods. Similarly, for the Flixster dataset, the social curiosity methods are able to achieve precision values an order of magnitude higher than the ranking methods. Hence, the experimental results show that by considering social curiosity, we are able to recommend a lot more items that users rated in the testing set, which are items that users felt interested in and indeed experienced with. This suggests that the recommended items are more likely to interest users.

One phenomenon worth noticing is that precision gets generally worse as more training data is used. The possible reason is as follows. Though when more training data is used, the accuracy for predicting interesting items should improve, it also means that less testing data is available, which makes it harder to match predicted items with the testing items. Therefore, precision becomes worse because the impact of less testing data is stronger than the impact of more training data.

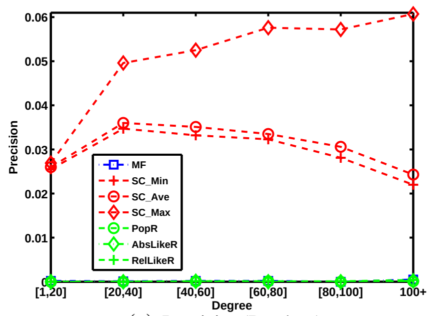
We next compare the performances of the three proposed strategies for measuring curiosity, **SC_Max**, **SC_Ave** and **SC_Min**. It can be observed that the conservative strategy consistently achieves the highest coverage value whereas the bold strategy consistently achieves the lowest coverage value. This observation can be explained as follows. At the system level, the bold strategy always chooses the most surprising items for recommendation. If the user who gives the highest surprising ratings to these items has a large number of friends, then these items will be repetitively recommended to all of his/her friends. Hence, due to the repetitive recommendation of the most surprising items, the coverage of the bold strategy tends to be smaller than that of the conservative strategy. On the other hand, the bold strategy consistently achieves the highest precision value and diversity value, whereas the conservative strategy consistently achieves the lowest precision value and diversity value. The reason for the bold strategy to achieve the best precision values is that more surprising items may be more interesting to the target users, and thus the users may be more likely to explore these items. Moreover, the surprising items tend to be very dissimilar to those that users preferred in the history, and the repetitive recommendation of the most surprising items may instill higher diversity into the recommendation lists.

D. Impact of Friend Degree

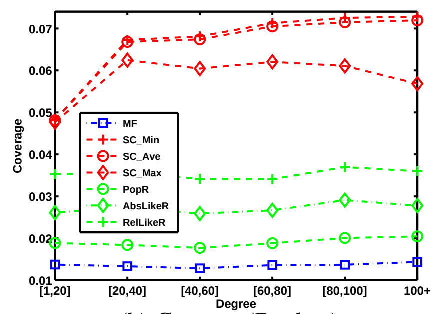
As the proposed method is closely related to the behavior of the target user's friends, it is useful to analyze the impact of the number of friends, i.e., friend degree, on the recommendation results. In our experiments, we separate users into 6 degree groups: the first group consists of users with degrees from 1 to 20, the second group from 21 to 40, the third group from 41 to 60, the fourth group from 61 to 80, the fifth group from 81 to 100, and the sixth group above 100. We do not continue dividing users with degree above 100 into groups because users with degree above 120 represent less than 1% of the total number of users in the dataset. Since the number of users decreases in larger degree groups, to make the comparisons fair, we randomly select 1000 users from each degree group for reporting the coverage value.

Table I: Performance Comparison on Douban and Flixster datasets.

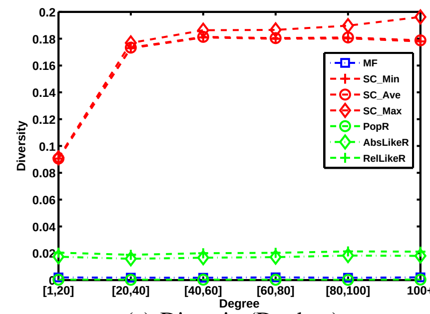
Dataset	Training	Metrics	Baseline	Ranking Methods			Social Curiosity Methods		
			MF	**PopR**	**AbsLikeR**	**RelLikeR**	**SC_Min**	**SC_Ave**	**SC_Max**
Douban	40%	Precision	0.0002	1×10^{-5}	0.0002	0.0006	0.0545	0.0554	**0.0618**
		Coverage	0.0167	0.0235	0.0306	0.0400	**0.0636**	0.0633	0.0615
		Diversity	0.0018	0.0003	0.0051	0.0093	0.1527	0.1531	**0.1555**
	60%	Precision	0.0001	1×10^{-5}	1×10^{-5}	1×10^{-5}	0.0347	0.0360	**0.0496**
		Coverage	0.0133	0.0184	0.0274	0.0356	**0.0673**	0.0668	0.0624
		Diversity	0.0018	0.0002	0.0158	0.0188	0.1729	0.1733	**0.1768**
	80%	Precision	0.0001	0.0001	0.0001	0.0001	0.0104	0.0110	**0.0157**
		Coverage	0.0163	0.0194	0.0264	0.0317	**0.0891**	0.0885	0.0839
		Diversity	0.0095	0.0013	0.0257	0.0333	0.1649	0.1656	**0.1706**
Flixster	40%	Precision	0.0124	0.0078	0.0081	0.0098	0.0202	0.0204	**0.0205**
		Coverage	0.0232	0.0279	0.0287	0.0280	**0.0301**	0.0301	0.0301
		Diversity	0.1096	0.0821	0.0831	0.1048	0.1217	0.1217	**0.1217**
	60%	Precision	0.0098	0.0065	0.0066	0.0080	0.0164	0.0164	**0.0171**
		Coverage	0.0207	0.0255	0.0265	0.0264	**0.0315**	0.0315	0.0313
		Diversity	0.1285	0.0995	0.1007	0.1263	0.1536	0.1538	**0.1538**
	80%	Precision	0.0052	0.0037	0.0037	0.0044	0.0103	0.0104	**0.0104**
		Coverage	0.0156	0.0206	0.0210	0.0209	**0.0297**	0.0296	0.0293
		Diversity	0.1229	0.0956	0.0967	0.1236	0.1625	0.1625	**0.1626**



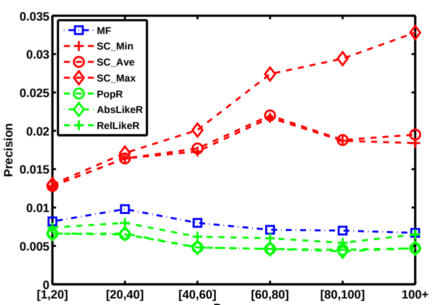
(a) Precision(Douban)



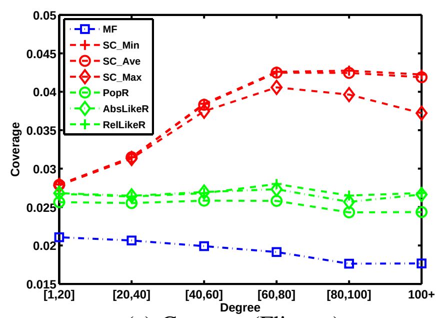
(b) Coverage(Douban)



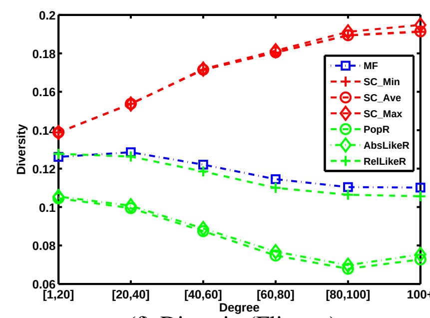
(c) Diversity(Douban)



(d) Precision(Flixster)



(e) Coverage(Flixster)



(f) Diversity(Flixster)

Figure 4: Impact of Friend Degree on Algorithm Performance

The experimental results are shown in Figure 4. These results are obtained using 60% of data for training. Similar trends were observed with the other two training settings (40% and 80%). Due to space limitation, we present here the results for the 60% training data setting. It can be observed from Figure 4a and 4d that the social curiosity methods consistently outperform all the other methods for precision across all the degree groups for both datasets. It can be clearly seen from Figure 4d that the ranking methods even perform worse than the baseline method for precision in Flixster dataset. It confirms the superior performance of the social curiosity methods on recommending interesting items that users may want to explore and finally give ratings to. Figure 4b and 4e show that the social curiosity methods consistently outperform all the other methods for coverage across all the degree groups for both datasets. Figure 4c and 4f show that the social curiosity methods consistently outperform all the other methods for diversity across all the degree groups for both datasets. By comparing Figure 4c and 4f, it can be also seen that the performance of ranking methods on diversity depends on the datasets being used. For example, the **AbsLikeR** and **RelLikeR** perform better than the baseline method in terms of diversity in Douban dataset but perform worse in Flixster dataset. In summary, the experimental results demonstrate that the social curiosity methods show overall robust performance for all three metrics in different degree groups.

Let us analyze the impact of degree on the recommendation results. It can be observed from Figure 4 that there is a clear trend for **SC_Max**: a larger degree tends to achieve a higher value for precision and diversity. However, no such trend is observed for the baseline method or the ranking methods. The performance of the baseline method and the ranking methods stays similar for all the degree groups. This is due to the fact that the baseline method and the ranking methods do not explicitly consider social information during recommendation. Another interesting phenomenon that can be observed is that for all the three metrics, the conservative strategy **SC_Min** and the average strategy **SC_Ave** always yield similar results, whereas the bold strategy **SC_Max** diverges further from the other two strategies as degree increases. This is due to the fact that the high-value surprises are rare. The effect of these high-value surprises is diluted when averaged with the more abundant low-value surprises through the average strategy, leading to a result similar to the conservative strategy. However, these rare high-value surprises are more likely to spread as the degree increases, thus magnifying the effects of these high-value surprises on the recommendation results. Similarly, when degree is large, the rare high-value surprises may be diluted by the abundant low-value surprise information available for the target user, which explains why **SC_Ave** and **SC_Min** tend to show a downward trend when degree increases in Figure 4a and 4d.

The experimental results empirically show that the proposed approach greatly enhances recommendation precision, diversity and coverage. From an intuitive point of view, the proposed approach can achieve such improvements because it not only relies on the similarity information as in traditional recommendation techniques but also considers the "surprise" element in the recommendation. The incorporation of such surprise information allows items that may not best match the user's usual preferences but arouse the user's interest to be ranked higher. Therefore, the recommendation precision, diversity, and coverage can be significantly improved.

VI. Conclusion

This paper takes a psychologically inspired view to recommend interesting items to users. Motivated by the close relation between curiosity and interest, we propose a social curiosity inspired recommendation model. This recommendation model evaluates the interestingness of an item through the combination of both users' personal preferences and their social curiosity. The proposed recommendation model takes into consideration the target user's different possible responses to different friends' surprising unexpected ratings. Three strategies were proposed for evaluating the target user's curiosity when multiple friends give surprising ratings for the same item. We also explored the impact of social curiosity on enhancing various evaluation metrics with large-scale real world datasets. The experimental results show that the proposed recommendation model significantly enhances precision, coverage and diversity in comparison with other state-of-the-art methods.

VII. Acknowledgments

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative.

References

[1] P. Adamopoulos and A. Tuzhilin. On unexpectedness in recommender systems: Or how to expect the unexpected. *RecSys*, 11:11–18, 2011.

[2] G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, 2012.

[3] A. Ashkan, B. Kveton, S. Berkovsky, and Z. Wen. Optimal greedy diversity for recommendation. *IJCAI*, pages 1742–1748, 2015.

[4] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez. Recommender systems survey. *Knowledge-Based Systems*, 46:109–132, 2013.

[5] K. Bradley and B. Smyth. Improving recommendation diversity. *In Proceedings of International Conference on Artificial Intelligence and Computer Science*, pages 75–84, 2001.

[6] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. *RecSys*, pages 257–260, 2010.

[7] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering. *SIGIR*, pages 230–237, 1999.

[8] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[9] Y. Hijikata, T. Shimizu, and S. Nishida. Discovery-oriented collaborative filtering for improving user satisfaction. *IUI*, pages 67–76, 2009.

[10] L. Iaquinta, M. De Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino. Introducing serendipity in a content-based recommender system. *HIS*, pages 168–173, 2008.

[11] M. Jamali and M. Ester. A transitivity aware matrix factorization model for recommendation in social networks. *IJCAI*, pages 2644–2649, 2011.

[12] N. Kawamae. Serendipitous recommendations via innovators. *SIGIR*, pages 218–225, 2010.

[13] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[14] G. Lekakos and P. Caravelas. A hybrid approach for movie recommendation. *Multimedia Tools and Applications*, 36(1-2):55–70, 2008.

[15] M. Li, B. Dias, W. El-Deredy, and P. J. Lisboa. A probabilistic model for item-based recommender systems. *RecSys*, pages 129–132, 2007.

[16] H. Liu, Z. Hu, A. Mian, H. Tian, and X. Zhu. A new user similarity model to improve the accuracy of collaborative filtering. *Knowledge-Based Systems*, 56:156–166, 2014.

[17] X. Luo, Y. Xia, and Q. Zhu. Incremental collaborative filtering recommender based on regularized matrix factorization. *Knowledge-Based Systems*, 27:271–280, 2012.

[18] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King. Recommender systems with social regularization. *WSDM*, pages 287–296, 2011.

[19] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. *On the Move to Meaningful Internet Systems*, pages 492–508, 2004.

[20] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. *CHI*, pages 1097–1101, 2006.

[21] L. Qin and X. Zhu. Promoting diversity in recommendation by entropy regularizer. *IJCAI*, pages 2698–2704, 2013.

[22] B. Renner. Curiosity about people: The development of a social curiosity measure in adults. *Journal of Personality Assessment*, 87(3):305–316, 2006.

[23] P. J. Silvia. Interest – The curious emotion. *Current Directions in Psychological Science*, 17(1):57–60, 2008.

[24] P. Symeonidis, E. Tiakas, and Y. Manolopoulos. Product recommendation and rating prediction based on multi-modal social networks. *RecSys*, pages 61–68, 2011.

[25] Q. Wu and C. Miao. Curiosity: From psychology to computation. *ACM Computing Surveys*, 46(2):18, 2013.