Research on Image Tagging Recommendation based on Relevance and Diversity

Hong GUO
Hunan City University, Yiyang, 413000, China

Abstract: The traditional recommendation algorithms of image tagging ignore the diversity between the visual content information and the tags recommended, which causes the recommended results have the problem of tag ambiguity, tag redundancy and so on. Therefore, this paper proposes the recommendation algorithm of image tagging based on relevance and diversity. The algorithm defines the relevance and diversity of a label set, and selects a label set which can reasonably balance the relevance and diversity to recommend to the user. The experimental results show that this algorithm improves the relevance between the recommended results and the image, and makes the recommended results be able to reflect the image content thoroughly at the same time.

Keywords: Relevance; Tagging; Vision distance; Topic coverage

1. Introduction

The number of the images on the Internet presents an explosive growth. In order to effectively organize and control such massive scale of the image resources, the image retrieval technology emerges at this historic moment, and has been widely studied. Since the 1990s, the content-based image retrieval has been developed constantly, but due to the existence of the “semantic gap” between the image’s low-level visual features and the high-level semantic concepts, the retrieval performance of CBIR is difficult to be satisfactory [1-3]. Therefore, the current commercial image retrieval engines (Google Image, Bing Image) still adopt the Text-based Image Retrieval (TBIR) approach, which creates index through the text information of the image, and uses the mature text retrieval algorithm to provide image retrieval service to the user, its retrieval performance is dependent on the quality of the image’s relevant text [4].

In recent years, the image sharing sites represented by Flicker flourished. In Flicker [5], users can define the semantic keywords of the image, and these keywords are called image tags. The image tags are used by users to describe the image’s semantic content, which provide reliable retrieval basis for TBIR. At the same time, the image sharing sites often classify and organize the images according to the image tags, which makes the users be willing to add tags to the images, because by doing so can make it easier for others to find the images [6-7]. Thus, how to help users to add tags to the images rapidly and accurately becomes a very important problem, while the image tag recommendation system is an important algorithm to solve the problem.

As shown in Figure 1, the image tag set recommendation means that in the process when the users are adding tags to the image, it find some new tag candidates for the users to choose from, according to the image content and the preliminary tags, that is the tags already added by users. The image tag recommendation system can provide helps to the image annotation and the subsequent image retrieval from the following three aspects. (1) Prompt the users to add more tags. In the process of adding tag to the image, the users often cannot come up with a large number of tags in a short period of time, while the tag recommendation system can provide image tag candidates for them, which reduces the workload of the users, and makes them be willing to add more tags. (2) Help the users to use more accurate and professional tags. Statistics show that in Flicker, the number of the tags that frequently used accounts for only about 6% of the total number of the tags. Many tags which can more accurately and professionally describe the certain object or scene are ignored by the users due to their less usage in the daily life. While the high-quality tags recommendation system can provide more accurate and professional tags and rich the vocabulary of the image annotation, according to the image content. (3) Reduce the occurrence of the noise emission labels. Noise emission labels refer to the label words with some spelling mistakes or being meaningless. The tag recommendation system transforms the process of label adding from typing into selection, which effectively avoids the occurrence of the noise emission labels.

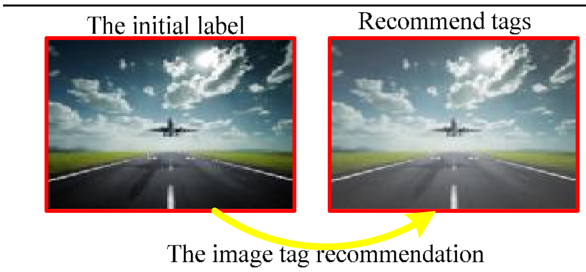
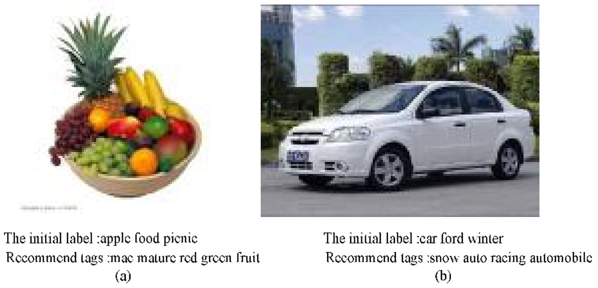


Figure 1. Simple graph of image tag recommendation

The previous image tag recommendation algorithm often makes use of the tag co-occurrence to recommend the tags that show a high co-occurrence of similarity with the preliminary image tag set to the users. Figure 2 presents two cases that use this algorithm to obtain the recommended results. It is thought that this kind of recommendation algorithm based on the tag co-occurrence has two following questions:

The problem of tag ambiguity. The performance of the algorithm is easily to be affected by the ambiguous labels, due to there is no consideration of the correlation between the tag and the image content. As shown in Figure 2 (a), because of the ambiguity of the preliminary tag "apple", and under the condition that only take the tag co-occurrence into consideration, the recommendation algorithm cannot make sure the true meaning that the image expressed, which will therefore lead to the existence of the tags that have nothing to do with the image content being recommended, such as "Mac".

The problem of tag redundancy. The tags recommended are often the synonyms and near synonyms of the preliminary tag, or the key words that describe the same concept with the preliminary tag, which cannot bring new information to the users. As shown in Figure 2 (b), although the recommended tags "auto", "automobile" have high correlation with the image content, they cannot provide new information to describe the image content, because the tag "car" has already been included in the preliminary tags. While the users want to get the tags that can describe the image content from different angles, such as "tree", "sky" and so on, when they are adding tags to the image.



The initial label :apple food picnic
Recommend tags :mac mature red green fruit
(a)

The initial label :car ford winter
Recommend tags :snow auto racing automobile
(b)

Figure 2. Results of the Recommendation Algorithm Based on Tag Co-occurrence

In order to solve the above problems, this paper proposes an image tag recommendation algorithm based on correlation and diversity. Given an image and its preliminary tag set, the algorithm hopes to find out a set of tags that satisfy the following two conditions: (1) Relevance. The tag has semantic relevance with the content that the image described. (2) Diversity. The tag is able to reflect the content information of the image from different aspects [15].

First of all, use the visual language model to respectively calculate the relevance between the tag and the image, and the visual distance between the tags. Based on this, the relevance and diversity of a label set are defined. The goal of the recommendation algorithm proposed in this paper is to find a label set with the specified size, making the set achieve a reasonable balance between the relevance and diversity, and recommends the label set to the users. The experimental results in the real data set show that the algorithm of this paper is superior to the current representative algorithm in the aspects of precision rate, topic coverage and F1 measure.

2. Relevance and Visual Distance

Use the visual language model to respectively calculate the relevance between the tag and the image, and the visual distance between the tags. First of all, learn the visual language model of each label by using the data set, and express the visual concept that the tag represented through that model. Then combine the co-occurrence similarity between the tag and the initial tag set with the visual similarity between the tag and the image, to calculate the relevance between the tag and the image. Finally calculate the visual distance between, through the Jensen-Shannon divergence between the visual language models of the two tags.

2.1. The visual language model

Using the visual language model to express the visual concept the tag represented. VLM is the expansion of the traditional statistical language model, which is shown by the Bag-of-Visual-Words based on images. VLM thinks that the visual words in the images are interdependent on the space, the arrangement of the adjacent words abides by some kind of visual grammar, and that a visual concept can be expressed by specific visual grammar.

Given a tag t, and sets the image set that contains the tag t in the data set to be St. Figure 3 shows the process that t creates VLM. Divide each image in St into a lot of patches with the same size and without occlusion, extract the feature description vectors with the same dimension from each patch, and using the clustering algorithm to encode the features into a visual word. VLM assumes that the visual words in the image are generated in the order from left to right and top to bottom, therefore, an image is represented as a visual word sequence, and the appearance condition of each visual word depends on its previous visual words. VLM of tag t obtains the dependence relationship between the visual words by estimating the conditional probability distribution of the visual words appeared in St, while this dependence relationship reflects the visual concept that the tag expressed.

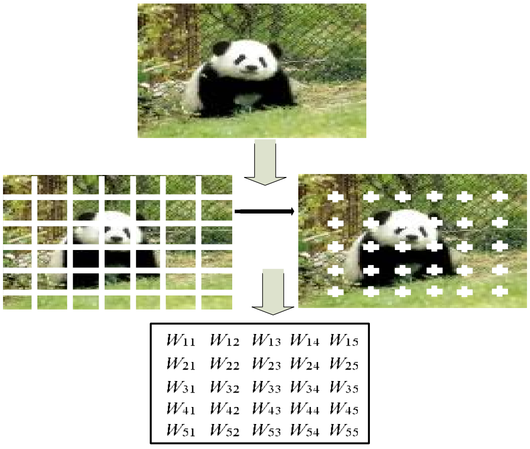


Figure 3. Generation Process Diagram of the Bigram Visual Language Model

When estimating the conditional probability, the model of the foregoing N visual words are being considered, which is called the N-gram Visual Language Model. For the comprehensive consideration of preformance and efficiency, this paper adopts the Bigram Visual Language Model (Bigram VLM), which holds that the appearance of the current visual word only relies on its left visual words.

$$p\left(p_{ij}\middle|p_{11},p_{12},...,p_{mn}\right)=p\left(p_{ij}\middle|p_{i,j-1}\right) \qquad (1)$$

p_{ij} Refers to the visual word in the ith row and jth column, $\left(p_{11},p_{12},...,p_{mn}\right)$ is the visual word sequence before p_{ij} .

Estimate the simplest algorithm of $p\left(p_{ij}\middle|p_{i,j-1}\right)$ is the Maxi- mum Likelihood Estimation (MLE), and set $count\left(p_{i,j-1},p_{ij}\right)$ to present the occurrences of the bigram grammar $p_{i,j-1},p_{ij}$, p presents the set of the different visual words.

$$s\left(p_{ij}\middle|p_{i,j-1}\right)=\frac{count\left(p_{i,j-1},p_{ij}\right)}{\sum_{s\in p}count\left(p_{i,j-1},p\right)} \qquad (2)$$

Due to the data sparsity, the training set may not be able to cover all the bigram grammars, and the direct using of MLE will lead to the happening of $p\left(p_{ij}\middle|p_{i,j-1}\right)=0$, therefore, the smoothing process is needed. This paper adopts the following smoothing algorithm, which combines the probability fallback technology with the probability discount technology.

$$d=1-\frac{n_1}{R} \qquad (3)$$

In formula (3), if the bigram grammar $p_{i,j-1},p_{ij}$ falis to appear in the training set, then use the probability fallback technology to calculate its conditional probability through the distribution of the unigram p_{ij} in which b is the fallback factor. And if the bigram grammar $p_{i,j-1},p_{ij}$ appears in the training set, then use the probability discount technology toreduce the estimated value of the conditional probability, in which d is the linear discount factor. As shown in formula (3), n_1 presents the number of the visual words whose occurrence number is 1, R refers to the total number of different visual words. Many experimental results show that the VLM with linear discount can achieve better performance.

2.2. Relevance between the tag and the image

Given an image i and its initial tag set m_i , and for a tag m , separately calculate the co-occurrence similarity between m and m_i , and the visual similarity between m and i , which commonly measure the relevance between t and i .

The Calculation of the Tag’s Co-occurrence Similarity

When the users are adding tags to the images, they always tend to use the tags that can reflect the image content. If there are two tags which always are added to the image at the same time, then it shows that the concepts the two tags represented are more likely to appear together. Therefore, if there is a high co-occurrence similarity between m and m_i , then t is more likely to reflect the content of i . The co-occurrence between the two tags m_i and m_j is defined as follow:

$$r\left(m_i,m_j\right)=\frac{\left|m_i\cap m_j\right|}{\left|m_i\right|} \qquad (4)$$

$\left|m_i\right|$ represents the number of the images which contain the tag m_i in the data set. Intuitively, $r\left(m_i,m_j\right)$ represents the image’s possibility to obtaining the tag m_j after the obtaining of tag m_i . Based on this definition, the co-occurrence similarity $s\left(m_i,m\right)$ between the tag m and the initial tag set m_i is defined as the sum of the co-occurrence similarities between t and each initial tag.

$$s\left(m_i,m\right)=\sum_{m_{i\in}m_j}s\left(r\left(m_i,m\right)\right) \qquad (5)$$

$s\left(.\right)$ is a monotonic increasing smooth function.

2.3. The visual distance between the tags

The previous image tag recommendation algorithm only considers the relevance between the recommended tag and the image, ignoring the relationship between them, which makes the recommended tags often represent the same or similar concepts. While an image always contains a variety of concepts, such as different objects and so on, thus the recommended results obtained through the previous algorithm may not be able to thoroughly reflect the content information of the image.

3. The Image Tag Recommendation Algorithm

Combine the above relevance between the tag and the image with the visual distance between the tags, this section introduces the image tag recommendation algorithm that combines the relevance and the diversity. The relevance and diversity of a label set are defined firstly, and then use the greedy search algorithm to find the tag set that can reasonably balance the relevance and the diversity. At the end, treat the tag set as the final recommended result.

3.1. The relevance and diversity of the tag set

In the previous image tag recommendation algorithm, the problem of the tag recommendation tends to be converted into the problem of tag ranking according to the relevance between the tag and the image, and the algorithm recommends the tag with a high ranking to the users. While the image tag recommendation algorithm proposed in this paper takes the interrelation between the recommended tags, thus the goal of the algorithm is to recommend a tag set with a specified size.

3.2. Algorithm description and time complexity

In the process of image tag recommendation, the algorithm proposed in this paper hopes to find a tag set that can reasonably balance the relevance and the diversity. Given the target image I and its initial tag set m_i , the algorithm chooses the tag set with a highest score of the balance degree between the relevance and the diversity in the remaining tags, and recommends it to the users. And it is as follows:

$$k_m^* = \arg\max h(s_t), s_t \subset M / M_i \tag{6}$$

M represents the collection of all tags in the data set. The solution of the formula (6) is a typical problem of non-linear integer programming, which belongs to the problem of optimization combination of NP-Hard class, and there is no accurate algorithm within the polynomial time. Thus, the greedy search algorithm is used to find out the near-optimal solution to the problem, and the solving process is shown in algorithm 1.

Initially, k_m^* is initialized to an empty set (line 1). First of all, the algorithm finds out the tag m_i with the highest relevance with the image in the remaining tags except of tag m_i, and treats m_i as the first tag to join in k_m^* (line 2 –line 3). Then, the algorithm iteratively finds out the remaining $m-1$ tags. In each round of the iteration, finds out the tag m_r in the remaining tags except of m_i and k_m^*, which is the tag that can make the current k_m^* become the tag with the highest score after its join, adds m_r into k_m^* (line 4 – line 7). Finally, the set k_m^* contains m tags, and return the k_m^* as the recommended result.

Before the start of the recommendation algorithm, first of all, train out the VLM of each tag in the data set offline, and calculate the co-occurrence similarity and visual distance between any two tags. The time complexity of algorithm 1 is $o(mn_2)$. In which, m is the expected number of the recommended tags, and m is the total number of the tags in the data set. In the actual calculation, the value of n is generally small ($N=10$ in the experiment), thus the running time of the algorithm mainly depends on the total number of the tags in the data set. The algorithm that can effectively improve the running efficiency is the one that will rank all the tags according to its relevance with the image in the first place when computing, and then in the basis of the performance requirements, select a number of tags which are in the top of the ranking, to continue the calculation in algorithm 1.

Algorithm 1 Tag Recommendation Algorithm Based on Greedy Search

Input: all the tags T in the training set, an image I , the initial tag set

Output: the recommendation tag set k_m^* with a size of m

```
1) Initialize $k_m^* = \Omega$;
2) $T_i$ of $i$, the expected number $N$ of the recommended tags
3) $k_m^* = k_m^* \cup \{t_i\}$
4) For $i = 2$ to $N$ do
5) Select tag $T / \{T_i \cup k_m^*\}$ from $t_i, t_j$ satisfy:
   $t_r = \arg\max_{t_r} R(i, t_r)$;
6) $k_m^* = k_m^* \cup \{t_r\}$;
7) End for
8) Return $k_m^*$.
```

4. Simulation Test and Analysis

4.1. Expeirmental environment and settings

In order to verify the effectiveness of the algorithm proposed in this paper, the NUS-WIDE data set is used as the experimental data set. The data set are 269648 images and 425059 different tags provided by about 5000 users from Flicker, the image centents contain a rich variety of objects and scenarios, which reflect the real situation of the massive images in the Web. Because the NUS-WIDE data set contains a lot of noise emission labels, the filter operation is firstly made to the tags in the data set. Remove the tags that miss the index of the WordNet or with a ocuurrence less than 50 times, and stems the remaining tags, ultimately, there are will be 4377 different tags retained.

Figure 4 provides the statistics of the number each tag occurs in the data set. From which it is known that they present the approxiamte features of the long-tailed distributions. Among them, the tags with a occurrence more than 5000 are less than 1%, which always represent the relatively commom and universal concepts, such as “nature” , “color” and so on. While the tags with a occurrence more than 500 are only 20%, and the tags with a occurrence less than 100 are more than half. Many tags that with a less occurrence always can accurately describe a particular scene or object, such as “purple”, “puss” etc.

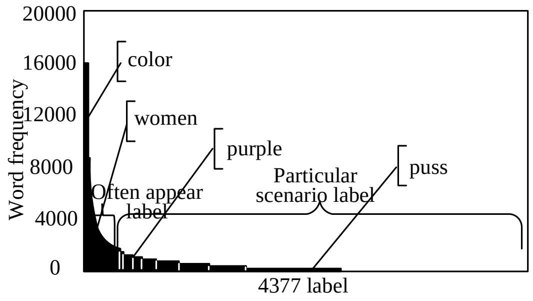


Figure 4. Statistics of tag’s occurrence number in the data set

In the experiment, to reduce the effects of the image’s size changes on the results, all the images are adjusted to the size of 320×320 pixel. Each image is evenly divided into multiple image blocks with a pixel of 8×8, and extracts the 8D texture gradient histogram from each image as the feature description vectors. This kind of feature has the characteristics of low dimension and scale invariance, by using it the VLM can achieve better performance. When establishing the visual dictionary, the size of the the dictionary is set to 300.

Respectively and randomly select 500 images as the validation set and the test set, in which the validation set is used to determine the optimal values of the parameters in the algorithm, while the test set is used to evaluate the performance of the algorithm. Use all the remaining images to train the VLM of the tags and calculate the co-occurrence similarity and visual distance between the tags. The smoothing functions in formula (6) are defined as the standard sigmoid functions, and the smoothing function in formula (6) is defined as the logarithmic linear smoothing function.

For each image in the validation set and the test set, different recommendation algorithms all produce 10 recommended tags. There are three volunteers independently judge the relevance of the tags, finally, the voting algorithm is used to determine whether if the tags are related to the image content. In the experiment, the Cohen’s Kappa statistics between each two volunteers is counted, the calculation results show that the average Cohen’s Kappa coefficient of the three volunteers is 0.77, which is more than the conforming optimal boundary of 0.75, indicating that the volunteers gain better consistency in judging the relevance between the recommended tags, and proves that the artificial judging of the experiment is reliable.

4.2. Results analysis

The Influences of Parameter Settings on the Algorithm Performance

The influences of the two parameters involved in the algorithm on the performance will be inspected through the experiment. The two parameters are respectively the parameter h in formula (6) and the parameter l in formula (6).

When calculating the relevance between the tag and the image, h is used to adjust the weight between the co-occurrence similarity and the visual similarity. First of all, the value of l is set to 0.5, then observe the performance of the algorithm in the data set when respectively set different values to l . Figure 5 shows the results of the experiment. It can be seen from the figure that, when the value of h is 0.5 or 0.6, the performance of the algorithm is the best. Which states that when calculating the relevance between the tag and the image, the proportion of the co-occurrence similarity and the visual similarity should be more balanced distributed. In the experiment, when the value of l varies within the range of [0.42, 0.81], the optimal value of h is not obviously affected. Thus, in the latter experiments, the value of h is set to 0.5.

In formula (6), l is used to adjust the proportion of the relevance and the diversity. In order to clearly understand the impacts of l , the experiment results of the algorithm obtained in the validation set are calculated when setting different values to l . Intuitively, if the value of l is too small, there will be irrelevant tags introduced into the recommended results; instead, if the value is too large, there probably will have the tags with semantic redundancy in the recommended results. In both cases, the algorithm all cannot obtain the optimal performance. The influences of the value changes of l on algorithm performance are shown in figure 6, and the same conclusion can be got from it. When the value of l is 0.6 or 0.71, the performance of the algorithm is the best. The main reason is that when evaluating the algorithm, the diversity of the relevant tags is only considered, and the algorithm will obtain the best performance in the situation of the guaranteeing of a high relevance of the recommended results. In the latter experiments, the value of l is set to 0.6.

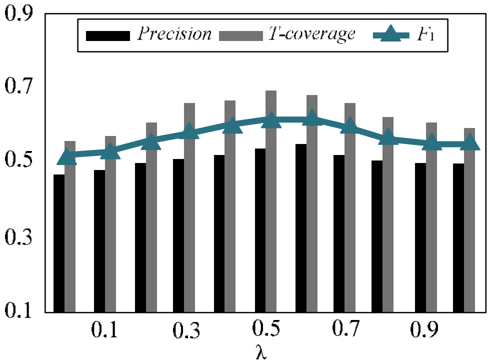


Figure 5. The influences of the value changes of η on algorithm performance

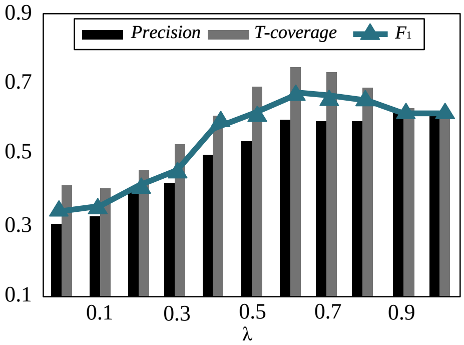


Figure 6. The Influences of the value changes of l on algorithm performance

The Comparison and Analysis of the Relevant Algorithms

In this experiment, the effectiveness of the image tag recommendation algorithm combined with the relevance and the diversity proposed in this paper is verified, by comparing with several other algorithms. The algorithms involved in here include TC, the image tag recommendation that uses the co-occurrence of the tag, MRR, the image tag recommendation based on the modal relevance, CR, the image tag recommendation based on image synergy, RD, the image tag recommendation combined with the relevance and the diversity proposed in this paper.

Figure 7 shows the results of the four algorithms in the test set. It can be seen that MRR wins the highest precision. The advantage of this algorithm lies in its considering of the modal relevance between the tag and the image, and using the Rank boost algorithm to put them together. RD is slightly lower than MRR in the aspect of precision, but is still increased by 6% compared with TC. That is mainly because RD combines the co-occurrence similarity with the visual similarity when calculating the relevance. The precision of CR is lower, probably due to the images in the data set are rich and varied, making it is unable to accurately find out the images with similar semantic meanings. Compare with the other three kinds of algorithms, RD achieves the best performance in the aspect of topic coverage, and is respectively 16%, 11% and 14% beyond, which proves that RD can better ensure the diversity of the recommended results. It can be seen that the algorithm proposed in this paper better balance the relevance and the diversity of the recommended results and it also gains the highest value of F1 in the four kinds of algorithms.

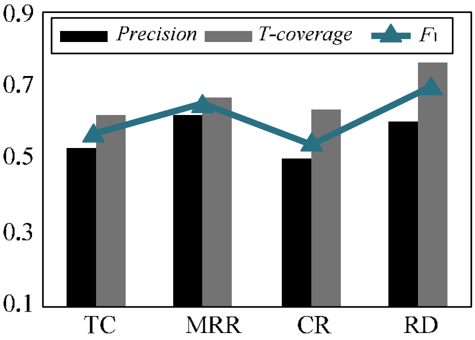


Figure 7. The performance comparison in the test set

In order to further observe the recommended tags produced by differnet algorithms, the number of the different relevant tags in the recommended results of each algorithm is counted, and the proportion of the total number that the occunrences of the most commom 50 relevant tags made up is calculated. Table 1 shows the comparison results. It can be seen that compared with the other three algorithms, RD uses a more rich vocabulary, and the number of the different relevant tags in the recommended results is nearly twice over that of the other three algorithms. In the relevant tags which are obtained by using TC, the occurrences of the most commom 50 tags accounts for about 60% of the total number, which states that TC tend to be concentrated on using a small amount of tags. And this tags often represent some general concepts, such as “nature” and “landscape”. Although there are a lot of images that are associated with these concepts, due to the content of the image is rich and varied, these tags always cannot accurately describe the specific information that the image reflected. While in the results of RD, the distribution of the relevant tags is more even, the occurrences of the most commom 50 tags accounts for only 15% of the total number, being the lowest among the four algorithms.

Table 1. Data Statistics of the Recommended Results in Different Algorithms

Algorithm	Number of different relevant tags/piece	The proportion of the most commom 50 tags
TC	325	60.37
MRR	362	44.01
CR	289	35.09
RD	670	16.03

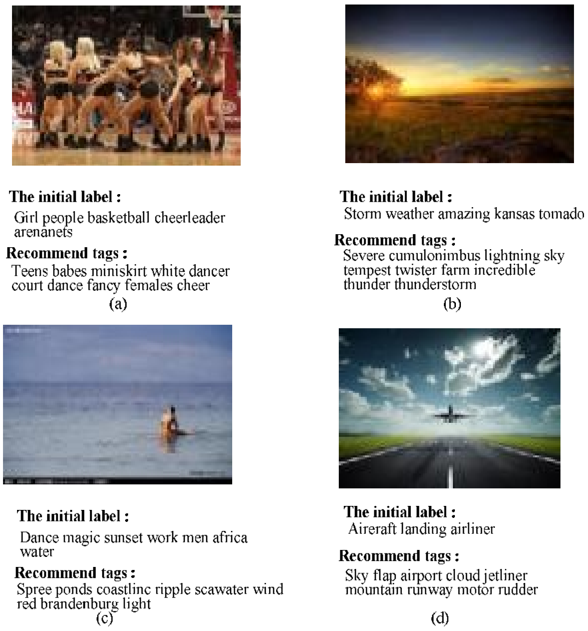

The initial label :
Girl people basketball cheerleader arenanets
Recommend tags :
Teens babes miniskirt white dancer court dance fancy females cheer
(a)
The initial label :
Storm weather amazing kansas tomado
Recommend tags :
Severe cumulonimbus lightning sky tempest twister farm incredible thunder thunderstorm
(b)
The initial label :
Dance magic sunset work men africa water
Recommend tags :
Spree ponds coastlinc ripple scawater wind red brandenburg light
(c)
The initial label :
Aireraft landing airliner
Recommend tags :
Sky flap airport cloud jetliner mountain runway motor rudder
(d)

Figure 8. The recommendation results of the algotithm in this paper

Figure 8 shows the recommended results of RD. For each image, its initial tags and recommended tags are listed in the figure. It can be seen that on the one hand, the recommended results in this paper is comprehensive, the recommended tags can more specifically describe the concepts that the initial tags represented. As shown in figure 8(a), the recommended tag “dancer” is the further description of the initial tags “girl” and “people”. On the other hand, when the number of the initial images is less, the recommended tags can express the objects or scenes that the initial tags failed to reflect, as the recommended tags “sky” and “grass” shown in figure 8(f). In conclusion, the recommendation results of the RD algorithm try to provide the users with new choices of the image annotation, from the different angle with the initial tags, and based on the aspects of the co-occurrence probability between the image’s visual features and the initial tags, the semantic diversity and so on.

5. Conclusion

For the traditonal image tag recommendation algorithm ignores the diversity between the visual content information of the image and the recommended tags, which leads to the recommendation results have the problem of tag ambiguity, tag redundancy and so on, the image tag recommendation algorithm based on the relevance and diversity is proposed in this paper. The algorithm solves the problem of tag ambiguity and tag redundancy in the traditional algorithm, defines the relevance and the diversity of a tag set, and selects a tag set which can reasonably balance the relevance and the diversity to recommend to the users. The experimental results show that the algorithm proposed in this paper improves the relevance between the recommended results and the image on the one hand, and on the other hand makes the recommended results be able to reflect the image content thoroughly.

References

[1] S. Li, Y. Geng, J. He, K. Pahlavan,Analysis of Three-dimensional Maximum Likelihood Algorithm for Capsule Endoscopy Localization, 2012 5th International Conference on Biomedical Engineering and Informatics (BMEI), Chongqing, China Oct. 2012 (page 721-725)

[2] D. Xu, Z. Y. Feng, Y. Z. Li, et al. Fair Channel allocation and power control for uplink and downlink cognitive radio networks. IEEE, Workshop on mobile computing and emerging communication networks, 2011 pp. 591-596

[3] Tong Zhao, Bingbing Qian, Yimin Li. Hybrid Adaptive Fuzzy Control Based on the Biological Adaptation Strategies. Journal of Networks, Vol 8, No 10 (2013), 2255-2262

[4] W. Q. Yao, Y. Wang, T. Wang. Joint optimization for downlink resource allocation in cognitive radio cellular networks. IEEE, 8th Annual IEEE consumer communications and networking conference, 2011 pp. 664-668

[5] Li Yimin, Wang Xiaomei, Application of fuzzy matching based on Type-2 in ecology, Mathematics in Practice and Theory, Vol.20, No.40, pp. 74-82, 2010

[6] S. H. Tang, M. C. Chen, Y. S. Sun, et al. A spectral efficient and fair user-centric spectrum allocation approach for downlink transmissions. IEEE., Globecom.,2011 pp. 1-6

[7] Li Yimin, Hao Yunli, Indirect T-S fuzzy adaptive control based on niche, Journal of Systems Engineering and Electronics, Vol. 10, No. 33, pp. 2282-2288, 2011