World Wide Web Journal: Internet and Web Information Systems manuscript No.
(will be inserted by the editor)

A Hybrid Recommendation System Considering Visual Information for Predicting Favorite Restaurants

Wei-Ta Chu · Ya-Lun Tsai

Received: Aug. 29, 2016

Abstract Restaurant recommendation remains as one of the most interesting recommendation problems because of its high practicality and rich context. Many works have been proposed to recommend restaurants by considering user preference, restaurant attributes, and socio-demographic behaviors. In addition to these, many customers review restaurants in blog articles where text-based subjective comments and various photos may be available. In this paper, we especially investigate the influence of visual information, i.e., photos taken by customers and put on blogs, on predicting favorite restaurants for any given user. By considering visual information as the intermediate, we will integrate two common recommendation approaches, i.e., content-based filtering and collaborative filtering, and verify the effectiveness of considering visual information. More particularly, we advocate that, in addition to text information or metadata, restaurant attributes and user preference can both be represented by visual features. Heterogeneous items can thus be modeled in the same space, and thus two types of recommendation approaches can be linked. Though experiments with various settings, we verify that visual information effectively aids favorite restaurant prediction.

Keywords Restaurant recommendation · Visual information · Content-based filtering · Collaborative filtering

1 Introduction

Recommender systems have been studied for years [1], with the targets from products [17], hotels [8], to friends [33]. Restaurant recommendation continuously is one of the most appealing topics because its wide applications in travel, as well as its strict demand on personalization [7][35]. In previous restaurant recommender

Wei-Ta Chu
National Chung Cheng University
E-mail: wtchu@ccu.edu.tw

Ya-Lun Tsai
National Chung Cheng University
E-mail: allen80311@yahoo.com.tw

(a) Indoor scene (b) Food

(c) Indoor scene (d) Food

Fig. 1 (a) Indoor scene and (b) food in American restaurants; (c) indoor scene and (d) food in Japanese restaurants.

systems, rich context like location, time, user profile, and restaurant attributes are widely used to estimate the degree of interest of one user on a restaurant [7]. Elegant methods have also been proposed to model spatial-temporal-historical features in restaurant recommendation [35]. Among the rich context being able to be used in recommendation, visual information associated with restaurants is just emerging. The study in [9] reported that visual information, e.g., food photos or restaurant images, influence users to produce food journaling. This motivates us to explore how visual information can be utilized to improve current restaurant recommender systems.

Restaurants of the same types may have similar visual appearance. Fig. 1 shows different types of restaurants in the representation of indoor scenes and food images. Fig. 1(a) shows that in American style restaurants, the decoration is bright, colorful and warm, while Fig. 1(c) shows that Japanese restaurants' decoration style is neat, simple, and clean. It is also obvious to see that food types and plate presentations are quite different. American food in Fig. 1(b) shows big amount of food and lots of fries, and Japanese food in Fig. 1(d) shows refined and elegant plate presentation. As can be seen in Fig. 1, different styles of restaurants usually have distinct visual appearance in scene and food, and in this work we conjecture that such visual characteristics can be utilized to improve restaurant recommender.

Currently recommendation methods can be roughly categorized into two types: content-based (CB) approach and collaborative filtering (CF) approach. In the CB approach [19], items are described by designated attributes or keywords, and user profiles are analyzed to describe user's preference. Items are ranked based

on the similarity between items and user preference. Two challenges remained to be studied in this approach: limited content analysis and overspecialization. Content analysis is limited to items, and it is hard to find representative attributes to describe items. When designing attributes of items, user's properties are not considered, and thus two different items with same set of item attributes are hard to be discriminated. On the other hand, all users are viewed to be independent, and the CB approach only recommends items to a user by checking this user's preference, leading to the overspecialization problem.

The second one is the collaborative filtering (CF) approach [28]. This approach computes the relationship between users and items, and recommends items to a given user based on others who present similar preference. Information about how users pick/rate items and user's preference are collected as a matrix, and techniques like matrix factorization [14] are used to discover latent factors of user selection, which is later utilized to recommend items to other users. There are also two issues remained to be addressed. The first issue is the sparsity problem. When the matrix is too sparse, too little information can be used to explore similar user behaviors or items. The items that are rarely rated are less likely to be recommended to other users. The second issue is the cold start problem. No rating or selection records are available for new users or new items, and thus there is little clue for the CF approach to appropriately recommend these new items to new users.

To simultaneously take the advantages of the CB approach and the CF approach, various hybrid recommender systems have been proposed [5]. Although many elegant methods were proposed based on text-based information or metadata, few studies considered state-of-the-art visual information to aid recommendation. Chu and Huang [8] verified that information from hotel's cover photos can be used to aid hotel rating prediction. For restaurant recommendation, we are wondering if the visual characteristics shown in Fig. 1 can be explored and utilized to construct a hybrid restaurant recommender system.

We utilize data collected from a restaurant-dedicated social platform[1]. This platform keeps crawling blog articles related to restaurants in Taiwan to build a restaurant database. Each restaurant has at least one associated blog article, which was written by users who ever consumed in this restaurant. Rich information is available in blog articles, like user comments and ratings, the restaurant's basic information, and photos users took in this restaurant. Other users who like a restaurant, after he/she visited the restaurant or after he/she browsed the webpage on the platform, can collect this restaurant into his/her *pocket*. The pocket collection thus presents favorite restaurants of users, and is the main clue to show user preference. Overall, we have the following information for study: 1) Restaurant name, ID, and varied numbers of associated blog articles; 2) user ID, user's pocket information; and 3) all text and photos in these blog articles.

Given the data collection mentioned above, we summarize contributions of this paper in the following.

- Seamlessly and finely integrate visual information into the recommendation framework: The essential idea to consider visual information is that, in addition to text information, we propose to further use photos in blog articles to represent restaurant attributes as well as user preference. We will utilize the

[1] http://hungry.9ifriend.com/main/

state-of-the-art visual features to represent photos and verify its effectiveness. Moreover, photos are actually classified into four groups, i.e., indoor, outdoor, food, and drink photos, from which visual features are extracted separately and are well fused. We will also verify the benefit brought by image classification to restaurant recommendation.
- Hybrid recommendation aided by visual information. We build hybrid restaurant recommender systems by combining the content-based approach and the collaborative filtering approach. We will design mechanisms to consider visual information in such integration, in order to build hybrid recommender systems based on three techniques, i.e., factorization machine (FM) [22], matrix factorization (MF) [14], and Bayesian Personalized Ranking Matrix Factorization (BPRMF) [23]. We will first use FM to demonstrate that visual information really aids restaurant recommendation, and then extend this idea to MF and BPRMF.

Through these designs, we investigate how visual information mitigates the sparsity problem and the cold start problem. We adopt visual features as implicit attributes to avoid limited content analysis, and reduce overspecialization by considering user preference.

The rest of this paper is organized as follows. In Sec. 2, restaurant recommender systems and related works are surveyed. Dataset and preprocessing for restaurant attributes and user preference are described in Sec. 3. Details of the baseline models and the proposed methods are given in Sec. 4. Sec. 5 presents experimental results showing the effectiveness of the proposed hybrid recommender systems, followed by the conclusion described in Sec. 6.

2 Related Works

2.1 Recommender Systems

Recommender systems have been studied for years since large amounts of items, users, and their relationships are available on the internet. Techniques for recommendation can be roughly categorized into three classes: collaborative filtering, content-based filtering, and hybrid methods.

The idea of content-based filtering (CB) is measuring the similarity between items and user profile (preference) to achieve recommendation [21][19]. In other words, content-based recommender systems try to recommend items that are similar to the ones a user liked before. Depending on the targeted domain, different features or attributes were designed to describe content, so that content-based similarity can be more accurately measured [32][20][30].

The idea of collaborative filtering (CF) is analyzing information of user's picking behaviors and discovering similar users' preference to predict what items a user may like [4]. From that, measuring similarity between users (or users' opinions) plays the key part of collaborative filtering, and some algorithms like the k-nearest neighbor approach [24] was adopted. An alternative to further consider relationships between items and discover latent factors for recommendation is the latent factor models. Currently, the state-of-the-art collaborative filtering recommendation methods are basically based on matrix factorization and its variants [14][23][34][18].

Because two approaches mentioned above separately have advantages and weakness, researchers attempt to combine them and build hybrid recommender systems [5]. Also depending on the working domain, different schemes were proposed to build hybrid methods. Shih and Liu [25] utilized customer demands derived from frequent purchased items as valuable content information. Users with similar customer demands were clustered together first, and the associate rule mining technique was adopted to extract recommendation rules. An interactive hybrid recommender system predicting items from multiple social web resources was presented in [3]. They designed an interface showing recommendation context and enabling interactive parameter setting, and demonstrated that visual representation of explanation and interaction for a hybrid system is important. Currently, deep learning techniques were also used to build hybrid recommender systems [27].

2.2 Restaurant Recommendation

Among various recommendation targets, restaurant recommendation is continuously a hot topic. Gupta and Singh proposed a location-based recommendation system that jointly considers user's location and profile to recommend restaurants [12]. Similarly, Chu and Wu also proposed to utilize mobile context in restaurant recommendation [7]. According to historical dining pattern, socio-demographic characteristics, and restaurant attributes, Zhang et al. [35] proposed a system based on conditional random field (for novelty seeking) and hidden Markov model (for non-novelty seeking) to predict a user's next dining.

In addition to location data, other context such as user ratings, reviews, and booking behaviors were also investigated to facilitate restaurant recommendation. Fu et al. [10] discovered user ratings, and proposed a generative probabilistic model to describe restaurants in multiple aspects. Geographical proximity, customer attributes, and restaurant attributes were also integrated into this model to further improve performance. In [11], multiple aspects of user reviews were also discovered by a topic model, and a regression model was built to estimate the relationship between users and restaurants. Instead of explicit user ratings, Kuo et al. [16] relied on users' restaurant booking logs to recommend restaurants. Sun et al. [29] integrated multiple sources of information, i.e., users' tasks, their friends' preferences, and mobility patterns, into the matrix factorization framework for personalized restaurant recommendation.

3 Data Preprocessing

3.1 Dataset

The evaluation dataset is collected from a social platform dedicated to restaurants. It consists of user information, restaurant information, relevant blog articles written by consumers who have ever been these restaurants. More importantly, the restaurants a user is interested in and collects into his/her pocket are also included. This information directly represents user's preference, and is actually the final target of our work – we would like to predict whether a restaurant will be picked into a user's pocket or not.

Table 1 Detailed information of the evaluation dataset.

Type	Number
Number of restaurants	30,692
Number of blog articles	120,489
Number of photos	193,935
Number of users	1,081
Number of food photos	135,966
Number of drink photos	10,665
Number of indoor photos	33,171
Number of outdoor photos	14,133

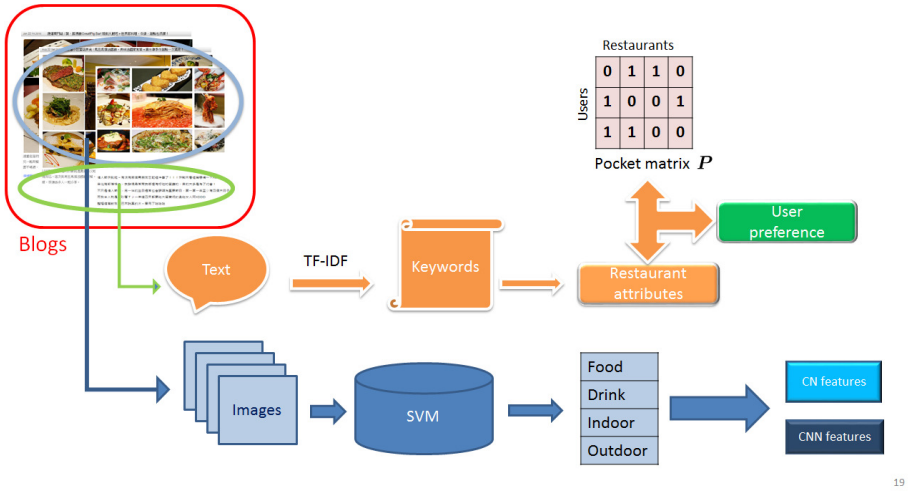


Fig. 2 The framework of collecting text and photos from blogs.

Table 1 shows detailed information of the evaluation dataset. This dataset has over 30,000 restaurants and 1000 users. Totally more than 120,000 relevant blog articles are available, and from these articles totally around 190,000 photos are embedded. These photos are further classified into four types, i.e., food, drink, indoor scene, and outdoor scene, and numbers of these four types of photos are shown in the bottom part of Table 1.

One of the main ideas of this work is that we can well utilize information from relevant blog articles. Fig. 2 illustrates the idea. From the text in articles related to a restaurant, we can extract keywords to describe this restaurant's (text-based) attributes. More importantly, we propose to utilize images embedded in articles to implicitly represent restaurant's (visual) attributes as well as user's preference. If a user collects restaurants A, B, and C as his favorites, we take images from the blog articles related to these restaurants, and extract visual features from these images to be the representation of this user's preference. This idea differs from conventional user preference presentation, and we will verify its effectiveness. In the following, we describe details of these ideas.

3.2 Text Information

From text in blog articles, we calculate term frequency-inverse document frequency (tf-idf) of each word in order to find keywords. The tf-idf value is commonly used in information retrieval and text mining. A word's *term frequency* is the number of times this word occurs in a document (blog article). A word's *document frequency* is the number of documents that contain this word. If a word appears frequently in a document but seldom in others, this word is more important. Such importance is measured by dividing term frequency by document frequency, i.e., tf-idf. Overall, important words have larger tf-idf values.

By picking words that are with high tf-idf values and are related to restaurants (manually examined), we finally keep 2,118 words as keywords. For a restaurant, we check its relevant blog articles, and represent (text-based) restaurant attributes by a 2118-dimensional binary vector $\boldsymbol{a} = (a_1, a_2, ..., a_{2118})$, which indicates whether a keyword appears in relevant articles or not. That is, if the ith keyword w_i appears one or more times in these articles, $a_i = 1$, and $a_i = 0$ otherwise.

In this work, we advocate that user preference can be described from various perspectives. In addition to restaurant attributes, here we can use keyword information to represent user preference as well. For a user u, assume that the set of restaurants this user collects is $R_u^+ = \{r_1, ..., r_m\}$, the sum of (text-based) restaurant attribute vectors is calculated as u's user preference:

$$\boldsymbol{Ft}_u = \sum_{i=1}^{m} \boldsymbol{a}^{(i)}, \tag{1}$$

where $\boldsymbol{a}^{(i)} = (a_1^{(i)}, a_2^{(i)}, ..., a_{2118}^{(i)})$ is the binary vector showing keyword appearance of the ith restaurant's related articles.

3.3 Visual Information

Another information we can get from blog articles is images. The images embedded in articles visually represent restaurants, and can also be used to describe restaurant attributes and user preference.

As image descriptors, we study to extract a state-of-the-art visual representation, i.e., convolutional neural network (CNN) features [15], as well as a conventional visual feature, i.e., color name, and will investigate their performance variations. In addition, we are also interested in whether different categories of images give different clues and thus yield different performance in restaurant recommendation. Therefore, in the following we will investigate performance variations obtained with and without image classification.

For image classification, we use a pre-trained deep network called vgg-f from the MatConvNet toolbox [31] to extract CNN features [26], and use a support vector machine (SVM) [6] to categorize images into four classes: food, drink, indoor, and outdoor. Different types of images may have different influences on restaurant recommendation. For example, food images show cuisine and ingredients, indoor images show decoration styles, outdoor images show the type of restaurants, and drink images may show the grade of restaurants, e.g., red wine only appears in expensive restaurants.

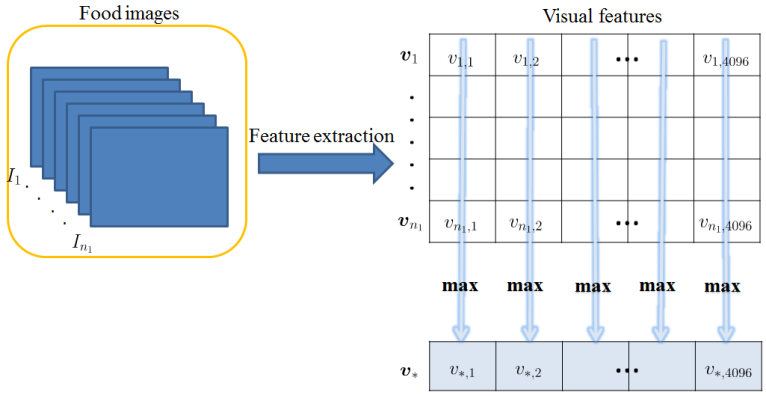


Fig. 3 An example of computing the representative visual vector for a restaurant.

For a restaurant r, images embedded in its related blog articles are collected and classified into four classes. Images in the representation of CNN features then can also be viewed as (visual) restaurant attributes. Suppose that the restaurant r have n_1 food images. From each image we extract CNN features as mentioned above, and thus n_1 feature vectors are obtained, denoted as $\boldsymbol{v}_1, ..., \boldsymbol{v}_{n_1}$. To obtain the representative food visual vector for the restaurant r, we conduct maximum pooling for each dimension of the CNN feature vector, as shown in Fig. 3. The representative vector $\boldsymbol{v}_*$ is

$$\boldsymbol{v}_* = [\max(v_{1,1}, ..., v_{n_1,1}), \cdots, \max(v_{1,4096}, ..., v_{n_1,4096})], \tag{2}$$

where $\boldsymbol{v}_i = (v_{i,1}, v_{i,2}, ..., v_{i,4096})$ is the 4096-dimensional CNN feature vector extracted from the output of the first fully-connected layer of the vgg-f model.

While Fig. 3 takes food images as the example, we can do the same process for drink, indoor, and outdoor images, and represent restaurant attributes from different perspectives. CNN features from four types of images can also be concatenated to jointly consider all perspectives. In the evaluation section, we will demonstrate performance variations based on different settings.

In order to study whether different visual features yield different performances in describing restaurant attributes, we also extract color name (CN) features [36][13] from images. CN features represent fourteen primary colors, based on eleven basic colors proposed in [2] (black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow), plus flesh, cyan, and pastel. Color of each pixel is quantized into one of the fourteen colors, and the number of pixels belonging to each color is calculated. After normalization, CN features of an image form a 14-dimensional vector, indicating the ratios of pixels in these fourteen colors, respectively.

Similarly, we can extract CN features from four types of images and represent restaurant attributes from four different perspectives, or concatenate them to jointly consider all perspectives.

4 Restaurant Recommendation

In this section, we first briefly introduce three widely used recommendation methods as the baseline approaches. Visual information will be further considered to achieve hybrid restaurant recommendation.

4.1 Baseline Approaches

Matrix Factorization (MF). Matrix Factorization (MF) [14] is a widely adopted method in recommender systems. It is a point-wise method based on user's implicit preference. Given a matrix storing how users select/collect items, the MF approach discovers latent relationships between users and items, and then predicts how likely a user would select an item that has never seen before. In our work, the pocket matrix $\boldsymbol{P}$ shown in Fig. 2 is the input of the basic MF approach. Given the matrix $\boldsymbol{P}_{N\times M}$, where N is the number of users and M is the number of restaurants, the prediction task is to find two matrices $\boldsymbol{X}_{N\times K}$ and $\boldsymbol{Y}_{M\times K}$ such that their product approximates $\boldsymbol{P}$:

$$\boldsymbol{P} \approx \boldsymbol{X} \times \boldsymbol{Y}^T = \hat{\boldsymbol{P}}. \tag{3}$$

This process maps users and restaurants to a K-dimensional latent factor space. The matrix $\boldsymbol{X}$ represents the relationship between users and latent factors, and the matrix $\boldsymbol{Y}$ represents the relationship between restaurants and latent factors.

Bayesian Personalized Ranking Matrix Factorization (BPRMF). In contrast to MF that is a point-wise method, the BPRMF approach [23] is a pair-wise method and is the state-of-art method for personalized ranking. For a user u, suppose that he/she collects a set of restaurants R^+, and the set of restaurants not collected by this user is denoted as R^-. By randomly selecting one item from R^+ and R^-, respectively, a tuple (u, i, j) denotes that the ith restaurant ($i \in R^+$) is collected by the user, and the jth restaurant ($j \in R^-$) is not. The main idea is that the user u prefers the item i over the item j, and thus the predicted score of the ith item should be larger than that of the jth item. This target is formulated as an optimization problem and solved by the gradient descent algorithm. Details of problem formulation and parameter learning please refer to [23].

Factorization Machines (FM). The MF and BPRMF approaches are basically collaborative filtering methods. On the other hand, factorization machine [22] is a method that can be adopted to combine the content-based recommendation approach and the collaborative filtering approach. We can concatenate heterogeneous feature vectors to describe relationships between users and restaurants. The relationship between a user and a restaurant can be represented by concatenating user identification, restaurant identification, restaurant attributes, and user preference (in various forms). The relationships between all considered users and restaurants are thus form a matrix. Given this matrix, relationships between different dimensions are discovered, and such relationships can be used to predict how likely a user is interested in a restaurant, by giving user identification, restaurant identification, and restaurant attributes.

4.2 Enhanced Factorization Machine

We advocate that visual information can be solely used to describe restaurant attributes or user preference, or can be viewed as an intermediate to integrated the content-based approach and the collaborative filtering approach. In the following, we will adopt factorization machine as the first instance to show how visual information can be considered in restaurant recommendation, as FM is inherently designed to jointly consider heterogeneous information. We will then design a scheme to consider visual information in MF and BPRMF approaches, so that the collaborative approaches can be enhanced by content-based ideas.

As the flexibility of FM, we can try different heterogeneous attribute combinations to constitute the input matrix. The heterogeneous attributes we consider in this work include:

- Binary vector $\boldsymbol{u}$ showing which user is indicated. Only one dimension of $\boldsymbol{u}$ can be unity. For example, $\boldsymbol{u} = (0, 0, 1, 0, ..., 0)$ indicates the third user in the dataset.
- Binary vector $\boldsymbol{r}$ showing which restaurant is indicated. Only one dimension of $\boldsymbol{r}$ can be unity. For example, $\boldsymbol{r} = (0, 1, 0, 0, ..., 0)$ indicates the second restaurant in the dataset.
- Text-based restaurant attributes $\boldsymbol{a}$ (binary vector), as mentioned in Sec. 3.2. For a restaurant, $a_i \in \boldsymbol{a}$ is equal to 1 if the ith keyword appears in this restaurant's related blog articles, and $a_i = 0$ otherwise. More than one dimensions can be unity. For example, $\boldsymbol{a} = (1, 0, 0, 1, ..., 0)$ indicates that the first and the fourth keywords were used in related blog articles.
- Visual restaurant attributes $\boldsymbol{v}$ (real-valued vector), as mentioned in Sec. 3.3. For a restaurant, CNN features and CN features are extracted from images in this restaurant's related blog articles. After pooling, visual features are used to describe restaurant attributes. We will investigate several variations here: (1) Images from articles may or may not classified into four classes (food, drink, indoor, outdoor); (2) images of different classes may be separately or jointly considered; (3) CNN features or CN features may be used.
- Text-based user preference (real-valued vector). For a user, suppose that the text-based restaurant attributes of his/her collected restaurants are $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, ..., $\boldsymbol{a}_k$, the text-based user preference of this user is set as $\boldsymbol{a}_1 + \boldsymbol{a}_2 + \cdots + \boldsymbol{a}_k$. That is, the aggregated restaurant attributes of a user's favorite restaurants are used to describe this user's preference.

Fig. 4 shows various settings that form different input matrices. For example, the first setting is jointly considering users, restaurants, and text-based restaurant attributes, while the last setting is jointly considering users, restaurants, text-based restaurant attributes, and text-based user preference.

Given the training data, matrices describing relationships between users and restaurants based on various settings are formed. For each record in the matrix, the corresponding target value is a binary number showing whether a user collects a specific restaurant. For example, following the first setting, if a record $(target; \boldsymbol{u}; \boldsymbol{r}; \boldsymbol{a})$ is $(1; 1, 0, ..., 0; 0, 1, 0, ..., 0; \boldsymbol{a})$, it indicates that the first user ($\boldsymbol{u} = (1, 0, ..., 0)$) collects ($target = 1$) the second restaurant ($\boldsymbol{r} = (0, 1, 0, ..., 0)$) in his pocket, and this restaurant's text-based attributes are represented as $\boldsymbol{a}$.

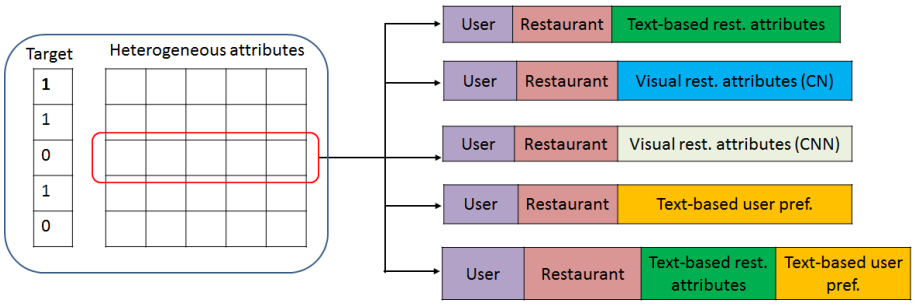


Fig. 4 The factorization machines with various heterogeneous attribute combinations.

4.3 Enhanced MF and BPRMF

Taking visual information as the intermediate to construct a hybrid recommendation system based on FM is relatively easier because of FM's inherent hybrid characteristics. For matrix factorization and its variants, jointly considering visual information is not intuitive and needs special design.

The idea to enhance MF and BPRMF (collaborative filtering approaches) is utilizing content-based attributes to update the input matrix first. The latent factor discovery processes are then adopted to recommend restaurants. More particularly, we first represent user preference by the visual features (CNN or CN) extracted from images in blog articles related to a user's favorite restaurants. To integrate visual features from multiple images, we try both average pooling and maximum pooling of these visual features as user preference. With this visual user preference, we can compute the similarity between this user preference and a restaurant's visual attributes (the vector $\boldsymbol{v}$ mentioned above). To integrate the content-based concept into the collaborative filtering approach, restaurants that have similar characteristics to user's preference are updated to the pocket matrix. Given the updated pocket matrix, we recommend restaurants to users by the MF approach or the BPRMF approach.

Fig. 5 illustrates the idea of enhancing MF or BPRMF with the content-based approaches. Two important parts are included: user preference description and matrix updating. Details of each part are described in the following.

User Preference Description. Preference of a user can be directly described as a binary vector representing which restaurants are collected by this user. In this work, because we have the third-party materials, i.e., blog articles, we advocate that the preference of a user can be more finely described by the aggregated keyword appearance or visual features extracted from the articles related to his collected restaurants.

Fig 6 illustrates the process of describing user preference. Let $\boldsymbol{P}$ denote the matrix describing which user collects which restaurant. One of the entry $P_{i,j} = 1$ indicates that the ith user collects the jth restaurant as one of his favorites, and $P_{i,j} = 0$ otherwise. In the example shown in Fig. 6, the second user's (u_2) collected restaurants are $R_2^+ = \{r_2, r_4, r_5\}$, i.e., $P_{2,2} = P_{2,4} = P_{2,5} = 1$. From each restaurant, we have calculated its representative CNN features from images in related blog articles, followed by maximum pooling (Sec. 3.3). Based on representative CNN features of r_2, r_4, and r_5, we again perform average pooling or maximum

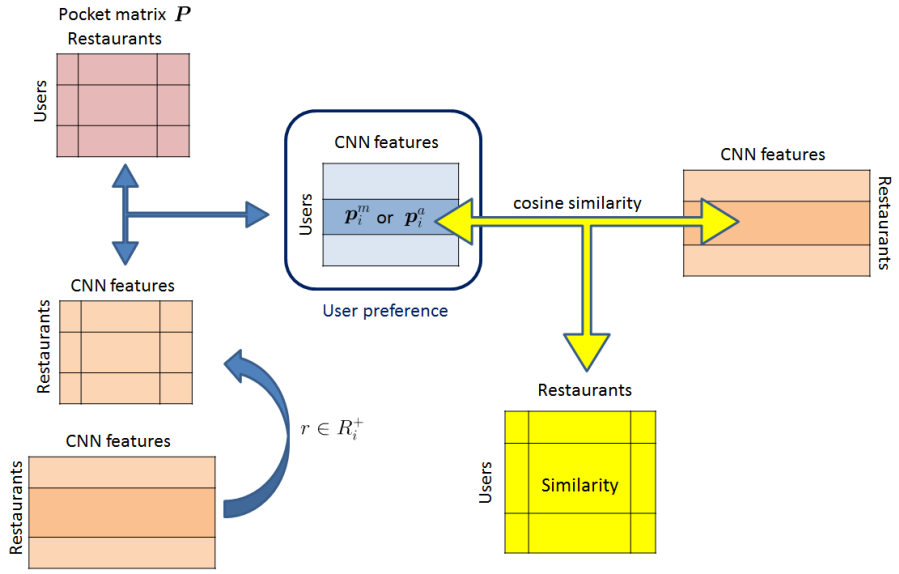


Fig. 5 Illustration of the idea to enhance MF or BPRMF with content-based ideas.

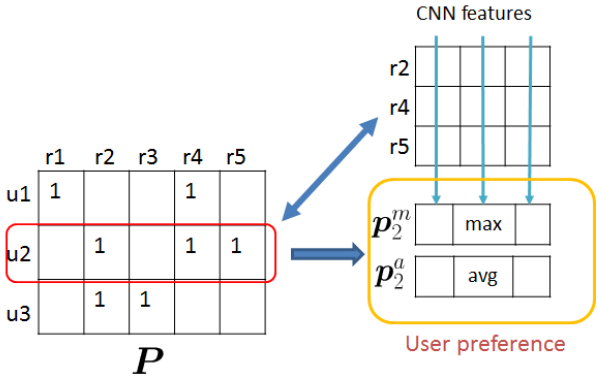


Fig. 6 Illustration of the process to describe user preference.

pooling to get the final pooling vector (denoted as $\boldsymbol{p}_2^a$ and $\boldsymbol{p}_2^m$, respectively), which is used to (visually) represent the preference of the user u_2.

Matrix Updating. The user preference vector can be the clue to connecting users with restaurants. Taking u_2 in Fig. 6 as the instance again, he has indicated r_2, r_4, and r_5 as favorite restaurants before. Based on $\boldsymbol{p}_2^a$ (or $\boldsymbol{p}_2^m$), we can calculate the similarity between u_2's preference and each of the restaurants not in the set R_2^+. For example, we calculate the cosine similarity between $\boldsymbol{p}_2^a$ and r_3:

$$Sim_{u_2,r_3}^a = \frac{\boldsymbol{p}_2^a \cdot \boldsymbol{v}_3}{\|\boldsymbol{p}_2^a\|\|\boldsymbol{v}_3\|}, \tag{4}$$

where $\boldsymbol{v}_3$ is the representative CNN features of r_3 (Sec. 3.3). Note that Fig. 6 just illustrates a very simple and small matrix. In fact, we have a 1081 (users) $\times$ 30692

(restaurants) big matrix. For any pair of user and restaurant, we can calculate the similarity. Many restaurants not being visited by a user can be checked. For those restaurants with similar visual attributes to the user preference, we conjecture that they will also be collected the user, and update the corresponding matrix entries. Particularly, if $Sim^a_{u_i,r_j}$ is larger than a threshold ψ, the matrix entry $P_{i,j}$ is updated as 1:

$$P_{i,j} = \begin{cases} 1 & \text{if } Sim^a_{u_i,r_j} \geq \psi, \\ 0 & \text{if } Sim^a_{u_i,r_j} < \psi \text{ and the original } P_{i,j} = 0. \end{cases} \tag{5}$$

With the updated input matrix $\boldsymbol{P}$, the standard MF and BPRMF methods are adopted to estimate whether a restaurant would be collected by a user as his favorites.

Overall, we use visual features to connect user preference with restaurant attributes. Through image categorization, we consider various visual aspects to avoid limited content analysis in content-based approaches. We compute similarity between user preference and restaurants to reduce overspecialization. By increasing selected items according to the content-based approach, we also implicitly mitigate the sparsity problem and the cold start problem in the collaborative filtering approach.

After all, we have the following methods to recommend restaurants. They will be compared in the evaluation section.

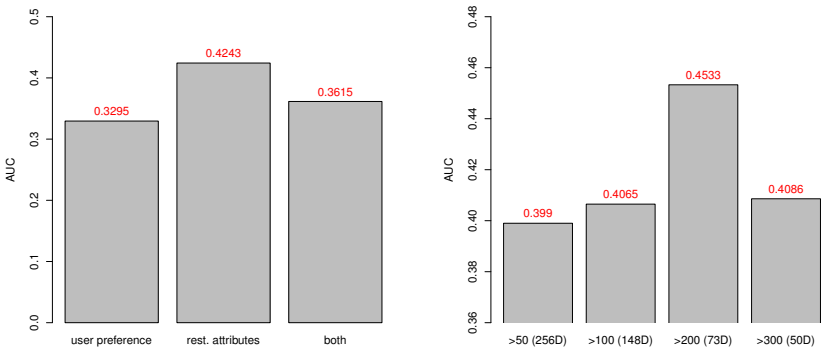
- Content-based approach: This is the simplified version of the aforementioned process. If the similarity between a user's preference and a restaurant is higher than a threshold, we recommend this restaurant to this user.
- Collaborative filtering approaches: We take the original pocket matrix $\boldsymbol{P}$ as the input of the MF method and the BPRMF method.
- Hybrid approaches: (1) The proposed enhanced factorization machine; (2) The enhanced MF method and the enhanced BPRMF method with the updated pocket matrix as the input.

5 Experimental Results

In this section we show results of different approaches with various settings. For each user, we randomly select two-thirds restaurants collected by the user and two-thirds restaurants not collected by the user, and use them as the training data. The rest one-thirds of collected restaurants and non-collected restaurants are used for testing. In real implementation, taking MF as the example, the rest one-thirds of two parts are simply set as unknown at first. These unknown values will be approximated and shown in the reconstruction matrix $\hat{\boldsymbol{P}}$ (eqn. 3). The approximated values are compared with original values, and experimental results are measured in terms of the AUC (Area Under the ROC curve) value.

5.1 Performance of Enhanced FM

In the following, we investigate enhanced FMs with various settings. Fig. 7(a) shows recommendation performance obtained solely based on text-based restaurant attributes, solely based on text-based user preference, and based on both, i.e.,



(a) (b)

Fig. 7 (a) Recommendation performance obtained solely based on text-based restaurant attributes, solely based on text-based user preference, and based on both. (b) The influence of the number of keyword on recommendation performance for pocketing.

the first, fourth, and fifth settings shown in Fig. 4, respectively. This experiment is designed to compare the influence of user preference and restaurant attributes on restaurant collection. From Fig. 7(a), we find that whether a restaurant is collected by a user or not more depends on restaurant attributes (AUC value larger than 0.4). When a user surfs on the web, the user can only see the name and thumbnail of restaurants, and thus he just chooses a restaurant based on his initial preference. But after entering the restaurant's page he can see more information and comments from the related blog articles. He can then make decision to pocket a restaurant or not based more on restaurant attributes. The performance of using both user preference and restaurant attributes is not better than only using restaurant attributes. This may be because user preference may change and not correspond to restaurant attributes. For example, if someone ate fast food frequently in this week, he may not want to eat fast food in the next week.

Fig. 7(a) shows recommendation performance when all 2,118 keywords are used. Here we want to further investigate influence of the number of keywords on recommendation performance. Fig. 7(b) shows recommendation performance variations when different numbers of keywords are used to represent restaurant attributes and user preference. The item 50(256D), for example, means that there are 256 keywords occurring more than 50 times, and these 256 keywords are used to represent text-based attributes. As can be seen in this figure, performance is worse when too many or too few keywords are used. The best performance can be obtained when 73 keywords are used.

We next demonstrate the performance of using visual features as the clues to estimate whether a restaurant will be collected by a user or not. The second and the third settings shown in Fig. 4 are compared. Fig. 8 shows recommendation performance based on CN features and CNN features respectively extracted from images without pre-classification or with pre-classification. The "with classification" shows performance based on concatenation of features respectively extracted from four categories of images. The "without classification" item shows performance

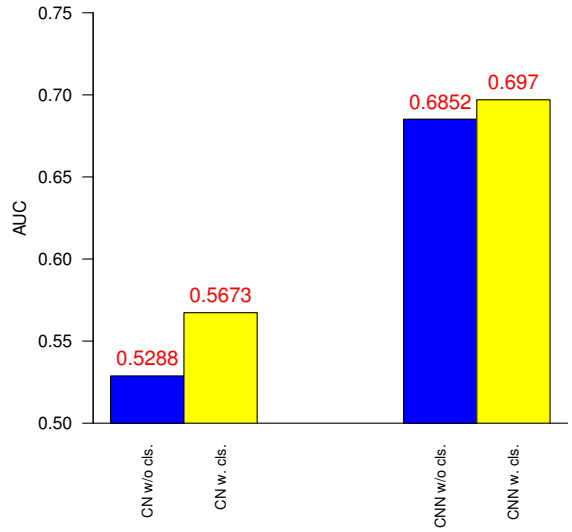


Fig. 8 Recommendation performance based on CN features and CNN features extracted from images with and without classification for pocketing.

based on features extracted from all images that are not pre-classified. As can be seen in the figure, based both on CN features and CNN features, concatenating visual features extracted from different image categories yield better performance. Another clear observation is that CNN features provide much better performance than CN features, which conform to the recent research trends showing superior performance brought by deep learning.

Overall, Fig. 9 juxtaposes results shown in Fig. 7 and Fig. 8, and shows performance comparison between text-based features and visual features as restaurant attributes and user preference. The red bars and green bars represent performance obtained based on text-based information, and the blue bars represent performance obtained based on visual information. We clearly see that visual features consistently outperform text-based features, and with image pre-classification, the best performance (around 0.7 AUC value) can be achieved. The stands for our claim that visual information provide benefits in restaurant recommendation.

5.2 Performance of Enhanced MF and BPRMF

In this section, we demonstrate the proposed MF and BPRMF methods enhanced by the content-based method. The ideas of MF and BPRMF methods are to decompose the given matrix into two matrices describing relationships between users (restaurants) and latent variables, such that the reconstruction errors are minimized. Standard learning algorithms adopt an iterative process to learn the decomposition. Previous works show that, with more learning iterations, better recommendation performance can be achieved. We therefore experiment based on

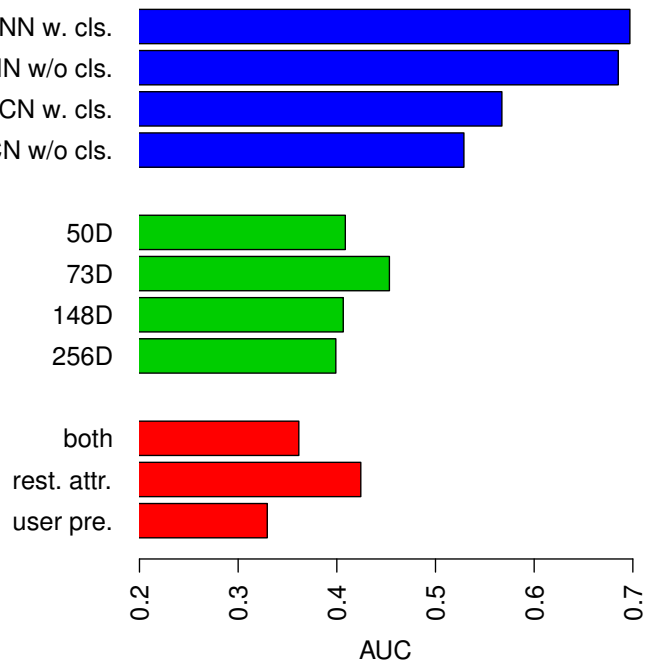


Fig. 9 Performance comparison between text-based features and visual features as restaurant attributes and user preference.

different numbers of iterations, and show performance evolution in Fig. 10. Several experiments are experimented. The curve "MF" is the conventional matrix factorization method without the aid from visual information. The curve "0.6max", for example, means that visual similarity threshold ψ is set as 0.6, and the maximum pooling scheme is adopted to find the representative visual vector (Sec. 4.3).

A few observations can be made from Fig. 10. First, with appropriate settings, MFs enhanced by visual information yield better performance than convention MF. Most of the enhancement settings consistently outperform the conventional MF. Second, better performance tends to be obtained with more iterations, but it depends on the enhancement settings. It seems that the settings "0.6max", "0.6avg", and "0.7max" keeps improved when more iterations are feasible. However, it is computationally expensive due to the huge input matrix. We don't show more iterations because of time limitation, but we expect the performance will get saturated after some point, just like other settings.

Fig. 11 shows recommendation performance vs. number of iterations based on the BPRMF approach. With the threshold $\psi = 0.7$ and the maximum pooling scheme, the best performance can be obtained. Comparing Fig. 11 with Fig. 10, performance of the BPRMF method is more robust to the number of iterations, and better performance can be achieved by the BPRMF method (the AUC value

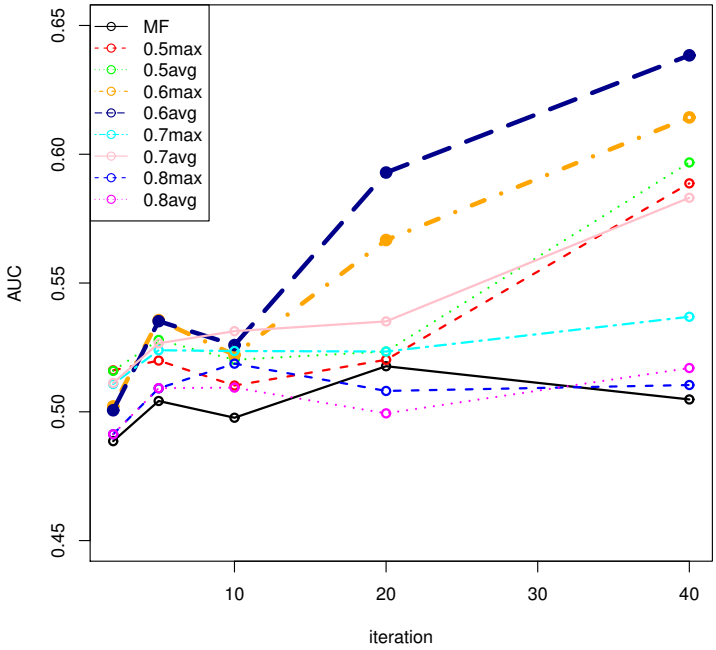


Fig. 10 Recommendation performance of MF with different settings of thresholds and pooling methods.

is larger than 0.7, comparing with the best 0.65 AUC value shown in Fig. 10). This verifies the state-of-the-art role of the BPRMF method.

5.3 Overall Performance Comparison

In this paper, we propose two categories of hybrid recommendation methods. One is taking visual features as restaurant attributes and user preference, and aggregates heterogeneous attributes/features for the factorization machine. Another is taking visual features as the intermediate to update pocket matrix, and the matrix factorization method and Bayesian personalized ranking matrix factorization method are adopted for recommendation. Each category is experimented based on various settings.

To clearly compare different approaches, we show best obtained performance of each category in Fig. 12. The first three bars show performance obtained by conventional FM, MF, and BPRMF methods. The BPRMF method is confirmed to be the current state of the art. The CB bar means performance obtained by the simple content-based method, i.e., restaurants with visual attributes similar to user preference are recommended (c.f. the end of Sec. 4). The last three bars (FM+, MF+, and BPRMF+) are performances obtained by enhanced FM, enhanced MF,

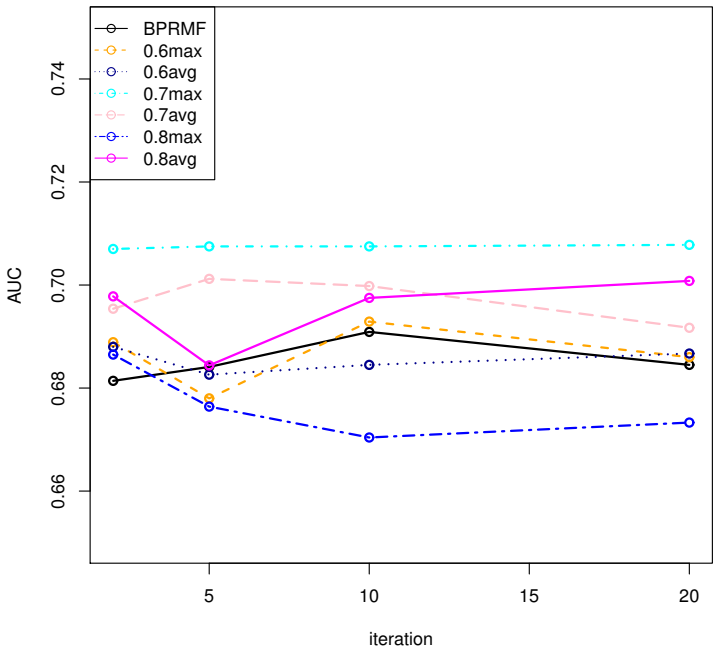


Fig. 11 Recommendation performance of BPRMF with different settings of thresholds and pooling methods.

and enhanced BPRMFM (all reported with their corresponding best settings), respectively. From that, we see that both enhanced BPRMF and enhanced FM outperform the current state of the art, and this verifies the effectiveness of the proposed ideas.

6 Conclusion

We have presented how to consider visual information in three recommendation approaches and verified the effectiveness of the proposed ideas based on several sets of experiments. By taking visual features extracted from images in related blog articles, we integrate visual information into different approaches, and construct hybrid recommender systems. With these designs, we avoid limited content analysis and reduce overspecialization by considering user preference in the content-based approach. We also mitigate the sparsity problem and the cold start problem in the collaborative filtering approach by further considering visual information. The evaluation results show performance superior to the state of the art can be obtained.

Although considering visual information improves performance, the degree of improvement of the BPRMF approach is much less than that of FM and MF. One

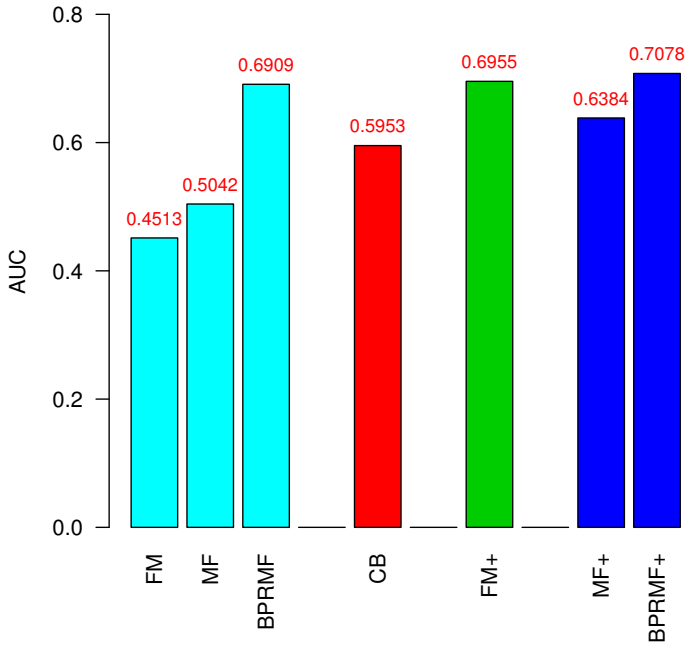


Fig. 12 All recommendation performances obtained with various settings.

of our future works is thus to further improve the BPRMF approach by integrating visual information more effectively. For example, we may design better similarity measure to describe relationships between restaurants, or learn feature weights from different types of images.

References

1. Adomavicius, G., Tuzhilin, A.: Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering **17**(6), 734–749 (2005)
2. Berlin, B., Kay, P.: Basic color terms: Their universality and evolution. University of California Press (1991)
3. Bostandjiev, S., O'Donovan, J., Hollerer, T.: Tasteweights: A visual interactive hybrid recommender system. In: Proceedings of ACM Conference on Recommender Systems, pp. 361–364 (2010)
4. Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of Conference on Uncertainty in Articial Intelligence, pp. 43–52 (1998)
5. Burke, R.: Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction **12**(4), 331–370 (2002)
6. Chang, C.C., Lin, C.J.: Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3), Article no. 27 (2011)

7. Chu, C.H., Wu, S.H.: A chinese restaurant recommendation system based on mobile context-aware services. In: Proceedings of IEEE International Conference on Mobile Data Management, pp. 116–118 (2013)
8. Chu, W.T., Huang, W.H.: Cultural difference and visual information on hotel rating prediction. World Wide Web Journal: Internet and Web Information Systems (2016)
9. Cordeiro, F., Bales, E., Cherry, E., Fogarty, J.: Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In: Proceedings of ACM Conference on Human Factors in Computing Systems, pp. 3207–3216 (2015)
10. Fu, Y., Liu, B., Ge, Y., Yao, Z., Xiong, H.: User preference learning with multiple information fusion for restaurant recommendation. In: Proceedings of SIAM International Conference on Data Mining, pp. 470–478 (2014)
11. Gao, Y., Yu, W., Chao, P., Zhang, R., Zhou, A., Yang, X.: A restaurant recommendation system by analyzing ratings and aspects in reviews. In: Database Systems for Advanced Applications, pp. 526–530 (2015)
12. Gupta, A., Singh, K.: Location based personalized restaurant recommendation system for mobile environments. In: Proceedings of International Conference on Advances in Computing, Communications and Informatics (2013)
13. Khan, F., Anwer, R., van de Weijer, J., Bagdanov, A., Vanrell, M., Lopez, A.: Color attributes for object detection. In: Proceedings of IEEE Conference onComputer Vision and Pattern Recognition, pp. 3306–3313 (2012)
14. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
15. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1106–1114 (2012)
16. Kuo, W.T., Wang, Y.C., Tsai, R.T.H., Hsu, J.Y.J.: Contextual restaurant recommendation utilizing implicit feedback. In: Proceedings of Wireless and Optical Communication Conference, pp. 170–174 (2015)
17. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing **7**(1), 76–80 (2003)
18. Liu, X., Aggarwal, C., Li, Y.F., Kong, X., Sun, X., Sathe, S.: Kernelized matrix factorization for collaborative filtering. In: Proceedings of SIAM International Conference on Data Mining (2016)
19. Lops, P., de Gemmis, M., Semeraro, G.: Content-based recommender systems: State of the art and trends. Recommender Systems Handbook pp. 73–105 (2011)
20. Musto, C.: Enhanced vector space models for content-based recommender systems. In: Proceedings of ACM Conference on Recommender Systems, pp. 361–364 (2010)
21. Pazzani, M., Billsus, D.: Content-based recommendation systems. The Adaptive Web pp. 325–341 (2007)
22. Rendle, S.: Factorization machines. In: Proceedings of IEEE International Conference on Data Mining, pp. 995–1000 (2010)
23. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Conference on Uncertainty in Artificial Intelligence, pp. 452–461 (2009)
24. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Application of dimensionality reduction in recommender system: A case study. In: Proceedings of ACM WebKDD Workshop (2000)
25. Shih, Y.Y., Liu, D.R.: Hybrid recommendation approaches: Collaborative filtering via valuable content information. In: Proceedings of Hawaii International Conference on System Sciences (2005)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of International Conference on Learning Representations (2015)
27. Strub, F., Mary, J., Gaudel, R.: Hybrid recommender system based on autoencoders. In: arXiv:1606.07659 (2016)
28. Su, X., Khoshgoftaar, T.: A survey of collaborative filtering techniques. Advances in Artificial Intelligence **2009** (2009)
29. Sun, J., Xiong, Y., Zhu, Y., Liu, J., Guan, C., Xiong, H.: Multi-source information fusion for personalized restaurant recommendation. In: Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 983–986 (2015)
30. van den Oord, A., Dieleman, S., Schrauwen, B.: Deep content-based music recommendation. In: Proceedings of Advances in Neural Information Processing Systems (2013)

31. Vedaldi, A., Lenc, K.: Matconvnet: Convolutional neural networks for matlab. In: Proceedings of ACM International Conference on Multimedia, pp. 689–692 (2015)
32. Wang, Y., Stash, N., Aroyo, L., Hollink, L., Schreiber, G.: Semantic relations for content-based recommendations. In: Proceedings of International Conference on Knowledge Capture, pp. 209–210 (2010)
33. Wang, Z., Liao, J., Cao, Q., Qi, H., Wang, Z.: Friendbook: A semantic-based friend recommendation system for social networks. IEEE Transactions on Mobile Computing **14**(3), 538–551 (2015)
34. Yu, K., Zhu, S., Lafferty, J., Gong, Y.: Fast nonparametric matrix factorization for large-scale collaborative filtering. In: Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 211–218 (2009)
35. Zhang, F., Zheng, K., Yuan, N.J., Xie, X., Chen, E., Zhou, X.: A novelty-seeking based dining recommender system. In: Proceedings of International Conference on World Wide Web, pp. 1362–1372 (2015)
36. Zheng, L., Wang, S., Tian, Q.: Coupled binary embedding for large-scale image retrieval. IEEE Transactions on Image Processing **23**(8), 3368–3380 (2014)