Service Quality Evaluation by Exploring Social Users' Contextual Information

Guoshuai Zhao, Xueming Qian, *Member, IEEE*, Xiaojiang Lei, and Tao Mei, *Senior Member, IEEE*

Abstract—Nowadays, with the boom of social media and e-commerce, more and more people prefer to share their consumption experiences and rate services on review sites. Much research has focused on personalized recommendation. However, quality of service also plays an important role in recommender systems, and it is the main concern of this paper. An overall rating that indicates the popular view usually represents the evaluation. There are some challenges when we do not have enough review information to extract public opinion. Take, for example, a movie for which one user rates a two star rating, and another rates a five star rating. In this case, it is difficult to conduct a quality evaluation fairly. However, it is possible to be improved with the help of big social users' contextual information. In this paper, we propose a model to conduct service quality evaluation by improving overall rating of services using an empirical methodology. We use the concept of user rating's confidence, which denotes the trustworthiness of user ratings. First, entropy is utilized to calculate user ratings' confidence. Second, we further explore spatial-temporal features and review sentimental features of user ratings to constrain their confidences. Last, we fuse them into a unified model to calculate an overall confidence, which is utilized to perform service quality evaluation. Extensive experiments implemented on Yelp and Douban Movie datasets demonstrate the effectiveness of our model.

Index Terms—Data mining, recommender system, service quality evaluation, social networks

1 Introduction

Recently with the rapid development of mobile devices and ubiquitous Internet access, social network services have become prevalent. Users share their experiences, reviews, ratings, photos, videos, check-ins, and moods on the Internet. Recommender systems have been proposed to provide interesting services for users by exploring their preferences from these information. The first generation of recommender systems [1] with traditional collaborative filtering algorithms [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], and many social network based models [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] mostly focus on personalized recommendations by predicting user preferences and ratings. They neglect the significance of service quality. However, personalized recommendation with just considering user preference is imperfect, because quality of service is also important in recommender systems. High quality services should be recommended more easily. Thus, this paper focuses on how to evaluate the quality of service.

When we choose an item, we rely heavily on reviews and ratings provided by social users to find out the evaluation of this item. Generally, an overall rating represents the evaluation from 1 to 5, which indicates the popular view. The more user ratings to this item there are, the more confidence in the overall rating there is. For example, for a given movie, its overall rating is 4.5 star rating given by hundreds of users. We are convinced that it is a great film based on these results. However, there would be a lack of confidence if there were very few users who have rated the item, such as just two. Then other audiences, who rely on ratings shown on a website to make their choices, will make decisions informed by a miniscule amount of data. In addition, it will confuse audiences if there are only two contrary ratings for the same item. For example, there exists a new movie titled *The Best Offer*, which has just two cumulative reviews and ratings. One user rated a two star rating, and another rated a five star rating. Which one we should trust? Usually, we average the ratings, and set it as the overall rating. It is an apposite approach for the items those have large number of ratings. However, for a new item, we cannot simply average the few ratings available and accept it as accurate. In addition, service providers can get feedback on their services from worldwide users, which are valuable for improving service quality. So it is urgent to address quality evaluation for services.

There are several challenges in quality evaluation. The first challenge is the sparsity of ratings. It has been represented in the above paragraph. The second challenge is user confidence bias. Users have different patterns of giving ratings of services. The third challenge is that the rating's confidence is not isolated. They are relevant with their spatial and temporal features. In addition, sometimes users give high ratings but there are many negative reviews for various reasons. Therefore, it is necessary to explore user rating's confidence by closely examining social users' contextual information, including spatial-temporal and sentimental information of reviews.

In this paper, we first utilize information entropy to calculate user ratings' confidence. Second, spatial-temporal features and review sentimental features of ratings from social users' contextual information are mined to constrain user rating's confidence. Last, they are fused into a unified

- *G. Zhao, X. Qian, and X. Lei are with the Key Laboratory for Intelligent Networks and Network Security, Ministry of Education, and the SMILES LAB, Xi'an Jiaotong University, Xi'an 710049, China. E-mail: {zgs2012, xiaolei3439}@stu.xjtu.edu.cn, qianxm@mail.xjtu.edu.cn.*
- *T. Mei is with Microsoft Research, Beijing 100080, China. E-mail: tmei@microsoft.com.*

Manuscript received 18 Feb. 2016; revised 31 Aug. 2016; accepted 2 Sept. 2016. Date of publication 8 Sept. 2016; date of current version 2 Nov. 2016. Recommended for acceptance by A. Gionis.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2016.2607172

probabilistic model to calculate the integrated confidence to improve the overall rating of services.

The biggest difference between with related works [2], [21], [22], [23], [24], [25], [26], [27] is that previous research has focused on personalized rating prediction and recommendation, while quality evaluation for services is our concern. The main contributions are shown as follows:

1) We address the issue of quality evaluation for services, and a probabilistic linear model is proposed by exploring social users' contextual information. It will benefit users and services providers to know the quality of the services with the help of ratings and reviews from worldwide users.
2) We use the concept of user rating's confidence to evaluate the quality of service. The basic idea is that different users have different levels of confidence in the evaluation. Furthermore, users' profiles are changing at different places and different times. From users' textual reviews, we can get more exact information, which always verifies and supports their ratings directly. Therefore, we adopt a probabilistic linear model with Gaussian observation noise to learn the weights of different features.
3) We find that the contextual information, including spatial-temporal features and review sentimental features of users' ratings, is helpful for constraining user rating's confidence. Several interesting findings are obtained. User rating's confidence is higher when a user is very far away from the rated item. User rating's confidence is increasing over time, and increasing with review sentiment.

The differences between this paper and our previous work [28] are: 1) more motivations and challenges are provided, 2) more related works are reviewed and more comparisons with existing works are given, 3) three factors, including spatial-temporal features and review sentimental features, are fused into our model, 4) more experiments and discussions are given. A new dataset Douban Movie is also utilized to test the effectiveness of our approach.

The remainder of this paper is organized as follows. In Section 2, we present related works on recommender systems. In Section 3, the proposed model is described thoroughly. In Section 4, we introduce our datasets in detail. Experiment results and discussions are given in Section 5, and conclusions are drawn in Section 6.

2 Related Works

Biases could represent users' rating habits, such as a scenario in which a user A's ratings are almost 4 and 5, while B's ratings are mostly 3. Koren [29] supposed customer preferences for products drift over time, and proposed a collaborative filtering model with temporal dynamics. He considered user and item time changing biases, and compared the ability of various suggested baseline predictors. Dror et al. [30] proposed a model that incorporates a rich bias model with terms that capture information from the taxonomy of items and different temporal dynamics of music ratings. We can use the idea of user biases and taxonomy biases for reference, and personalized rating prediction can be converted to service quality evaluation.

There are some more approaches to predict users' ratings. A typical model is the matrix factorization model. Many systems [6], [7], [21], [22], [23], [24], [25], [26], [27], [29], [30], [31], [32], [33] employ matrix factorization techniques to learn the latent features of users and items, and predict the unknown ratings using these latent features. Yang et al. [23] proposed using the concept of 'inferred trust circle' based on the domain-obvious of circles of friends on social networks to predict users' ratings. Meanwhile, besides interpersonal influence, Jiang et al. [24] proved that individual preference is also an important factor in social networks. In their Context Model, user latent features should be similar to his/her friends' according to preference similarity. Our previous works [21], [22], [25], [26], [27] consider more social factors in matrix factorization, including interpersonal influence, interpersonal interest similarity, personal interest, user rating behavior similarity and behavior diffusion, and geographical distances. These models can be deployed in the cloud by some cloud computing methods and data storage approaches [55], [56], [57].

Some relevant works address multimedia recommendation [19], [20], [34], [35], [36], [51], [52], [53], [54]. Lee et al. [34] proposed a recommender system that uses the concepts of experts to find both novel and relevant recommendations. Wang et al. [19] designed a joint social-content recommendation framework to suggest videos that users are likely to import or re-share in the online social network.

Existing works mainly focus on personalized rating prediction or recommendation. We focus on quality evaluation for services by exploring social users' contextual information. Matrix factorization also can be utilized to predict all users' ratings for each item as is utilized in personalized rating prediction [23], [24], [25], [26], [27], [37]. For instance, we can simply calculate the quality evaluation by the averaging all users' ratings for the item. Directly exploiting similarity between items is also an approach to predict evaluation. Sarwar et al. [2] proposed an item-based collaborative filtering algorithm. They focused on predicting a user's rating of an item based on the average ratings of similar or correlated items by the same user. It is one of the most popular algorithms in recommender systems.

There are various methods of sentiment analysis [38], [39], [40] that focus on social networks, public sentiment, and web queries. Zhang et al. [38] proposed fusing self-supervised emotion-integrated sentiment classification results into CF recommenders, by which the User-Item rating matrix can be inferred by decomposing item reviews that users give on items. Tan et al. [39] proposed a model that can be used to discover special topics or aspects in one text collection in comparison with another background text collection.

For some items, there are only a few ratings. Thus, in service quality evaluation, we face the classic cold start problem. Many researchers focused on solving this problem [41], [42], [43], [44], [45], [46]. Leroy et al. [41] focused on cold start link prediction. They leveraged some other information regarding available nodes to predict the structure of a social network when the network itself is totally missing. Jiang et al. [45] proposed a user topic based collaborative filtering approach for personalized travel recommendation. It is an improved version of traditional collaborative filtering by fusing the rich user information in social media.

3 The Approach

We propose using information entropy values to measure user ratings' confidence. Furthermore, social users' contextual information is explored from both spatial-temporal aspects and review sentimental aspect. These features are fused into a unified probabilistic model to constrain user rating's confidence. The basic idea is that users' profiles vary with time, places, and sentiments, i.e., user rating's confidence is different at different places, different time, and different sentiments. When we get the final confidence, the quality evaluation for services will be figured out.

3.1 User Rating's Confidence

Different users have different contributions to quality evaluation of services. In this paper, user rating's confidence is leveraged to conduct evaluation. If users' ratings are confident, their ratings must have little differences with the overall rating of services. As we know, entropy is a measure of uncertainty. The information entropy value of these differences can be used to represent the confidence value of user ratings. That is to say, we set the differences between user ratings and the overall rating of services as a difference system, and then the entropy of this system reflects his/her rating habits and stability. Additionally, we add a coefficient to distinguish weights of different values to enhance user ratings' confidence, because the entropy algorithm cannot make a difference in different values. The lower entropy value is, the more stable the system is, and the more confident the user's rating is. User ratings' confidence is represented as the reciprocal of entropy value. It can be calculated by:

$$E_u = -1 \Big/ \sum_i (|d_i| \times p(d_i) \log_2 p(d_i)) \tag{1}$$

$$d_i = r_{u,i} - r_i, \tag{2}$$

where E_u denotes user u's confidence value. d_i is the difference between user rating $r_{u,i}$ and the overall rating r_i. $p(d_i)$ indicates the probability of the value d_i. User ratings' confidence is leveraged to evaluate the overall rating of items by:

$$\hat{r}_i = \sum_{u=0}^{N} {E_u}^* \times r_{u,i}, \tag{3}$$

where ${E_u}^*$ is the normalized form of E_u satisfying $\sum_{u \in i} E_u^* = 1$. $u \in i$ is the set of users who have rated item i. Note that, u is starting from 0. An additional rating, the average rating, is used to avoid a situation where there is only one rating of the test item.

3.2 Contextual Features of User Ratings

The method of calculating user ratings' confidence by entropy is based on all ratings of a user. That is to say, each user ratings' confidence is a constant, whatever the item is. Users' profiles are changing constantly so that their rating's confidence may be different at different places and different time. Sometimes users give high ratings but there are many negative words in their reviews for various reasons. Thus, we further constrain each rating's confidence by its spatial-temporal features and review sentimental features.

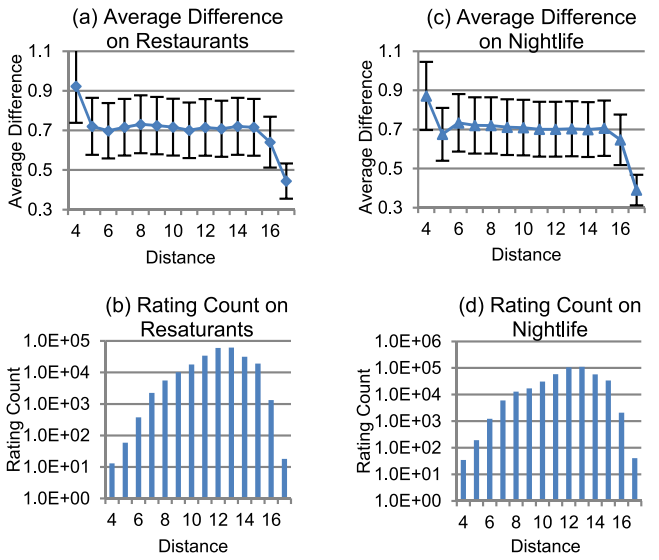


Fig. 1. The distributions of the average difference and the corresponding number of ratings in different user-item geographic location distances based on Yelp restaurants and nightlife datasets. In (a) and (c), the value of the *x*-axis denotes user-item geographic distance which has been normalized by a logarithm, and the value of the *y*-axis denotes the average value of differences between user ratings and the overall rating of services. We also show the proportionate standard deviation of each group.

3.2.1 Spatial Features

People live in a large social network, and may be influenced by others easily. Inevitably, there may be some unfair ratings and reviews on the Internet.

We start by analyzing the distribution of rating's confidence in different user-item geographic location distances. Fig. 1 shows the distributions of the average difference and the corresponding number of ratings in different user-item geographic location distances. The horizontal axis represents user-item geographic distance, which has been operated by the following logarithm:

$$x = lnD(u,i), \tag{4}$$

where $D(u,i)$ denotes the geographical distance between user u and item i. The ordinate axis represents the average difference between user ratings and the overall rating of services. It is an absolute value here.

From Fig. 1, the rating's confidence is low if users are very close to the rated items. We suppose that users may be influenced by their friends or some discounts for services. When users take a long distance travel, they may prefer to local specialties and well-known services [22] resulting the high rating's confidence. In terms of items, most of them have competitors. Inevitably, there may be some malicious evaluation given by their competitors on the Internet. Generally, competitors are mostly native and geographically close. They may be the reasons for this phenomenon. However, for different kinds of datasets, the spatial feature of rating's confidence may be different. Instead of being given some motivations, it can be regarded as a kind of bias in statistics considering that with the extensibility of our model, no matter what the latent reason for this result is.

Curve fitting is conducted to learn ratings' spatial features. Note that the curve fitting is based on the 4th degree Gaussian model. The curve fitting formula is:

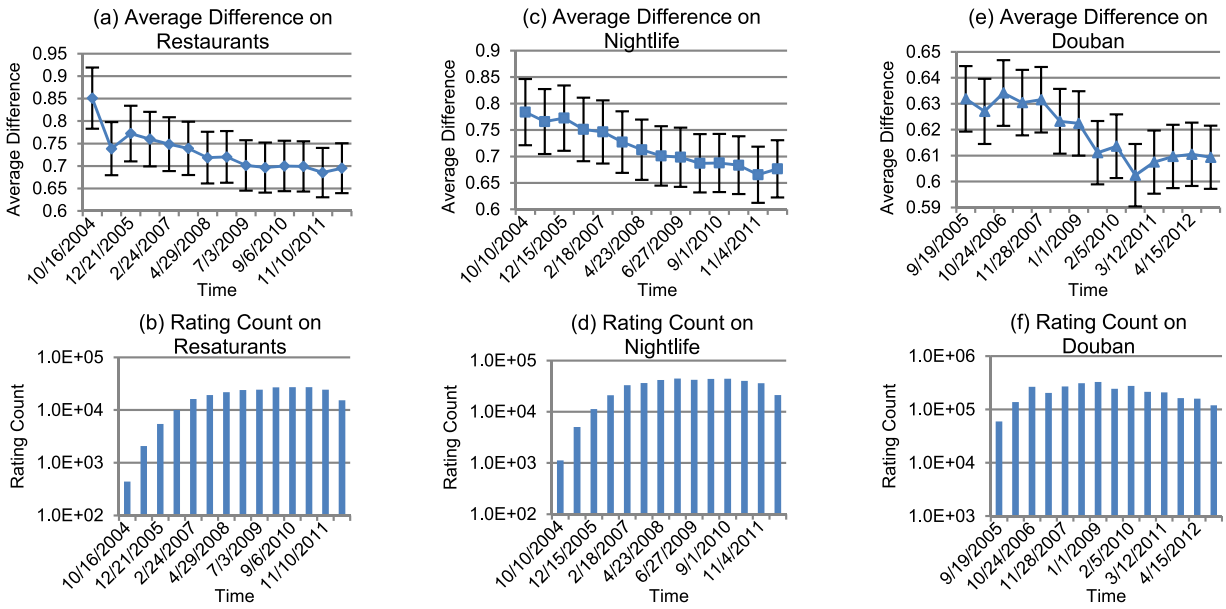


Fig. 2. The distributions of the average difference and the corresponding number of ratings in different periods based on Yelp restaurants, Yelp nightlife, and douban datasets. In (a), (c), and (e), the value of the *x*-axis denotes the time user rated item, and the value of the *y*-axis denotes the average value of differences between user ratings and the overall rating of services.

$$y = \sum_j a_j \times exp\Big(-\big((x-b_j)/c_j\big)^2\Big), \tag{5}$$

where a_i, b_i, and c_i are the coefficients that need to be learned by curve fitting. Rating's confidence is inversely proportional to y. Therefore, rating's confidence based on spatial features can be represented by:

$$G_{u,i} = 1 \Big/ \sum_j a_j exp\Big(-\big((lnD(u,i)-b_j)/c_j\big)^2\Big), \tag{6}$$

where $G_{u,i}$ denotes rating's confidence user u to item i. a_i, b_i and c_i are the coefficients learned by curve fitting. $D(u,i)$ denotes the geographical distance value between user u and item i. In addition, we discuss the performance of different fitting curves in our experiments.

3.2.2 Temporal Features

In the same way, we can get rating's confidence based on temporal features. Fig. 2 shows the distributions of the average difference and the corresponding number of ratings in different periods. In (a), (c), and (e), the value of the x-axis denotes the time of the rating, and the value of the y-axis denotes the average difference between user ratings and overall rating of services. It decreases over time. We suppose that there are more and more ratings and reviews for each item, resulting in users getting more and more useful information from former ratings and reviews, and then give a suitable rating. That is to say, when we search the Internet, we may be unconsciously influenced by the ratings and reviews, because the external circumstance can affect us, especially on fields we do not know well.

Curve fitting is conducted based on the 4th degree Gaussian model. Ratings' temporal features can be represented by:

$$T_{u,i} = 1 \Big/ \sum_j a_j exp\Big(-\big((Day(u,i)-b_j)/c_j\big)^2\Big) \tag{7}$$

where $T_{u,i}$ indicates rating's confidence user u to item i based on temporal features. $Day(u,i)$ denotes the rating time of user u to item i. a_i, b_i and c_i are the coefficients that need to be learned by curve fitting.

3.2.3 Sentimental Features

On most review sites, users cannot only rate the commodity, but also share their experiences and attitudes by reviewing. From their textual reviews, we can get more exact information, which always verifies and supports their ratings directly. Therefore, it is necessary to analyze the relevance between user confidence and textural review sentiment. First, the method of sentiment analysis proposed in [38] is leveraged to calculate sentiment scores. Second, the relevance between user rating's confidence and review sentiment is mined. Last, we learn sentimental features to constrain users' confidence.

Fig. 3 shows the distributions of the average difference and the corresponding number of ratings in different sentiment scores. In (a) and (c), the value of the x-axis is the normalized review sentiment score. The value of the y-axis is the average difference between ratings and overall rating of services. It decreases with the sentiment score. That is to say, user confidence increases with review sentiment score. The sentimental features can be represented by:

$$S_{u,i} = 1 \Big/ \sum\nolimits_j a_j \times (RS(u,i))^j \tag{8}$$

where $S_{u,i}$ denotes rating's confidence user u to item i according to review sentimental features. $RS(u,i)$ is the normalized sentiment score user u to item i.

3.3 Service Quality Evaluation Model

The overview of our service quality evaluation model is shown in Fig. 4. We fuse user's confidence with contextual features, including spatial-temporal features and review

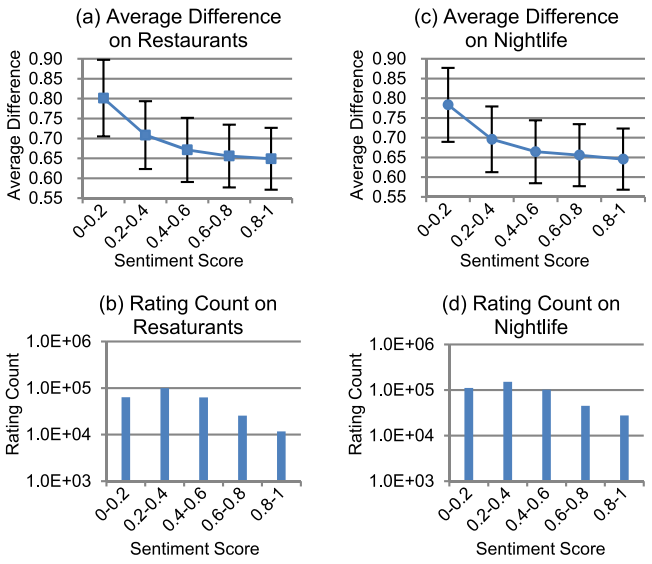


Fig. 3. The distributions of the average difference in different sentiment scores based on Yelp restaurants and Yelp nightlife datasets. In (a) and (b), the value of the *x*-axis is the normalized textural review sentiment score. The value of the *y*-axis is the average difference between ratings and the overall rating of services.

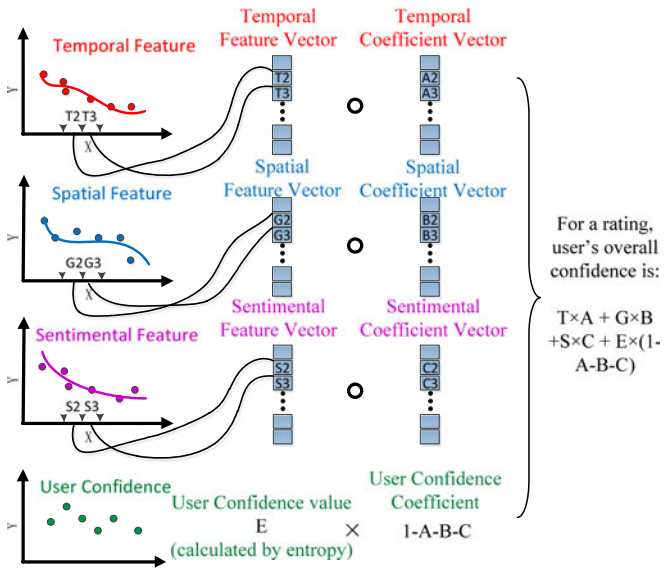


Fig. 4. Overview of user's overall confidence calculation. We explore user rating's confidence by considering spatial-temporal features and review sentimental features. ∘ is the symbol of the Hadamard product.

sentimental features, to calculate an overall confidence value of a rating. Note that we define the confidence coefficient in an effective interval under a condition that the sum of coefficients is one. Now our purpose is to learn spatial-temporal and review sentimental coefficient vectors of user ratings by training them in a unified probabilistic model. As shown in Fig. 4, the dimensions of the statistical chart are set as the dimensions of feature vectors and coefficient vectors. In order to simplify our formulas, we define the overall confidence of the rating that user u to item i as follows:

$$\Phi_{u,i} = A_{u,t(u,i)}\, T_{t(u,i)} + B_{u,g(u,i)} G_{g(u,i)} + C_{u,s(u,i)} S_{s(u,i)} + D_{u,t(u,i),g(u,i),s(u,i)} E_u, \tag{9}$$

where:

$$D_{u,t(u,i),g(u,i),s(u,i)} = 1 - A_{u,t(u,i)} - B_{u,g(u,i)} - C_{u,s(u,i)}, \tag{10}$$

where $t(u,i)$ denotes the time user u rated item i. $g(u,i)$ indicates the geographic distance between user u and item i. $s(u,i)$ implies the sentimental value of the review that user u given item i. $T_{t(u,i)}$ is the rating's confidence based on temporal features calculated by (7), and $G_{g(u,i)}$ indicates the rating's confidence based on spatial features calculated by (6). $S_{s(u,i)}$ is the rating's confidence based on sentimental features calculated by (8). E_u is the user ratings' confidence calculated by (1). $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ are the corresponding coefficient matrixes. The sizes of these coefficient matrixes are all $M \times k$, where M is the number of users and k is the dimension of feature vectors.

3.3.1 Model Inference

A probabilistic linear model with Gaussian observation noise is adopted as [23], [25], and [31]. Here we define the conditional probability of the observed ratings as follows:

$$p(\mathbf{R}|\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{G},\mathbf{E},\mathbf{T},\mathbf{S},\sigma_R^2) = \prod_i \mathcal{N}\left(R_i \middle| \sum_{u=0}^{n_i} \left(\frac{\Phi_{u,i}}{\sum_{u=0}^{n_i}\Phi_{u,i}} r_{u,i}\right), \sigma_R^2\right), \tag{11}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ denotes the probability density function of Gaussian distribution with mean μ and variance σ^2. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$ is users' temporal, spatial, sentimental and confidence coefficient matrix respectively. n_i denotes the number of users who have rated item i. R_i is the overall rating of service i, i.e., it is the ground truth. If there is only one user having rated item i, quality evaluation for this service cannot be performed. To avoid this case, we manually set average rating value as a new rating from user $u = 0$ to item i.

According to [31], zero mean Gaussian priors are assumed for users' spatial-temporal and sentimental coefficient vectors:

$$p\left(\mathbf{A}|\sigma_A^2\right) = \prod\nolimits_u \mathcal{N}\left(A_u|0,\sigma_A^2\right) \tag{12}$$

$$p\left(\mathbf{B}|\sigma_B^2\right) = \prod\nolimits_u \mathcal{N}\left(B_u|0,\sigma_B^2\right) \tag{13}$$

$$p\left(\mathbf{C}|\sigma_C^2\right) = \prod\nolimits_u \mathcal{N}\left(C_u|0,\sigma_C^2\right). \tag{14}$$

The posterior distribution over these coefficient matrices is given by:

$$\begin{aligned} &p(\mathbf{A},\mathbf{B},\mathbf{C}|\mathbf{R},\mathbf{T},\mathbf{G},\mathbf{S},\mathbf{E},\sigma^2) \\ &= \frac{p(\mathbf{R},\mathbf{T},\mathbf{G},\mathbf{S},\mathbf{E}|\mathbf{A},\mathbf{B},\mathbf{C},\sigma^2)P(\mathbf{A},\mathbf{B},\mathbf{C}|\sigma^2)}{P(\mathbf{R},\mathbf{T},\mathbf{G},\mathbf{S},\mathbf{E},\sigma^2)} \\ &\propto p(\mathbf{R}|\mathbf{A},\mathbf{B},\mathbf{C},\sigma^2)P(\mathbf{A}|\sigma^2)P(\mathbf{B}|\sigma^2)P(\mathbf{C}|\sigma^2) \\ &= \prod_i \mathcal{N}\left(R_i \middle| \sum_{u=0}^{n_i}\left(\frac{\Phi_{u,i}}{\sum_{u=0}^{n_i}\Phi_{u,i}} r_{u,i}\right), \sigma_R^2\right) \\ &\times \prod_u \mathcal{N}\left(A_u|0,\sigma_A^2\right) \times \prod_u \mathcal{N}\left(B_u|0,\sigma_B^2\right) \times \prod_u \mathcal{N}\left(C_u|0,\sigma_C^2\right). \end{aligned} \tag{15}$$

Then the log of the posterior distribution is given by:

$$\begin{aligned} & lnp\left(\mathbf{A},\mathbf{B},\mathbf{C}|\mathbf{R},\mathbf{T},\mathbf{G},\mathbf{S},\mathbf{E},\sigma^2\right) \\ & = -\frac{1}{2\sigma_R^2}\sum_i\left(R_i-\sum_{u=0}^{n_i}\left(\frac{\Phi_{u,i}}{\sum_{u=0}^{n_i}\Phi_{u,i}}r_{u,i}\right)\right)^2 \\ & -\frac{1}{2\sigma_A^2}\sum_u A_u^TA_u-\frac{1}{2\sigma_B^2}\sum_u B_u^TB_u-\frac{1}{2\sigma_C^2}\sum_u C_u^TC_u \\ & -\frac{1}{2}\left(N\times ln\sigma_R^2+(M\times k)ln\sigma_A^2+(M\times k)ln\sigma_B^2\right. \\ & \left.+(M\times\ k)ln\sigma_C^2\right)+c, \end{aligned} \tag{16}$$

where c is a constant that does not depend on the parameters. M is the number of users. N is the number of items, and k indicates the dimension of the latent space. Keeping the parameters (observation noise variance and prior variance) fixed, maximizing the posterior distribution is equivalent to minimizing the sum-of-squared errors objective function with quadratic regularization terms:

$$\begin{aligned} & \Psi(\mathbf{R},\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{T},\mathbf{G},\mathbf{S},\mathbf{E}) \\ & =\frac{1}{2}\sum_i\left(R_i-\sum_{u=0}^{n_i}\left(\frac{\Phi_{u,i}}{\sum_{u=0}^{n_i}\Phi_{u,i}}r_{u,i}\right)\right)^2 \\ & +\frac{\lambda_A}{2}\mathbf{A}_F^2+\frac{\lambda_B}{2}\mathbf{B}_F^2+\frac{\lambda_C}{2}\mathbf{C}_F^2, \end{aligned} \tag{17}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm.

3.3.2 Model Training

Once we get the objective function, it can be minimized by the gradient decent approach as [23], [25], [31], and [32]. The gradients of the objective function with respect to the variables $A_{u,t(u,i)}$, $B_{u,g(u,i)}$, and $C_{u,s(u,i)}$ are respectively shown as (18), (19), and (20):

$$\begin{aligned} \frac{\partial\Psi}{\partial A_{u,t(u,i)}} & =(-1)\left(R_i-\sum_{u'=0}^{n_i}\left(\frac{\Phi_{u',i}}{\sum_{u'=0}^{n_i}\Phi_{u',i}}r_{u',i}\right)\right) \\ & \times\left(\frac{\left(T_{t(u,i)}-E_u\right)\left(\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}r_{u,i}\right. \\ & \left.-\frac{\sum_{u'\neq u}^{n_i-1}\left(\Phi_{u',i}\left(T_{t(u',i)}-E_{u'}\right)r_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}\right) \\ & +\lambda_A A_{u,t(u,i)} \end{aligned} \tag{18}$$

$$\begin{aligned} \frac{\partial\Psi}{\partial B_{u,g(u,i)}} & =(-1)\left(R_i-\sum_{u'=0}^{n_i}\left(\frac{\Phi_{u',i}}{\sum_{u'=0}^{n_i}\Phi_{u',i}}r_{u',i}\right)\right) \\ & \times\left(\frac{\left(G_{g(u,i)}-E_u\right)\left(\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}r_{u,i}\right. \\ & \left.-\frac{\sum_{u'\neq u}^{n_i-1}\left(\Phi_{u',i}\left(G_{g(u',i)}-E_{u'}\right)r_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}\right) \\ & +\lambda_B B_{u,g(u,i)} \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\partial\Psi}{\partial C_{u,s(u,i)}} & =(-1)\left(R_i-\sum_{u'=0}^{n_i}\left(\frac{\Phi_{u',i}}{\sum_{u'=0}^{n_i}\Phi_{u',i}}r_{u',i}\right)\right) \\ & \times\left(\frac{\left(S_{s(u,i)}-E_u\right)\left(\sum_{u'\neq u}^{n_i-1}\Phi_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}r_{u,i}\right. \\ & \left.-\frac{\sum_{u'\neq u}^{n_i-1}\left(\Phi_{u',i}\left(S_{s(u',i)}-E_{u'}\right)r_{u',i}\right)}{\left(\sum_{u'=0}^{n_i}\Phi_{u',i}\right)^2}\right) \\ & +\lambda_C C_{u,s(u,i)}. \end{aligned} \tag{20}$$

Once we get the gradients, we update coefficient matrices as follows:

$$A_{u,t(u,i)}=A_{u,t(u,i)}\ -\alpha\frac{\partial\Psi}{\partial A_{u,t(u,i)}} \tag{21}$$

$$B_{u,g(u,i)}=B_{u,g(u,i)}\ -\alpha\frac{\partial\Psi}{\partial B_{u,g(u,i)}} \tag{22}$$

$$C_{u,s(u,i)}=C_{u,s(u,i)}\ -\alpha\frac{\partial\Psi}{\partial C_{u,s(u,i)}}, \tag{23}$$

where α is the learning rate.

Lastly, after several iteration computations, we conduct service quality evaluation by the learned coefficient matrices as follows:

$$\hat{r}_i=\sum_{u=0}^{n_i}\left(\frac{\Phi_{u,i}}{\sum_{u=0}^{n_i}\Phi_{u,i}}r_{u,i}\right). \tag{24}$$

The whole procedure of our algorithm is summarized in Algorithm 1. Moreover, the time complexity is $O(T\times M^2\times N\times k)$, where M and N are the number of users and items, T is the number of iterations, and k is the dimension of feature vectors. The space complexity is $O(M\times N+M+3k+3M\times k)$. Since the rating matrix is usually sparse and $M,N\gg k$, the time complexity is $O(T\times M\times k\times L)$, and the space complexity is $O(M\times N)$, where L is the number of links between users and items.

Algorithm 1. Service Quality Evaluation (SQE) Model

Input: The rating matrix **R** in training dataset, user confidence **E** calculated by Equation (1), spatial bias **G** calculated by Equation (6), temporal bias **T** calculated by Equation (7), sentimental bias **S** calculated by Equation (8).
Output: Quality evaluation of test services.

```
Initialize coefficient matrices A, B, C, set learning rate α.
for t = 1:T do
    for each element of coefficient matrices A, B, C, do:
        A_{u,t(u,i)} ← A_{u,t(u,i)} − α ∂Ψ/∂A_{u,t(u,i)};
        B_{u,g(u,i)} ← B_{u,g(u,i)} − α ∂Ψ/∂B_{u,g(u,i)};
        C_{u,s(u,i)} ← C_{u,s(u,i)} − α ∂Ψ/∂C_{u,s(u,i)};
    end for
end for
for each test item do
    for each rating of this item do
        Calculate the overall confidence by Equation (9);
    end for
    Calculate the overall rating r̂_i of this item by Equation (24);
end for
Return: The overall rating r̂ of services.
```

TABLE 1
Statistics of Our Datasets

Dataset	Restaurants		Nightlife		Douban	
Num. users	4,138		11,152		8,226	
Num. items	62,221	52,071 (training)	21,647	14,066 (training)	14,715	12,286 (training)
		10,150 (test)		7,581 (test)		2429 (test)
Num. ratings	263,124	244,205 (training)	436,301	420,790 (training)	2,968,648	2,961,176 (training)
		18,919 (test)		15,511 (test)		7472 (test)
Average rating	3.646		3.5893		3.7867	

4 Datasets Introduction

In this section, we introduce the Yelp and Douban datasets and the preprocessing approach. The datasets are extended from our previous works [21], [22], and [25]. Our dataset can be downloaded from website of SMILES LAB.[1]

4.1 Yelp Dataset

Yelp is a local directory service with social networks and user reviews [21], [22], [25]. It is the largest service review site in America. Users can rate businesses, submit comments, communicate shopping experiences, etc. It combines local reviews and social networking functionality to create a local online community. In this paper, the utilized Yelp dataset consists of two categories: Restaurants and Nightlife. Table 1 shows the statistics of our datasets.

4.2 Douban Dataset

Douban is one of the most popular social networks in China. It includes several parts: Douban Movie, Douban Read and Douban Music, etc. We crawled the ratings from the Douban Movie website. The dataset consists of 2,968,648 ratings from 8,226 users who have rated 14,715 movies. Note that there is no geographic location information and reviews in Douban dataset. We perform our model on Douban dataset by fusing user ratings' confidence and temporal features.

4.3 Preprocessing

The issue proposed in this paper is quality evaluation for services with very few ratings. That is to say, the entity of our dataset is service. Thus, we must handle our dataset to extract appropriate test data. As shown in Table 1, we preselect some items to be used for training and others to be used for testing. The ratings in our dataset are split according to the preselected items. For instance, a rating of a training item will be selected as a training data. Note that every tested item has no more than five ratings. Fig. 5 shows the distributions of items in our three test sets according to the number of ratings. The y-axis represents the count of items. The x-axis represents the number of ratings under each item. From Fig. 5, it can be observed that none of these test items has more than five ratings.

1. http://smiles.xjtu.edu.cn/Download/Download_SQE.html

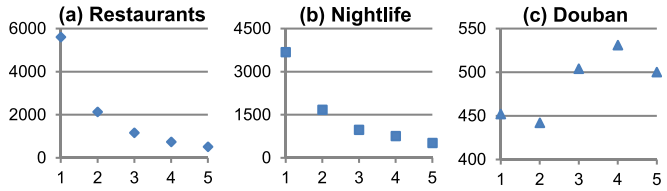


Fig. 5. The distributions of items in the test set according to the number of ratings on Yelp restaurants, nightlife, and douban test set, respectively.

5 Experiments

In this section, we implement a series of experiments on our datasets to evaluate the performance of our model. We compare the performance with some related methods, and some discussions are given.

5.1 Performance Measures

When we get the predicted overall rating of services, the performance of methods will be embodied by the errors. Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are the most popular accuracy measurements [23], [24], [25], [29], [30], [31], [32], [33]. They are defined by:

$$RMSE = \sqrt{\sum_{i \in \Re_{test}} (r_i - \hat{r}_i)^2 / |\Re_{test}|} \tag{25}$$

$$MAE = \sum_{i \in \Re_{test}} |r_i - \hat{r}_i| / |\Re_{test}|, \tag{26}$$

where r_i is the overall rating of service i. $\hat{r}_i$ is the predicted overall rating. $\Re_{test}$ denotes the set of test items. $|\Re_{test}|$ indicates the number of test items.

The differences between the prediction and the overall rating of services can be leveraged to measure our model. However, in our datasets, the real overall ratings of services are discrete as [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0], while our predictions are decimals. The predicted decimals can be rounded into discrete quantities. Then Precision, Recall and AUC (Area under Curve) measures [47], [48], [49] are utilized to evaluate the proposed model. It is discussed in Section 5.3.6.

5.2 Performance Evaluation

5.2.1 Compared Algorithms

a) *BM (Basic Method).* Intuitively, a basic method is utilizing the average ratings to the item to represent its quality evaluation:

$$\hat{r}_i = \frac{1}{n} \sum_{u=1}^{n} r_{u,i}, \tag{27}$$

where $\hat{r}_i$ is predicted evaluation of item i. n is the number of ratings to item i, and $r_{u,i}$ denotes the rating user u to item i.

b) *Biases (Basic Biases).* Biases could represent users' rating habits. Koren [29] considered user and item time changing biases, and compared the ability of various suggested baseline predictors. Basic biases could represent users' rating habits. In order to overcome different rating criteria, users' rating biases can be considered into rating prediction as follows:

$$\hat{r}_i = \frac{1}{n} \sum\nolimits_{u=1}^{n} \left(r_{u,i} + b_u\right) \tag{28}$$

$$b_u = \mu - \bar{r}_u, \tag{29}$$

where b_u denotes user u's rating bias and μ indicates the average of all ratings. $\bar{r}_u$ is user u's average rating. This method overcomes users' different rating criteria simply.

c) *BT (Biases Based on Taxonomy).* Compared with basic biases, we utilize the idea of biases based on taxonomy [30] to explore users' rating criteria with more refinements. Biases are detailed into many categories. That is to say, one user may have different rating criteria in different categories. Thus, we have:

$$\hat{r}_i = \frac{1}{n} \sum\nolimits_{u=1}^{n} \left(r_{u,i} + b_{u,c_i}\right) \tag{30}$$

$$b_{u,c} = \mu_c - \bar{r}_{u,c}, \tag{31}$$

where $b_{u,c}$ denotes user u's rating bias in category c, μ_c implies the overall average rating in category c, $\bar{r}_{u,c}$ indicates user u's average rating in category c. c_i is the category that item i belonging to.

d) *BaseMF.* The BaseMF is the basic probabilistic matrix factorization approach [32], which aims at reducing the error of the predicted rating values using $\mathbf{R}$ to real rating values. The latent features of users and items are learned on the observed rating data by minimizing the objective function:

$$\Psi(\mathbf{R}, \mathbf{U}, \mathbf{P}) = \frac{1}{2} \sum_{u,i} \left(R_{u,i} - \hat{R}_{u,i}\right)^2 + \frac{\lambda}{2} \left(\|\mathbf{U}\|_F^2 + \|\mathbf{P}\|_F^2\right) \tag{32}$$

$$\hat{R}_{u,i} = \mu + U_u^T P_i, \tag{33}$$

where $\hat{\mathbf{R}} \in \mathbf{R}^{M \times N}$, M is the number of users and N is the number of items as the rating matrix $\mathbf{R}_{M \times N}$. $R_{u,i}$ denotes the real rating value user u to item i. $\hat{R}_{u,i}$ is the predicted rating value user u to item i. μ is an offset value, which is empirically set average rating value of the training data. Matrixes $\mathbf{U}$ and $\mathbf{P}$ are latent feature matrices of users and items. $\|\mathbf{X}\|_F$ is the Frobenius norm of matrix $\mathbf{X}$, and $\|\mathbf{X}\|_F = \sqrt{\sum_{i,j} x_{i,j}^2}$. This term is utilized to avoid over-fitting. This objective function can be minimized efficiently using the gradient descent method. Once the low-rank matrices $\mathbf{U}$ and $\mathbf{P}$ are learned, rating values can be predicted according to (33) for any user-item pairs. Then predicted ratings are leveraged to evaluate the quality of services by:

$$\hat{r}_i = \frac{1}{M} \sum\nolimits_{u=1}^{M} \hat{R}_{u,i}. \tag{34}$$

e) *CircleCon Model.* The CircleCon model [23] has been found to outperform BaseMF and SocialMF [31] with respect to accuracy of recommender systems. It focuses on the factor of interpersonal trust in the social network and infers the trust circle. The basic idea is that user latent feature U_u should be similar to the average of his/her friends' latent features with weight of trust value in category c. Once the model is trained in c, the rating value in c can be predicted according to (33). Then we can utilize predicted ratings to evaluate items by:

$$\hat{r}_i = \frac{1}{M} \sum\nolimits_{u=1}^{M} \hat{R}_{u,i}^c. \tag{35}$$

f) *ContextMF.* Besides the factor of interpersonal influence, Jiang et al. [24] proposed another important factor: the individual preference. The results demonstrate the significance of social contextual factors (including individual preference and interpersonal influence) in their model. The factor of interpersonal influence is similar to the trust values in the CircleCon model [23]. Moreover, another factor of interpersonal preference similarity is mined from the topic of items adopted from the receiver's history. The basic idea is that user latent feature U_u should be similar to his/her friends' latent feature with the weight of their preference similarity. Once the model is trained, the rating can be predicted according to (33). Then we can utilize predicted ratings to evaluate items by (35).

g) *PRM.* In previous works [25], [26], we considered more social factors to constrain user and item latent features, involving interpersonal influence, interpersonal interest similarity, and personal interest. The proposed new factor personal interest denotes user's interest vector has similarity to item's topic vector which user interest in. The factor of personal interest can recommend items to meet users' individualities, especially for experienced users. Once we get the learned user and item features, items can be evaluated by (35).

h) *Item-based Collaborative Filtering.* Item-based collaborative filtering recommendation [2] is one of the most popular algorithms. It produces the rating from a user to an item based on the average ratings of similar or correlated items by the same user. It gets better performance by computing the similarity $sim(i, j)$ between items as follows:

$$sim(i,j) = \frac{\sum_{u \in U} \left(r_{u,i} - \bar{r}_u\right)\left(r_{u,j} - \bar{r}_u\right)}{\sqrt{\sum_{u \in U} \left(r_{u,i} - \bar{r}_u\right)^2} \sqrt{\sum_{u \in U} \left(r_{u,j} - \bar{r}_u\right)^2}}, \tag{36}$$

where $r_{u,i}$ denotes the rating user u to item i. $r_{u,j}$ denotes the rating user u to item j. $\bar{r}_u$ denotes user u's average rating. Authors aimed at personalized rating prediction, but if we would like to predict quality evaluation, we can utilize the similarity $sim(i, j)$ as follows:

$$\hat{r}_i = \frac{1}{\left|\Re_{training}\right|} \sum\nolimits_{j \in \Re_{training}} r_j \times sim(i,j)^*, \tag{37}$$

where r_j is the overall rating of services. $sim(i, j)^*$ is the normalized similarity value. $\Re_{training}$ is the set of items in the training set.

i) *MART-SQE.* We add a non-linear model to the proposed approaches by using multiple-additive regression trees (MART) [50]. When performing MART, we set the proposed features as the predictors, including entropy based user confidence, spatial features, temporal features, review sentimental features, and user personalized rating. That is to say, five predictors are used in this experiment, and individual trees have six

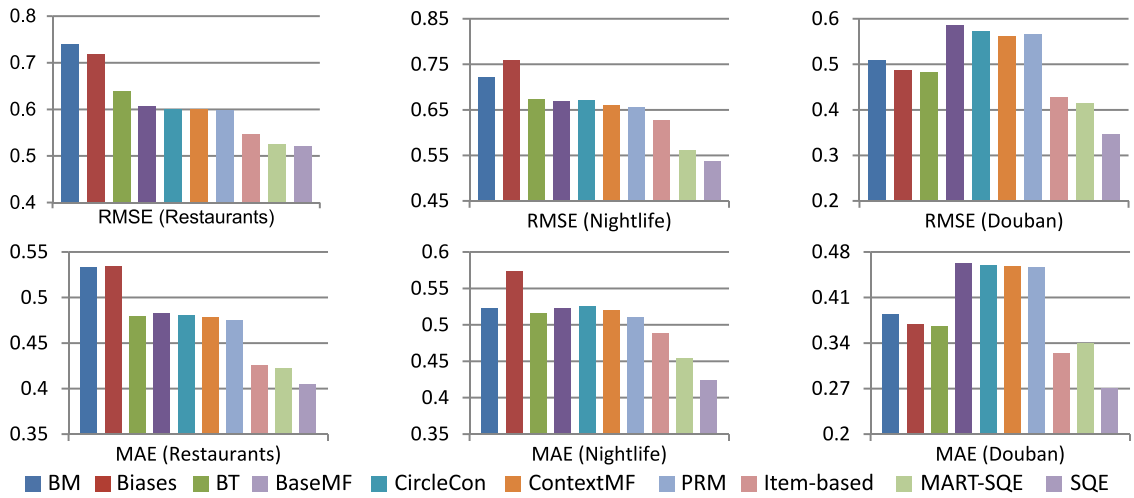


Fig. 6. Performance comparison based on Yelp restaurants, Yelp nightlife, and douban datasets.

terminal nodes, 200 trees are grown. We call the method MART-SQE (Multiple-additive regression trees based service quality evaluation).

j) *SQE*. This is our model proposed in this paper. We explore user rating's confidence by considering spatial-temporal and sentimental features to conduct service quality evaluation.

5.2.2 Performance Comparison

Here, we compare the performance of our SQE model with other methods, including BM, Biases, BT, BaseMF, CircleCon, ContextMF, PRM, item-based collaborative filtering, and MART-SQE on Yelp Restaurants, Yelp Nightlife, and Douban dataset respectively.

Fig. 6 shows the performance on Yelp Restaurants, Nightlife, and Douban datasets. The accuracy of our SQE model is much better than other approaches. Additionally, matrix factorization models, including BaseMF, CircleCon, ContextMF, and PRM, have few differences in terms of performance. Actually, matrix factorization models are not suitable to solve quality evaluation for services, because matrix factorization models aim at personalized ratings prediction [24], [25], and [26]. These models focus on calculating users and items' latent feature vectors. However, in this paper we utilize them to predict users' personalized ratings, and then average these personalized ratings. It seems inconsistent. Thus, their performance is not satisfactory. Additionally, when we average these personalized ratings as (34), denominator M, which denotes the number of users, is so large that the final evaluations have little diversity. Most of the evaluations we predict by matrix factorization models are ranged from star level 3.4 to 3.8. Performance is barely affected by data sparsity and ground truth because of the overly temperate and smooth results. Then we conclude that matrix factorization models are not suitable to solve quality evaluation of services.

Besides matrix factorization models, the performance on the Douban dataset is much better than the Yelp dataset. This is caused by the characteristics of dataset. We deem that the overall confidence of ratings on Douban dataset is better than Yelp. In our opinion, Douban Movie focuses on the content of movies, while Yelp focuses on the quality of services. In other words, all users of Douban Movie will rate the same movie, which is constant. However, users of Yelp will rate the same item from different aspects, such as service attitude, environment, and the taste of food. Different users may meet different waiters, and taste different foods. Thus, many different external factors can affect users' ratings on Yelp Restaurants and Yelp Nightlife datasets. Conversely, ratings on Douban Movie are not affected by many other external factors except the content of movies. Table 2 supports our supposition. We set each rating as an evaluation to service, and then perform all ratings' confidence by computing errors between user ratings and the overall rating of services. RMSE and MAE are utilized to describe the overall ratings' confidence. It can be concluded that ratings on Douban have more confidence than Yelp.

5.3 Discussions

Besides the performance comparison, here we discuss six aspects in our SQE model: 1) the impact of data sparsity, 2) the impact of review count, 3) the impact of different curves fitting approaches in spatial-temporal and sentimental feature constrained user confidence measurement, 4) the impact of each feature, 5) the impact of less training data on performance, and 6) the impact of the type of prediction. At the last part, when we quantify our prediction, the new measurements Precision, Recall, and AUC are utilized to demonstrate the improvements of our model.

5.3.1 The Impact of Data Sparsity

As mentioned before, the number of ratings for each item in the test set is no more than five. Then we conduct a series of experiments to discuss the impact of data sparsity. Fig. 7

TABLE 2
The Overall Rating's Confidence Represented by Errors on Three Datasets

Dataset	RMSE	MAE
Yelp Restaurants	0.9377	0.7142
Yelp Nightlife	0.9242	0.7035
Douban	0.7763	0.6192

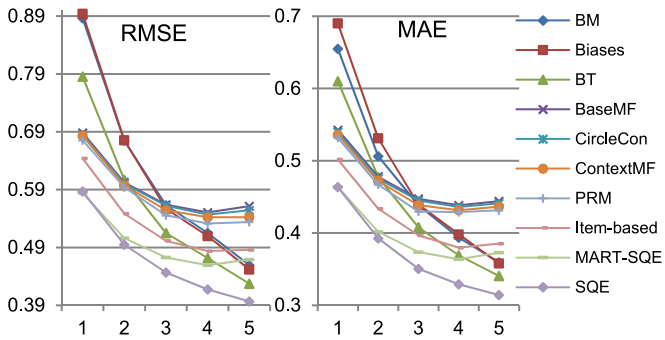


Fig. 7. The average impact of data sparsity on restaurants and nightlife datasets.

shows the average impact of data sparsity on performance on Restaurants and Nightlife datasets. According to Fig. 5, our test set is classified into five groups, with each group just contains the items which have the same number of ratings. That is to say, it is according to different data sparsity. Then in Fig. 7, the performance improves with the increase of data density. It can be observed that our model is better than other methods in terms of performance, no matter what the data sparsity is.

5.3.2 The Impact of Review Count

The goal of this paper is to predict service quality evaluation. However, it is difficult to get the ground truth of the overall rating of services, because the ground truth relies heavily on the review count. For example, if the real review count is too small, the overall rating we crawled will be a lack of trustworthiness. Thus, we discuss the impact of review count by grouping test items. As shown in Fig. 8, our test set is classified into five groups: the real review count of items is greater than 0, 5, 10, 20, and 50 respectively. It shows the average impact of review count on performance on Restaurants and Nightlife datasets. We deem that performance will become better with the increasing number of real ratings. This assumption is supported by the experiment result shown in Fig. 8.

5.3.3 The Impact of Different Curves Fitting Approaches

As mentioned in Section 3.2, we conduct curve fitting based on the 4th degree Gaussian model. A series of experiments are conducted according to different fitting curves as shown in Fig. 9. It shows the average impact of different fitting curves on performance on Restaurants and Nightlife

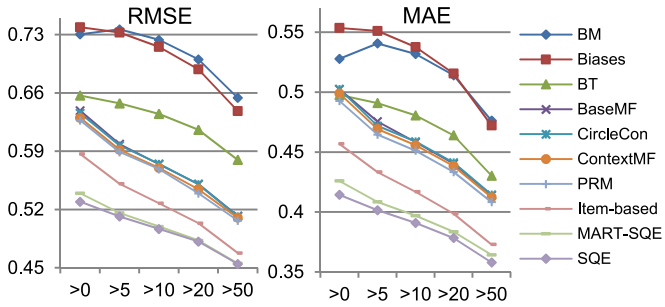


Fig. 8. The average impact of review count on restaurants and nightlife datasets.

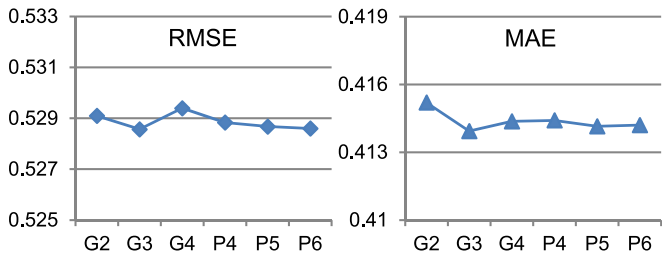


Fig. 9. The average impact of different fitting curves on restaurants and nightlife datasets.

datasets. P4, P5, P6 denotes fitting curve based on the 4th, 5th and 6th degree polynomial respectively. G2, G3, and G4 denotes fitting curve based on the 2nd, 3rd and 4th degree Gaussian model respectively. It can be observed that different curves fitting approaches have little influence on the final performance. It demonstrates the robustness of our SQE model.

5.3.4 The Impact of Different Features

Here, we discuss the impact of different features on the performance. Spatial-temporal features and sentimental features are fused into our SQE model. However, we would like to know the effectiveness of each feature. Therefore, user ratings' confidence (UTC) calculated by entropy in Section 3.1 is set as the baseline. Then we conduct a series of experiments by fusing the confidence respectively with ratings' temporal features (denoted by UTC+TF), ratings' spatial features (denoted by UTC+SF), review sentimental features (denoted by UTC+RF), spatial-temporal features (denoted by UTC+ TF+SF), both temporal features and review sentimental features (denoted by UTC+TF+RF), both spatial features and review sentimental features (denoted by UTC+SF+RF), and both spatial-temporal features and review sentimental features, i.e., SQE method. The performance is shown in Fig. 10, which shows the average impact of each feature on performance on Restaurants and Nightlife datasets. Both the spatial-temporal features and review sentimental features are significant to the final performance. In Fig. 10, the temporal features decrease the prediction error by 5.9 percent on RMSE and 5.0 percent on MAE. The spatial features decrease the prediction error by 5.5 percent on RMSE and 4.9 percent on MAE. The sentimental features decrease the prediction error by 5.2 percent on RMSE and 4.6 percent on MAE. The proposed model SQE combining with spatial-temporal features and review sentimental features decreases the prediction error by 10.1 percent on RMSE and 9.8 percent on MAE.

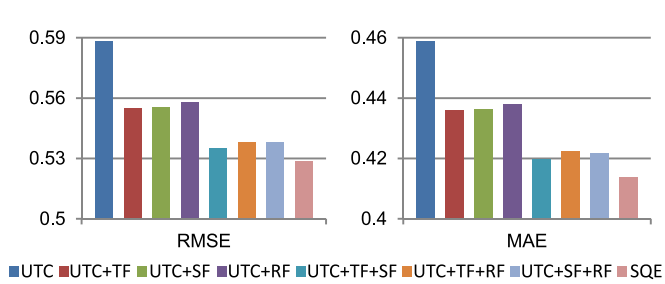


Fig. 10. The average impact of each feature on restaurants and nightlife datasets.

TABLE 3
The Impact of Less Training Data on the Performance of Our Model Based on Douban Dataset

Measure	1% training data	2% training data	10% training data	20% training data	50% training data	100% training data
RMSE	0.3598	0.3539	0.3509	0.3508	0.3506	0.3464
MAE	0.2814	0.2760	0.2737	0.2734	0.2733	0.2701

TABLE 4
Accuracy in Terms of Precision (P) and Recall (R) on Yelp Nightlife Dataset

	BM		Biases		BT		BaseMF		CircleCon		ContextMF		PRM		Item_based		MART-SQE		SQE	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
@Top-1	0.333	0.333	0.309	0.309	0.323	0.323	0.297	0.297	0.297	0.297	0.296	0.296	0.296	0.296	0.327	0.327	0.349	0.349	**0.382**	**0.382**
@Top-3	0.773	0.258	0.742	0.247	0.780	0.260	0.767	0.256	0.767	0.256	0.767	0.256	0.767	0.256	0.785	0.262	0.827	0.276	**0.833**	**0.278**
@Top-5	0.949	0.190	0.940	0.118	0.955	0.119	0.941	0.118	0.941	0.118	0.941	0.118	0.941	0.118	0.949	0.119	**0.975**	**0.195**	0.972	0.194

TABLE 5
Accuracy in Terms of Precision (P) and Recall (R) on Yelp Restaurants Dataset

	BM		Biases		BT		BaseMF		CircleCon		ContextMF		PRM		Item_based		MART-SQE		SQE	
	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P
@Top-1	0.285	0.285	0.283	0.283	0.316	0.316	0.322	0.322	0.321	0.321	0.321	0.321	0.321	0.321	0.376	0.376	0.371	0.371	**0.384**	**0.384**
@Top-3	0.735	0.245	0.734	0.245	0.770	0.257	0.807	0.270	0.811	0.270	0.811	0.270	0.812	0.272	0.844	0.282	**0.856**	**0.285**	0.855	**0.285**
@Top-5	0.932	0.186	0.939	0.188	0.954	0.191	0.965	0.193	0.967	0.193	0.967	0.193	0.968	0.194	0.969	0.194	**0.979**	**0.196**	0.978	**0.196**

TABLE 6
Area Under Curve (AUC) Comparison on Our Datasets

Measure	Dataset	BM	Biases	BT	BaseMF	CircleCon	ContextMF	PRM	Item_based	MART-SQE	SQE
AUC	Restaurants	0.808	0.815	0.832	0.843	0.843	8.844	0.845	0.864	0.869	**0.871**
	Nightlife	0.830	0.822	0.837	0.821	0.821	0.821	0.821	0.837	0.857	**0.866**

5.3.5 The Impact of Less Training Data

For the impact of less training data on the performance of our model, the Douban dataset is used for this experiment. Table 3 shows the impact of less training data on the performance of our model on Douban dataset. In the step of model training, we randomly select some data from the complete dataset. 10 percent training data denotes that only 10 percent of our training ratings are selected for experiment, and we randomly select one of each 10 pieces of our data as the training data. It can be observed that there is little impact on performance. In addition, the performance of our model becomes worse with less training data.

5.3.6 The Impact of the Type of Prediction

In our datasets, the real overall ratings of services are discrete as [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0], while our predictions are decimals. Therefore, some experiments are conducted to discuss the impact of the type of prediction. Due to the interval of real overall ratings are 0.5, we calculate the approximation of prediction to fit the type of the overall rating. For example, given a prediction 4.3935, the approximation we get is 4.5, not 4. We leverage this method to implement experiments to discuss the impact of the type of prediction on Yelp Restaurants and Nightlife datasets. The performance of approximating prediction is shown in Fig. 11. It can be concluded that our SQE model is better than other algorithms, combining with Fig. 6 and Fig. 11, whatever the type of prediction is.

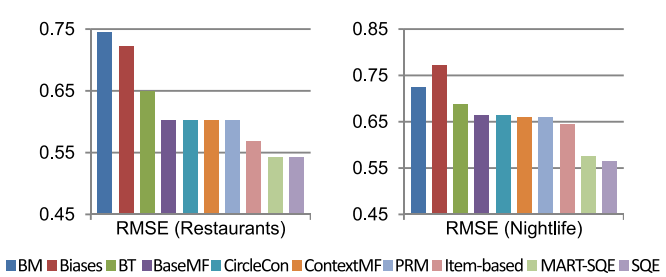


Fig. 11. The impact of the type of prediction on performance on Yelp Restaurants and Yelp Nightlife datasets.

If we quantify our prediction into [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0], there are three other measurements Precision, Recall and AUC can be used. We set the prediction sequence as the records retrieved. For example, given a prediction 3.186, the Top-5 prediction sequence is [3.0, 3.5, 2.5, 4.0, 2.0], which is ordered by the quantization distance. The sequence can be seen as retrieved records. Tables 4, 5, and 6 respectively shows the accuracy in terms of Precision, Recall, and AUC. It can be observed that SQE model and MART-SQE are better than other algorithms. Additionally, our SQE model has performed the best in terms of AUC. Moreover, the accuracy of our SQE model in terms of precision and recall @Top-1 is less than 0.4. Thus, there are many works to do for improving the accuracy of service quality evaluation.

6 Conclusions

Much research has focused on personalized recommendation and rating prediction. However, it is important to

conduct service quality evaluation, especially for the new services with few ratings. Additionally, local urban services providers can get feedback on their services from worldwide users, which are valuable for them to improve their qualities of services. In this paper, we proposed a model to solve service quality evaluation by exploring contextual information of social users. We focused on exploring user rating's confidence, which denotes the trustworthiness of this rating. Entropy is utilized to calculate user ratings' confidence. We further explored the spatial-temporal features and the sentimental features of user ratings by fusing them into a unified model to calculate overall confidence. Through our model, we can use a few ratings to predict the overall rating of services. Note that for different domains or datasets, the method of confidence calculation by entropy, which is not based on empirical observations, has wide applicability.

In future work, different aspects reflected by user reviews will be the emphasis. Usually the review text can reflect users' thoughts and the different confidences in different aspects such as color, taste and price. It can offer the more detailed quality evaluation for services.

Acknowledgments

This work was supported in part by Program 973 under Grant 2012CB316400, in part by Program of Guangdong Science and Technology under Grant 2016A010101005, in part by the National Science Foundation of China under Grant 60903121, Grant 61173109, and Grant 61332018, and in part by Microsoft Research Asia. Xueming Qian is the corresponding author.

References

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, 2001, pp. 285–295.

[3] M. Jahrer, A. Toscher, and R. Legenstein, "Combining predictions for accurate recommender systems," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 693–702.

[4] T. Ma, et al., "Social network and tag sources based augmenting collaborative recommender system," *IEICE Trans. Inf. Syst.*, vol. E98-D, no. 4, pp. 902–910, 2015.

[5] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, 2010.

[6] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 426–434.

[7] Y. Koren, "Collaborative filtering with temporal dynamics," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 447–456.

[8] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.

[9] N. Liu, M. Zhao, and Q. Yang, "Probabilistic latent preference analysis for collaborative filtering," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 759–766.

[10] Z. Fu, K. Ren, J. Shu, X. Sun, and F. Huang, "Enabling personalized search over encrypted outsourced data with efficiency improvement," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 9, pp. 2546–2559, Sep. 2016.

[11] Y. Chen and J. Canny, "Recommending ephemeral items at web scale," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 1013–1022.

[12] M. Harvey, M. Carman, I. Ruthven, and F. Crestani, "Bayesian latent variable models for collaborative item rating prediction," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 699–708.

[13] X. Yang, Y. Guo, and Y. Liu, "Bayesian-inference based recommendation in online social networks," in *Proc. IEEE INFOCOM*, 2011, pp. 551–555.

[14] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King, "Recommender systems with social regularization," in *Proc. 4th ACM Int. Conf. Web Search Data Mining*, 2011, pp. 287–296.

[15] L. Yu, R. Pan, and Z Li, "Adaptive social similarities for recommender systems," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 257–260.

[16] P. Bedi, H. Kaur, and S. Marwaha, "Trust based recommender system for Semantic Web," in *Proc. 20th Int. Joint Conf. Artifical Intell.*, 2007, pp. 2677–2682.

[17] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, and L. Sun, "Who should share what? Item-level social influence prediction for users and posts ranking," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 185–194.

[18] S. Scellato, A. Noulas, and C. Mascolo, "Exploiting place features in link prediction on location-based social networks," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1046–1054.

[19] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu, "Joint social and content recommendation for user-generated videos in online social network," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 698–709, Apr. 2013.

[20] Y. Chen, A. Cheng, and W. H. Hsu, "Travel recommendation by mining people attributes and travel group types from community-contributed photos," *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1283–1295, Oct. 2013.

[21] G. Zhao, X. Qian, and X. Xie, "User-service rating prediction by exploring social users' rating behaviors," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 496–506, Mar. 2016.

[22] G. Zhao, X. Qian, and C. Kang, "Service rating prediction by exploring social mobile users' geographic locations," *IEEE Trans. Big Data*, to be published. Doi: 10.1109/TBDATA.2016.2552541.

[23] X. Yang, H. Steck, and Y. Liu, "Circle-based recommendation in online social networks," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1267–1275.

[24] M. Jiang, et al., "Social contextual recommendation," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 45–54.

[25] X. Qian, H. Feng, G. Zhao, and T. Mei, "Personalized recommendation combining user interest and social circle," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 7, pp. 1487–1502, Jul. 2014.

[26] H. Feng and X. Qian, "Recommendation via user's personality and social contextual," in *Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage.*, 2013, pp. 1521–1524.

[27] X. Lei, X. Qian, and G. Zhao, "Rating prediction based on social sentiment from textual reviews," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1910–1921, Sep. 2016.

[28] G. Zhao and X. Qian, "Service objective evaluation via exploring social users' rating behaviors," in *Proc. IEEE Int. Conf. Multimedia Big Data*, 2015, pp. 228–235.

[29] Y. Koren, "Collaborative filtering with temporal dynamics," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 447–456.

[30] G. Dror, N. Koenigstein, and Y. Koren, "Yahoo! music recommendations: Modeling music ratings with temporal dynamics and item taxonomy," in *Proc. 5th ACM Conf. Recommender Syst.*, 2011, pp. 165–172.

[31] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proc. 4th ACM Conf. Recommender Syst.*, 2010, pp. 135–142.

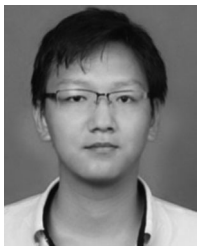
[32] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. NIPS*, 2007, pp. 1257–1264.

[33] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[34] K. Lee and K. Lee, "Using dynamically promoted experts for music recommendation," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1201–1210, Aug. 2014.

[35] X. Yang, T. Zhang, and C. Xu, "Cross-domain feature learning in multimedia," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 64–78, Jan. 2015.

[36] X. Wang, et al., "Semantic-based location recommendation with multimodal venue semantics," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 409–419, Mar. 2015.
[37] L. Hu, A. Sun, and Y. Liu, "Your neighbors affect your ratings: On geographical neighborhood influence to rating prediction," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 345–354.
[38] W. Zhang, G. Ding, L. Chen, C. Li, and C. Zhang, "Generating virtual ratings from chinese reviews to augment online recommendations," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, 2013, Art. no. 9.
[39] S. Tan, et al., "Interpreting the public sentiment variations on Twitter," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1158–1170, May 2014.
[40] S. Chelaru, I. Altingovde, S. Siersdorfer, and W. Nejdl, "Analyzing, detecting, and exploiting sentiment in web queries," *ACM Trans. Web*, vol. 8, no. 1, 2013, Art. no. 6.
[41] V. Leroy, B. Cambazoglu, and F. Bonchi, "Cold start link prediction," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 393–402.
[42] P. Lou, G. Zhao, X. Qian, H. Wang, and X. Hou, "Schedule a rich sentimental travel via sentimental POI mining and recommendation," in *Proc. IEEE 2nd Int. Conf. Multimedia Big Data*, 2016, pp. 33–40.
[43] D. Quercia, N. Lathia, F. Calabrese, G. Lorenzo, and J. Crowcroft, "Recommending social events from mobile phone location data," in *Proc. 2010 IEEE Int. Conf. Data Mining*, 2010, pp. 971–976.
[44] Y. Moshfeghi, B. Piwowarski, and J. Jose, "Handling data sparsity in collaborative filtering using emotion and semantic based features," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2011, pp. 625–634.
[45] S. Jiang, X. Qian, J. Shen, Y. Fu, and T. Mei, "Author topic model-based collaborative filtering for personalized POI recommendations," *IEEE Trans. Multimedia*, vol. 17, no. 6, pp. 907–918, Jun. 2015.
[46] S. Jiang, X. Qian, Y. Fu, and T. Mei, "Personalized travel sequence recommendation on multi-source big social media," *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 43–56, Mar. 2016.
[47] X. Qian, X. Tan, Y. Zhang, R. Hong, and M. Wang, "Enhancing sketch-based image retrieval by re-ranking and relevance feedback," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 195–208, Jan. 2016.
[48] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 233–240.
[49] A. Yates and O. Etzioni, "Unsupervised methods for determining object and relation synonyms on the web," *J. Artif. Intell. Res.*, vol. 34, 2009, pp. 255–296.
[50] J. Friedman and J. Meulman, "Multiple additive regression trees with application in Epidemiology," *Stat. Med.*, vol. 22, no. 9, pp. 1365–1381, 2003.
[51] J. Zhuang, T. Mei, S. Hoi, X. Hua, and S. Li, "Modeling social strength in social media community via kernel-based learning," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 113–122.
[52] T. Mei, B. Yang, X. Hua, and S. Li, "Contextual video recommendation by multimodal relevance and user feedback," *ACM Trans. Inf. Syst.*, vol. 29, no. 2, 2011.
[53] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput Surveys*, vol. 46, no. 38, 2014, Art. no. 38
[54] D. Lu, X. Liu, and X. Qian, "Tag based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.
[55] Y. Ren, J. Shen, J. Wang, J. Han, and S. Lee, "Mutual verifiable provable data auditing in public cloud storage," *J. Internet Technol.*, vol. 16, no. 2, pp. 317–323, 2015.
[56] Z. Fu, X. Sun, Q. Liu, L. Zhou, and J. Shu, "Achieving efficient cloud search services: Multi-keyword ranked search over encrypted cloud data supporting parallel computing," *IEICE Trans. Commun.*, vol. E98-B, no. 1, pp. 190–200, 2015.
[57] Z. Xia, X. Wang, X. Sun, and Q. Wang, "A secure and dynamic multikeyword ranked search scheme over encrypted cloud data," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 2, pp. 340–352, Feb. 2016.



Guoshuai Zhao received the BE degree from Heilongjiang University, Harbin, China, in 2012, and the MS degree from Xi'an Jiaotong University, Xi'an, China, in 2015. He is currently working toward the PhD degree at Xi'an Jiaotong University, Xi'an, China. He is a member of the SMILES LAB. He is mainly engaged in the research of social media big data analysis and recommender systems.

Xueming Qian (M'10) received the BS and MS degrees from the Xi'an University of Technology, Xi'an, China, in 1999 and 2004, respectively, and the PhD degree from the School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2008. From 1999 to 2001, he was an assistant engineer with Shaanxi Daily, Xi'an, China. He was a visiting scholar with Microsoft Research Asia, Beijing, China, from August 2010 to March 2011. From 2008 to 2011, he was an assistant professor in the School of Electronics and Information Engineering, Xi'an Jiaotong University, where he has been a full professor since 2014. He is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, and he is the director of the SMILES LAB, Xi'an, China. He has authored or coauthored more than 70 journal and conference papers. His research interests include social mobile multimedia mining learning and search. He was a TPC member of ICME, Multimedia Modeling, ICIMCS, and is the session chair/organizer of VIE '08, ICME '14, and MMM '14. He is a member of the IEEE and ACM, and a senior member of CCF. He was awarded a Microsoft Fellowship in 2006.

Xiaojiang Lei received the BE degree from Chang'an University, Xi'an, China, in 2013. He received the MS degree from Xi'an Jiaotong University, Xi'an, China, in 2016. He is a postgraduate at the SMILES LAB. He is mainly engaged in the research of social multimedia mining and recommendation.

Tao Mei (M'07-SM'11) received the BE degree in automation and the PhD degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively. He is a researcher with Microsoft Research Asia, Beijing, China. His research interests include multimedia information retrieval and computer vision. He has authored or co-authored more than 150 papers in journals and conferences, eight book chapters, and edited two books. He holds 15 U.S. granted patents. He was the recipient of several paper wards from prestigious multimedia journals and conferences. He is an editorial board member of the *IEEE Transactions on Multimedia* (TMM), the *ACM Transactions on Multimedia Computing, Communications, and Applications* (TOMM), *Machine Vision and Applications (MVA), and Multimedia Systems (MMSJ)*, and was an associate editor of *Neurocomputing*, and a guest editor of seven international journals. He is the general co-chair of ACM ICIMCS 2013, the program co-chair of ACM Multimedia 2018, IEEE ICME 2015, IEEE MMSP 2015, and MMM 2013, and the area chair for a dozen international conferences. He is a senior member of the IEEE and the ACM.