Research Article

Research on Cross-Platform Image Recommendation Model Fusing Text Information

Heyong Wang, Ming Hong, and Canxin Lin

Department of E-Business, South China University of Technology, Guangzhou, China

Correspondence should be addressed to Ming Hong; ming@scut.edu.cn

Received 9 April 2022; Accepted 21 July 2022; Published 25 August 2022

Academic Editor: O. S. Albahri

Copyright © 2022 Heyong Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The user data from different types of network platforms are often presented in different modalities, such as text, image, or audio. Many researches have shown that fusing the data information displayed by users on different platforms can better reflect the interest characteristics of users. Hence, this paper proposes a cross-platform image recommendation model (FITIFCIR), which fuses text and image information to achieve cross-platform data recommendation. Furthermore, it realizes the semantic information fusion of text and image, so as to recommend the images collected by users on the image sharing platform to the text to be published. Compared with baseline image recommendation models, the experimental results indicate that the FITIFCIR outperforms baseline models. The proposed model is effective to recommend appropriate images for users to better illustrate their ideas.

1. Introduction

With the popularization and development of the Internet, large quantities of the multimodal user data, such as texts and images, are gathered in many online platforms. It means that the scale, source and modality of original data are increasing. Compared with the traditional single-modal data analysis, the analysis of cross-platform and multimodal data [1] contributes to generate more complete data features. However, it is difficult to find suitable methods and models to analyze multimodal data effectively. Therefore, multimodal data analysis has become a research hotspot in the field of electronic commerce [2].

With the explosive growth of online data, users are suffering from the problem of information overload. Under this background, personalized recommendation, the technique for recommending contents, services or products according to the information of users’ interest, has captured more and more attention. The application of personalized recommendation has benefits on both users and enterprises. On the one hand, the personalized recommendation improves user experiences by saving time of searching for what they need. On the other hand, it enhances the competition ability of e-commerce enterprises [3].

User data for personalized recommendation is usually cross-platform and multimodal. As shown in Figure 1, the major form of user-generated text has changed from the long text in news or blogs to the short text in messages, forums, or microblogs. Moreover, image has become a popular form of user-generated contents. Sharing images, which is easy to express, spread, and create, just adapt to the new era of user habits and information dissemination [4]. However, with the increasing demand of users for publishing images, it is often difficult for users to find an appropriate image that matches the text content. Although the appearance of image recommendation function has solved this problem, the method and effect of image recommendation still need to be improved. In order to achieve this goal, this paper aims to provide appropriate images for text according to users’ interests and preferences through cross-platform image recommendation by fusing the information of texts and images.
The multimodal fusion techniques are applied for different fields such as spatio-temporal big data, media retrieval, and emotion analysis. et al. [9] propose a fusion system for the data of multidimensions in the scenario of traffic management. Zhu et al. [10] summarize the major characteristics of multimodal spatio-temporal big data in descriptive, explanatory, and exploratory levels. Xiong et al. [11] propose a semantic correlation analysis model based on the structure of the multimodal knowledge maps. Tang et al. [12] combine the frequency domain features of electroencephalogram signals with the image features to propose an image emotion annotation method. Lin and Meng [13] use multimodal emotion analysis method based on attention neural network in order to solve the problem of information redundancy in multimodal emotion classification task. Yu et al. [14] proposed a deep framework which applied specific deep network architectures to analyze image, text, and label information. Yu et al. [15] introduced a semisupervised method for analyzing multimodal data.

Following the ideas of Hamid et al. [16] and Albahri et al. [17], the benchmarking checklist for comparing current methods and our method is proposed in Table 1. As shown in Table 1, there are two possible ideas for generating features for multimodals from their original representations. One of the two ideas (single feature idea (SFI)) is to map the original representations into the same features space, respectively, to generate single features for multimodals. Another idea (joint feature idea (JFI)) is to generate joint features for multimodals from their original representations.

The discussion on the benchmarking checklist is as follows. Firstly, compared with single features generated by using SFI, joint features of multimodals generated by using JFI are more convenient for analysis because specific structures are not required for processing various types of single features of multimodals. Secondly, semantic correlation is important for generating features for multimodals because it is capable of improving the interpretability of features. Current methods of JFI generate joint features by fusing original representations or deep features of multimodals. The semantic correlations between modals are not considered in the fusing processes in these methods, which causes the weak interpretability for the joint features. Moreover, some methods of SFI, that is, methods proposed by Rasivasia et al. [2] and Xiong et al. [11], generate single features according to semantic correlation. However, these methods do not further generate joint features according to the semantic correlation. Thirdly, single features or joint features generated by deep learning models are more effective to be applied in various tasks such as information retrieval, recommendation, and similarity comparing because this kind of features is more suitable for computational calculation.

According to the above discussion, there is a lack of consideration on generating joint features according to semantic correlation for multimodals. Our method is design by considering the advantages of analyzing multimodal data in current methods. In our method, single features of images and text are generated as semantic words. The single features of images and text are then fused as a semantic matrix according to their semantic correlation. The matrix is finally transformed to a word embedding matrix by using the deep learning model Word2Vec.
2.2. Image Recommendation. In the field of personalized recommendation, many researchers have designed and improved the image recommendation algorithms. Sejal et al. [18] propose an algorithm to recommend images based on ANOVA Cosine Similarity where text and visual features are integrated to fill the semantic gap. Sejaletal.[19]alsopresent an image recommendation algorithm with absorbing Markov chain to retrieve relevant images for a user’s input query. Zhu et al.[20]build a similar image recommendation system based on deep hash, which can effectively analyze images by integrating the deep information of image layout, color and hue. Bo and Peng[21] build an interactive image recommendation system by using K-means clustering to represent the semantic content of the images. Widisinghe et al.[22] develop a context-aware image recommendation service that exploits collaborative filtering to recommend appropriate images with respect to the context. Li et al.[23]propose to leverage web search engine users’ behavior data to perform image recommendation and label users’ preferences for images through crowdsourcing techniques.

Following the ideas of Hamid et al.[16] and Albahri et al. [17], the benchmarking checklist for comparing current methods and our method is proposed in Table 2. It is indicated that there are three types of information which are applied to image recommendation in current researches. Current methods perform recommendation by using one of the three types of information. Text information is important because it helps to improve the interpretability of recommendation. Moreover, deep features are more suitable for the similarity calculation in the matching tasks.

According to the above discussion, there is a lack of consideration on generating deep features for image recommendation from text information. Considering the advantages of text information and deep feature information, our method is designed to perform image recommendation according to the deep features of the fusion of image class labels and image description text.

3. Research Methodology

3.1. General Processing Flow. This paper aims to provide an effective image recommendation model for recommending appropriate images for the text generated by users. An example is illustrated as follows. A user displays a post to describe a song on social media. He wants to add some images for the post but does not know what types of images are suitable to match the description text in his post. At this moment, our model can recommend appropriate images for the user according to the similarity between his post and the posts displayed by other users.

The general processing flow to collect multimodal data for cross-platform data recommendation is shown in Figure 2(a). Firstly, the various platforms for collecting user data are selected. User data of different modalities is usually distributed in different platforms. Secondly, the data processing methods are performed on the collected data to
<table>
<thead>
<tr>
<th>Checklist issues references</th>
<th>Text information</th>
<th>Deep feature information</th>
<th>Image information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sejal et al. [18]</td>
<td>(1) Recommendation based on the match of queried text and image description text.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(2) A synonym term dictionary is used for extending text queries to improve recommendation.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sejal et al. [19]</td>
<td>Recommendation based on the match of queried text and image annotation text.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Zhu et al. [20]</td>
<td>—</td>
<td>(1) Recommendation based on the match of deep features of queried images and candidate images.</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(2) Deep features of images are generated by using convolutional neural network.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Bo and Peng [21]</td>
<td>—</td>
<td>—</td>
<td>Recommendation based on the match of color histogram features or Gabor texture features of queried images and candidate images.</td>
</tr>
<tr>
<td>Widisinghe et al. [22]</td>
<td>(1) The user-generated tag information is applied to collaborative filtering for top-n recommendation.</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Our method</td>
<td>(1) Single features of images are the image class labels generated by using convolutional neural network.</td>
<td>(1) Recommendation based on the match of word embeddings of queried text and fusion matrix.</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(2) Single features of text are keyword vectors generated by using keyword extraction method.</td>
<td>(2) The fusion matrix and queried text is transformed to the word embedding matrices by using Word2Vec.</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>(3) Single features of images and text are fused to generate a fusion matrix with elements of “image class label — keyword” pairs.</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
generate features for training the data fusion model. Finally, based on the outputs of the trained data fusion model, the data item will be recommended across platforms by using recommendation algorithms. The methodology of the general cross-platform data recommendation has universal applicability and it is applicable for the multimodal data from different network platforms.

In order to verify the effectiveness of general processing flow, this paper conducts experiments for cross-platform image recommendation based on data from two major social platforms in China: Petal Net and Sina Microblog (see Figure 2(b)). Petal Net is a social sharing platform according to the user interests. Users can collect, store, classify, and label their favorite images. Sina Microblog is a social media platform based on user relationships. Users are allowed to share multimodal contents such as text, images, and videos. Under the cooperation of the Petal Net and the Sina Microblog, the accounts owned by the same users on the two platforms can be linked. Users can store the images from Sina Microblog for the Petal Net and share images from Petal Net to Sina Microblog. Therefore, when users create posts on Sina Microblog, the FITIFCIR can recommend suitable images from the Petal Net for their posts.

3.2. Technical Processing Flow. This section illustrates the overview of our model for image recommendation. The cross-platform image recommendation involves a large amount of text and images from different sources, techniques for processing text and images, and data fusion algorithms for integrating text and image information, so the technical processing flow will be composed of three stages. As shown in Figure 3, the first stage includes data acquisition and preprocessing to generate inputs for the next stage. The last two stages constitute our model FITIFCIR. The second stage performs data fusion. Different from current methods illustrated in Tables 1 and 2, our method firstly generates semantic text matrix to fuse information of images and text according to semantic correlation, and then transforms the matrix to word embedding matrix by using deep learning model Word2Vec, in order to achieve better interpretability in fusion process and more effectiveness in calculating similarity for recommendation. The third stage completes collaborative filtering (CF) recommendation for images and recommendation effectiveness evaluation. We use user-based CF, item-based CF, and hybrid CF which integrates user-based CF and item-based CF to perform image recommendation.

3.3. Data Acquisition and Preprocessing. In the first stage, we need to search for users with accounts in both Petal Net and Sina Microblog. These users will post images and text description in Petal Net and post text in Sina Microblog at the same time. These text and images are acquired and preprocessed to generate the inputs for FITIFCIR. Image transformation is conducted to transform the images into a suitable size and image coding is adopted to further compress transformed images, which will improve the efficiency of image processing methods. In addition, image classification is performed to assign class label for each image according to their features. In this paper, the CNN [24] (convolutional neural network) is used for image classification because CNN has the characteristics of high classification accuracy. The structure of the CNN is shown in Figure 4.

For the texts, word segmentation and stop word removal are conducted to achieve a bag of words. Furthermore, keyword extraction is employed to select keywords. The keywords from Petal Net mainly describe
Figure 3: Technical processing flow of cross-platform image recommendation.

Figure 4: Schematic diagram of CNN.
the information related to images, while the keywords from Sina Microblog describe the main contents of the text.

3.4. FITIFCIR. The second stage completes the fusion of image and text information. In the third stage, the implementation and evaluation of image recommendation algorithms are realized. These two processes constitute the FITIFCIR.

3.4.1. Data Fusion. The stage-based data fusion [25, 26] is the simplest method of data fusion, which only inputs the data of different modalities at different stages. It processes multimodal data independently, so it is not adopted in this paper. The feature-based data fusion [27, 28] and the semantics-based data fusion [29, 30] are adopted in second stage to fuse the information of image and text more effectively. The schematic diagram of these two methods is shown in Figure 5.

In FITIFCIR, the feature-based data fusion is used to fuse image labels and keywords of image description. We assume that there are $n$ users and user $1$ has posted $m$ images in Petal Net, so $m$ image labels ($L_{11}, L_{12}, \ldots, L_{1m}$) and $m$ groups of keywords for image description ($W_{11}, W_{12}, \ldots, W_{1m}$) will be obtained after the first stage. Any one of the image labels corresponds to a particular keyword group. For example, $L_{11}$ corresponds to $W_{11}, L_{12}$ corresponds $W_{12}$, and so on. As a result, we can get a word matrix with $n$ rows and $m$ columns, which contain the information of images and their text description in Petal Net. Significantly, all the images posted by selected users are greater than or equal to $m$, and we will randomly keep $m$ images among users with more than $m$ images. Besides, the number of keywords in each image description is also set to $w$. Only by keeping the same dimension can the similarity be calculated successfully in the next step.

The word matrix covers the word features of images and text, but these words are not related to each other semantically. Therefore, semantics-based data fusion is used to fuse data image labels and keywords to achieve semantic representation. Word2Vec [31, 32], based on CBOW and Skip-Gram algorithms [33], is adopted to convert the keywords into word vectors with the characteristic of good semantics. For example, $L_{11}$ and $W_{11}$ will be transformed into word vectors $A_{11}$ with $w + 1$ dimensions. Each user has $m$ word vectors and all the word vectors of $n$ users make up the word vector matrix. The final word vector matrix has $n$ rows and $m$ columns.

3.4.2. Image Recommendation. The third stage mainly includes implementation and evaluation of cross platform image recommendation. Based on the word vector matrix from the second stage, we use three different neighborhood-based recommendation algorithms [34] to achieve the image recommendation: the user-based collaborative filtering recommendation [35], the item-based collaborative filtering recommendation [36, 37], and the hybrid collaborative filtering recommendation [38, 39]. In the later section, we name these algorithms as the user-based recommendation, the item-based recommendation, and the hybrid recommendation for short. Introduction of the three recommendation algorithms is presented as follows.

The user-based recommendation (shown in Figure 6) aims at recommending the items of other users to the target user who shares similar interests. In FITIFCIR, we need to calculate the similarity between users, so that the images from the users with the highest similarity will be recommended to the target user. In Figure 5, each row of the word vector matrix needs to be combined into a vector, such as $A_1$ represents $[A_{11}, A_{12}, \ldots, A_{1m}]$ in Figure 6. It means that user interest features are described by all the features of images which user is interested. In addition, $A_1$ has $m \times (w + 1)$ dimensions, so the number of keywords in text from Sina Microblog is set to $m \times (w + 1)$ so that $B_1$ has $m \times (w + 1)$ dimensions as well. Taking user $v$ as an example, we need to calculate cosine similarity between all users and user $v$, so that user $u$ with the highest similarity $S_{uv}$ (see formula 1) can be obtained. Eventually, any image from user $u$ will be recommended to user $v$.

$$ S_{uv} = \frac{A_u \cdot B_v}{A_u \times B_v} \quad \text{(1)} $$

The item-based recommendation (shown in Figure 7) conducts recommendation according to the similarity between items of other users and items of the target user. In FITIFCIR, the item of Petal Net and Sina Microblog is image and text, respectively. We need to recommend images with the highest similarity of semantic features to the target user who writes text. Unlike user-based recommendation, each column of the word vector matrix in Figure 5 needs to be combined into a vector such as $[A_{11}, A_{21}, \ldots, A_{m1}]$ is expressed as $A'_1$ in Figure 7. $A'_1$ covers all the user interest features of image 1 and has $n \times (w + 1)$ dimensions, so $B'_1$ is a $n \times (w + 1)$ dimensional vector correspondingly. Taking user $v$ as an example, cosine similarity between the text from user $v$ and all the images is calculated and the highest similarity $S_{iv}$ (see formula 2) can be obtained. We will recommend image $i$ to user $v$.

$$ S_{iv} = \frac{A'_i \cdot B_v}{A'_i \times B_v} \quad \text{(2)} $$

The hybrid recommendation (shown in Figure 8) integrates the advantages of the user-based recommendation and the item-based recommendation by considering both user similarity and item similarity. In FITIFCIR, the recommended image has the high similarity of semantic features to the target text and in the meanwhile, the user who posts this image shares similar interests with target user. Furthermore, weight parameters will be added to the calculation of cosine similarity in hybrid recommendation. We suppose that $\theta_1$ and $\theta_2$ are two weight parameters and their sum is 1. Taking user $v$ as an example, we can calculate cosine similarity between the text from user $v$ and all the images from every user by adopting weighted sums of the user similarity and item similarity. If $S_{imv}$ (see formula 3) is the maximum of hybrid similarity, image $m$ shared by user 1 will be recommended to user $v$. $S_{iv}$ and $S_{mv}$ can be calculated by formulas 1 and 2, respectively.
For $\theta_1$ and $\theta_2$, 0.5 is the default value of the weight parameters, which means that the importance of the user-based recommendation and the item-based recommendation is equal. In addition, if $\theta_1$ is 1 and $\theta_2$ is 0, the hybrid recommendation is converted into the user-based recommendation. If $\theta_1$ is 0 and $\theta_2$ is 1, the hybrid recommendation is converted into the item-based recommendation. In fact, the weighting parameter needs to be adjusted for specific applications. Therefore, this paper conducts an experiment on hybrid recommendation parameter selection in order to compare recommendation performance of different parameter values.

After the implementation of image recommendation, we need to evaluate the recommendation effectiveness [38]. The evaluation methods are mainly divided into three categories: offline experiment, online experiment, and user survey. Their advantages and disadvantages are listed in Table 3. In the case of image recommendation, the result of image recommendation cannot be evaluated by objective indexes and needs real users to judge whether the recommended images match the given text description. Therefore, user survey is more suitable for evaluating effectiveness of image recommendation. Some volunteers are recruited to test recommendation performance and provide feedback accordingly. At the end of the survey, we will analyze the survey results and put forward some evaluation indexes that can reflect the effectiveness of image recommendation.

### 4. Experiment

#### 4.1. Experiment Data

The Petal Net provides 26 interest categories for users (see Table 4). Users are allowed to label their images directly into these categories. Meanwhile, this website will recommend users with images and other users from every category, so the recommended images and users are significantly representative. In this paper, we have collected 737 users who own accounts both in Petal Net and Sina Microblog in total. For Petal Net, we select 500 latest images per user on average. For Sina Microblog, we also acquire 30 latest texts per user on average. In addition, there are a lot of duplicate images in some interest categories of Petal Net. Therefore, this paper only reserves 9 interest categories for users.
categories of Petal Net as the labels of image classification. In image classification model, 500 images of each user are prefetched, of which 450 are used as training data and 50 are used as test data.

For the users selected in this article, their Petal Net accounts and Sina Microblog accounts are, respectively, associated with each other. Table 5 gives the examples of five users.

4.2. Experiment Process

4.2.1. Image Classification Model. The image classification model is used to classify the images into specific category labels according to image features on the basis of the trained model. In this paper, there are 9 category labels selected from Petal Net including children, animation, photography, pet, art, sport, food, star, and travel. After dividing 9 image sets into training set and test set, the classification accuracy of each category after image classification by CNN is shown in Table 6.

It can be seen from the table that the classification accuracy of CNN for each category is higher than 75%. Moreover, the classification performance of Pet, Children, and Food is relatively better. Generally speaking, CNN can classify images reasonably according to the semantic features of images, so the category label of each image will accurately reflect the image content.
4.2.2. Petal Net Data Processing. Figure 9 is an example of images posted by a user on Petal Net. The text is replaced by “x.” As can be seen from Figure 9, users are required to provide text title and description for images. Therefore, the data from Petal Net contains both image and text information. The collected images are input to CNN to achieve image labels. The fusion of image labels and text contents is based on feature-based data fusion. In addition, the key-word extraction technique is used to obtain keywords to represent the semantics of text. Table 7 is an example of images, text contents, and classification labels from a user in Petal Net. The Chinese text has been translated into English.

After obtaining image description and image category, we need to extract keywords from these texts to get a keyword matrix composed of a series of keywords. However, the generation of keyword matrix for the user-based recommendation and item-based recommendation is different. In the user-based recommendation, the keyword matrix reflects the most frequent categories and description keywords of all images that the user is interested in. In the item-based recommendation, the keyword matrix contains the category and description keywords of every image. Table 8 is an example of some users after data processing under different recommendation algorithms. The Chinese text has been translated into English.

4.2.3. Sina Microblog Data Processing. Figure 10 is an example of a microblog published in Sina Microblog. The user name and contents are replaced by “x.” Moreover, the other text is translated into English for better illustration. The microblogs that are not presented along with images will have the demands for image recommendation. The microblog text is processed to achieve keywords through the keyword extraction method. Table 9 is an example of texts published by a user in Sina Microblog and generated keywords. The Chinese text is translated into English.

4.3. Experiment Result

4.3.1. Cross-Platform Image Recommendation Evaluation Experiment. For the semantic-based fusion method, the keyword matrix is transformed into a word vector according to its semantic information by combining the user preference models on Petal Net and Sina Microblog. It is convenient to facilitate different recommendation algorithms to calculate the similarity and generate the recommendation results according to the similarity ranking. Two models named iLike [40] and ACS (ANOVA Cosine Similarity) [18] are applied as the baseline for comparison. Both models utilize the information of text and image.

Table 10 is an example of recommendation results after random recommendation, iLike, ACS, user-based recommendation, item-based recommendation, and hybrid recommendation for the same microblog text. The Chinese text is translated into English.
Table 7: Example of Petal Net user data.

<table>
<thead>
<tr>
<th>User name</th>
<th>Old age leaves no memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>User link</td>
<td><a href="https://huaban.com/zhengyuexu6892/">https://huaban.com/zhengyuexu6892/</a></td>
</tr>
<tr>
<td>Image</td>
<td><img src="image.png" alt="Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Title</th>
<th>Description</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maybe it is a love relationship. – The innocent and artless Mr. Volcano – CNU Visual alliance</td>
<td>Located in prince William Bay, Columbia Glacier, Chuhachi, Alaska national Forest, USA</td>
<td>Beauty</td>
</tr>
<tr>
<td>Flower, wine, Qingming Festival and tomb robbing notes</td>
<td>Travel</td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Example of Petal Net keyword matrix.

<table>
<thead>
<tr>
<th>User name</th>
<th>User-based recommendation</th>
<th>Item-based recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old age leaves no memory</td>
<td>“Pet,” “humanities and arts,” “heart is like,” “breeze,” “beautiful”</td>
<td>“Architecture design,” “heart is like,” “breeze,” “Star,” “heart is like,” “breeze” “Pet,” “heart is like,” “planet”</td>
</tr>
<tr>
<td>sher7y</td>
<td>“Home furnishing/home decoration,” “humanities and arts,” “baseball,” “wechat,” “material”</td>
<td>“Home furnishing,” “amorous feelings,” “black and white” “Pet,” “home decoration,” “brief” “Sports,” “amorous feelings,” “black and white”</td>
</tr>
</tbody>
</table>

Figure 10: Example of a microblog published by Sina Microblog user.

Table 9: Example of Sina Microblog keywords.

<table>
<thead>
<tr>
<th>User name</th>
<th>Text</th>
<th>Keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wangwanger</td>
<td>I Really miss you very much. Miss you very much. You are by my side, but I cannot be with you. Uncle, you do not want me anymore. How many people said the same words with the same number of words, and how many people left at last, waving their sleeves. The setting sun is infinitely good but it’s not as good as the sun shining in the sky. I am really old. After a few days without the sun, my mucus runs crossly. A ray of sunshine, a light, step on the bluestone again, look up to the sky ~ ~ I do not think about it. I’m reading [dictation: Chinese women marry foreign men and divorce because they cannot bear it.] Share it with you and read it together!</td>
<td>“Miss you very much,” “by my side,” “really” “The same,” “many,” “waving their sleeves” “Shining in the sky,” “runs crossly,” “mucus” “Look up to the sky,” “bluestone,” “think about it” “Dictation,” “divorce,” “bear”</td>
</tr>
</tbody>
</table>
success rate represents the proportion of the number of users who determine that the recommendation is successful in all the users. Furthermore, the criterion for judging whether the recommendation is successful is that if two or more of five recommended images are marked to match the microblog.

After filtering out the texts with no keywords and actual meanings, we randomly selects 10 real microblogs published by users to adopt random recommendation, iLike, ACS, user-based recommendation, item-based recommendation, and hybrid recommendation to recommend images for microblogs. Among them, random recommendation will randomly recommend several images from all the collected images for the microblogs. The parameter weights \( \theta_1 \) and \( \theta_2 \) of hybrid recommendation are both 0.5. Generally, we provide evaluation of the six recommended image sets for the same microblog content. In this paper, three experts in related fields have been invited to complete the survey based on questionnaire. Each expert has passed the attention test and has been told that it is necessary to carefully compare the content of text and image. Eventually, experts need to determine the matching degree of all image sets. The evaluation results of the six recommendation algorithms are shown in Table 11.

It is indicated that the three recommendation methods of this paper are significantly better than the recommendation results of the random recommendation algorithm. The results of the three recommendation algorithms are ranked from high to low: hybrid recommendation, item-based recommendation, and user-based recommendation. Compared with the recommendation model proposed by other researches, item-based recommendation and hybrid recommendation have a better performance.

It can be concluded that in the field of image recommendation, the strategy of random recommendation algorithm is not feasible, and the cross-platform image recommendation algorithm for multimodal data is an effective method to improve the image recommendation effectiveness. It shows that the recommendation results of hybrid recommendation are optimal.

### 4.3.2. Hybrid Recommendation Parameter Selection Experiment

In the last experiment, parameter weights \( \theta_1 \) and \( \theta_2 \) of hybrid recommendation are artificially set to 0.5. Under this situation, the initial hybrid recommendation is
better than the user-based recommendation and item-based recommendation. The recommendation success rate of item-based recommendation is slightly higher than user-based recommendation, so we can speculate: if the hybrid recommendation is more biased towards item-based recommendation (i.e., increasing the $\theta_2$ value), it will further improve the effect of hybrid recommendation.

This experiment will select user-based recommendation, item-based recommendation, and initial hybrid recommendation as reference groups. Their ratio values of $\theta_1$ to $\theta_2$ are "1:0," "0:1," and "0.5:0.5." In addition, eight experimental groups with gradient distribution are set in this experiment. Their ratio values of "$\theta_1 : \theta_2$" are "0.1:0.9," "0.2:0.8," "0.3:0.7," "0.4:0.6," "0.6:0.4," "0.7:0.3," "0.8:0.2," and "0.9:0.1." In addition, 200 volunteers are invited to subjectively evaluate the image recommendation results. The evaluation results are shown in Table 12.

It clearly indicates that as the ratio of two parameter weights increases, the matching rate of the recommendation algorithm and the recommendation success rate will fluctuate. In other words, if the item-based recommendation has a higher weight in hybrid recommendation algorithm within limits, the performance of hybrid recommendation will improve, which is consistent with our hypothesis based on the cross-platform image recommendation evaluation experiment.

From the table, we can also find that the best recommendation result is achieved at "0.2:0.8." At "0.3:0.7" and "0.1:0.9," the hybrid recommendation results are also good. Furthermore, when $\theta_2$ increases from 0.9 to 1, both the matching rate and recommendation success rate decrease significantly. It shows that the contribution of image similarity is greater than that of user similarity in hybrid recommendation. However, if only the image similarity is considered, the performance of recommendation algorithm will be worse.

### 5. Conclusion and Discussion

Addressing the problem of cross-platform image recommendation based on multimodal data, this paper proposes the model FITIFCIR to recommend images for the texts without any image. Our model applies the multimodal data fusion algorithms to fuse the information of text and images and adopts different collaborative filtering algorithms for recommendation. The main experimental results of this paper are as follows. (1) In FITIFCIR, the evaluation indexes of user-based recommendation, item-based recommendation, and hybrid recommendation are obviously better than baseline models. The hybrid recommendation that has considered both the user similarity and image similarity performs best. (2) For the hybrid recommendation, the recommendation performance will be improved when the parameter weights bias towards the item-based recommendation within limits. The above conclusions show that the model FITIFCIR is effective and the cross-platform image recommendation based on multimodal data is feasible.
The research of this paper still has some deficiencies. The direction of future discussion can be considered on the following two aspects. (1) More data of other modalities, such as audios and videos, can be taken into account to reflect users' interest characteristics to a certain extent. (2) Cross-platform recommendation for more types of social media platforms can be considered to make full use of user interests to better promote the development of the e-commerce industry. For example, recommending products to users on online shopping platforms according to the mining results of users' interests on video sharing platforms.

Data Availability

The data used to support the findings of this study are collected from Petal Net and Sina Microblog.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was supported by Project of National Nature Science Foundation of China, Grant no. 71731006, the Fundamental Research Funds for Guangdong Natural Science Foundation, Grant no. 2022A1515011848, and Guangzhou Philosophy and Social Science, Grant no. 2020GZYB04.

References


