Learning to Tag

Lei Wu*
MOE-MS KeyLab of MCC
University of Science and Technology of China
leiwu@live.com

Linjun Yang
Microsoft Research Asia
49 Zhichun Road,Beijing
100190, China
linjuny@microsoft.com

Nenghai Yu
MOE-MS KeyLab of MCC
University of Science and Technology of China
ynh@ustc.edu.cn

Xian-Sheng Hua
Microsoft Research Asia
49 Zhichun Road,Beijing
100190, China
xshua@microsoft.com

ABSTRACT

Social tagging provides valuable and crucial information for large-scale web image retrieval. It is ontology-free and easy to obtain; however, irrelevant tags frequently appear, and users typically will not tag all semantic objects in the image, which is also called semantic loss. To avoid noises and compensate for the semantic loss, tag recommendation is proposed in literature. However, current recommendation simply ranks the related tags based on the single modality of tag co-occurrence on the whole dataset, which ignores other modalities, such as visual correlation. This paper proposes a multi-modality recommendation based on both tag and visual correlation, and formulates the tag recommendation as a learning problem. Each modality is used to generate a ranking feature, and Rankboost algorithm is applied to learn an optimal combination of these ranking features from different modalities. Experiments on Flickr data demonstrate the effectiveness of this learning-based multi-modality recommendation strategy.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing-indexing methods; H.2.8 [**Database Applications**]: Image databases

General Terms

Algorithms, Theory, Experimentation

Keywords

Tag recommendation; Learning to tag; multi-modality Rankboost; social tagging

1. INTRODUCTION

With the advance of Web2.0 technology, multimedia content creation and distribution are much easier than ever before [6]. Along with the proliferation of images on the World-Wide-Web, effective image search approaches to obtain targeted images have gradually become an urgent demand. Currently, the performance of Web image search mainly depends on the quality of the image annotations or keywords (tags). Some methods automatically generate metadata by analyzing the image content, or the surrounding text on the webpages; while others generate these textual metadata by manual tagging. Most recently, social tagging has become a popular means to annotate Web images.

Although the automatic creation of metadata costs little human effort, the result of these statistical model based automatic methods are generally unsatisfying [14][1]. Especially on web images, which are quite noisy. To improve the performance of the automatic annotation, some approaches combine both image content analysis and the surrounding text on the image's webpages, e.g., [11][20][16][19]. These methods obtain some improvements over the purely content based methods, but they are still unacceptable for practical use.

The manual metadata generation is relatively more accurate and practical than the automatic annotation. The manual metadata generation is mainly based on the idea of ontology based labeling, which firstly defines an ontology and then let users label the web resources using the semantic markups in the ontology. There are also some work to mitigate the manually labeling work by semi-automatic annotation [5]. Although these ontology based annotation work is successful in some applications, e.g. bioinformatics and knowledge management, there are several limitations. Firstly, to build a semantic ontology that covers sufficient descriptions for multimedia content itself is expensive, time consuming and often requires domain knowledge [15]. Secondly, ontology based annotation usually requires users familiar with the ontology, which is too complicated for anyone without specialized training and knowledge.

Recently, a promising approach for manual metadata generation is social tagging, which requires all the users in the social network label the web resources with their own keywords and share with others. This labeling operation is named "tagging". Different from ontology based annotation; there is no pre-defined ontology or taxonomy in social tagging. Thus this task is more convenient for ordinary users. Social tagging has currently attracted huge amount of web

*This work was performed when Lei Wu was visiting Microsoft Research Asia as a research intern.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

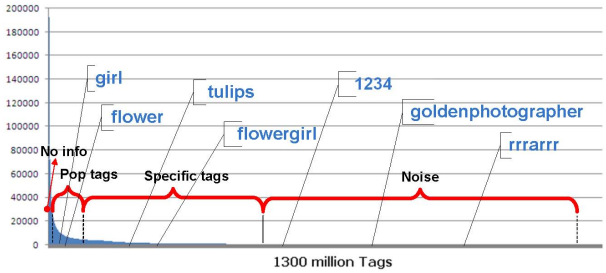


Figure 1: Tag distribution over a collection of 640 million images from Flickr.com. There are totally 1,300 million tags. Around 1% of the tags appearing more than 20,000 times, which contain little information. Around 5.82% of the tags have appeared more than 5,000 in the collection, which are considered as popular tags. 33.21% of the tags appears more than 50 and less than 5,000 times, which are defined as specific tags. 60% of the tags have appeared less than 50 times

users and effectively helps the organization of web resources. This strategy is adopted by some famous websites (e.g. Delicious, Flickr). This organic system of organization is also called "folksonomy".

Although social tagging is easy to perform, there are also some drawbacks. Firstly, it suffers polysemy and synonyms problem. As the users can use their own words to tag the images, different users may tag similar images with different words. So when querying "sea", one may not find images tagged "ocean" which represents the same concept. On the other hand, it is difficult for the users to input all the tags of the equivalent meaning. For this reason, lots of images may not be effectively retrieved. Secondly, ambiguity is also a problem. Users may use a general tag to represent different things. For example, when an image is tagged "apple", maybe it refers to the fruit "apple", or it could refer to the corporation or the product. In general, it is also quite difficult for the web users to realize the existence of ambiguity when tagging if they did not think of or even know the other meanings of the query. With these ambiguous tags, lots of irrelevant images may be retrieved.

To tackle the above problems, some researchers proposed the query expansion and suggestion [9][23], which extend the query to some related words to make the intention more clear. However, it does not completely eliminate the synonymy and tag ambiguity problems. The information in the query is limited, and the query expansion frequently cannot compensate the semantic loss in the tagging process, when users may ignore some semantic objects in the images. Recently, Xirong et al. [10] proposed the neighbor voting algorithm for image retrieval, which tried to predict the relevance of the user contributed tags. However the similarity between individual images is itself an open and complex problem. In this paper, we propose to tackle the semantic loss problem during the tagging process by combining both visual correlation in concept level and tag co-occurrence information. The semantically or visually related tags are recommended to the users to improve the tagging quality. The recommendation system will remind the users of the alternative tags and it can also help clarify the true semantic of the images. For example, when the user tags an image with word "sea", the recommendation system will list more rich and precise tags based on the input tags, such as "ocean", "water", "wave", etc. These recommendations will help users clarify the image content as well as reminding them of related semantics which may otherwise be ignored.

The quality of tag recommendation is quite important to social tagging and the consequent performance of image search. Firstly, high quality tag recommendation will motivate users to contribute more useful tags to an image [13]. The average number of tags for each image on Flickr is relatively small [2]. One of the reasons for that the users did not make large amount of tags is that they generally cannot think of too many words [17] in a short moment and few people would like to spend much time thinking about the alternative tags or more precise tags. With the help of high quality tag recommendation, users can provide a lot of useful tags in a short time. Also the spelling errors can be effectively avoided. Thus the average number of correct tags for each image is expected to increase. Secondly, tag recommendation will remind the users of more rich and specific tags. The distribution of tags on Flickr follows a power law distribution (1). Most of the users only use the popular keywords, which are only 5.82% of the whole tag collection. These tags are popular because they are common vocabulary and easily come to mind. Another 33.21% of the tags which appear 50-5,000 times are also informative while generally ignored by most users, because these words are more professional terms or only used for specific object or situations. The tag recommendation will help remind the user to use both popular and specific tags for social tagging. This reminder also helps create more precise tags. Thirdly, tag recommendation can depress the noise in social tagging system. It shows in the tag distribution that there are around 60% of tags in the tag corpus are misspelling or meaningless words. With the help of tag recommendation, users can tag an image by choosing rather than typing, which effectively avoids these spelling errors.

Existing tag recommendation approaches are performed by ranking the related tags based on the tag co-occurrence information. Much information is ignored in these methods, such as the visual correlation between tags, and the image content. A better choice is to use correlation from multi-modalities, such as tag co-occurrence, correlation between tag related images, the content of the target image, etc. However, it is not easy to combine these multi-modality correlations, since these modalities should be weighted differently for different samples. The basic idea of this paper is to learn an optimal combination of the multi-modality correlations to generate a ranking function for tag recommendation. Given the image and one or more initial tags, the algorithm will rank and sort the rest of the tags based on the tag correlation from each modality. Each is taken as a weak ranker. Then Rankboost[7] is adopted to combine weak rankers and form a better ranking function. Users can click the tags on the ranking list to annotate the image. After each click, the algorithm will update the ranking function as well as the tag recommendation function. Since the recommendation is based on the multi-modality correlations and is dependent on the ever-increasing tags in the database, it seems the users are using an selected ontology for tagging. The proposed method actually regularizes the

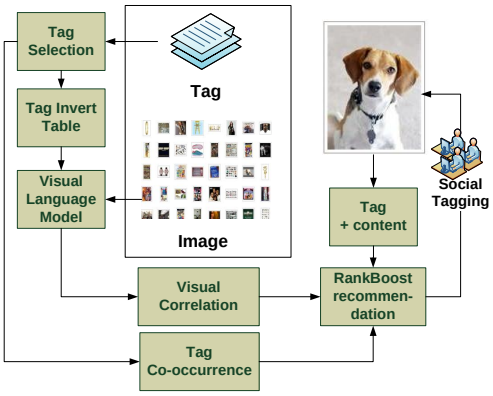


Figure 2: The flowchart of the social tagging recommendation system.

folksonomy with this dynamic ontology and makes the tagging converge to the underlying taxonomy.

The rest of the paper is organized as follows. Section 2 briefly introduces the related work on tag recommendation. Section 3 gives an overview of the annotation recommendation framework. Section 4 discusses the tag co-occurrence measurement. Section 5 discusses the tag content correlation measurement. 6 elaborates on the image content conditioned tag correlation. The hybrid information based RankBoost algorithm is discussed in Section 7. The experimental results and discussion are given in Section 8. Section 9 concludes this paper.

2. RELATED WORK

Recently, the tag recommendation based on the collective knowledge [18] is proposed. The authors measured the similarity between tags by their co-occurrence information in the data collection, and used the top similar tags as recommendations. This work has achieved some exciting results on the Flickr data, however, this kind of similarity measurement is greatly influenced by the synonyms and polysemy tags. For example, the similarity between "player" and "football" may be underestimated, since the co-occurrence of "player" and "soccer" should also have been taken into account to calculate the similarity between "player" and "football". In the other case, the similarity between "apple" and "ipod" is overestimated, since the "apple" here only indicates the corporation, and the cases where it represents the fruit should not be taken into account. While these problems can be better handled if the multi-modality visual similarity between the tags is used.

3. OVERVIEW OF THE TAG RECOMMENDATION FRAMEWORK

Given an image and one or several initial tags, we would like to recommend more tags which may have a semantical or visual correlation to the image. We rank these tags by their correlation to the target image, and list the top N keywords as the recommendation for further tagging.

We use a combination of three kinds of correlations to rank the tags: tag co-occurrence, tag visual correlation, and image conditioned tag correlation. The web users are very likely to tag the images with semantically related words, like"ipod" together with "apple". Based on this assumption, this semantic relationship can be somewhat captured by the tag co-occurrence in a large online photo sharing website with great number of independent users. However, it does not capture all relations between tags, such as the "tyre" in a photo of car, or the "eye" in a photo of "face". The photos containing both "tyre" and "car" may be tagged "car" only, and it is the same with "eye" and "face". To tackle these problems, visual correlation of the tags are applied. We build a visual language model (VLM) [22] for each tag and then use the inverse of the distance between these visual language models to measure the tag visual similarity. These two kinds of correlations only use the relation between tags, and the content of the target image is ignored. Further more, the image conditioned tag correlation is proposed to capture the tag similarity with respect to the target image.

We also formulate the recommendation as a learning to rank problem and combine these three kinds of correlation to generate the ranking. Since different types of correlation are independent measurements, it does not make sense to linearly combine them. In this paper, we consider these different correlations as different ranking features and combine them in the Rankboost framework, which uses the order of instances rather than the absolute distance. The flowchart of the system is shown in Fig. 2.

4. TAG CO-OCCURRENCE (TC)

Concept co-occurrence in daily life contains useful information to measure their similarity in the semantic domain. The semantic about the concepts is related to human cognition. Since 80% of the human cognition is formed from the visual information in daily life, the occurrence of concepts in daily life contributes a lot to their semantics. For example, the "monkey" is semantically related to "trees" because we often see monkeys living on the trees. This visual co-occurrence information contribute in forming the "monkey - tree" semantic relationship. It is also the same with lots of other kinds of semantic relationships, such as "flower - fruit", "fish - sea", "football - soccer","bird - sky" etc.

Tag co-occurrence (TC) on Flickr can partially capture the conceptual relationship in daily life. We assume that if two tags are frequently assigned to the same image, the corresponding concepts also have a high probability to co-occur in daily life. This statement makes sense for the following points. First, the Flickr dataset contains a huge amount of daily life photos generated by individual users. It seems that there are hundreds of millions of cameras capturing the object co-occurrence in daily life. Secondly, users are supposed to label the images according to their content with good reasons. There are studies [2], which show that the motivation of the users to tag the images is for social incentives. In other words, the users would like to make themselves known by contributing to the tagging task. Based on this conclusion, the users would more likely to tag the images truly according to the content rather than making noises.

The calculation of the tag co-occurrence on Flickr has already been investigated by the recent work [18]. Here we adopt the similar method to calculate the tag co-occurrence over a large dataset of 6 million images from Flickr. This dataset is sufficiently large for generating the statistics about the tag co-occurrence. Generally, according to different applications, the tag relevance is normalized into asymmetric

and symmetric forms, which are briefly represented as follows.

Asymmetric relevance measure. The relevance measurement between two tags is defined as follows.

$$R^a_{tag}(t_i, t_j) = \frac{|t_i \bigcap t_j|}{|t_i|} \quad (1)$$

where t_i and t_j are any two tags in the database. $|t_i|$ represents the number of times of the tag t_i that appears in the database.This relevance measure is asymmetric, which makes sense. Given tag t_i the probability of tag t_j may not be the same from that given tag t_j the probability of tag t_i.

Symmetric relevance measure. Although the Asymmetric relevance measure makes sense for tag recommendation, in some cases, the symmetric relevance measurement is more convenient to use.

$$R^s_{tag}(t_i, t_j) = \frac{|t_i \bigcap t_j|}{|t_i \bigcup t_j|} \quad (2)$$

5. TAG CONTENT CORRELATION (TCC)

Tag co-occurrence ignores some kinds of correlation. For example, if a user has tagged the image with "football", he/she may not tag it with "soccer". In other words, the correlation between synonym tags is not well estimated by tag co-occurrence. For another case, like the previously mentioned "tyre - car" and "eye - face" cases, users usually ignore some appearing concepts unconsciously. This drawback can be compensated by calculating the tag content correlation (TCC).

To represent the content correlation of the tags, visual language model is adopted to model each tag. For the i^{th} tag t_i, we collected a set of images marked with the tag. Then we build VLM based on these images to represent the *content* of the tag, and the differences of these quantified models are used to measure the tag visual correlation.

5.1 Data Source

In multimedia domain, the content of the images generally contain the semantic information of the tag. We intend to use the images as the tag content. Since a single image contains insufficient information to describe the tag, we collected a set of images to represent a tag. We aim to generate a semantic representation for the tags using these sets of images. In order to sample sufficient images to describe the tag, we generated a large image pool consisting of 1,000,000 images related to more than 50,000 popular tags by random walk sampling [4] from the popular photo sharing website Flickr.

5.2 Visual Representation

Visual language model (VLM) [22] is adopted to model the content of the tags in visual domain. This type of model can be generated very fast, which is appropriate for large scale datasets. It captures both the frequency of the visual features related to the tag, but also considers the spatial relationship between the neighboring features. This additional spatial information used in this model makes it more discriminative to characterize different tags.

The generation of the VLM is briefly described as follows. Each image is firstly divided into uniformly distributed equal-sized patches. Then some type of local appearance features, such as the texture histogram, color moment, etc, are generated for each patch. To depress the noise as well as the consequent training process, these local appearance features are sometimes coded into a visual word by k-means clustering or hash coding method. Afterwards, the VLM assumes that there are some visual grammar constraints on the arrangement of these visual words. According to different constraints, the VLM can be divided into unigram, bigram and trigram models. The unigram model assumes that the visual words are independent to each other, the output of the unigram model is the conditional distribution of the visual words given the tag. The bigram model assumes that the visual words are related to one of its neighboring words (usually left neighbor), and the output of the bigram model is the conditional distribution of the visual words given both the tag and one of their neighboring words. Accordingly, the trigram model assumes that the visual words are correlated to two of its neighboring words. So the output of the trigram model is the conditional distribution given the tag and two of the neighboring words. The dependency assumption can be described in the following equations.

$$P(x) = \Pi_{ij} P(w_{ij}) \quad (3)$$

$$P(x) = \Pi_{i=1}^{n-1} P(w_{i0}) \Pi_{j=1}^{n-1} P(w_{ij}|w_{i,j-1}) \quad (4)$$

$$\begin{aligned} P(x) =& P(w_{00}) \Pi_{j=1}^{n-1} P(w_{0j}|w_{0,j-1}) \Pi_{i=1}^{n-1} P(w_{i0}|w_{i-1,0}) \\ & \Pi_{i,j=1}^{n-1} P(w_{ij}|w_{i-1,j} w_{i,j-1}) \end{aligned} \quad (5)$$

where x is an image, and $w_{i,j}$ is the visual word of the i,j^{th} patch in the image.

The commonly used back-off and smoothing methods in the statistical language model are also adopted to estimate the conditional distribution of $P(w_{ij})$, $P(w_{ij}|w_{i,j-1})$, and $P(w_{i0}|w_{i-1,0})$. More details about this modeling method are discussed in previous work [22].

For the unigram model, the VLM is the visual word distribution over each tag. $P(w_i), w_i \in V, i = 1, \cdots, |V|$ where V is the vocabulary of size $|V|$. For bigram and trigram models, the VLM are the conditional distribution of visual words, $P(w_i|w_j)$,$w_i, w_j \in V, i, j = 1, \cdots, |V|$ and $P(w_i|w_j, w_k)$, $w_i, w_j, w_k \in V, i, j, k = 1, \cdots, |V|$. In summary, in all types of VLM, the content representation of a tag is a conditional distribution.

5.3 Similarity Measurement

Similarity measurement is critical in many applications [24]. It is also true in the tag recommendation system, where the core is the measurement of the visual relationship between tags. This similarity measurement directly influences the recommendation results. The direct measurement of the tag relationship is hard to achieve. As a result, we intend to represent the tags with the related images by context modeling and dictionary learning, which are successful in object category recognition [3][25][12]. The tag similarity is measured by the differences between the VLM of these tags, which is easily calculated using a generalized Flickr Distance [21]. If two tags are semantically correlated, their images are more probably to share some objects or scenes. Since the VLM captures the visual statistics of the tags and each VLM is a type of conditional distribution, we can calculate the square root of JS divergence between these distributions to measure the visual distance among the tags. Although there are some other kinds of distance measurements [8], the measurement is simple and effective.

The content correlation is divided into symmetric and asymmetric measurements. The symmetric measurement considers the two tags are equally important in the measurement. However, sometimes the two tags are not equally considered, i.e. given the tag "apple tree" the concept "apple" is more probably to appear than given the tag "apple" there exists an "apple tree", since "apple" have many other meanings and may exist in many occasions. Thus the asymmetric relevance measurement makes sense. Generally, these two kinds of measurements makes tiny different. Since the symmetric relevance can reduce half of the computational burden, it is widely used in common cases.

Asymmetric relevance measure.

$$D^a_{TCC}(t_i, t_j) = KL(L(t_i)||L(t_j)) \tag{6}$$

$$R^a_{TCC}(t_i, t_j) = \frac{1}{D^a_{TCC}(t_i, t_j)} \tag{7}$$

where $L(t_i)$ represents the VLM for tag t_i and $KL(\cdot)$ is the KL distance between two visual language models. D^a_{TCC} is defined as the asymmetric tag correlation measurement.

Symmetric relevance measure.

$$D^s_{TCC}(t_i, t_j) = \frac{1}{2}[KL(L(t_i)||M) + KL(L(t_j)||M)] \tag{8}$$

$$M = \frac{1}{2}[L(t_i) + L(t_j)] \tag{9}$$

$$R^s_{TCC}(t_i, t_j) = \frac{1}{D^s_{TCC}(t_i, t_j)} \tag{10}$$

where M is the average language model between the two tags. D^s_{TCC} is the symmetric content distance between the two tags. In fact considering the VLM as a probabilistic distribution, this distance equals to the well-known Jensen-Shannon divergence, the square of which is demonstrated to be a strict metric. The symmetric relevance measure between the content of the tags $R^a_{TCC}(t_i, t_j)$ is defined as the reciprocal of their symmetric visual distance.

6. IMAGE CONDITIONED TAG CORRELATION MEASUREMENT (ITC)

Although the tag visual similarity adopted the visual information of the images related to the respective tag to calculate the relevance, it does not use the information about the test image. We believe that for different test images, the similarity measurement should be different. For example, the tag "apple" and "pear" is of high relevance, however, given the test image of an ipod, the "pear" should not be a proper recommendation. In order to explore this kind of conditional correlation, the image conditioned tag correlation (ITC) measurement is proposed.

We assume that if the two tags are similar to each other, the likelihood of generating the target image should be similar. We represent the tag by the likelihood of generating the target image given unigram, bigram and trigram models respectively. These three likelihoods can be taken as location of the tag in the likelihood space.

Let x be a novel test image (for user to tag), and x_k, $k = 1, \cdots, L$ be L related images to a tag $t \in T$, where T is the tag corpus. First, we measure the similarity between the target image x and all the tag associated images. To achieve this task, the target image x is also represented in

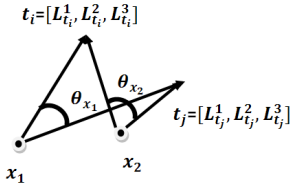


Figure 3: The target image conditioned tag distance measurement. Given different target image x_1,x_2, the distance between tags t_i,t_j is different.

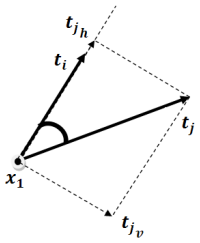


Figure 4: The geometric interpretation of the asymmetric relevance measurement.

the matrix of visual words $x = [w_{ij}]_{i,j=1,\cdots,n}$ representation as discussed in Section 5. Then the similarity is defined by the likelihood $\mathcal{L}$.

$$\mathcal{L}^m_t(x) \propto P(x|VLM^m_t) \tag{11}$$

Also we apply the visual context constraints, the similarity measurement can be further represented in the unigram, bigram and trigram forms.

$$\mathcal{L}^m_t(x) \propto \Pi_{ij} P(w_{ij}|VLM^m_t) \tag{12}$$

$$\begin{aligned} &\mathcal{L}^m_t(x) \\ &\propto \Pi^{n-1}_{i=1} P(w_{i0}|VLM^m_t)\Pi^{n-1}_{j=1} P(w_{ij}|w_{i,j-1}, VLM^m_t) \end{aligned} \tag{13}$$

$$\begin{aligned} &\mathcal{L}^m_t(x) \\ &\propto P(w_{00}|VLM^m_t)\Pi^{n-1}_{j=1} P(w_{0j}|w_{0,j-1}, VLM^m_t) \\ &\Pi^{n-1}_{i=1} P(w_{i0}|w_{i-1,0}, VLM^m_t)\Pi^{n-1}_{i,j=1} P(w_{ij}|w_{i-1,j}w_{i,j-1}, VLM^m_t) \end{aligned} \tag{14}$$

where VLM^m_t is the m-gram visual language model for the t^{th} tag. It is worth noting that the m-gram used here could be the unigram model, the bigram model, or the trigram model. For simplicity, we denote these three kinds of likelihood functions with the brief index $m = 1, 2, 3$ respectively, where $m = 1$ corresponds to the unigram model; $m = 2$ is the bigram model; $m = 3$ denotes the trigram model.

Given the image x, each tag t is represented by three likelihood values corresponding to the unigram, bigram and trigram models respectively, as shown in Fig. 3. Thus the distance between the two tags t_i and t_j can be defined in this likelihood space as the inter product of the two likelihood vectors. This kind of distance also consists of symmetric and asymmetric forms, which are formulated in Eq. (15)

and (16) respectively.

$$D^s_{ITC}(t_i, t_j, x) = \frac{\mathbf{L}_{t_i}(x) \cdot \mathbf{L}_{t_j}(x)}{\|\mathbf{L}_{t_i}(x)\|\|\mathbf{L}_{t_j}(x)\|} \tag{15}$$

$$\mathbf{L}_{t_i}(x) = [\mathcal{L}^1_{t_i}, \mathcal{L}^2_{t_i}, \mathcal{L}^3_{t_i}] \tag{16}$$

and the asymmetric distance is defined as

$$D^a_{ITC}(t_i, t_j, x) = \|\frac{\mathbf{L}_{t_i}(x) \cdot \mathbf{L}_{t_j}(x)}{\|\mathbf{L}_{t_i}(x)\|} - \mathbf{L}_{t_i}(x)\| \tag{17}$$

where t_i is usually the initial tag related to the target image and t_j is the novel tag.

Correspondingly, the similarity measurement between tags given the target image is the inverse of the distance metric.

$$R^s_{ITC}(t_i, t_j, x) = \frac{1}{D^s_{ITC}(t_i, t_j, x)} \tag{18}$$

$$R^a_{ITC}(t_i, t_j) = \frac{1}{D^s_{ITC}(t_i, t_j, x)} \tag{19}$$

This asymmetric distance measurement (Eq. (17)) is revisited here. In another aspect, this relevance measurement can also be interpreted from the relevance decomposition view, which is illustrated as Fig.4. In the likelihood space, the novel tag t_j is represented as a 3-dimensional vector, which is further decomposed into two components, the relevant component t_{j_h} that is parallel to the initial tag t_i, and irrelevant component t_{j_v} which is vertical to the initial tag t_i. The correlation is inverse to the distance between relevant component t_{j_h} and the initial tag t_i.

In the target image conditioned tag similarity measurement, the relevance between both the initial tag and the target image is considered, which makes sense for image tagging recommendation.

7. MULTI-DOMAIN RELEVANCE FOR TAG RECOMMENDATION

In this section, we discuss to combine multi-domain relevance for tag recommendation (MRR). Here the multi-domain relevance refers to the tag co-occurrence, tag content correlation, and image conditioned tag correlation. The basic idea is to produce an accurate ranking function by combining many "weak" learners. Different from traditional training procedure, these "weak" learners are trained based on cross domain relevance of the semantic targets.

7.1 Recommendation framework

Given an image and one or more tags, the task is to recommend more related tags to label the image. This recommendation is performed by ranking based on the three types of similarity, tag co-occurrence, tag to tag similarity, and image conditioned tag similarity. The tag co-occurrence (TC) measures how likely two different tags are labeled together. The tags are generated by web users, and their co-occurrence can reflect the tag relationship in human cognition. The tag to tag similarity (TTS) measures the correlation of the visual content of the tags. In this sense, we can also measure the similarity of the tags by their VLMs. The image based tag similarity (ITS) measures the similarity between the tags conditioned on the target image.

These different types of correlation can be combined together to generate a more robust recommender. However, as these similarity metrics are generated from different domains, the linear combination of these similarity measurements does not make sense. To tackle the multi-domain measurement problem, we propose to combine these multi-domain metrics into the Rankboost framework to generate a reasonable combined recommender. Since the Rankboost algorithm only considers the relative order of the samples and it does not actually use the distance metric, it can generate a fair ranking based on the multi-domain information.

7.2 Rankboost

In this framework, each sample tag is considered as an instance. All the tags in the dataset form the instance space $\mathcal{X}$. For each weak ranker, we aim to generate a function f_i, which maps an instance x_i from the instance space $\mathcal{X}$ to the *preference space* or *ranking space* R. These given rankings of the instances are called *ranking feature*, which is taken as a kind of feature to train the weak rankers. In fact, the ranking feature is any kinds of measurement indicating the relative order between two instances. This measure does not have to be metric, since only the relative order is useful and the distance is not used. For example, the $f_i(x_1) > f_i(x_2)$ means the ranking feature of x_1 is superior to that of x_2. The distance $f_i(x_1) - f_i(x_2) > f_i(x_2) - f_i(x_3)$ is meaningless. Based on this property, the ranking features can be generated using different kinds of information or from different domains. The requirement is that they can be used to measure the relative order between any two instances. In this paper, these ranking features are generated from both the textual domain and visual domain. So the weak rankers are also called multi-domain weak rankers.

In order to learn the ranking, we define the ranking loss as follows.

$$rloss_D(H) = \sum_{x_0, x_1} D(x_0, x_1)\delta(H(x_1) \leq H(x_0)) \tag{20}$$

$$D(x_0, x_1) = c \cdot \max(0, \Phi(x_0, x_1)) \tag{21}$$

where H_i is the sum combination of all weak rankers, and $\delta(\pi)$ is 1 if π holds and 0 otherwise. $\Phi(\cdot)$ is the feedback function, $\Phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. If x_1 is more relevant to the tag than x_0, then $\Phi(x_0, x_1) > 0$. This feedback function is usually generated by user interaction.

As is demonstrated in [7], the upper-bound for this ranking loss is Z_t

$$Z_t = \sum_{x_0, x_1} D_t(x_0, x_1) exp(\alpha_t(h_t(x_0) - h_t(x_1))) \tag{22}$$

where h_t is the output of the t^{th} weak ranker. The weak learner is only needed to minimize this upper-bound.

7.3 Learning to tag

Given the image and some of its initial user created tags, we would like to recommend a list of related tags which may be also applicable to the image. We denote the set of initial tags as $\mathcal{OT}$, and the set of remaining tags as $\mathcal{UT}$. The relevance of the tags is represented in two domains. The average tag co-occurrence on other Flickr images is deemed as one domain, and their content correlation as another domain. Then for these domains, we generate several ranking features $\{f_l\}_{l=1}^{3n}$ (n is the number of initial tags). The first n ranking features are generated based on tag co-occurrence;

Algorithm 1 Cross domain Rankboost training process

Input: Given tags $t_1, \cdots, t_n \in \mathcal{OT}$, and $t_1, \cdots, t_m \in \mathcal{UT}$, and distribution D over $\mathcal{UT} \times \mathcal{UT}$.
where
$\mathcal{OT}$ is the set of the initial tags.
$\mathcal{UT}$ is the set of the remaining tags in the database.
Initialize $D_1 = D$.
Generate ranking features $\{f_l\}_{l=1}^{3n}$; $\forall t_i \in \mathcal{UT}, t_l \in \mathcal{OT}$
$f_l(t_i, t_l) = R^s_{TC}(t_i, t_l),\ l = 1, \cdots, n$
$f_{n+l}(t_i, t_l) = R^s_{TCC}(t_i, t), \ l = 1, \cdots, n$
$f_{n+l}(t_i, t_l) = R^s_{ITC}(t_i, t), \ l = 1, \cdots, n$
where $t_i \in \mathcal{UT}$, $t_l \in \mathcal{OT}$
for $k = 1, \cdots, K$. **do**
• Select pair $(t_i, t_j) \in \mathcal{UT} \times \mathcal{UT}$ with distribution D.
• Get weak ranking h_k from ranking features of selected pairs
• Update: $\alpha_k = \frac{1}{2}\ln(\frac{1+r}{1-r})$,
where $r = \sum_{t_i,t_j} D_k(h_k(t_i) - h_k(t_j))$
• Update: $D_{k+1}(t_i, t_j) = \frac{D_k(t_i,t_j)\exp(\alpha_k(h_k(t_i)-h_k(t_j)))}{Z_k}$.
where Z_k is a normalization factor.
$Z_k = \sum_{t_i,t_j} D_k(t_i, t_j)\exp(\alpha_k(h_k(t_i) - h_k(t_j)))$
end for
Output the final ranking: $H(t) = \sum_{k=1}^{K} \alpha_k h_k(t)$.

the next n ranking features are generated by TTS; and the last n are generated by ITS.

$$f_l(t_i, t_l) = R^s_{TC}(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{23}$$

$$f_{n+l}(t_i, t_l) = R^s_{TCC}(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{24}$$

$$f_{2n+l}(t_i, t_l) = R^s_{ITC}(t_i, t_l),\ t_l \in \mathcal{OT},\ t_i \in \mathcal{UT},\ l = 1, \cdots, n \tag{25}$$

These ranking features of the two domains are combined in the Rankboost framework. Eq. 22 shows that the weak ranker depends only on the relative ordering information of the samples rather than the specific scoring information. Since here we adopted similarity measurements from multiple domains, and the score may represent different semantics, it does not make sense to use the scoring information directly. For these reasons, it is reasonable to adopt the Rankboost framework as described in [7] in this paper.

In the training process, the algorithm generates some random pairs over $\mathcal{UT} \times \mathcal{UT}$ with the distribution D. Then based on these random pairs, some weak learners are trained using both the tag co-occurrence relevance and the content correlation between the tags in each pair. The distribution D is also updated to minimize the ranking loss. A high weight α is assigned to the pairs indicating importance of making that pair correct.

We adopt the same type of weak ranker as [7]. In the recommendation process, the top N relevant tags, according to the ranking result, are recommended to the user for further tagging.

8. EXPERIMENT

In this section, we compare the proposed approach with the commonly used tag co-occurrence based recommendation approach in the real world image tagging task.

8.1 Setting

We have collected 1,000,000 images and associated 200,000 tags from Flickr as the database. Since these photos are uploaded by large amount of independent users, there are diverse categories of topics and can reflect the real situation of web images. We randomly select 500 images to perform the manually tagging task. The rest of the images with tags are used to train the tag co-occurrence measurement as well as the content correlation measurement.

Some frequency filtering methods are adopted to remove the noise in tag co-occurrence. In the real world, the web users may make mistakes during the tagging process. Some of the tags may be misspelled. They will be low frequent tags. To depress this kind of noise, we adopt a simple and effective method by removing the tags that appear less than 50 times in the dataset. There are also some high frequency stop words, such as "bravo", "image", "photo", etc. These tags contain little information. We also consider them as irrelevant tags, which can be suppressed by removing the tags that appear more than 10,000 times in the dataset.

To initialize the feedback function D for the Rankboost algorithm, we adopt the pseudo feedback function generated from WordNet, which contains the human created knowledge about word correlations. Using WordNet for initialization is less expensive than the human labeling. However, since there are lots of tags out of the scope of WordNet, we can just generate the feedback function among the tags within the WordNet corpus. To extend the coverage of the feedback function, human interaction is unavoidable. Given some initial tags, if tag x_1 gets a higher similarity score to the initial tags than tag x_0 in the WordNet, the feedback function $\Phi(x_0, x_1) > 0$; otherwise $\Phi(x_0, x_1) < 0$.

For each test image, we give 5 initial tags generated by users on Flickr. For the baseline method, we adopt the commonly used recommendation by tag co-occurrence. For the second method, we use only tag co-occurrence as the ranking feature for the Rankboost recommendation. For the third method, we use both tag co-occurrence and content correlation in the Rankboost recommendation. To provide real-time recommendation, we calculate the co-occurrence similarity and content similarity of all tags in the database offline, and the Rankboost algorithm is only performed on the top 100 relevant tags for each image.

8.2 Evaluation

Each recommendation approach will generate an ordered list of relevant tags for each image. Then a volunteer is required to evaluate these recommended tags. If a tag is relevant to the image, it will be marked true; otherwise false. The average precision of the topN (N=5) recommendations and the coverage over all correct recommendations are adopted to qualitatively measure the performance of each recommendation method. The coverage is defined as the proportion of correct tags (including all correct tags by both methods and initial tags) that are recommended by the specific method. We adopt the coverage rather than the recall, because the recall is unapplicable for the recommendation task.

$$Coverage(m_i) = \frac{\sum_{x \in TopN} \delta(x|m_i)}{\sum_i \sum_{x \in TopN} \delta(x|m_i)} \tag{26}$$

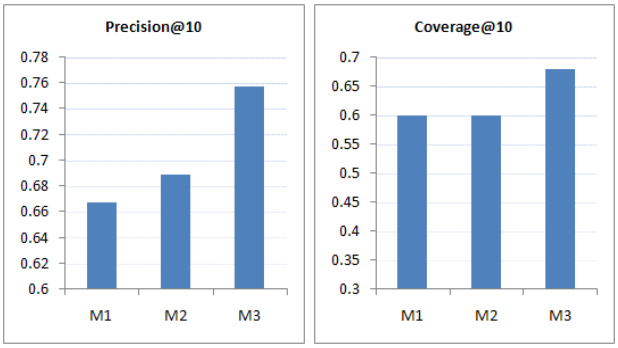
$$\delta(x|m_i) = \begin{cases} 1, & \text{given method i, x is related tag;} \\ 0, & \text{otherwise.} \end{cases} \quad (27)$$

8.3 Experiment A: Multi-domain relevance V.S. tag cooccurrence

In this experiment, we aim to compare the Multi-domain Relevance based Rankboost recommendation (MRR) with the Tag Co-occurrence based tag recommendation (TC). For the MRR method, one weak ranker is trained based on the tag co-occurrence information; another one is trained by the tag content correlation; and a third one is trained by the image conditioned tag correlaion. These weak rankers are combined together into the Rankboost framework. The comparison between MRR and TC shows how much of the improvement is gained by incorporating the tag content correlation and image conditioned tag correlation. To reveal how much is gained by the multi-domain features, we further compared the MRR method with the Tag Co-occurrence based Rankboost (TCR).

We randomly choose 20% of sample images as test data, and the rest images and associated tags are used as the training data. To train the TCR model, we count the co-occurrence frequency between every pair of tags from the collection, and then normalize them into range 0 to 1. For the MRR method, we train the tag based weak ranker the same way as the purely tag co-occurrence based method. In the visual domain, we generate the ranking features based on the visual similarity of these tags, which are discussed in the previous sections. Then we combine these weak rankers with the Rankboost algorithm.

Figure 5 and Figure 6 compare the performance of the two methods under 5 initial tags. In Figure 5, the left figure shows average precision at top 10 tags. "M1" represents the TC method [18]. "M2" denotes the TCR. "M3" denotes the MRR method. Comparing M2 and M1, we find the supervised recommendation outperforms the unsupervised method by 3.3% in precision and 8.8% in coverage. Comparison of M3 with M2 shows that after the combination of content correlation, the precision and coverage gain 5.5% and 10.5% separately. In total, the proposed hybrid information based Rankboost has gained 9.0% in precision and 20.3% in coverage over the commonly used tag co-occurrence approach. These results demonstrate the effectiveness of the supervised recommendation as well as the usefulness of the correlation of the image content.



(a) Precision@5 (b) Coverage@5

Figure 5: Performance of different methods.

Init Tags	Cruise party boat purple spectrum	Travel sea seaweed water colors	Travel family sea sun beach
Method1	friends fun birthday art girls summer Florida winter snow flower	vacation Asia trip holiday nature city cannon tree Europe building	vacation holiday Europe nature city water trip building Asia light
Method2	friends girls music fun night love art holiday vacation trip	vacation holiday trip Asia Europe nature city fun music friends	vacation trip Asia holiday water Europe nature tree friends sun
Method3	friends dance fun girls night music love men happy laugh	trip ocean sky island nature landscape blue umbrella red men	vacation fun water kids ocean sky holiday sand wave blue

Figure 6: Tag recommendation examples.

Figure 6 gives an illustration of the tag recommendation results. The texts below each image are its tags. The first row is the initial tags assigned by the Flickr users. The second row is the results of the simple tag co-occurrence method [18]. The third row is the result for tag domain Rankboost method, and the last row is multi-domain Rankboost method. The proposed method provides more relevant tags.

8.4 Experiment B: Rankboost V.S. linear combination

To combine the visual correlation with the tag cooccurrence, we can also use the linear combination rather than the Rankboost algorithm. For the multi-domain relevance linear combination method (MRL), we generate a new similarity score in the following form.

$$R^s_{linear} = R^s_{TC} + \eta_1 R^s_{TCC} + \eta_2 R^s_{ITC} \quad (28)$$

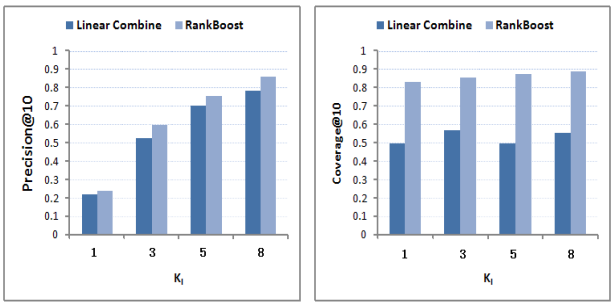
$$R^a_{linear} = R^a_{TC} + \zeta_1 R^a_{TCC} + \zeta_2 R^a_{ITC} \quad (29)$$

where R^s_{linear} is the combination of the symmetric similarity measurement, and R^a_{linear} is the combination for asymmetric similarity. $\eta_1, \eta_2, \zeta_1, \zeta_2$ are the combination coefficients, which are determined by experiments.

This kind of linear combination also takes multi-domain correlation information, but there are two disadvantages for this simple method. Firstly, the combinational coefficients is difficult to determine automatically. Secondly, the linear combination method combines the similarity score in different domains, which is not reasonable. In the Rankboost method, we combine the results of the weak rankers rather than the similarity scores.

In this experiment, we intend to demonstrate the advantage of Rankboost based multi-domain recommendation over the linear combination. For each test data, we give K (K=1,3,5,8) initial tags, and a ranking of the rest tags are generated by both Rankboost and the linear combination method. For the linear combination method, the set both η_1, η_2 and ζ_1, ζ_2 to 1. Then the precision at top 10 tags are evaluated manually. The results are shown in Fig. 7.

From these comparisons, we find that both precision@10 and coverage@10 rise with increasing the number of initial tags. The Rankboost algorithm outperforms the simple lin-



(a) (b)

Figure 7: Comparison with linear combinations

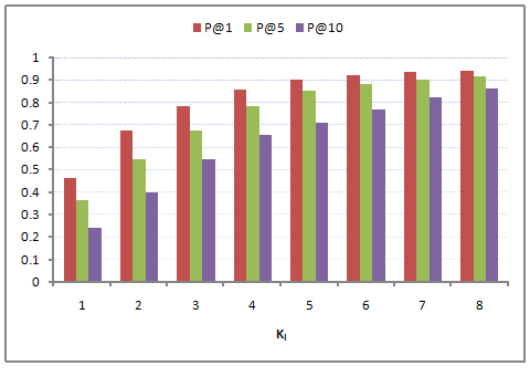


Figure 8: precision@N with K_I initial tags.

ear combination of the ranking features. This advantage is more obvious when the number of initial tags is larger. This is because given more initial tags, the Rankboost algorithm will have more information for ranking as well as more information to determine how to more effectively combine the two kinds of ranking features. While for the linear combination algorithm, since the combination coefficients are fixed, it is not likely for this method to optimize the combination. The increase of performance in the linear algorithm is purely based on more information for ranking.

8.5 Experiment C: Parameter influence

In this experiment, we would like to study the influence of the number of initial tags K_I. In the real world, there may not be large amount of tags for each image. The number of initial tags may greatly limit the application of the method. In this study, we selected a collection of images which have more than 10 initial tags. Then we set K_I from 1 to 8, and calculate the precision@1,5,10 under respective initial tag setting. The results are shown in the following figure.

From this results, we find the precision of the recommendation is improved with the increasing of K_I. This is because with more initial tags, the algorithm can use more information to predict the content of the image. It is also found that the precision with 5 initial tags is already acceptable for a common recommendation system. Fortunately, with respect to the statistic generated from the 10,000 random sampling from Flickr, the average number of correct tags for an image is above 5. This means that in general cases the users can label more than 5 correct tags, and then it is possible to apply this recommendation scheme on Flickr to help users tagging more.

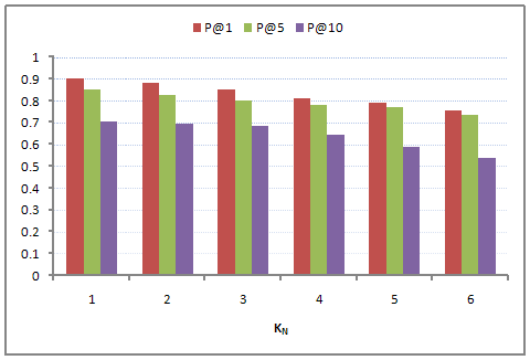


Figure 9: precision@N with K_N irrelevant tags

8.6 Experiment D: Noise resistance

Noise exists in the real world data. We can not assume that the tags on the web are all correct. For the real application, noise resistance is a critical feature. In this experiment, we would like to show this property of the proposed recommendation approach. 5 initial correct tags are given for each test image. In this experiment, we also generate some noise (irrelevant tags), which are combined with the correct tags. For a detailed analysis of the influence of noise on the recommendation results, we gradually increase the number of irrelevant tags K_N from 1 up to 5. The final precisions at topN (N=1,5,10) under each irrelevant rate are shown in Figure 9.

From the figure, we find the precision of the recommendation drops when the number of irrelevant tags increase. The reason is that the irrelevant tags have mislead the ranking function. The irrelevant tags may not relate to each other, while the correct tags of the same image are semantically related to each other. Thus the ranking algorithm will still take more consideration on tags related to the correct ones. This is also reflected in the result. Even with 5 irrelevant tags, that is 50% of the number of tags, the precision@1,5,10 still remain above 75%,73% and 53% respectively. This means the algorithm is applicable with 50% of irrelevant tags, which is actually worse than the real world situation. By sampling 10,000 images from Flickr, the statistic shows that on average, there are about 24% of tags for each image are irrelevant.

8.7 Experiment E: Computational cost

Computational cost is also one of the main considerations for the web scenarios. The main computational time for the proposed methods lies on the visual similarity measurement between tags. Fortunately, this process can be generated offline. We first generate a dictionary of tags, and then generate both co-occurrence based similarity measurement as well as the content based tag similarity measurement offline. For the online recommendation, we only need to rank the tags according to the initially labeled tags. This process is very fast. The details of the computational time for each sample under different number of initial tags are shown in the following table. This experiment is performed by Intel(R) Core(TM)2 Duo CPU@2.00GHz, 1G memory. For each image, the multi-domain Rankboost method takes only 0.02 second.

Table 1: Computational cost between different methods.

	TC	TCR	MRL	MRR
Training	0.178	0.183	2.190	2.195
Test	0.001	0.002	0.001	0.002

In Table 8.7, the training time shows the cost for calculating the similarity between each concept pair. For recommendation, the time shows the cost for generating a recommendation list for each target image. From this comparison, we find the computational cost for the proposed method lies mainly on the computation of the visual information. However, the building of VLM can be performed offline in the training process. So the computational time for the online recommendation process is almost the same for all the methods.

9. CONCLUSIONS

In this paper, we have introduced the learning based tag recommendation approach. It generates ranking features from multi-modality correlations, and learns an optimal combination of these ranking features by the Rankboost algorithm. This recommendation will update each time when a new tag is added, the efficiency makes this recommendation method suitable for real time applications. With this learning based recommendation, better quality of recommendations is achieved, and the users are reminded of more diverse and correlated tags. Experiments also demonstrate that this learning to tag framework is more effective than current approaches and the combination of multi-modality relevance is helpful to tag recommendation.

10. ACKNOWLEDGMENTS

The research is supported in part by the National Natural Science Foundation of China (60672056), the 863 Program (2008AA01Z117), and USTC Postgraduate Innovation Foundation (KD2007049).

11. REFERENCES

[1] E. Akbas and F. Yarman Vural. Automatic image annotation by ensemble of visual descriptors. *CVPR '07.*, June 2007.

[2] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07*, 2007.

[3] J. Amores, N. Sebe, and P. Radeva. Context-based object-class recognition and retrieval by generalized correlograms. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1818–1833, 2007.

[4] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *WWW '06 Proceedings*, 2006.

[5] J. Blythe and Y. Gil. Incremental formalization of document annotations through ontology-based paraphrasing. In *WWW '04*, 2004.

[6] S. Boll, P. Sandhaus, A. Scherp, and U. Westermann. Semantics, content, and structure of many for the creation of personal photo albums. In *Proceedings of ACM Multimedia '07*, 2007.

[7] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. In *Proceedings of ICML'98,*, 1998.

[8] G. Koloniari, Y. Petrakis, E. Pitoura, and T. Tsotsos. Query workload-aware overlay construction using histograms. In *Proceedings of CIKM '05*, 2005.

[9] H. Li, Y. Wang, D. Zhang, M. Zhang, and E. Y. Chang. Pfp: Parallel fp-growth for query recommendation. In *ACM Recommendation Systems, Lausanne,*, 2008.

[10] X. Li, C. G. Snoek, and M. Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of MIR '08*, 2008.

[11] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma. Dual cross-media relevance model for image annotation. In *Multimedia'07*, 2007.

[12] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning, 2008.

[13] M. Naaman and R. Nair. Zonetag's collaborative tag suggestions: What is this person doing in my phone? In *IEEE Multimedia,*, 2008.

[14] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proceedings of ACM Multimedia'07*, 2007.

[15] Y. Qi, K. S. Candan, J. Tatemura, S. Chen, and F. Liao. Supporting olap operations over imperfectly integrated taxonomies. In *SIGMOD'08 Conference*, 2008.

[16] X. Rui, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Bipartite graph reinforcement model for web image annotation. In *Multimedia'07*, 2007.

[17] S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, and J. Riedl. tagging, communities, vocabulary, evolution. In *CSCW '06*, 2006.

[18] B. Sigurbjörnsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *WWW '08*, 2008.

[19] C. G. M. Snoek, B. Huurnink, L. Hollink, M. D. Rijke, G. Schreiber, and M. Worring. Adding semantics to detectors for video retrieval. *IEEE Transactions on Multimedia*, 9, 2007.

[20] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang. Content-based image annotation refinement. *Proceedings of CVPR '07*, 2007.

[21] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. *Proceedings of ACM Multimedia'08*, 2008.

[22] L. Wu, M. Li, Z. Li, W.-Y. Ma, and N. Yu. Visual language modeling for image classification. *Proceedings of MIR'07*, 2007.

[23] R. Yan and A. Hauprmann. Query expansion using probabilistic local feedback with application to multimedia retrieval. In *Proceedings of CIKM '07*, 2007.

[24] J. Yu, J. Amores, N. Sebe, P. Radeva, and Q. Tian. Distance learning for similarity estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):451–462, 2008.

[25] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian. Visual synset: towards a higher-level visual representation. In *Proceedings of CVPR'08*, 2008.