

Refining Image Annotation by Integrating PLSA with Random Walk Model

Dongping Tian^{1,2}, Xiaofei Zhao¹, and Zhongzhi Shi¹

¹ Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² Graduate University of the Chinese Academy of Sciences, Beijing, 100049, China
{tiandp, zhaoxf, shizz}@ics.ict.ac.cn

Abstract. In this paper, we present a new method for refining image annotation by integrating probabilistic latent semantic analysis (PLSA) with random walk (RW) model. First, we construct a PLSA model with asymmetric modalities to estimate the posterior probabilities of each annotating keywords for an image, and then a label similarity graph is constructed by a weighted linear combination of label similarity and visual similarity. Followed by a random walk process over the label graph is employed to further mine the correlation of the keywords so as to capture the refining annotation, which plays a crucial role in semantic based image retrieval. The novelty of our method mainly lies in two aspects: exploiting PLSA to accomplish the initial semantic annotation task and implementing random walk process over the constructed label similarity graph to refine the candidate annotations generated by the PLSA. Compared with several state-of-the-art approaches on Corel5k and Mirflickr25k datasets, the experimental results show that our approach performs more efficiently and accurately.

Keywords: Refining Image Annotation, PLSA, EM, Random Walk, Image Retrieval.

1 Introduction

With the rapid development of multimedia information technology, image retrieval has become more and more important in Internet and other multimedia platforms. As we known, image annotation is a previous and vital step when it comes to the semantic based image retrieval. Traditional method for image annotation is to let people manually annotate the images by some keywords. However, this method is onerous and time-consuming. Furthermore, the annotating result is subjective to different people. To address these limitations, automatic image annotation (AIA) has become a focus and received extensive investigation, whose purpose is to automatically assign some keywords to an image that can well describe the content of it. Subsequently many methods have been developed for AIA, and most of them can be roughly classified into two categories, i.e. classification-based methods and probabilistic modeling methods.

The representative works of the former are automatic linguistic index for pictures [1] and content-based annotation method with SVM [2] etc. The probabilistic modeling methods include the translation model (TM) [3], the cross-media relevance model (CMRM) [4], the continuous-space relevance model (CRM) [5], the multiple-Bernoulli relevance model (MBRM) [6] and the latent aspect model PLSA [7], etc. Unfortunately, all the mentioned annotation methods, to some extent, can achieve relative success compared to the manual annotation, but they are still far from satisfaction due to the well-known semantic gap problem.

In recent years, some researchers propose to refine the image annotation by taking the word correlation into account. As a pioneer work, Jin et al. [8] implement image annotation refinement based on WordNet by pruning the irrelevant annotations. In their work, however, only global textual information is employed and the refinement process is independent of the target image, which means that different images with the same candidate annotations would obtain the same refinement results. Subsequently, Wang et al. [9] apply random walk with restarts model to refine candidate annotations by integrating word correlations with the original candidate annotation confidence together. Followed by they propose a content based approach by formulating the annotation refinement as a Markov process [10]. Recently Liu et al. [11] rank the image tags according to their relevance with respect to the associated images by tag similarity and image similarity in a random walk model. Xu et al. [12] come up with a new graphical model termed as regularized latent Dirichlet allocation (rLDA) for tag refinement. In addition, Zhu et al. [13] put forward an efficient iterative approach for image tag refinement by pursuing the low-rank, content consistency, tag correlation and error sparsity, which constitute a constrained yet convex optimization problem and an efficient accelerated proximal gradient method is utilized to resolve it. More recently, Zhuang et al. [14] propose a two-view learning approach for image tag ranking by effectively exploiting both textual and visual contents of social images to discover the complicated relationship between tags and images.

Most of these approaches can achieve state-of-the-art performance and motivate us to explore image annotation with the help of their excellent experiences and knowledge. So in this paper, we present a new method for refining image annotation by means of combining PLSA and random walk model (PLSA-RW). To begin with, a PLSA model with asymmetric modalities is constructed to estimate the scores (i.e. posterior probabilities. For simplicity, we use the terminologies score and posterior probability interchangeably in the rest of this paper) of all the annotating keywords, and this can be seen as the initial annotation for the image. And then a label ¹ similarity graph is constructed by a weighted linear combination of label similarity and visual similarity. Followed by a random walk process over the label similarity graph is implemented to further mine the words correlation. Once the random walk reaches the steady-state probability distribution, the top several candidates with the highest probabilities can be seen as the refining annotation. Our method can boost the annotating performance by introducing a two-stage annotation refinement process. We evaluate our method

¹ Here label means the initial annotation generate by the PLSA.

on Corel5k and Mirflickr25k datasets and their experimental results compare favorably with several state-of-the-art approaches. To the best of our knowledge, this is the first study to try to integrate PLSA with random walk in the task of refining image auto-annotation.

The rest of the paper is organized as follows. Section 2 presents how to apply PLSA to model annotated images. In section 3, the construction of label similarity graph is first introduced, and then a random walk over the graph is elaborated. Experimental results on Corel5k and Mirflickr25k datasets are reported and analyzed in section 4 respectively. Finally, we end this paper with some important conclusions and future work in section 5.

2 PLSA Model

PLSA [15] is a statistical latent class model which introduces a hidden variable (latent aspect) z_k in the generative process of each element x_j in a document d_i . Given this unobservable variable z_k , each occurrence x_j is independent of the document it belongs to, which corresponds to the following joint probability:

$$P(d_i, x_j) = P(d_i) \sum_{k=1}^K P(z_k|d_i)P(x_j|z_k) \quad (1)$$

The model parameters of PLSA are the two conditional distributions: $P(x_j|z_k)$ and $P(z_k|d_i)$. $P(x_j|z_k)$ characterizes each aspect and remains valid for documents out of the training set. On the other hand, $P(z_k|d_i)$ is only relative to the specific documents and cannot carry any prior information to an unseen document. An EM algorithm is used to estimate the parameters through maximizing the log-likelihood of the observed data.

$$L = \sum_{i=1}^N \sum_{j=1}^M n(d_i, x_j) \log P(d_i, x_j) \quad (2)$$

where $n(d_i, x_j)$ is the count of element x_j in document d_i . The steps of the EM algorithm can be succinctly described as follows.

E-step. The conditional distribution $P(z_k|d_i, x_j)$ is computed from the previous estimate of the parameters:

$$P(z_k|d_i, x_j) = \frac{P(z_k|d_i)P(x_j|z_k)}{\sum_{l=1}^K P(z_l|d_i)P(x_j|z_l)} \quad (3)$$

M-step. The parameters $P(x_j|z_k)$ and $P(z_k|d_i)$ are updated with the new expected values $P(z_k|d_i, x_j)$:

$$P(x_j|z_k) = \frac{\sum_{i=1}^N n(d_i, x_j)P(z_k|d_i, x_j)}{\sum_{m=1}^M \sum_{i=1}^N n(d_i, x_m)P(z_k|d_i, x_m)} \quad (4)$$

$$P(z_k|d_i) = \frac{\sum_{j=1}^M n(d_i, x_j)P(z_k|d_i, x_j)}{\sum_{j=1}^M n(d_i, x_j)} \quad (5)$$

If one of the parameters ($P(x_j|z_k)$ or $P(z_k|d_i)$) is known, the other one can be inferred by using fold-in method, which updates the unknown parameters with the known parameters kept fixed, so that it can maximize the likelihood with respect to the previously trained parameters. In this paper, we construct a PLSA model with asymmetric modalities since the textual modality is more appropriate to learn a semantically meaningful latent space [7], and the joint probability between an image and the semantic concepts is calculated from two linked PLSA models sharing the same distribution over aspects. Given an unseen image visual features $v(d_{new})$, the conditional probability distribution $P(z_k|d_{new})$ can be inferred with the previously estimated model parameters $P(v|z_k)$, then the posterior probability of words can be computed by the following equation.

$$P(w|d_{new}) = \sum_{k=1}^K P(w|z_k)P(z_k|d_{new}) \quad (6)$$

3 Random Walk-Based Refining Annotation

As a latent aspect model, PLSA has been successfully applied in automatic image annotation, such as the representative PLSA-WORDS and PLSA-FEATURES [7] as well as the PLSA-FUSION proposed by Li et al. [16], which uses two linked PLSA models to learn the mixture of aspects from both visual and textual modalities. However, since all the annotations are calculated independently in PLSA model and the relations among them are not exploited, which inevitably results in some ambiguity and inconsistency in the process of image annotation. In order to combine the prior confidence of candidate annotations and word correlations together, we present a two-stage image annotation refinement framework displayed in Figure 1. More details of it will be described in the following subsections.

3.1 Label Graph Construction

To construct the label graph, i.e. the initial annotation graph, each candidate is transformed to a vertex, and the pair-wise label similarity is used as the weight of the corresponding edge. For now we focus on how to reasonably estimate the similarities between pair-wise concepts related to an image, which is still a tough problem in multimedia information processing. The mostly used methods include WordNet [17] and normalized Google distance (NGD) [18]. From their definitions, we can easily see that NGD is actually a measure of the contextual relation while WordNet focuses on the semantic meaning of keyword itself. What is more, both of them build word correlations only based on textual descriptions, and the visual information of images in the dataset is not utilized for refinement, which also plays a key role in precise image annotation. So in this paper, the

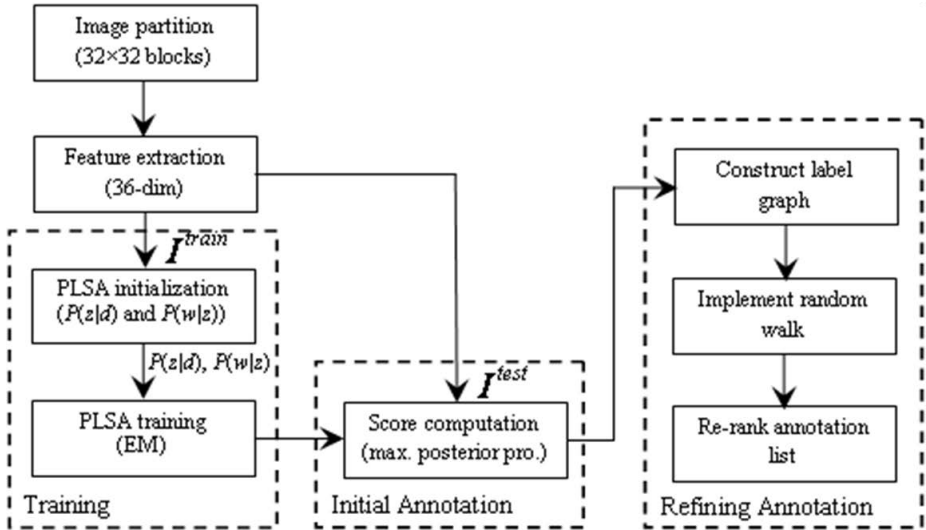


Fig. 1. The proposed refining annotation framework in this paper

pair-wise annotation similarity is calculated by a weighted linear combination of label similarity and visual similarity, which can effectively avoid the phenomenon that different images with the same candidate annotations would obtain the same refinement results. The label similarity between w_i and w_j is defined as follows:

$$s_l(w_i, w_j) = \exp(-d(w_i, w_j)) \quad (7)$$

where $d(w_i, w_j)$ represents the distance between two labels w_i and w_j and it is defined similarly to NGD as:

$$d(w_i, w_j) = \frac{\max(\log f(w_i), \log f(w_j)) - \log f(w_i, w_j)}{\log G - \min(\log f(w_i), \log f(w_j))} \quad (8)$$

where $f(w_i)$ and $f(w_j)$ are the numbers of images containing labels w_i and w_j respectively, and $f(w_i, w_j)$ is the number of images containing both w_i and w_j , G is the total number of images in the dataset. Similar to [11], for a label w associated with an image x , we collect the K nearest neighbors from the images containing w , and these images can be regarded as the exemplars of the label w with respect to x . Thus from the point view of labels associated with an image, the visual similarity between labels w_i and w_j is given as follows:

$$s_v(w_i, w_j) = \exp\left(-\frac{1}{K \times K} \sum_{x \in \Gamma_{w_i}, y \in \Gamma_{w_j}} \frac{\|x - y\|^2}{\sigma^2}\right) \quad (9)$$

where Γ_w is the representative image collection of label w , x and y denote image features corresponding to the respective image collections of label w_i and w_j , σ

is the radius parameter of the Gaussian kernel function. To benefit from each other of the two similarities described above, a weighted linear combination of label similarity and visual similarity is defined:

$$s_{ij} = s(w_i, w_j) = \lambda s_l(w_i, w_j) + (1 - \lambda) s_v(w_i, w_j) \quad (10)$$

where $\lambda \in [0, 1]$ controls the weights for each measurement and the corresponding performance with different λ values is to be discussed in section 4.

3.2 Random Walk over Label Graph

Implementing random walk over the graph structure at least needs two important parameters, i.e. the importance of nodes and the probability transition matrix. Suppose that a label graph constructed in subsection 3.1 with n nodes, we use $r_k(i)$ to denote the relevance score of node i at iteration k , P denotes a $n \times n$ transition matrix, whose element p_{ij} indicates the probability of the transition from node i to node j and it is computed as follows:

$$p_{ij} = s_{ij} / \sum_k s_{ik} \quad (11)$$

where s_{ij} is the pair-wise label similarity (defined in Eq.10) between node i and node j . Then the random walk process is formulated as:

$$r_k(j) = \alpha \sum_i r_{k-1}(i) p_{ij} + (1 - \alpha) v_j \quad (12)$$

where $\alpha \in (0, 1)$ is a weight parameter to be determined, v_j denotes the initial annotation probabilistic scores calculated by the PLSA. In the process of refining annotation, random walk proceeds until it reaches the steady-state probability distribution and then the top several candidates with the highest probabilities can be seen as the final refining image annotation results.

4 Experimental Results and Analysis

For the purpose of comparison, we first conduct our experiments on the Corel5k dataset, which consists of 5000 images from 50 Corel Stock Photo CD's provided by [3]. Each CD contains 100 images with a certain theme, of which 90 are designated to be in the training set and 10 in the testing set, resulting in 4500 training images and a balanced 500-image test collection. Since the focus of this paper is not on image feature selection, we use similar features extracted by [6] to make a fair comparison with the state-of-the-art approaches. First of all, we simply decompose images into a set of 32×32 -sized blocks, then compute a 36 dimensional feature vector for each block, consisting of 24 color features (auto-correlogram) computed over 8 quantized colors and 3 Manhattan Distances, 12 texture features (Gabor filter) computed over 3 scales and 4 orientations. As a result, each block is represented as a 36-dim feature vector. Then each image is represented as a bag of features, i.e., a set of 36 dimensional vectors.

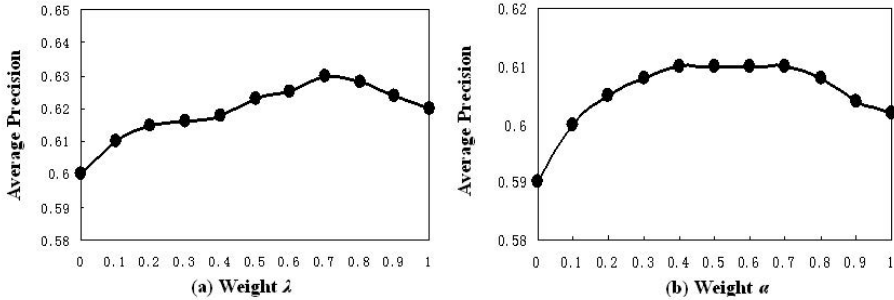


Fig. 2. Evaluation for weight parameters λ and α

4.1 Evaluation for the Weights

Since there are two variable weights λ and α to be determined, we should first fix one of them so as to observe the other’s varied trend and vice versa. Suppose that α is set to 0.5, then we range λ from 0 to 1. As shown in Figure 2(a), we can clearly see that the performance is better when $\lambda \in (0, 1)$ than $\lambda = 0$ or $\lambda = 1$ individually. Particularly, the best result is achieved when $\lambda = 0.7$, which demonstrates the complementary nature of label similarity and visual similarity. On the other hand, we set $\lambda = 0.7$ and range α from 0 to 1. From the curve in Figure 2(b), we note that the performance improves consistently before 0.5, followed by it almost keeps in a smooth state. The performance begins to reduce when α exceeds 0.7. Thus we choose $\alpha = 0.5$ as the optimal parameter in our experiment.

4.2 Refining Image Annotation on Corel5k

To show the effectiveness of our model (PLSA-RW) proposed in this paper, we make a direct comparison with several previous approaches [3,4,5,6,7,16]. Similar to [6], we compute the recall and precision of every word in the test set and use the mean of these values to summarize its performance. The experimental results listed in Table 1 are based on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. From table 1, it is easy to see that our model PLSA-RW outperforms all the others, especially the first three approaches. Meanwhile, it is also superior to MBRM, PLSA-WORDS and PLSA-FUSION.

Figure 3 shows some annotating results (only four cases are listed here due to the limited space) using PLSA-FUSION and PLSA-RW. It is worth noting that the annotations with the highest probabilities obtained in the last iteration of the random walk process are considered as the final annotation of the corresponding image. It is also important to note that the annotation order of the keywords for each image, which is very significant for semantic based image retrieval. Especially those with different annotating orders and enriched

Table 1. Performance comparison of AIA on Corel5k dataset

Models	Translation	CMRM	CRM	MBRM	PLSA-WORDS	PLSA-FUSION	PLSA-RW
#words with recall > 0	49	66	107	122	105	122	126
Results on 49 best words							
Mean per-word recall	0.34	0.48	0.70	0.78	0.71	0.76	0.78
Mean per-word precision	0.20	0.40	0.59	0.74	0.56	0.65	0.75
Results on all 260 words							
Mean per-word recall	0.04	0.09	0.19	0.25	0.20	0.22	0.27
Mean per-word precision	0.06	0.10	0.16	0.24	0.14	0.19	0.25





Images				
Ground Truth Annotation	tiger, forest, cat, trees	garden, flowers, trees, farm, plants	mountain, water, sky, clouds	polar, bear, snow, tundra
PLSA-FUSION Annotation	cat, tiger, trees, forest, leaves	flowers, trees, garden, farm, plants	sky, mountain, water, clouds, trees	snow, polar, bear, tundra, ice
PLSA-RW Annotation	<u>tiger</u> , trees, <u>leaves</u> , forest, <u>cat</u>	flowers, trees, garden, <u>plants</u> , <u>farm</u>	sky, mountain, water, clouds, <u>trees</u>	snow, <u>bear</u> , <u>polar</u> , tundra, <u>ice</u>

Fig. 3. Annotation comparison between PLSA-FUSION and PLSA-RW (Re-ranked and enriched annotations are underlined and italic)

annotating keywords compared to the PLSA-FUSION and the ground truth annotation are underlined and italic, respectively.

4.3 Refining Image Annotation on Mirflickr25k

To further demonstrate the effectiveness of PLSA-RW proposed in this paper, we also conduct experiment on Mirflickr25k dataset ², which contains 25000 images with 1386 labels. For the sake of fair comparison with the state-of-the-art approaches in [10] and [13], we use similar features to reference [13], that is, a 428-dimension feature vector is extracted from each image, including 225-dim block-wise color moment features generated from 5×5 fixed partition, 128-dim wavelet texture features and 75-dim edge distribution histogram features. At the same time, we evaluate the performance on 18 tags in Mirflickr25k where the ground-truth annotation of these tags has been provided. In addition, we remove those tags whose occurrence numbers are less than 50, thus 205 unique tags are obtained in total for Mirflickr25k in our experiment.

Table 2 summarizes the average performances measured by F-value for different refinement methods. As can be seen from Table 2, the F-value of our method is 0.475 which gives significant better result than the value obtained by the original user-provided tags (UT) [19]. Furthermore, it compares favorably with the

² Download from <http://press.liacs.nl/mirflickr/dlform.php>



Fig. 4. Four exemplars of image annotation refinement on Mirflickr25k

state-of-the-art approaches proposed by Wang et al. (RWR) [10] and Zhu et al. (LR-ES-CC-TC) [13], which further proves that the PLSA-RW is efficient in refining image annotation.

Table 2. Performance comparison of different methods on Mirflickr25k

Methods	UT	RWR	LR-ES-CC-TC	PLSA-RW
F-value	0.221	0.338	0.477	0.475

Alternatively, some exemplars of image annotation refinement are depicted in Figure 4 (only four cases are listed here due to the limited space). It can be observed that our method PLSA-RW can generate more accurate annotation results compared with the original annotations as well as the ones provided in [13]. Taking the second image of the first row for example, there exists only one tag ‘girl’ in the original annotation. However, after refinement by PLSA-RW, its annotation is enriched by other three keywords ‘face’, ‘child’ and ‘portrait’, which are very appropriate and reasonable to describe the visual content of the image. Overall, the experiment on Mirflickr25k indicates that PLSA-RW is fairly stable and efficient with respect to its parameters setting.

5 Conclusions

In this paper, we have proposed a novel refining image annotation method by combining PLSA with random walk to enhance the annotating performance. We

first construct a PLSA model with asymmetric modalities to estimate posterior probabilities of each annotating keyword for one image, and then employ a random walk process to mine the correlations of the keywords so as to capture the final refining annotation results. A weighted linear combination of label similarity and visual similarity is employed to calculate the pair-wise similarities between two candidate annotating keywords. Experimental results on Corel5k and Mirflickr25k datasets show that our model outperforms several state-of-the-art approaches. In the future, we intend to introduce semi-supervised learning into our approach and employ different image datasets to detect its performance comprehensively.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No.61035003, No.61072085, No.60933004, No.60903141), the National Program on Key Basic Research Project (973 Program) (No.2013CB329502), the National High-tech R&D Program of China (863 Program) (No.2012AA011003) and the National Science and Technology Support Program of China (2012BA107B02).

References

1. Li, J., Wang, J.: Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(9), 1075–1088 (2003)
2. Cusano, C., Ciocca, G., Schettini, R.: Image annotation using svm. In: *Proceedings of Internet imaging IV*. SPIE, vol. 5304, pp. 330–338 (2004)
3. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002, Part IV*. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
4. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp. 119–126. ACM (2003)
5. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: *NIPS* (2003)
6. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, vol. 2*, pp. 1002–1009. IEEE (2004)
7. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(10), 1802–1817 (2007)
8. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: *Proceedings of the 13th Annual ACM International Conference on Multimedia*, pp. 706–715. ACM (2005)
9. Wang, C., Jing, F., Zhang, L., Zhang, H.: Image annotation refinement using random walk with restarts. In: *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 647–650. ACM (2006)

10. Wang, C., Jing, F., Zhang, L., Zhang, H.: Content-based image annotation refinement. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
11. Liu, D., Hua, X., Yang, L., Wang, M., Zhang, H.: Tag ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 351–360. ACM (2009)
12. Xu, H., Wang, J., Hua, X., Li, S.: Tag refinement by regularized lda. In: Proceedings of the 17th ACM International Conference on Multimedia, pp. 573–576. ACM (2009)
13. Zhu, G., Yan, S., Ma, Y.: Image tag refinement towards low-rank, content-tag prior and error sparsity. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 461–470. ACM (2010)
14. Zhuang, J., Hoi, S.: A two-view learning approach for image tag ranking. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp. 625–634. ACM (2011)
15. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1), 177–196 (2001)
16. Li, Z., Liu, X., Shi, Z., Shi, Z.: Learning image semantics with latent aspect model. In: IEEE International Conference on Multimedia and Expo, ICME 2009, pp. 366–369. IEEE (2009)
17. Fellbaum, C.: *Wordnet. Theory and Applications of Ontology: Computer Applications*, 231–243 (2010)
18. Cilibrasi, R., Vitanyi, P.: The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering* 19(3), 370–383 (2007)
19. Huiskes, M., Lew, M.: The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 39–43. ACM (2008)