第44卷第1期 中国科学技术大学学报 Vol. 44,No. 1
2014年1月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Jan. 2014

Article ID:0253-2778(2014)01-0067-07

Image annotation by searching semantically related regions

DAI Lican, YU Nenghai

(*Department of Electronic Engineer and Information Science*, USTC, *Hefei* 230027, *China*)

Abstract: Based on abundant partially annotated images on the web, a novel framework for image annotation was proposed. By utilizing both the visual and textual knowledge of public available image database Image-Net, the proposed framework first learnt a set of weakly labeled visual concept classifiers, and then used the outputs of these learnt classifiers on image regions as descriptors to conduct the region-based search in a large scale image database for a query image. After that, search results mining and clustering was introduced to generate annotations to the query image. Compared with image-level representation, the proposed region-based semantic representation performs better at capturing image's multi-objects/semantics. The proposed framework takes advantage of both traditional classification-based approaches and large scale data-driven approaches. Experimental results conducted on 2.4 million web images and challenging image database have demonstrated the effectiveness and efficiency of the proposed approach.

Key words: image annotation; region-based search; large scale data-driven; classeme learning

CLC number: TP391 **Document code**: A doi:10.3969/j.issn.0253-2778.2014.01.008

Citation: Dai Lican, Yu Nenghai. Image annotation by searching semantically related regions[J]. Journal of University of Science and Technology of China, 2014,44(1):67-73.

基于语义相关区域搜索的图像标注

戴礼灿,俞能海

(中国科学技术大学电子工程与信息科学系,安徽合肥 230027)

摘要:探讨了如何有效地利用互联网上大规模的图像和文本信息以数据驱动的方式来实现图像的自动标注,并提出了一种基于语义相关区域搜索的图像自动标注框架.该框架首先利用人工建立的视觉和文本知识库Image-Net来训练一组弱分类器;然后将学习好的弱分类器作用于分割后的图像区域块生成Region-level的语义特征表示用以在大规模的图像数据库中进行相关图像区域的搜索,最后从搜索结果的文本描述中通过聚类挖掘的方式产生最终的图像标注结果.对比于image-level的底层特征表示,基于分类学习的区域模块具有更强的语义表达能力和更好的鲁棒性,更容易抓住图像本身包含的多个目标的多重语义;从而使得该框架兼具了大规模数据驱动和传统基于分类算法的优点.大量web图像和公认的测试数据集上进行的实验结果证明了本文提出框架的有效性.

关键词:图像标注;区域搜索;大规模数据驱动;分类学习

Received:2013-03-19;**Revised**:2013-04-28

Foundation item:National Natural Science Foundation (60933013).

Biography:DAI Lican, male, born in 1985, PhD student. Research filed: Information retrieval. E-mail: licand@mail.ustc.edu.cn

Corresponding author:YU Nenghai, PhD/professor. E-mail: ynh@ustc.edu.cn

0 Introduction

The number of digital images has exploded with the prevalence of digital cameras and the Internet which necessitates effective image retrieval techniques. Currently, there are mainly two image retrieval frameworks: Text-based image retrieval (TBIR) and content-based image retrieval (CBIR). Due to the semantic gap between low-level image visual features and high-level semantic concepts, most users prefer textual queries instead of abstract image content such as color or texture. However, manually annotating large quantities of images for the text-based searches is a tedious and time-consuming task, not scalable to large scale image databases. Therefore, automatic image annotation becomes a highly desired feature for large scale image retrieval and management applications.

Image annotation has been extensively studied in recent years. Many different approaches have been proposed to solve the problem in the computer vision and multimedia communities[1-4]. Although great improvements have been achieved, image auto-annotation is still far from practical, especially when we face an unlimited number of images and an unlimited vocabulary. By treating Web as a huge repository of weakly labeled images, data-driven approaches have demonstrated great potential for image annotation. As an example, search-based image annotation (SBIA)[5-6] is a representative work in this domain. Compared with traditional classification-based or probabilistic modeling-based methods, search-based image annotation utilizes rich media information (filename, URL and surrounding text) which is publically available on the web, and therefore effectively avoids two challenging problems: Limited images for training and limited lexicon for annotation. Though this framework has made good progress in image annotation, how to effectively measure the semantic similarity between images remains a key problem for further improving the annotation accuracy. In Ref. [5], low-level visual distance (64 dimensional color and texture feature) is directly employed to search for visually similar images. However, because of the well-known semantic gap problem and different concept distributions in different visual spaces[7], this simple similarity measure in one visual space inevitably limits the system's performance. Although the search result mining can bridge the semantic gap to some degree, it is obvious that the mining stage would benefit a lot if the search results could be improved. Recently, to overcome the semantic problem, a few researchers[8-10] have proposed image representations based on a set of attributes. Rather than directly represent visual content by low-level visual features, such approaches use the techniques in machine learning to generate a middle-level representation and have shown promising progress in image relevance ranking.

Actually, images usually have multiple objects/regions and different regions may have totally different contents, representing different semantic meanings. Therefore, it is intuitive to divide an image into regions and extract region-based features for semantic representation of images. Motivated by this, a framework is proposed to annotate uncaptioned images by searching semantically related regions in a large scale image database. Firstly, 200 visual concepts which have object bounding boxes are selected from Image-Net ontology[11]. These concepts are used to construct a semantic space which we call classeme space in this paper. The corresponding classeme classifiers are trained using these object regions' visual features. Secondly, millions of images with rich textual information are crawled from the Web and segmented into multiple stable regions. And each region is represented by the output scores of the trained classeme classifiers. With the supervised information and the fusion of multiple types of visual features, the learned classeme features can better represent the image regions' semantics. Furthermore, as shown in Ref. [12], these trained object category classifiers are capable of expressing not only those handpicked concepts but also thousands of concepts which share more or less common visual characteristics with the classemes. This generalization capability makes it possible for our classeme features to remarkably improve image search results as well as image annotation accuracy.

The framework of the proposed region-based image search for annotation is depicted in Fig. 1. To conduct image annotation for a query image, we first compute its stable segmentations and extract their region-level visual features to compute the classeme features by feeding the visual features to the trained 200 concept classifiers. Then, we use the classeme features as region descriptors to search for images with semantically related regions. Finally, search results mining is conducted to rank and generate the final annotations.

1 Search semantically related regions for annotation

1.1 Classeme learning

To construct a classeme space, a set of C category labels was drawn from the concepts in Image-Net[10] which have object bounding boxes. Taking both effectiveness and efficiency into consideration, we chose the number of 200 as an attempt. To pursue large coverage in real-world semantic space, we utilize a greedy strategy and use the minimal spanning tree algorithm to choose the final concepts which comprise our classeme space with largest diversities. For each category $c \in \{1 \cdots C\}$, 150 images with their bounding boxes were collected from Image-Net. To capture the visual characteristics of each region, the following four visual features were computed: Color Histogram (CH), pyramid of HOG (pHOG), color SIFT descriptors (cSIFT)[13] and self-similarity descriptors (SSIM)[14].

For CH, we used LAB color space. Histograms with 23 bins for each channel were computed and concatenated into a 69 dimensional descriptor. pHOG descriptors were computed as in Ref. [15], using three pyramid levels with 8 bins. cSIFT descriptors were computed at interest points detected with the Hessian-Affine detector. These descriptors were then quantized using a vocabulary of 20 000 code words. The SSIM descriptors were computed at the same locations with the detected cSIFT descriptors, and then quantized into a vocabulary of 400 code words. We normalized each descriptor to sum 1 and then employed the kernels to define the χ^2 distance between feature vectors, i. e. $k(x, x') = \exp(-\chi^2(x, x')/\gamma)$.

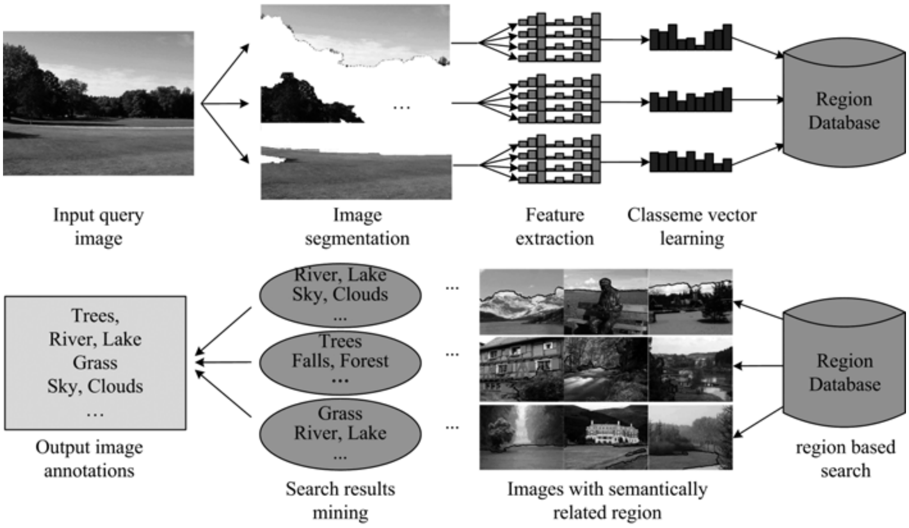


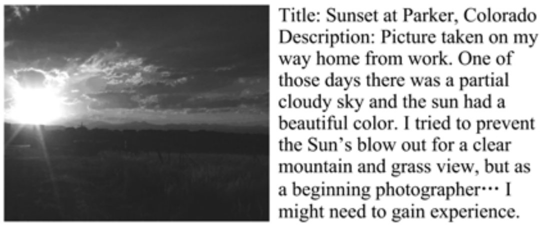
Fig. 1 The flowchart of our image auto-annotation process which mainly contains five steps: (1) Image segmentation, (2) Feature extraction, (3) Classeme vector learning, (4) Region-based search, (5) Search results mining. The final annotations are ranked according to the clusters' relevant scores

Finally, binary one-versus-all SVM classifiers were trained using multiple kernel learning (MKL)[16] for each category, and then the corresponding posteriors was obtained by Ref. [17]. These classifiers' outputs are all real-valued, and the expression $\phi_c(x) > \phi_c(y)$ implies that x is more similar to class c than y. After the classeme learning, given an image region x_u, the region-level feature called classeme feature vector is computed as follows to conduct later similarity search.

$$f(x_u) = [\phi_1(x_u), \cdots, \phi_C(x_u)] \qquad (1)$$

1.2 Data collection

To build a large scale image region database, we crawled about 2.4 million images from several photo forum websites, e. g. Photosig (http://www.photosig.com). Images of such websites are usually of high-resolution and have rich textual information, such as the title, photographer's descriptions and some semantically related comments. As shown in Fig. 2, the textual information roughly reflects the content of the image and the corresponding objects' semantics. For each image in the database, we employ the algorithm called JSEG[18] to segment it into multiple stable regions. Then multiple visual features are extracted from each image region, and used to compute the classeme feature vector as region descriptor for indexing the large scale region collection D.



Title: Sunset at Parker, Colorado
Description: Picture taken on my way home from work. One of those days there was a partial cloudy sky and the sun had a beautiful color. I tried to prevent the Sun's blow out for a clear mountain and grass view, but as a beginning photographer… I might need to gain experience.

Fig. 2 An example image with its textual description

1.3 Region based search and search results mining

Our annotation algorithm is generally based on the k-nearest neighbor prediction. The proposed framework is motivated by the following observation. If the region representation $x_{i,j}$ of image x_i is one neighbor of our query region x_u from image x in the constructed classeme feature space, the region $x_{i,j}$ may share common semantics with the query region x_u. Moreover, the closer $x_{i,j}$ is to the query region x_u in the feature space, the more likely x_i contains the labels of the shared semantics. This observation naturally leads to a k-nearest neighbor search for annotation procedure.

In the search stage, neighbors are retrieved by using the χ^2 distance in the classeme feature space:

$$D(x_u, x_{i,j}) = \sum_{k=1}^{C} \frac{(\phi_k(x_u) - \phi_k(x_{i,j}))^2}{\phi_k(x_u) + \phi_k(x_{i,j})} \qquad (2)$$

In order to improve the retrieval efficiency, as in Ref. [19], we first employ the Multi-Index algorithm to index database images and retrieve the top 50 000 images as candidate results for the query image. And then we conduct region-based search within the set of resulting images. After obtaining nearest neighbors for each query region, we adopt Search Result Clustering (SRC)[20] as our search result mining algorithm to generate candidate annotations for this region. The SRC algorithm is an effective clustering technique which can generate clusters with highly readable names. And distinct from other clustering approaches, it clusters documents by ranking salient phrases. It is worth noting that, due to the lack of region-level annotations, SRC is performed on the image-level textual information of the nearest neighbor regions. This simple operation will inevitably introduce some noises due to the mismatching between the image-level description and the regions' semantic meaning. However, this is a compromising solution given that there are not enough labeled regions and we leave this problem to our future study. After obtaining textual clustering results for each query region, we calculate a relevant score for each cluster based on the average member score criterion which has been proven effective in Ref. [5]. The average similarity of the members to the corresponding query region is computed for each cluster. Finally, the image annotation is generated by merging and ranking the clusters for all the query regions according to their relevant scores.

To summarize, the proposed auto-annotation for a test image is processed in five steps:

① Query image I is partitioned into stable segments X. Regions that are too small are discarded.

② Multiple types of visual features are computed for each image region.

③ For each $x_u \in X$, we apply the learned classeme classifiers $x_u \rightarrow f(x_u) = [\phi_1(x_u), \cdots, \phi_C(x_u)]$ to compute its classeme feature vector.

④ The k-nearest neighbors $N \subset D$ of $f(x_u)$ are retrieved and then used to generate candidate annotation keywords for this query region.

⑤ Post-processing is employed to merge and rank the candidate annotation keywords for all the query regions to obtain the final image annotations.

2 Experiments

2.1 Dataset

In our experiments, two query datasets are used to evaluate our annotation performance. The first one is 150 Google images of 15 categories selected in the same way as Ref. [5]. The second one is U. Washington dataset (UW) which is a content-based image retrieval database and has been widely used in CBIR and SBIA. Images in this dataset have about five manually labeled ground truth annotations on average. And for many images, not all objects are annotated. In our evaluation, we stick to the UW ground truth which means synonyms or correct annotations that do not appear in UW annotations are treated as incorrect.

2.2 Evaluation criterion

Since no ground truth is available for our crawled Google images, we simply asked the labelers to judge a suggested annotation as positive or not, based on which the annotation precision was evaluated for this query set. Ten volunteers were involved in the evaluation and each of them evaluated all the results. An annotation was assumed as positive only if seven out of ten labelers marked it positive.

$$\text{Precision@}\,m = \frac{1}{|J_q|}\sum_{I_i \in J_q}\frac{\text{correct}(I_i)}{\text{automatic}(I_i)} \quad (3)$$

where m is the number of returned top-ranked image annotations, J_q represents the image query set. The correct(I_i) is the number of correctly annotated words for the query image I_i, and automatic(I_i) is the number of automatically annotated words for image I_i. For the 1 109 images in UW, we introduce the recall to measure the annotation performance as well.

$$\text{Recall@}m = \frac{1}{|J_q|}\sum_{I_i \in J_q}\frac{\text{correct}(I_i)}{\text{groundtruth}(I_i)} \quad (4)$$

where groundtruth(I_i) is the number of the given ground truth annotations for the testing image I_i.

2.3 Results

In the evaluation, another two SBIA methods are conducted to compare with our approach based on the Google dataset. The first one (AS) is the approach adopted in Ref. [5] and the second one (RF) utilized the random forest based on multiple features for image annotation[3]. Fig. 3 shows the experimental results where $P@1$, $P@3$ and the overall precision (i. e. all the annotation results were included in evaluation) were measured.

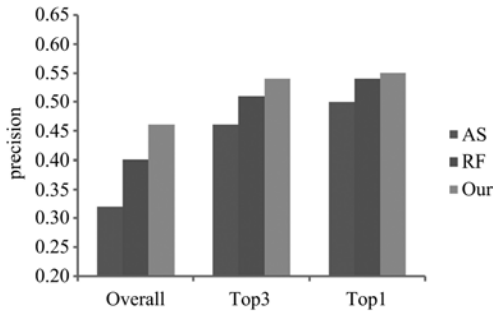


Fig. 3 Annotation precision of different approaches

From the results, we can derive the following observations:

(Ⅰ) RF and our approach have a great improvement compared with AS, which demonstrates the effectiveness of supervised learning and multiple feature fusion for image semantic representation.

(Ⅱ) Compared with DF, our approach can still obtain a remarkable gain especially when more than one annotation is generated, which demonstrates the effectiveness of proposed region-level image representation in capturing images' multi-objects and semantics.

To better understand the effect of the two components, i. e. classeme learning and region-level representation to the final annotation results, another three comparative experiments were conducted on the UW dataset, namely "SBIA-pure" (i. e. the same technique as in Ref. [5]), "SBIA-region" (i. e. use the same visual feature as SBIA-pure which is conducted at region-level for image retrieval), and "SBIA-region-classeme" which uses the learned classeme features at region-level for the related image search. Fig. 4 shows the corresponding experimental results in which both annotation precision and recall were evaluated.

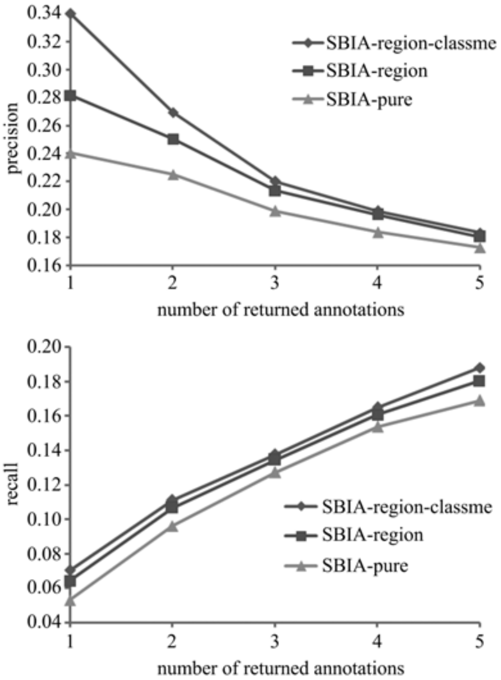


Fig. 4 Effects of different components

Performance gap between "SBIA-pure" and "SBIA-region" suggests the effectiveness of region-level image representation in capturing multiple semantics; whereas the performance gap between "SBIA-region-classeme" and "SBIA-region" indicates the effectiveness of our classeme feature learning. From the results, several observations can be drawn:

(Ⅰ) The precision of SBIA-region whose image search is conducted at the region-level has a considerable improvement compared with the baseline. Moreover, as the number of returned annotations increases, the obtained improvements are more obvious, which demonstrates once again that the region-based feature is a better representation in capturing an image's multiple semantic meanings.

(Ⅱ) Based on the learned classeme feature, the annotation precision is further improved, which demonstrates that the human-edited knowledge base is very useful for object semantic representation.

(Ⅲ) Compared with the above precision, the annotation recall is relatively low. This is partially caused by the incomplete annotations in the ground truth. In fact, it is well-known that one image is worth a thousand words, but no one knows which thousand words should be chosen. Nevertheless, we can still obtain a remarkable improvement via our region-level semantic representation model.

We list some exemplar image annotation results (top 5 are displayed) in Fig. 5 to illustrate the above problem. This demonstrates that the real performance of our approach on UW is much better than that shown in Fig. 4. As the evaluation did not take the synonyms into consideration and at the same time the given UW ground truth may ignore some contents of an image, many semantically relevant annotations generated by our approach were treated as incorrect.

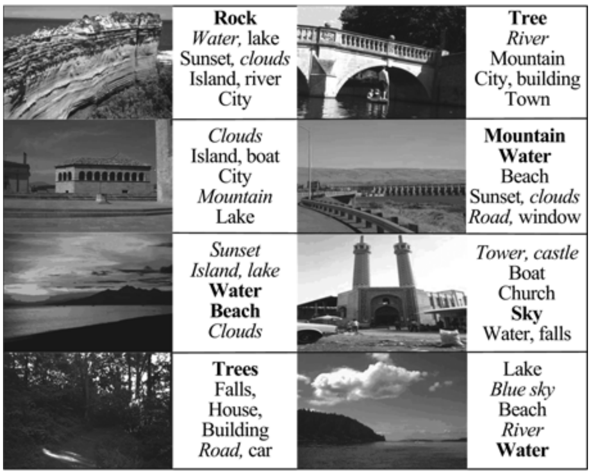


Fig. 5 Some annotation examples, and the italic words are those relevant ones not in the ground truth and treated as incorrect

3 Conclusion

In this paper, we have presented an image auto-annotation framework by searching semantically related regions. To build the database, millions of images together with rich textual information were crawled from the Web and segmented into multiple stable regions for experiments. To reduce the impact of semantic gap to SBIA, we built a classeme feature space for similarity measure and conducted image representation at region-level for searching. Experimental results demonstrate the effectiveness and efficiency of our proposed approach.

References

[1] Li J, Wang J Z. Automatic linguistic indexing of pictures by a statistical modeling approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(10): 1 075-1 088.

[2] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Toronto, Canada: ACM Press, 2003: 119-126.

[3] Fu H, Zhang Q, Qiu G P. Random forest for image annotation[C]// Proceedings of the 12th European Conference on Computer Vision. Florence, Italy: Springer, 2012: 86-99.

[4] Zhang D S, Islam M M, Lu G J. A review on automatic image annotation techniques[J]. Pattern Recognition, 2012, 45(1): 346-362.

[5] Wang X J, Zhang L, Li X R, et al. Annotating images by mining image search results[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11): 1 919-1 932.

[6] Wang X J, Zhang L, Liu M, et al. ARISTA - image search to annotation on billions of web photos[C]// IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: IEEE Press, 2010: 2 987-2 994.

[7] Lu Y J, Zhang L, Liu J M, et al. Constructing concept lexica with small semantic gaps[J]. IEEE Transactions on Multimedia, 2010, 12(4): 288-299.

[8] Li L J, Su H, Xing E P, et al. Object bank: A high-level image representation for scene classification & semantic feature sparsification[C]// Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 1 378-1 386.

[9] Mahajan D K, Sellamanickam S, Nair V. A joint learning framework for attribute models and object descriptions [C]// International Conference on Computer Vision. Barcelona, Spanish: IEEE Press, 2011: 1 227-1 234.

[10] Yu F X, Ji R R, Tsai, M H, et al. Weak attributes for large-scale image retrieval[C]// Proceedings of IEEE International Conference on Computer Vision and Patten Recognition. Portland, USA: IEEE Press, 2012: 2 949-2 956.

[11] Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database [C]// IEEE Conference on Computer Vision and Pattern Recognition. Miami, USA: IEEE Press, 2009: 248-255.

[12] Torresani L, Szummer M, Fitzgibbon A W. Efficient object category recognition using classemes[C]// Proceedings of the 11th European Conference on Computer Vision: Part I. Heraklion, Greece: IEEE Press, 2010: 776-789.

[13] van de Sande K E A, Gevers T, Snoek C G M. Evaluating color descriptors for object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1 582-1 596.

[14] Shechtman E, Irani M. Matching local self-similarities across images and videos[C]// IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE Press, 2007: 1-8.

[15] Bosch A, Zisserman A, Muñoz X. Representing shape with a spatial pyramid kernel[C]// Proceedings of the 6th ACM International Conference on Image and Video Retrieval. New York: ACM Press, 2007:401-408.

[16] Varma M, Babu B R. More generality in efficient multiple kernel learning[C]// Proceedings of the International Conference on Machine Learning. Montreal, Canada: ACM Press, 2009: 1 065-1 072.

[17] Platt J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods[J]. Advances in Large Margin Classifiers, 1999: 61-74.

[18] Deng Y N, Manjunath B S, Shin H. Color Image segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Fort Collis: IEEE Computer Society, 1999: 1-6.

[19] Li X R, Chen L, Zhang L, et al. Image annotation by large-scale content-based image retrieval [C]// Proceedings of the 14th Annual ACM International Conference on Multimedia. Santa Barbara, USA: ACM Press, 2006: 607-610.

[20] Zeng H J, He Q C, Chen Z, et al. Learning to cluster web search results[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK: ACM Press, 2004: 210-217.