Neurocomputing 169 (2015) 180–184

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Robust latent semantic exploration for image retrieval in social media

Liujuan Cao [a], Fanglin Wang [b,*]

[a] Department of Computer Science, School of Information Science and Engineering, Xiamen University, China
[b] School of Computing, National University of Singapore, 13 Computing Drive, Singapore

ARTICLE INFO

Article history:
Received 1 May 2014
Received in revised form
6 January 2015
Accepted 12 February 2015
Available online 3 June 2015

Keywords:
Image retrieval
$l_{2,1}$ norm
Word-to-vector

ABSTRACT

With the speedy development of social media, more and more multimedia data are generated by users with tags associated. The tag information provides the extra cue to link multimedia data in addition to the multimedia content itself. However, the manually added tags are always with noise and not correct enough. Moreover, the semantically similar tags exist massively but cannot be accounted for well. This paper proposes a new algorithm to robustly combine multimedia content and associated tags by mining the latent semantic which takes into account the semantically similar tags. The $l_{2,1}$ norm is proposed to employ in latent semantic indexing for a more robust latent space, and a word-to-vector based clustering method is proposed to address the massive tags with similar meaning. The experiments on extensive data demonstrate the proposed method. Compared to the existing latent semantic based methods, the algorithm proposed a more robust model to deal with noise.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

Recent years have witnessed the boom of social media with the rapid development of smart phones and convenient internet access. People can easily acquire and share information and communicate with others on these platforms. One distinctive feature of social media is that it is more easier for users to create and share their own contents. Among the massive social media platforms, many image and video sharing web sites have become very popular. On these platforms, users not only produce and store their own contents but also can view and comment on others' contents. These platforms provide the social network characteristics which make users get connected and communicate with one another in multiple forms. Users can view the contents from one another and are provided the ability to comment on these contents by using tags. These tags can be viewed as annotation multimedia contents.

However, since users do not have the obligation to add tags accurately and thoroughly, directly using tags as annotation will bring problems due to being noisy and incomplete. Users sometimes give tags not from the same angle. There are many ways to describe one multimedia content, so it is impossible to add all possible tags completely for users. For example, for the landscape images taken at the same place, different users would pay attention to different aspects and hence prefer to add different tags to describe this place. Consequently, although these two images are visually similar, the difference introduced by tags will cause that these two images are not "similar" during searching. On the other hand, there is also the possibility that the visually different images are declared "similar" based on the associated tags.

As a reference, in the dataset proposed in [1], it has been shown that the average precision of user tags is about 0.5 and the average recall (completeness rate) of the user tags is 0.5 as well. Hence it means half of the tags created by users are noise and half of the true labels are missing.

To address the noisy and missing tags, a lot of works have been proposed recently. Tag refinement is one potential solution to improve the quality of user-generated tags associated to the multimedia data [2–5]. The relevance between tag and image is explored and further refined in these works. In [2] tag refinement is formulated as a tag ranking problem and a probabilistic framework was proposed to rank the associated tags. In [3], content consistency, tag consistency and low rank properties are studied in the mean time to generate a refined tag set. Tang et al. [4] used a robust graph and employed semi-supervised learning technique to learn a tag ranking model to do tag refinement.

In this paper, we propose an automatic image annotation algorithm by introducing a new latent semantic space to discover the semantic structure hidden in image and its tags. During the latent space construction, we use $l_{2,1}$ norm as the regularizor which has been demonstrated more robust to noise. To address those tags in similar meanings but regarded as different words, a word to vector based clustering method is proposed to build connection with similar tags. In addition, visual feature learnt from

* Corresponding author.
E-mail address: hardegg@gmail.com (F. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.02.082
0925-2312/© 2015 Elsevier B.V. All rights reserved.

a deep convolutional neural network is invoked to compute the visual similarity between images. Based on these improvements, the proposed image retrieval method is demonstrated superior with using extensive dataset.

The rest of this paper is organized as follows. Section 2 reviews the foundation of the proposed algorithm, i.e., the latent semantic indexing method and the low rank based latent semantic discovery means. Section 3 is devoted to presenting the proposed robust latent semantic mining method. The experimental results are provided in Section 4. We conclude this paper in Section 5.

2. Semantic modeling of multimedia

In this section, we review the existing latent semantic based indexing methods which are the basis of our algorithm.

2.1. Latent semantic indexing

Latent semantic indexing (LSI) can be used to discover patterns in the latent relationships between tags. The method employs singular value decomposition (SVD) to identify patterns. Given N images and the overall number of tags in database M, a context link $N \times M$ matrix A can be constructed, where each row of A_i represents a tag vector denoting the tags existing in image i. Each element of the matrix A_{ij} is either 1 or 0 to represent whether a tag is associated with this image. As an image usually has very few tags hence A_i is a sparse vector and A a sparse matrix. LSI employs SVD on matrix A and get $A = U\Sigma V^T$. U and V are two orthogonal matrices, and Σ is a diagonal matrix containing the singular values at the diagonal positions. By removing the smaller singular values in Σ and approximate them as zeros, a new diagonal matrix $\tilde{\Sigma}$ containing fewer singular values is obtained. Then the new image–tag link matrix $\tilde{A}$ is reconstructed by

$$\tilde{A} = U\tilde{\Sigma}V \tag{1}$$

After LSI, the reconstructed matrix $\tilde{A}$ depicts a more compact relationship among tags where the relevance is removed. Via LSI, a set of new tag vector associated to each image is obtained. In addition, a new feature representation $F = U\tilde{\Sigma} = [F_1, F_2, ..., F_n]$ is yielded and each row of F is a lower dimensional feature representing each image.

However, there are two problems with LSI. As A is quite sparse the latent semantic space will cause overfitting because the small number of nonzero elements cannot reflect the underlying structure. On the other hand, the image content is not taken into account which cause the loss of more information.

2.2. Low rank based latent semantic

As discussed, there is usually noise existing in link matrix A, the matrix can be represented as a clean link matrix H and a noise matrix ϵ:

$$A = H + \epsilon. \tag{2}$$

Similar to LSI, in this case the objective is to keep a low rank of matrix and in the mean time minimize the noise. The objective function is

$$\min \|A - H\|_F^2 + \gamma \operatorname{rank}(H), \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and γ is the parameter to control the weight of the two terms and $\operatorname{rank}(\cdot)$ is the function to calculate the rank of a matrix.

As stated in [6], to account for the lowest rank of a matrix is a NP-hard problem. A good approximation is to use nuclear norm to replace the computation of matrix rank. We use $\|A\|_*$ to denote the nuclear norm of matrix A then $\|A\|_* = \operatorname{tr}\left(\sqrt{A^* A}\right)$. An advantage of using nuclear norm is it is convex and hence more easier for optimization. With the nuclear norm approximation, Eq. (3) becomes

$$\min \|A - H\|_F^2 + \gamma \|H\|_*. \tag{4}$$

For Eq. (4), it has the analytical solution

$$H = U\left(\Sigma - \frac{\gamma}{2} I\right)_+ V^T, \tag{5}$$

where U, Σ, V consist of the SVD of A, i.e., $A = U\Sigma V^T$, and $(As)_+$ is an element wise operation of a matrix A so that each element $A_{ij} = \max(0, A_{ij})$. It is obvious to see that the singular values are subtracted by a value of $\gamma/2$ and thresholded by 0 while LSI keeps the original singular value. Suppose the resultant H is of rank k, similar with LSI, the row vectors $X = \left(X_1^T, X_2^T, \ldots, X_N^T\right) = U\Sigma_k$ can work as the latent representation of an image in latent space of tags.

2.3. Low rank latent semantic combined with content relevance

So far the image content has not been taken into account yet and it is obviously crucial for better performance. The tag information itself is very sparse and it easily causes overfitting problem. Adding an image content related term will make the optimization more reliable. Suppose the content term is $C(H)$, then the objective now becomes to optimize

$$\min \|A - H\|_F^2 + \lambda C(H) + \gamma \|H\|_*, \tag{6}$$

where λ is a parameter to control the weight of image content.

So the next step is to model the content regularization term properly. Suppose we have a matrix $R \in \mathbb{R}^{n \times n}$ where each element R_{ij} is the visual similarity of image i and image j. The greater the R_{ij} is the more similar the two images are in terms of visual content. According to the previous subsection, we have discussed the latent representation of row vectors $X = \left(X_1^T, X_2^T, \ldots, X_N^T\right)$ obtained from tag latent space can be used to represent the image object. Now the image content can be used to construct a more robust latent semantic space. Here the image content and the latent semantic can be combined according to

$$C(X) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} R_{ij} (X_i - X_j)(X_i - X_j)^T. \tag{7}$$

It is apparent that Eq. (7) imposes the penalty on visually similar image pairs. That means if R_{ij} is larger X_i and X_j should be closer, otherwise would cause larger gap. This term can be further derived as

$$\begin{aligned} C(X) &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} (X_i X_i^T + X_j X_j^T - X_i X_j^T - X_j X_i^T) \\ &= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} (X_i X_i^T + X_j X_j^T) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} (X_i X_j^T + X_j X_i^T) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} X_i X_i^T - \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} X_i X_j^T \end{aligned} \tag{8}$$

According to [7], Eq. (8) can be further simplified as

$$C(X) = \operatorname{tr}(H^T L H). \tag{9}$$

So far we can model the latent semantic by combining tag information and visual content information. According to [7], Eq. (6) now turns into

$$\min \mathcal{F}(H) = \|A - H\|_F^2 + \lambda \operatorname{tr}(H^T L H) + \gamma \|H\|_* \tag{10}$$

3. The proposed method

3.1. Robust latent semantic mining

In [7], the latent semantic model depicted by Eq. (10) has demonstrated its superiority. In this paper, our objective is to make it more robust to noise. In Eq. (10), the first term is Frobenius norm which is well known that it is sensitive to noises since the squared error of noises may dominate this term. While $l_{2,1}$ norm has shown its better robustness to noises [8–10], in our work we change the first term from Frobenius norm to $l_{2,1}$ norm:

$$\min_H \mathcal{F}(H) = \|A-H\|_{2,1}^2 + \lambda \operatorname{tr}(H^T L H) + \gamma \|H\|_*, \tag{11}$$

where $l_{2,1}$ norm of a matrix $A \in \mathbb{R}^{N\times M}$ is defined as

$$\|A\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} A_{ij}^2} = \sum_{i=1}^{N} \|A_i\|_2, \tag{12}$$

where A_i and A_{ij} denote the i-th row vector and the element at i-th row and j-th column. Note the objective function in Eq. (11) is non-smooth. Inspired by iterative re-weighted least square (IRLS) method to solve non-smooth data reconstruction problems [10], we solve the $l_{2,1}$-norm minimization problem in an iteratively re-weighting manner. Suppose we have the solution $H^{(t)}$ at the t-th iteration, the solution of the original problem in Eq. (11) can be obtained by iteratively solving the following re-weighted problem:

$$\begin{aligned} H^{(t+1)} &= \arg\min_H \left(\sum_i \frac{\|A_i - H_i\|_2^2}{\|A_i - H_i^{(t)}\|_2} \right)^2 + \lambda \operatorname{tr}(H^T L H) + \gamma \|H\|_* \\ &= \arg\min_H \|\Lambda^{1/2}(A-H)\|_2^4 + \lambda \operatorname{tr}(H^T L H) + \gamma \|H\|_*, \end{aligned} \tag{13}$$

where $\Lambda = diag(1/(\|A_i - H_i^{(t)}\|_2))$.

Let $\mathcal{G}(H) = \|\Lambda^{1/2}(A-H)\|_2^4 + \lambda \operatorname{tr}(H^T L H) + \gamma \|H\|_*$. As there is no closed-form solution for $\mathcal{G}(H)$, we employ the proximal gradient method [11] to use a sequence of quadratic approximations of $\mathcal{G}(H)$ in order for the optimal solution. We define $f(H) = \|\Lambda^{1/2}(A-H)\|_2^2 + \lambda \operatorname{tr}(H^T L H)$ and can see that it is a differentiable function. Hence $\mathcal{G}(H)$ is summation of a differentiable function and the nuclear norm. According to [11], $\mathcal{G}(H)$ can be approximated when given $H^{(t)}$:

$$\begin{aligned} \mathcal{G}(H) &\approx f\left(H^{(t)}\right) + < \nabla f(H^{(t)}), H - H^{(t)} > + \frac{\tau}{2}\|H - H^{(t)}\|_2^2 + \gamma \|H\|_* \\ &= \frac{\tau}{2}\|H - G^{(t)}\|_2^2 + \gamma \|H\|_* + f(H^{(t)}) - \frac{1}{2\tau}\|\nabla f(H^{(t)})\|_2^2, \end{aligned} \tag{14}$$

where $G^{(t)}$ is defined as

$$G^{(t)} = H^{(t)} - \frac{1}{\tau}\nabla f\left(H^{(t)}\right)$$

$$= H^{(t)} - \frac{4}{\tau}\left(\Lambda^{1/2}(\Lambda^{1/2}H^{(t)} - \Lambda^{1/2}A)^3 + \lambda L^T H^{(t)}\right). \tag{15}$$

To meet the requirement of Lipschitz condition that $\|\nabla f(x) - \nabla f(y)\|_2 \le \tau \|x-y\|_2$, we set τ as the largest singular value of matrix $(I + \lambda L^T)$.

In Eq. (14), the rightmost two terms do not depend on $H^{(t+1)}$ and are ignored during minimizing with respect to $H^{(t+1)}$. Then we have

$$H^{(t+1)} = \arg\min_H \frac{\tau}{2}\|H - G^{(t)}\|_2^2 + \gamma \|H\|_* \tag{16}$$

3.2. Visual and textual features

In this paper, instead of using the frequent bag of visual words feature, we adopted the 4096-D Caffe generic visual feature [12]. It is the input of the 6-th layer of the deep convolutional neuro network (CNN) [13] trained in a fully supervised fashion with the objective of image classification on data from ImageNet [14]. This feature has been demonstrated effective on various tasks like [15]. After getting the 4096-D feature, we normalized the visual feature vector into a zero-mean and unit-variance Gaussian distribution, i.e., $\mathbf{f} \sim \mathcal{N}(\mathbf{0}|\mathbf{1})$. Accordingly, the visual similarity between two images are defined as

$$R_{ij} = \exp(-\|\mathbf{f}_i - \mathbf{f}_j\|_2^2) \tag{17}$$

For the textual tags, instead of using the tag words directly, we transform each tag word into a new feature to account for the extreme sparsity and being noisy. As we discussed before, the tags themselves cannot deal with the semantic similarity. For example, "car" and "automobile" can appear at two visually similar images but will be treated as two different words. Recently, a new method so-called Word2Vect has been proposed [16] to convert words into vectors. Based on this work, each word is represented by a vector and its similarities with other words can be calculated in a straightforward way. To make the visual tag feature less sparser and have textual semantic taken into account, we take advantage of Word2Vec method. First, we transform each tag word to a 640-D vector by directly using the Word2Vec model trained in [16]. Then we do clustering using k-means where Euclidean distance is used and the number of clusters is set as 1024. After this, each tag word can be assigned to one of the 1024 cluster centers. After this transform, tags with similar semantic meaning will be grouped into one cluster.

3.3. Image retrieval with latent space

Given image set with tags associated, we first compute its visual and textual features via Caffe and Word2Vec detailed in previous subsection. With the 1024-D Word2Vec based textual feature vector, we compute the $l_{2,1}$ norm based low rank matrix H

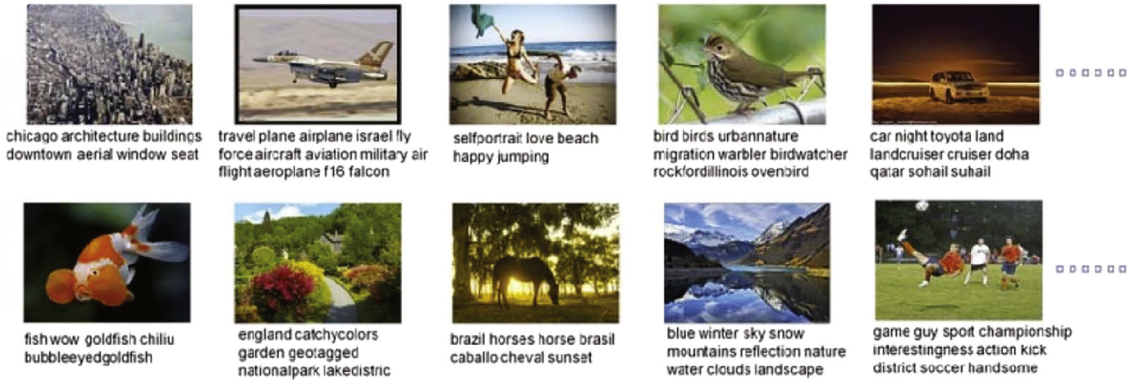


Fig. 1. Examples of NUS-WIDE images and associated tags.

depicted by Eq. (16). Before conducting image retrieval, we convert each image into latent space. Suppose the rank of H is k, then the SVD of H can be obtained via $U\Sigma_k V^T = H$. Here Σ_k is a $k \times k$ diagonal matrix. Then we use the row vectors of

$$X = U\Sigma_k \tag{18}$$

to represent each image in latent space.

In this work, we follow the manner of [1] and set the goal of image retrieval as to retrieve a list of images which are relevant to the objective concepts. All the retrieved images are ranked according to the prediction scores in a decent order. The relevant images are expected to be ranked higher in the resultant list. To conduct image retrieval, we first extract the latent space feature according to Eq. (18) for each image, then for each concept we employ SVM to train a classifier by using the images with associated tags belonging to this concept. To rank resultant images, the classification score is set as measurement metric.

4. Experimental results

In this section we evaluate the proposed robust latent method and its application in image retrieval. The experiments are conducted on a public dataset with a large number of images with noisy tags associated. To show the effectiveness, we set the so-called context-and-content-based multimedia retrieval (C2MR) method proposed in [7] as the baseline method, and we call our method Robust C2MR (RC2MR).

4.1. Dataset

We conduct experiments on the popular NUS-WIDE dataset [1]. The lite version is chosen which includes a subset of 55 615 images and their associated tags randomly selected from the full NUS-WIDE dataset. The images were crawled from the image sharing web site Flickr.com. The tag distribution is quite sparse and most of images

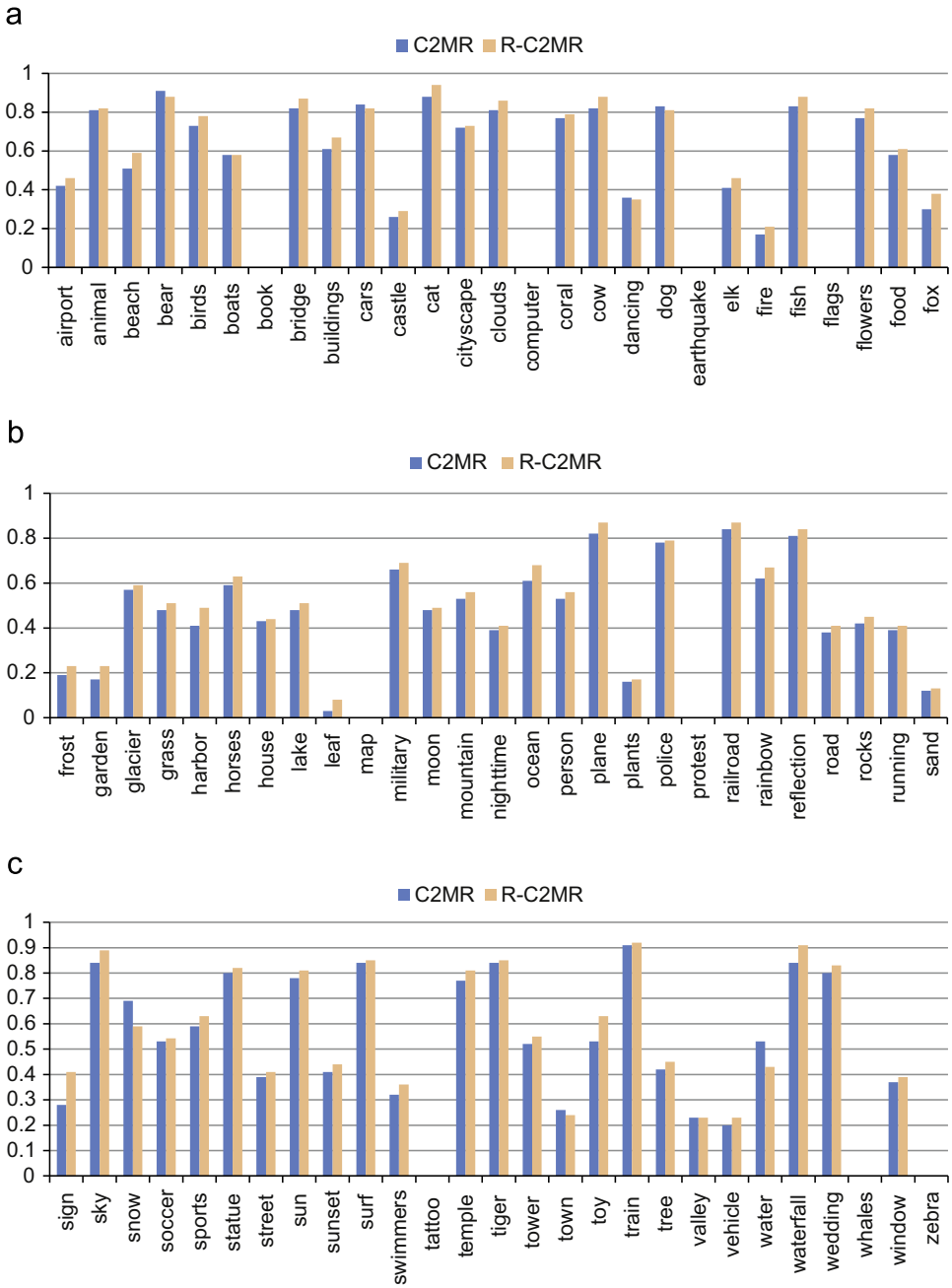


Fig. 2. AP comparison of different algorithms over 81 concepts from the NUS-WIDE.

have fewer than 10 tags and the average number of tags per image is 7.3. The dataset covers 81 concepts. Another phenomenon is the tags associated with images are full of noise, as shown in Fig. 1.

4.2. Performance evaluation

In this work, we do image retrieval in terms of concepts. That is, given an input image, we classify whether it belongs to a concept. In this definition, the objective of image retrieval is to rank the relevant images higher than the irrelevant ones. To evaluate the ranking performance, we employ average precision (AP) to evaluate the performance. Let Ω be the number of true positive images in the test set and Ω_i be the number of relevant image in the top i images in the resultant rank list. Then AP is defined as

$$\frac{1}{\Omega}\sum_i \frac{\Omega_i}{i} J_i, \tag{19}$$

where $J_i=1$ if i-th image is relevant and 0 otherwise. In this paper, we compute AP for each concept defined in NUS-WIDE dataset.

4.3. Result analysis

Here we do the concept retrieval task provided in NUS-WIDE dataset. For both C2MR and R-C2MR, we extract latent space features for each images and train SVM classifiers for each concept based on the positive and negative sample sets provided in NUS-WIDE. Then the retrieval is conducted on the testing sets and the AP metric defined by Eq. (19) is calculated for each concept.

The experimental results in terms of AP are shown in Fig. 2. For these two methods, C2MR achieves an overall 0.50 AP, and R-C2MR achieves an overall 0.59 AP. As a comparison, R-C2MR achieves 18% improvement. The proposed R-C2MR takes advantage of the robustness to noise of $l_{2,1}$ norm and can get a more robust latent space. It means that the impact of noisy tags will be significantly accounted for. As there are 50% noisy tags in NUS-WIDE dataset [1] which is too much for most of existing methods, Fig. 2 demonstrates the superiority of proposed R-C2MR. One more factor corrupting the performance of image retrieval is that tags are with similar meaning cannot be well accounted for. In our work, we make use of Word2Vec as textual features rather than the raw tags to account for similar semantic meaning. The effectiveness of the new textual feature is also demonstrated by the experimental results.

5. Conclusions

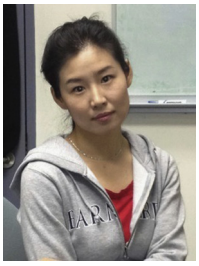
This paper has presented a latent semantic based image retrieval method. As the state-of-the-art latent semantic mining method works on sparse and noisy tags, we tried to improve the existing methods from two aspects. To increase the robustness of the latent semantic modeling, we adopted $l_{2,1}$ norm rather than the popular Fobenius norm and presented the optimization method. Then to deal with the sparse tags associated with each images, we proposed to use a new representation of word tags in a lower dimensional space based on Word2Vec method. The proposed method has been demonstrated effective on the popular NUS-WIDE set and its superiority to the state-of-the-art methods have been shown.

Acknowledgment

This work is supported by the National Natural Science Foundation of China Under Grant Number 61402388.

References

[1] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: a real-world web image database from National University of Singapore, in: Proceedings of ACM Conference on Image and Video Retrieval, CIVR'09, Santorini, Greece, July 8–10, 2009.
[2] D. Liu, X.-S. Hua, L. Yang, M. Wang, H.-J. Zhang, Tag ranking, in: Proceedings of the 18th International Conference on World Wide Web, ACM, 2009, pp. 351–360.
[3] G. Zhu, S. Yan, Y. Ma, Image tag refinement towards low-rank, content-tag prior and error sparsity, in: Proceedings of the International Conference on Multimedia, ACM, 2010, pp. 461–470.
[4] J. Tang, S. Yan, R. Hong, G.-J. Qi, T.-S. Chua, Inferring semantic concepts from community-contributed images and noisy tags, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, 2009, pp. 223–232.
[5] H. Xu, J. Wang, X.-S. Hua, S. Li, Tag refinement by regularized lda, in: Proceedings of the 17th ACM International Conference on Multimedia, ACM, 2009, pp. 573–576.
[6] E.J. Candes, Y. Plan, Matrix completion with noise, Proc. IEEE 98 (6) (2010) 925–936.
[7] G.-J. Qi, C. Aggarwal, Q. Tian, H. Ji, T.S. Huang, Exploring context and content links in social media: a latent space method, IEEE Trans. Pattern Anal. Mach. Intell. 34 (5) (2012) 850–862.
[8] C. Ding, D. Zhou, X. He, H. Zha, R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 281–288.
[9] F. Nie, H. Huang, X. Cai, C. Ding, Efficient and robust feature selection via joint l2, 1-norms minimization, Adv. Neural Inf. Process. Syst. 23 (2010) 1813–1821.
[10] R. Chartrand, W. Yin, Iteratively reweighted algorithms for compressive sensing, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, ICASSP 2008, IEEE, 2008, pp. 3869–3872.
[11] K.-C. Toh, S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems, Pac. J. Optim. 6 (615–640) (2010) 15.
[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, arxiv: hepth/1408.5093.
[13] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. (2012) 1097–1105.
[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009, IEEE, 2009, pp. 248–255.
[15] H. Zhang, Y. Yang, H. Luan, S. Yang, T.-S. Chua, Start from scratch: towards automatically identifying, modeling, and naming visual attributes, in: Proceedings of the ACM International Conference on Multimedia, ACM, 2014, pp. 187–196.
[16] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. (2013) 3111–3119.



Liujuan Cao is currently an Assistant Professor at the Department of Computer Science, School of Information Science and Engineering, Xiamen University. Before that, she obtained her Ph.D. degree from Harbin Engineering University. Her research is in the field of multimedia analysis, geo-science and remote sensing, and computer vision. She has published extensively at CVPR, Neurocomputing, Signal Processing, ICIP, VCIP, etc.

Fanglin Wang is currently a Research Fellow in School of Computing, National University of Singapore. He received the B.S. and M.S. degrees both from Harbin Institute of Technology, Harbin, China, in 2003 and 2005, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, China, in 2009. He had worked as a Senior Researcher, Software Researcher and Senior Researcher at Sharp Laboratories China, Autodesk China Research and Development, Carestream Inc., respectively, from 2009 to 2012. His research interests include object detection, visual tracking and medical tomographic reconstruction.