World Wide Web (2011) 14:133–156
DOI 10.1007/s11280-010-0099-8

Mining multi-tag association for image tagging

Yang Yang · Zi Huang · Heng Tao Shen · Xiaofang Zhou

Received: 29 July 2010 / Revised: 17 October 2010 /
Accepted: 26 October 2010 / Published online: 22 December 2010

Abstract Automatic media tagging plays a critical role in modern tag-based media retrieval systems. Existing tagging schemes mostly perform tag assignment based on community contributed media resources, where the tags are provided by users interactively. However, such social resources usually contain dirty and incomplete tags, which severely limit the performance of these tagging methods. In this paper, we propose a novel automatic image tagging method aiming to automatically discover more complete tags associated with information importance for test images. Given an image dataset, all the near-duplicate clusters are discovered. For each near-duplicate cluster, all the tags occurring in the cluster form the cluster's "document". Given a test image, we firstly initialize the candidate tag set from its near-duplicate cluster's document. The candidate tag set is then expanded by considering the implicit multi-tag associations mined from all the clusters' documents, where each cluster's document is regarded as a transaction. To further reduce noisy tags, a visual relevance score is also computed for each candidate tag to the test image based on a new tag model. Tags with very low scores can be removed from the final tag set. Extensive experiments conducted on a real-world web image dataset—NUS-WIDE, demonstrate the promising effectiveness of our approach.

Keywords image tagging · tag completion · tag denoising ·
weighted association rule mining

Y. Yang · Z. Huang · H. T. Shen (✉) · X. Zhou
School of Information Technology & Electrical Engineering, The University of Queensland,
Brisbane, Australia
e-mail: shenht@itee.uq.edu.au

Y. Yang
e-mail: yang.yang@itee.uq.edu.au

Z. Huang
e-mail: huang@itee.uq.edu.au

X. Zhou
e-mail: zxf@itee.uq.edu.au

1 Introduction

With the rapid development of Internet and Web 2.0 technology, a large number of community contributed multimedia contents have been produced and shared on the Web. Quite a few representative Web 2.0 websites, such as Flickr,[1] YouTube,[2] etc., not only provide users interfaces of image or video sharing, but also allow users to collaboratively describe the resources with their own tags (or annotations) through social tagging services. This kind of user annotated multimedia contents is so called community contributed dataset, based on which, a bottom-up and self-organized classification system is formed, namely *folksonomy*.[3]

From the perspective of critical web applications such as keyword-based image search engines, image tags are indispensable for image indexing and retrieval. Currently, the performance of image search engines mainly relies on the quality of image tags, which are generated automatically or manually. Duo to the well-known semantic gap between low-level features and high-level semantics, annotations or tags are regarded as a natural bridge for narrowing the gap between text-based query and visual features of multimedia objects. Automatic keywords assignment is implemented by analyzing the multimedia content or the surrounding text on the web pages [23, 29, 33]. Manually tagging is often performed by experienced experts based on a predefined ontology. It is much more accurate than automatic annotation methods. However it is highly labor- and time-consuming. In recent decades, image annotation [19, 20, 22, 30, 44] has been attracting significant research attention in multimedia and computer vision area. It is usually formulated as a classification problem over a predefined concept set and a well-established training data set. Although these approaches enjoy relatively high performance in terms of *precision* and *recall*, they usually suffer from the lack of training set and the limited descriptive ability of the “small” concept set. Moreover they are hard to be extended to general cases due to the model-driven property. Most recently the emergence of manually social tagging [8, 10, 25, 36] provides a good opportunity to collect image tags from users, yet it suffers from several intrinsic problems. The first problem is tag ambiguity [21, 34], which means that one tag may have different meanings. For instance, tag “apple” can either refer to a kind of fruit or a computer brand. In reality, it is difficult for users to be conscious of the existence of tag ambiguity if they do not know the other senses of the ambiguous tags. Tag noise [7, 28] is another severe problem existing in community contributed image datasets. Existing studies reveal that some tags provided by Flickr users are inaccurate and only about 50% tags truly reflect the content of the images [12]. Moreover, the original tags associated with the images in the dataset are expectedly incomplete [7, 15] due to the knowledge and terminology limit of users. Based on such noisy and incomplete user provided tags, existing tagging methods can hardly acquire satisfying results.

[1] http://www.flickr.com/

[2] http://www.youtube.com/

[3] http://en.wikipedia.org/wiki/Folksonomy

Our work presented in this paper aims to overcome the above problems for improving the performance of image tagging. Figure 1 depicts the overall framework which consists of an offline process and an online tagging process. In the offline process, three major components are designed: (1) Near-duplicate image discovery is applied to find near-duplicate clusters from the image dataset such that incomplete tag set of an image is compensated by the tags of its near-duplicate images; (2) Weighted multi-tag association mining is proposed to discover multi-tag correlations from the derived clusters' documents. We mine multi-tag correlations in a weighted scenario because traditional association rule mining only considers counts as support, which may lead to the loss of important tag correlations; (3) Tag modeling are developed to find multiple latent semantic meanings of each individual tag represented by multiple image groups. The purpose of proposing this component is to clarify the ambiguity of each tag. In the online process, given a test image, we firstly initialize its tag set as its near-duplicate cluster's document (Step 1). We apply near-duplicate cluster to perform tag initialization because we believe near-duplicate images should contain same semantic meanings [45]. The initial tag set is then expanded by using the weighted association rules (Step 2) since we cannot expect all potentially meaningful tags are included in the first place. In the last step, because noisy tags are inevitably introduced in the above two steps we further perform tag denoising by computing the visual relevance between the test image and the tag's multiple image groups (Step 3).

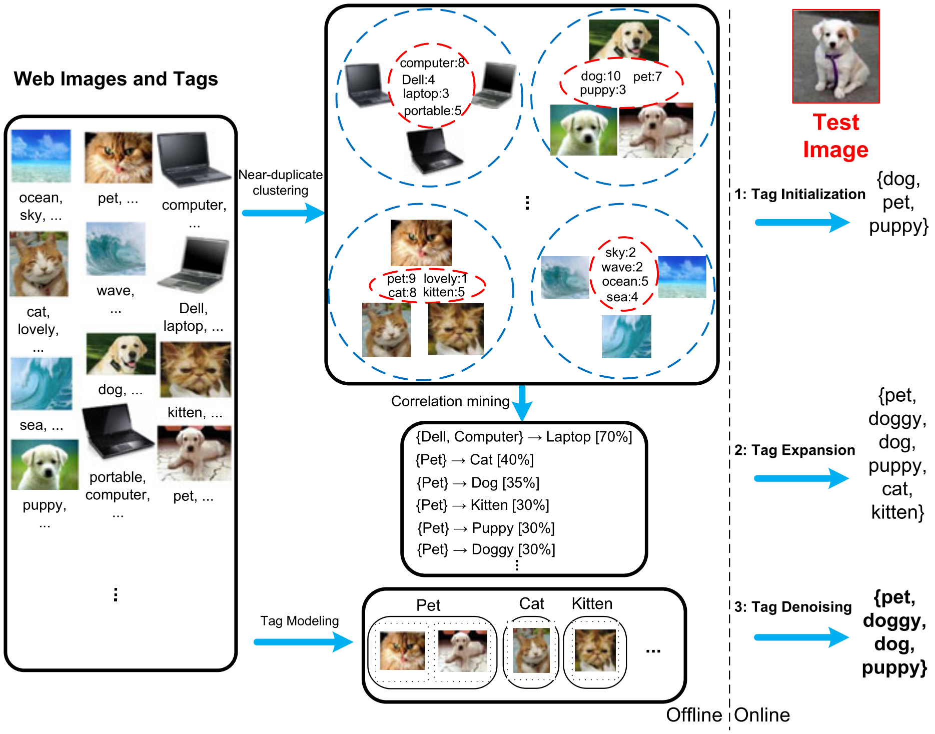


Figure 1 The overall tagging framework.

More details will be revealed shortly. Our main contributions are summarized as below:

1. Tag incompletion problem in community contributed image datasets always limits the performance of existing tagging methods. Instead of dealing with individual images, we group near-duplicate images together and aggregate their individual tags to form a document for each image cluster.
2. Compared with existing methods using pairwise-tag correlation, we introduce a new concept of multi-tag association to discover the correlation among multiple tags, based on which new tags can be further expanded from the candidate tag set. By considering each cluster's document as a transaction and a tag as an item, multi-tag associations can be discovered by the weighted association rule mining model (WARM). The *weighted support* and *weighted confidence* measures are taken into account for measuring tag information importance with respect to the test image.
3. To further reduce noisy tags introduced during the tag initialization and expansion steps, we also introduce a new tag model. Given a tag, all the images which contain the tag are partitioned into groups using an effective clustering method to elicit its multiple latent semantic meanings if exist. A visual tag relevance is then defined to indicate the maximal visual similarity between the test image and the tag's latent semantic groups. Tags with low scores are expected to be irrelevant to the test image and thus can be removed from the results.
4. We conduct extensive experiments to confirm the effectiveness of our proposal by comparing with existing methods.

The rest of this paper is organized as follows. Related work will be reviewed in Section 2, followed by the detailed discussions on multi-tag correlation mining and tag aggregation in Section 3. Section 4 introduces three components of the online aumatic tagging framework, including tag initialization, tag expansion and tag denoising. Experimental results are reported in Section 5, followed by conclusions in Section 6.

2 Related work

In this paper, we aim to address the problem of multimedia tagging by exploiting knowledge mined from community contributed multimedia repository. In addition to reviewing research work related to multimedia tagging, we also focus on an extremely relevant topic, namely multimedia annotation, which has been attracting significant attention in multimedia and computer vision area. The basic difference between annotation and tagging is that most annotation methods are developed based on machine learning models (model-driven) and the words are limited in a small lexicon while tagging is usually built based on data-driven approaches and the words used in tagging are theoretically arbitrary. In this section we discuss state-of-the-art research in tagging and annotation areas respectively.

2.1 Image & video annotation

In the most recent years, multimedia annotation problem has been attracting significant research attention in multimedia and computer vision area. Duo to the well-known semantic gap between low-level features and high-level semantics, annotations are regarded as a natural bridge for narrowing the gap between text-based query and visual features of multimedia objects. Typically, multimedia annotation can be formulated as a classification problem in which semantic concepts are used as class labels. Hence, supervised learning techniques are suitable for building connection between annotations and low-level visual features in such scenarios. Amir et al. [3] simultaneously exploited several machine learning approaches, including Support Vector Machine, Gaussian Mixture Model, Maximum Entropy learning, modified nearest-neighbor classifier, and Multiple Instance Learning to model concepts. They also fused across various features and approaches to boost concepts modeling performance. In [22], Qi et al. formulated video annotation as a multi-label classification problem and proposed a Correlative Multi-Label framework to simultaneously model multi-concepts and take into account semantic correlations between concepts. They also illustrate that their model can be intuitively interpreted by Gibbs Random Field. By extending this idea, Hua et al. [11] further proposed a novel multi-label active learning framework which adopts an online learner instead of SVM-like classifier to deal with large-scale data. Wu et al. [35] focused on learning an optimal similarity metric by exploiting side information associated with social media, such as surrounding text and existing tags. After that, they used nearest-neighbors classifier to identify tags for test images. Similarly, Mei et al. [19] also defined and learned a semantic distance function to measure similarity containing more sematic information. Some researches focus on exploring more effective data representation rather than novel machine learning models. Liu et al. [18] introduced a tensor framework in which video samples are represented by three modalities, namely image, audio and text. Then, a generalized SVM—STM is used to classify the samples. Finally, an active strategy is added to refine the STM classifier. Cao et al. [5] proposed a Logistic Canonical Correlation Regression which first discovered canonical correlations between heterogeneous features and existing annotations, then exploited logistic regression to create more enhanced annotations for web images. Similarly, in [4] the authors also utilized CCA to fuse different types of visual features to generate a more descriptive feature for subsequent annotation task.

Most recently, sparse coding becomes extremely popular in computer vision research area and has been applied to annotation task to some extent. Zhang et al. [44] focused on investigating properties of features and exploited a regularization based feature selection algorithm to leverage both sparsity and clustering properties of features. The selected features were then incorporate into nearest-neighbor classification to predict annotations for test images. Han et al. [9] also proposed to use a structural group sparsity for feature selection and boost annotation performance by exploiting correlations among multiple tags. Liu et al. [17] proposed a bi-layer sparse coding for encoding regions and propagating labels at region level, and showed state-of-the-art performance in region-level image annotation. In their work, images in dataset are first over-segmented into basic patches followed by grouping spatially coherent patches into candidate regions which can be treated as potential semantic regions. Then the proposed bi-layer sparse coding which guarantees both image- and patch-level sparsity was applied for reconstructing candidate regions from segmented basic patches. The common tags of images containing target region and selected patches will be re-assigned to the region according to reconstruction coefficients. However, basic patches in the dictionary are implicitly assumed to be independent with each other in their work. Wang et al. [30] proposed to use sparse coding twice for image annotation. First, they applied sparse coding to reconstruct images in label space, which amounts to establish semantic relatedness between images. The label reconstruction coefficients were further used to perform dimensionality reduction over the feature representation derived from Gaussian Mixture Model. Then, sparse coding was used again to reconstruction images in the reduced feature space and coefficients were used to propagate labels from training images to query image.

Due to the scarcity of pre-labeled data, many researchers have turned to take advantage of semi-supervised learning which can simultaneously learn from labeled and unlabeled data. In [27], Tang et al. extended a semi-supervised learning method called linear neighborhood propagation to a non-linear kernel-mapped space and used the optimal propagation coefficients to reconstruct the annotations. Yuan et al. [42] adopted a graph-based semi-supervised method named manifold-ranking to conduct video annotation task.

It can be observed that learning-based annotation approaches usually suffer from lack of training data and a pre-defined annotation set. As a result, they are hardly extended to large-scale data set and the annotated labels has limited descriptive and indexing ability.

2.2 Collaborative multimedia tagging

Confronted with huge amount of emergent web media and tags, traditional machine learning based methods are mostly unapplicable. In such case, we have to seek for new tagging schemas. Automatic tagging is to automatically annotate an object with descriptive tags by mining knowledge from web media and their associated context, such as surrounding text, existing tags, etc. The existing tagging methods mostly focus on mining tag-to-tag relationships, object-to-object relationships and tag-to-object relationships from social media and its associated contextual information.

Wang et al. [29] first collected candidate tags (terms) from textual information (e.g. captions and surrounding text) by exploiting *tf-idf* weight, and then random walk with restart was use to re-rank these candidate tags based on visual similarity of images, finally only top candidates were selected as final tags. Moxley et al. [20] first searched visually similar videos based on multiple modalities and then proposed a graph reinforcement mining approach to filter out meaningful tags for test video. In [24], Siersdorfer et al. revealed the relationships between videos from the perspective of content redundancy, existing near-duplicate detection techniques were applied to identify redundant videos. They further proposed neighbor-based and context-based tag propagation strategies for assigning tags to test videos. Similarly in [14], Li et al. proposed a neighbor voting algorithm which can establish tag relevance with respect to images by accumulating votes from their visually similar neighbors. In [33], Wang et al. also used neighbor-based method. They first search visually and semantically similar images, then mined searching results by an SRC clustering model to identify latent terms which can be treated as the final tags. Compared with our approach, most of these method does not make use of tag corpus knowledge to mine tag-to-tag correlations. Most recently Liu et al. [15] proposed a novel retagging scheme which is similar to our framework. But the difference between our work and their approach is that we propose a more comprehensive framework that is suitable to handle the situation that images are associated with no tags.

Tag recommendation is to recommend more tags for a sematic object based on the existing clues, including tags, surrounding text and visual content information. Tag recommendation is similar the research topic of query expansion [43]. They are two ways of improving information retrieval but there are still differences between them in that tag recommendation aims to describe the content resources in a more complete and precise way while query expansion tries to boost retrieval performance from the perspective of clarifying query's ambiguity. Sigurbjörnsson et al. [25] first characterized users tagging behaviors in Flickr and then presented different tag recommendation strategies which are mainly based on tag co-occurrence statistics information. Wu et al. [36] proposed a multi-modality recommendation approach based on not only tag co-occurrence but also visual correlation among tags. A Rankboost algorithm was adopted to fuse different modalities into an optimal integration feature. In [13], Krestel et al. introduced an approach based on Latent Dirichlet Allocation which uses tagged resources to elicit latent topics. Based on these topics, other tags within the same topic can be recommended for the new resource. Xu et al. [37] utilized collaborative tagging information to recommend tags. Their recommendation algorithm aggregates tags from similar textual content and prefers tags used by a large number of people on the target document. Ames et al. [2] built a system called ZoneTag to make it easier for mobile-phone users to tag their photos based on geographical information and existing tags. However, these methods mainly focused on finding semantically similar tags based on tag co-occurrence, which only took into account the relationship between two individual tags and may lead to tag semantics loss. Tag ranking aims to rank the tags associated with a given semantic object. The key problem is how to evaluate tag relevance with respect to the object. The approach in [16] first estimated initial relevance scores for the tags based on probability density estimation, and then performed a random walk over a tag similarity graph to refine the relevance scores. In [21], the tag translation task is formulated as a network comparison in order to handle the disambiguation issue. Each tag and its translation candidates are represented as networks of co-occurring tags. Then the tag similarity can be obtained by computing network similarity. Yin et al. [41] exploited social tags as a bridge to connect web objects. An efficient algorithm was proposed to enrich the semantics of the objects and to infer the labels for unlabeled objects.

Compared with existing annotation methods, our work mainly focuses on exploiting data mining to infer potentially unlimited tags from large-scale social media and the associated contextual information rather than learning models for mapping annotations and low-level visual features from training set. Tag recommendation is able to complete tags for individual images, but our tag completion mainly focuses on how to preliminarily explore more hidden correlations from near-duplicate clusters and then mine them out. Therefore the basic purpose of our tag completion aims to provide more comprehensive correlations among tags for more accurate image tagging or tag recommendation. Most existing tagging schemas usually ignore tag semantic incompletion issue of community contributed social media data, which may omit or underestimate certain tag correlations.

3 Multi-tag correlation mining

In this section we aim to handle the problem of tag incompletion in existing web image datasets, and further propose to employ a weighted association rule mining algorithm to discover multi-tag correlations.

3.1 Tag aggregation

Though user-generated-content websites provide a great opportunity to easily collect enormous images with user-provided tags, unfortunately there many noisy tags exist while some necessary and meaningful tags are missing. It is a very common phenomenon that users always depict images with simple and insufficient tags. For example, if a user has tagged an image with word "iphone", hardly will he/she tag this image with extra words like "mobile", "handphone", or "apple", because from the personal perspective, it is not necessary to do so. In this case, those essential but implicit tags can not be embodied completely, such as synonymous relationship (*ocean-sea*), hierarchical relationship (*computer-dell*), etc. This kind of tag incompletion issue exists in most of the community contributed image datasets.

In this subsection we intend to address the problem of tag incompletion problem by grouping the tags of near-duplicate images. It is well understood that near-duplicate images should carry same semantic meanings [45]. Take Flickr as an example. It contains hundreds of millions tagged images. Obviously, it is possible that near-duplicate images are annotated with different tags. We next will describe the way to analyze the annotated image dataset for tag aggregation.

Given an image dataset $\mathcal{X} = \{x_1, x_2, \cdots, x_N\}$, by considering each image as a node, a near-duplicate graph can be built, where the weight on each edge connecting two nodes implies the visual similarity of the corresponding pair of near-duplicate images. Graph mining algorithms [32] can be further applied to identify cohesive subgraphs which are regarded as groups of near-duplicate images denoted as $\hat{\mathcal{X}} = \{X_1, X_2, \cdots, X_M\}$ $(M \ll N)$. For each $X_i \in \hat{\mathcal{X}}$, comprised of a group of near-duplicate images, the aggregation of tags associated with images in it forms a "document" of cluster X_i:

$$D_i = \{t_{i_1} : n_{i_1}, t_{i_2} : n_{i_2}, \cdots, t_{i_l} : n_{i_l}\} \tag{1}$$

where $t_{i_j}(j = 1...l)$ denotes the tag that occurs in cluster X_i and n_{i_j} is the frequency of tag t_{i_j} (i.e., the number of times t_{i_j} occurs).

After this preprocessing step we compensate tag incompletion in original image dataset to some extent. We believe more meaningful tag correlations are explicitly built in such way. In follow work, instead of X the new dataset $\hat{X}$ is used as the basis of the tag correlations mining and tag initialization.

3.2 Mining tag correlation

Tag co-occurrence is the key of measuring tag correlation. In a sizable annotated image dataset, two or more tags appearing together frequently can be considered as being highly relevant to each other. Different relationships can be derived, such as hierarchical relationship (*fruit-pear*), inclusion relationship (*car-tyre*) and some other relationships (*ocean-boat, sky-grass*, etc.). Most recently, tag co-occurrence based tagging methods [25, 36] have been well investigated. However, to the best of our knowledge, they only take into account pairwise tag correlation, which is defined as:

$$rel(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i|} \tag{2}$$

where t_i and t_j are any two tags in the corpus. $|t_i|$ indicates the number of images that are associated with tag t_i, and $|t_i \cap t_j|$ means the number of images that are associated with both tag t_i and t_j.

In reality, many strong correlations exist on a set of tags. For instance, tag *mac* usually has a stronger correlation with tag set {*computer,apple*} than with tag *computer* or tag *apple* only. In other words, given an image with tag set {*computer,apple*}, tag *mac* should be assigned to the image with a higher confidence. However, if we only consider pairwise correlation, neither *computer* ⇒ *mac* nor *apple* ⇒ *mac* might be confident enough to support the assignment of tag *mac* to the image, thereby underestimating the tag relevance score. In order to overcome the above drawback, we adopt weighted association rule mining model (WARM) to generalize tag co-occurrence method and explore multi-tag correlations from the derived near-duplicate image clusters $\hat{\mathcal{X}}$.

By considering a near-duplicate cluster as a transaction and its associated tags as the items in the transaction, it is very natural to discover tag correlations by applying association rules mining model [1], whose effectiveness has been proved in recommending tags for social bookmarking system [10]. However, the conventional association rule mining methods treat items and transactions with equal weight, yet ignore their individual information importance. Thus, under a certain minimum support requirement, some infrequent itemsets with significant information importance will be filtered out during the frequent itemset mining stage, thereby leading to the loss of some meaningful tag correlations. Therefore, it is vital to assign different weights to different transactions and items so as to reflect their different importance in supporting itemset. When there are not many supportive transactions for an itemset, the itemset should also be considered as a candidate for deriving useful association rules if the items contained in it or the transactions containing it are quite important. To this end, we intend to take into account tags information importance to enhance the conventional association rule mining model and promote the potential important itemsets with small support.

Given the tag corpus $\mathcal{C} = \{t_1, t_2, \cdots, t_n\}$, we initialize a weight w_i for each tag t_i, with $0 \leq w_i \leq 1$, where $i = 1, 2, \cdots, n$. Here we estimate w_i using normalized information importance:

$$w_i = \frac{h(t_i)}{\sum_{t_j \in \mathcal{C}} h(t_j)} \tag{3}$$

where $h(t_i)$ is the Shannon Information for tag t_i. Accordingly, the weight of a transaction D (i.e., the "document" of an image cluster X) is calculated as:

$$w_D = \sum_{t_i \in D} tn(t_i, D) \times w_i \tag{4}$$

where $tn(t_i, D)$ denotes the frequency of tag t_i occurring in D. In terms of aggregation of transaction weights, we define *weighted support* of an itemset A as follows:

$$w\ support(A) = \frac{\sum_{D:A\subset D\wedge D\in\hat{\mathcal{X}}} w_D}{\sum_{D\in\hat{\mathcal{X}}} w_D} \tag{5}$$

Accordingly, the *weighted support* and *weighted confidence* of an association rule $A \Rightarrow B$ can be obtained by:

$$w\ support(A \Rightarrow B) = w\ support(A \cup B) \tag{6}$$

$$w\ confidence(A \Rightarrow B) = \frac{w\ support(A \cup B)}{w\ support(A)} \tag{7}$$

Although the new weighted measures does not satisfy the downward closure property of conventional association rule mining model, as proved in [26], they follow a weighted downward closure property: when an itemset satisfies a pre-defined minimum *weighted support* threshold, all its subsets satisfy this minimum *weighted support* threshold as well. With this property, a modified Apriori algorithm can be applied to mine the weighted frequent tag sets (multi-tag correlations), as illustrated in Algorithm 1.

Algorithm 1: Weighted Apriori for finding multi-tag association.

```
Input  : $\hat{\mathcal{X}}$, near-duplicate clusters;
         $\mathcal{C}$, tag corpus;
         $min\_wsupp$, minimum weighted support.
Output: $L$, frequent itemsets in $\hat{\mathcal{X}}$
$L_1 = \{t_i | t_i \in \mathcal{C} \wedge w_i \geq min\_wsupp\}$;
for $k = 2; L_{k-1} \neq \emptyset; k++$ do
    $C_k =$ apriori_gen$(L_{k-1})$;
    foreach cluster document $D$ in $\hat{\mathcal{X}}$ do
        $C_D = subset(C_k, t)$;
        foreach candidate itemset $c \in C_D$ do
            $c.wsupport+ = w_D$;
        end
        $W+ = w_D$;
    end
    $L_k = \{c \in C_k | \frac{c.wsupport}{W} \geq min\_wsupp\}$;
end
return $L = \cup_k L_k$;
```

As we can see from this algorithm the support of tag sets are measured by the accumulated weights (information content) rather than counts. Thus, we are able to preserve those infrequent tag sets carrying rich information. More potentially meaningful tag correlations can be elicited from these tag sets.

4 Online automatic tagging

Based on near-duplicate image grouping results and the mined multi-tag correlations we are able to design our online automatic image tagging scheme. Three steps are developed in this online process, i.e. tag initialization, tag expansion and tag denoising.

4.1 Tag initialization

In this section, we introduce the first step for the online process of our automatic image tagging framework, i.e., tag initialization. Given a test image without any contextual information, the only clue that we can use to obtain the initialized tags is its own visual content.

Starting from this we employ the results of near-duplicate image groups to collect initial tags for test images. Given a test image x, we first find the near-duplicate cluster X_i that it belongs to, and then initialize the candidate tags as D_i. If x does not have any near-duplicates, the cluster which contains the most similar image to x is used. On the other hand, if x has near-duplicates from multiple clusters, the cluster which contains the most number of near-duplicates is chosen.

4.1.1 Corpus tag relevance

Nevertheless, the target of our automatic tagging framework is not only to find relevant tags for images, but also to provide the importance of each tag to the test image. Apparently this is of high significance in improving the searching and indexing effectiveness of image retrieval engines. In this subsection, an information theory based approach is developed to quantify a tag's information importance.

Intuitively, more common a tag is, the less information it contains. Based on the tag corpus knowledge, we can use the information volume of a tag t to quantify the degree of tag importance. Given an initial candidate tag set S for a test image x, first we can treat it as a bag-of-words document, and for each tag t, its initial corpus tag relevance w.r.t. image x is defined as follows:

$$r_c(t, x) = tn(t, S) \times h(t) \tag{8}$$

where $tn(t, S)$ is the number of times that tag t appears in image x's initial candidate set S. $h(t)$ denotes the quantity of tag t's information importance, which can be estimated using the *Shannon Information*:

$$h(t) = -\log p(t) \tag{9}$$

where $p(t)$ is the proportion of images associated with t in the whole image dataset. Denote $|t|$ as the number of images associated with t, N as the number of images in the dataset, we estimate $p(t)$ as follows:

$$p(t) = \frac{|t|}{N} \tag{10}$$

Finally, $r_c(t, x)$ is normalized by the sum of relevance scores of all tags that appear in x's candidate tag set

$$r_c(t, x) = \frac{tn(t, x) \times h(t)}{\sum_{t' \in S} tn(t', x) \times h(t')} \tag{11}$$

In fact, the underlying intuition of this estimation method is widely used in information retrieval and data mining, such as the idea of *tf-idf* weight. The relevance increases proportionally to the number of times a word appears in the document but is counteracted by the information volume of the word in the corpus. Also, this estimation method can be explained by a voting scheme. The candidate tag set is actually the aggregation of the tags of near-duplicates of the test image, and the tag occurrence number denotes the voting number. Thus, the corpus tag relevance score can be treated as the collective contribution of the image's near-duplicates Table 1.

4.2 Tag expansion

As mentioned, it is possible that the initial candidate tags are not comprehensive enough for describing the testing image. Tag expansion (Step 2 in the online process of our framework) plays an essential role to enrich the tag set. Given an image associated with a set of tags, we assume that a new tag which has strong correlation with the existing tags should be assigned to the given image with a high probability. In this section, we focus on how to exploit the mined multi-tag correlations to find more potentially meaningful tags for forming a more descriptive tag set for test images.

Starting from the weighted association rules derived from the annotated image dataset, we design a novel approach to find the maximal relevance of a tag t^* w.r.t. to the existing candidate tag set S of the image x, where $t^* \notin S$.

Definition 1 (support subset) Given a tag t^* and the weighted association rules set $R = \{r_1, r_2, \cdots, r_p\}$, if association rule $S' \Rightarrow t^*$ exists in R and $S' \subseteq S$, then S' is a *support subset* of S with respect to t^*.

It is possible that both S' and S'' are support subsets of S with respect to t^* where $S'' \subset S'$. In this case, only $S' \Rightarrow t^*$ will be considered for deriving new tags. In other words, *weighted confidence* of the association rule $S' \Rightarrow t^*$ is used for deriving tag t^*. We assume a "larger" subset (e.g., S') can reflect the correlation between S and t^* more precisely. Therefore, we introduce a new concept of *maximal support subset.*

Definition 2 (maximal support subset) A support subset S' of S w.r.t. t^* is a maximal support subset if there exist no proper superset S'' such that $S'' \subseteq S$, and rule $S'' \Rightarrow t^*$ exists in R.

We believe that the maximal support subsets reflect the genuine underlying relation between S and t^*. Meanwhile, considering the maximal support subsets only essentially reduces the computational cost by ignoring the redundant tags derived from their subsets. Recall the example in Section 3.2. If rules *computer* $\Rightarrow$ *mac*, *apple* $\Rightarrow$ *mac* and *computer, apple* $\Rightarrow$ *mac* all exist in R, {*computer*}, {*apple*} and {*computer, apple*} are the support subsets of S w.r.t. *mac*. If we consider all of

Table 1 Main notations.

Notation	Description
$\mathcal{X}$	a web image dataset
$\mathcal{C}$	the tag corpus contained in $\mathcal{X}$
$\hat{\mathcal{X}}$	the set of near-duplicate image clusters
S	the candidate tag set
R	the set of weighted association rules mined from $\hat{\mathcal{X}}$
$r_c(t, x)$	corpus tag relevance of tag t w.r.t. image x estimated by *Shannon information*
$r_v(t, x)$	visual tag relevance of tag t w.r.t. image x estimated by tag models

them, tag *mac* would be expanded redundantly. Thus, we only consider the maximal support subset {*computer*, *apple*} to generate tag *mac* once.

Definition 3 (maximal division) Given a set of subsets of S, denoted as $d = \{S_1, S_2, \ldots, S_m\}$, if $S_i \cap S_j = \emptyset (1 \leq i < j \leq m)$ and $\bigcup_i S_i = S$, then d is called a division of S. Further, if either S_i is a maximal support subset of S w.r.t. t^*, or any subset of S_i' is not a support subset, then d is a maximal division of S w.r.t. t^*.

We may merge the S_i of which any subset is not a support subset, and denote it as S_m. Thus, we define the relevance of t^* w.r.t. image x as follows:

$$r_c(t^*, x|d) = \sum_{i=1}^{m-1} \text{confidence}(S_i \Rightarrow t^*) \times r_c(S_i, x) \tag{12}$$

where S is the existing tag set of x, confidence($S_i \Rightarrow t^*$) is the confidence score of association rule $S_i \Rightarrow t^*$, and $r_c(S_i, x)$ is the relevance score of tag set S_i and x:

$$r_c(S_i, x) = \sum_{t \in S_i} r_c(t, x) \tag{13}$$

Then, the optimal relevance score of the extended tag t^* can be defined as:

$$r_c(t^*, x) = \max_{d \in \mathcal{D}} r_c(t^*, x|d) \tag{14}$$

where $\mathcal{D}$ is the set of all possible maximal divisions w.r.t. t^*.

So far, we have discussed multi-tag correlation to generate a more complete set of candidate tags for a testing image. This candidate set will be further refined in the tag denoising step which will be discussed in details in the next section.

4.3 Tag denoising

Note that in both tag initialization step and tag expansion step, noisy tags are inevitably brought in due to the inherent dirtiness of dataset and the existence of tag ambiguities. On one hand, in tag initialization step, existing noisy tags in near-duplicate clusters will be easily propagated to the test image. One another hand, in tag expansion step, more irrelevant tags can be derived from both noisy tags and ambiguous tags introduced in the first step. A quantitative analysis of noisy tags has been conducted in [7]. As noisy tags can severely affect the performance of tag-based search engine, it is essential to further refine the candidate tag set before using them to index images.

To handle noisy tags, we intend to build the connection between tags and images. Due to the tag ambiguity issue, an individual tag often has more than one meanings, which can be regarded as the underlying latent semantics of the tag (Figure 2). Starting from this, we first model the latent semantics for each candidate tag as multiple image groups based on content closeness, then measure the relevance between a candidate tag and a test image by finding the most relevant latent semantic with the test image.

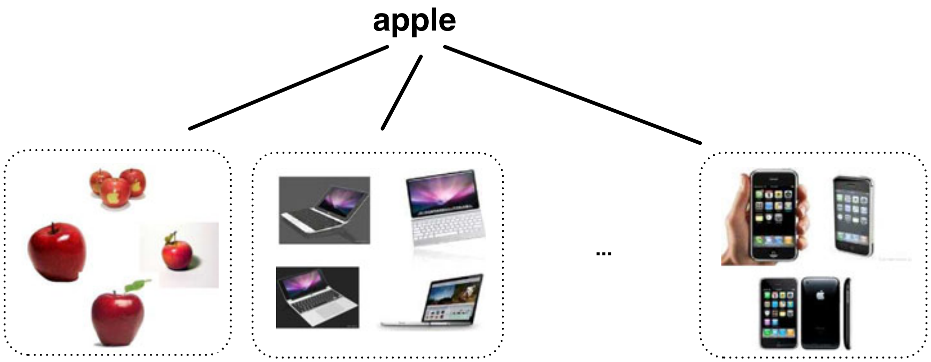


Figure 2 An example of multiple latent semantics.

4.3.1 Eliciting tag latent semantics

Given an image x and a candidate tag t, we first collect all the images associated with t. These images are treated as the representative samples of t in real world. If the latent semantics of t are mapped onto these representative images, then they will be spontaneously divided into different groups. Meanwhile, building the latent semantics for tags can also help isolate the images that are incorrectly tagged with t. To elicit the underlying latent semantics of tags more explicitly, we partition these images using a clustering process method. A lot of clustering models could be applied here, such as k-means and the image clustering proposed in [39], etc. Most recently, Wang et al. [31] proposed a novel Integrate KL (K-means—Laplacian) Clustering approach which coherently combines K-means clustering and normalized cut spectral clustering. The main advantage of this model is that it can take advantage of multiple data sources for harvesting better results. It has shown that simultaneously incorporate textual and visual features helps narrow the so-called semantic gap [38, 40] In this new tag modeling process, we assume each group represents a specific latent semantic meaning of tag t. To get closer to the underlying semantics, both visual and textual features are used to cluster the representative images.

4.3.2 Visual tag relevance

Given a tag t and a test image x, the visual relevance between them is defined as follows:

$$r_v(t, x) = \max_{g \in G_t} sim(g, x) \tag{15}$$

where G_t is the set of the latent semantic groups for tag t, and $sim(g, x)$ is the similarity between group g and x which is defined as:

$$sim(g, x) = \exp(-\frac{1}{|g|} \sum_{y \in g} \frac{\|x - y\|^2}{\sigma^2}) \tag{16}$$

The reason of choosing the most similar latent semantic cluster is that we assume usually an image only contains one sense of a specific tag. Recall Figure 1. After tag expansion, tags like *cat* and *kitten* are also included in the candidate set. Tag *pet* is modeled by two latent semantic groups while *cat* and *kitten* have one only. In tag denoising, *cat* and *kitten* will have very low visual relevance scores since both groups are not similar to the test image. However, *pet* will have a high visual relevance score since it has one very similar latent semantic group to the test image. As a result, although *cat* and *kitten* have strong correlation with *pet*, both can be removed from the final tag set due to their very low visual relevance scores.

For the purpose of tag ranking, the complementary nature of both corpus and visual tag relevance can be explored and linearly combined as:

$$r(t, x) = \beta \cdot r_c(t, x) + (1 - \beta) \cdot r_v(t, x) \tag{17}$$

where $\beta \in [0, 1]$. The combined tag relevance is treated as the final relevance score of tag t and the test image x.

5 Experiment

In this section, we test our proposed automatic tagging method on a real-world Flickr image dataset and present the evaluation results. We first introduce the characteristics of this dataset in details. Then, in order to compensate the lack of tagging ground truth in real-world datasets, we adopt a classification evaluation strategy rather than traditional *Precision* and *Recall* measures to illustrate the effectiveness of our tagging approach. Besides, we also invite users to directly judge the relevance of the tags generated by our tagging approach.

5.1 Dataset

5.1.1 Statistical information of images and tags

To evaluate our proposed approach, we conduct experiments on a real-world web image dataset—NUS-WIDE [7]. This dataset consists of 269,648 web images and the associated tags created by users from Flickr. The total number of the extracted unique tags is 425,059, and the power law distribution of tag frequency illustrates that most tail tags occur seldom in the dataset (usually less than 2 times), which could be caused by misspelling or used for specific name, etc. Although these tags contain more information content from the perspective of information theory, they are actually less helpful for general media tagging or indexing task. Hence, those tags appearing less than 100 times are first removed. Then, in order to further obtain more meaningful and refined tag list, those tags that does not appear in WordNet are further pruned as well, such as meaningless words 2008, 2009, and so on. At the end, a tag list of 5,018 unique tags are left for performing our tagging task. We denote the image set as $\mathcal{X}$ and the refined tag list as our tag corpus $\mathcal{C}$. Table 2 summarizes the basic information of NUS-WIDE dataset.

5.1.2 Low-level visual features

NUS-WIDE dataset provides five different types of global low-level features and one local visual feature to describe image content: 64-D Color Histogram in LAB color space, 144-D color correlogram in HSV in HSV color space, 73-D edge distance histogram, 128-D wavelet texture, 225-D block-wise LAB-based color moments

Table 2 Summary of NUS-WIDE dataset.

Feature	Description
Original tag set	425,059 unique tags
Refined tag set $\mathcal{C}$	5,018 unique tags
Image set $\mathcal{X}$	269,648 unique images
Color histogram	64-D
Color correlogram	144-D
Edge distance histogram	73-D
Color histogram	128-D
Wavelet texture	128-D
Bag of visual word	500-D

extracted over 5 × 5 fixed grid partitions, and 500-D bag of visual words based on SIFT descriptor. On one hand, global features provide a global view of image content in a high-dimensional feature space. It can capture different kinds of image content characteristics, such as object contour (e.g., shape feature), content distribution (e.g., color histogram), recurrent spatial layout (e.g., textual feature), etc. On the other hand, local features can describe more semantic content of image than global features. They capture the structural elements of the semantic content contained in the image by detecting a series of local interesting points or keypoints. Each interesting point is further described by a high-dimensional feature vector. Scale-invariant feature transformation (SIFT) and its variants such as PCA-SIFT and PSIFT are useful local descriptors for image near-duplicate detection and classification task. Usually, hundreds of or more local interest points can be detected from a single image. Hence the similarity computation could be very time-consuming. Thus, a practical way is to transfer local feature to a bag-of-visual-word representation by clustering interest points into "word bags". To balance the computational cost and semantic precision, both global and local features are together adopted in our paper.

5.2 Mining prior knowledge

In our proposed tagging method, we have to mine some essential prior knowledge from the dataset. We partition the whole NUS-WIDE dataset into two disjoint sets: a training set containing 200,000 randomly selected images and a test set comprised of the remaining images. The training set are first utilized to perform semantic aggregation task. For reducing the computational cost, k-means clustering is used to divide the training set into smaller subsets. For each subset, we run the near-duplicate clustering algorithm on a weighted image graph [32], where the weight on each edge implies the visual similarity of the corresponding image nodes.

There is a parameter—subgraph cohesion threshold γ which indicates how strongly the images are connected within a near-duplicate cluster. In fact, while the subgraph cohesion threshold γ increases (i.e., the subgraphs are more rigid), less images are included into a near-duplicate group and more near-duplicate groups can be produced, which means more aggregated tag transactions are available in the multi-tag correlation mining step. Hence it is easier to bring in false near-duplicate images. To control the false rate, we set γ with 0.8, 0.85, 0.9 and 0.95. Table 3 shows the effect of different γ values on the results of near-duplicate grouping and weighted association rules. For saving computational cost, we randomly select 100,000 images

Table 3 Effect of subgraph cohesion threshold γ on near-duplicate clusters and multi-tag correlation mining.

	ND clusters	Weighted association rules
$\gamma = 0.8$	95,961	1,460
$\gamma = 0.85$	97,437	993
$\gamma = 0.9$	98,528	656
$\gamma = 0.95$	99,469	499

from the training set to evaluate the effect of γ. We can see that as γ increases, the number of near-duplicate clusters becomes larger while that of weighted association rules decreases. In fact, when the subgraph cohesion constraint γ becomes larger, fewer near-duplicate images can be detected and fewer tags will be aggregated together. Obviously, the number of weighted association rules will significantly limit the effectiveness of our tag expansion. By default, we set γ as 0.8 to have more mined rules. Note that a very small γ value will potentially group many non-near-duplicate images into the same near-duplicate cluster.

Table 4 illustrates the comparison of weighted association rules mining (WARM) and conventional association rules mining (CARM) in terms of the number of rules. Apparently, WARM mines many more available rules than CARM, which can greatly benefit the tag expansion step. Figure 3 shows the statistical information about the number of tags expanded by the weighted association rules. As illustrated, most test images (76.83%) can be expanded with 4 to 10 additional tags which can provide more potentially meaningful semantics for the test images.

5.3 Classification evaluation

Currently, one of the main difficulties in automatic tagging evaluation is the lack of real tagging ground truth. Although social tags are manually provided by social users, they cannot be regarded as tagging grounding truth because of the semantic incompletion and noisy tag problems. This means that traditional information retrieval measures such as *Precision*, *Recall*, *F*1 score, etc., which mostly rely on ground truth are not applicable in such a scenario any more. Fortunately, NUS-WIDE dataset provides semi-manually annotation ground truth for 81 concepts which can be used to evaluate image annotation task. Therefore, we plan to design a classification-based experiment to present the effectiveness of our tagging method and the descriptive ability of the automatically generated tags.

5.3.1 Evaluation strategy

We compare our proposed method with different automatic tagging methods in terms of classification performance. Given the automatically generated tags, not only we can treat them as the descriptive keywords w.r.t. the image, but also can we apply bag-of-words model to formulate them as a high-dimensional textual feature vectors.

Table 4 The number of association rules mined by WARM and CARM.

min_w support	CARM	WARM
0.3	67	2,867
0.4	43	2,122
0.5	27	1,459

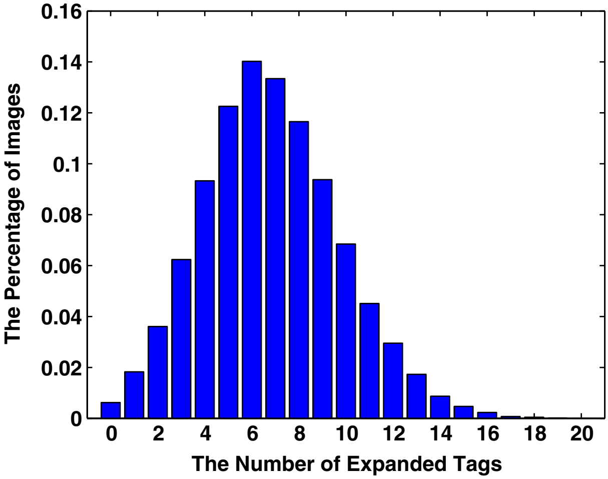


Figure 3 The number of expanded tags for the test images.

One reasonable criteria of evaluating the effectiveness of the tags is the classification performance by utilizing such tag feature. If the tag feature generated by a tagging method performs better than the other tagging method in terms of classification performance, we can say this tagging method can create more descriptive tags for images. In fact, many mature text classification techniques can be applied to conduct our evaluation task. In this paper, we adopt an implementation of Support Vector Machines (SVMs) [6] in our experiments as it can support sparse format input data and has shown its effectiveness in text classification task. The purpose of our classification experiments is to estimate the descriptive ability of the tags generated by our tagging method rather than the performance of an image annotation mode. Hence, we do not need to consider the distribution of the training dataset. The tag feature vectors are constructed by sorting the tag set according to the relevance computed by different tagging methods. In order to obtain reliable classification results, we select the concepts that are annotated at least 1,000 times. Table 5 illustrates the most frequent concepts in our test data set. For each ground truth concept, we randomly select a set of 500 images for binary classification training and a disjoint set of 500 images for classification testing from the tagged test images. As to the tag vector, we choose top K tags according to the tag relevance score as the tag feature, where $K = 100, 90, 80, 70$ and 60.

Table 5 The set of most frequent concepts.

Concept	Frequency	Concept	Frequency
Sky	4,812	Tree	1,725
Water	4,525	Animal	1,682
Clouds	3,533	Sun	1,613
Sunset	2,766	Ocean	1,553
Beach	2,116	Flowers	1,535
Reflection	1,902	Snow	1,432
Street	1,725	Lake	1,161

We compare our proposed automatic image tagging approach with three existing tagging methods:

1. Simple Neighbor-based Tagging (**SNTag**). SNTag method is proposed in [24]. The basic idea is to accumulate votes from image x's visually similar neighbors for tag t.
2. Tag Propagation (**TagRank**). TagRank is also proposed in [24], and it performs an iterative process over a neighbor image graph to compute the tag relevance.

$$rel(t, x) = \sum_{x' \in N_x} rel(t, x') \times sim(x, x') \tag{18}$$

3. Random Walk based Tagging (**RWTag**). Random Walk model has been widely used in tagging or annotation refinement task [16, 29]. We conduct random walk process over candidate tag graph of which edges are constructed based on symmetric tag co-occurrence:

$$rel(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \tag{19}$$

   The process can promote tags that have more co-occurring tags within the candidate tag set.

5.3.2 Parameter tuning

We first conduct experiments to analyze the effect of linear combination parameter β (17) on the classification accuracy. We test β from 0.0 to 1.0. Specifically, when $\beta = 0.0$ it indicates tag relevance totally depends on visual tag relevance. When $\beta = 1.0$, tag relevance is totally determined by corpus tag relevance. For each β value, we calculate combined tag relevances for all tags associated with test images and use top $K = 100, 90, 80, 70$ and 60 tags as tag features for classification. Average accuracy over 14 different concepts is used as the performance measure. The experimental results are shown in Figure 4. We can see that five average accuracy curves fluctuate in the range [0.0, 1.0]. Especially when $\beta = 0.6$, all accuracy curves reach their best performance, which is much higher than using either corpus tag relevance or visual tag relevance alone. When $\beta = 0.6$, it indicates that corpus tag relevance contributes slightly more than visual tag relevance in the final relevance aggregation. Meanwhile, a smaller K leads to a better accuracy due to less noises included in the results. Interestingly, the curves also have some local peaks (e.g., when $\beta = 0.2$), indicating the instability of β's effect. This experiment shows the importance of considering both visual tag relevance and corpus tag relevance in improving the tagging accuracy. However, the performance could be sensitive to different β values. By default, we set $\beta = 0.6$ since it achieves the best results.

5.3.3 Evaluation results

The criteria we use to evaluate the classification performance is the ratio of images correctly categorized, namely *accuracy*. Usually, accuracy is not a good metric in classification evaluation due to the imbalance of positive and negative samples in training set. Nevertheless, in our test we just use classification to illustrate the descriptive ability of our methods, and we randomly select even number of positive

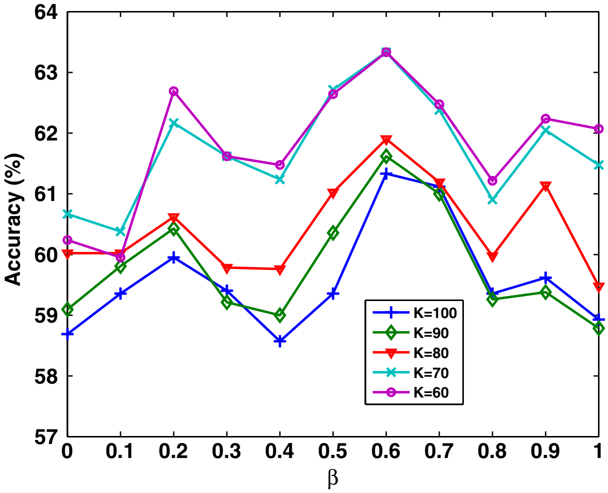


Figure 4 Effect of linear combination parameter β.

and negative samples. Therefore, we believe accuracy in such scenario is able to explain the performance of our methods. The average classification accuracy over all concepts are concluded in Table 6. For more comprehensive comparison, we also include the results for $\beta = 0.2$ in our method.

As we can see, our methods ($\beta = 0.2$ and $\beta = 0.6$) consistently outperform the other tagging methods in terms of average classification accuracy. Especially when $\beta = 0.6$ for large K (e.g., 100, 90, and 80), our method improves SNTag, RWTag, TagRank methods by around 3–5% (i.e., relative improvement close to 10%). We believe the reason is in that the tag expansion step really provides some meaningful and descriptive extra tags to the test images, which makes them more discriminative. For small K (e.g.,70 and 60), though the performance difference becomes smaller, still our method obtains about 1–3% improvement over existing methods. As the K decreases, the proportion of relevant tags generated by each tagging method is expected to increase because the refinement processes are capable of promoting the relevant tags and degrading those noisy and irrelevant tags. It is worth noting that rather than boosting descriptive ability, more tags actually do not bring in more information for depicting the content. We believe the extra tags contain more noisy tags that degrade the classification performance. As a result, the differences of final tag results among various tagging methods may become smaller as K decreases, which

Table 6 Average Classification Accuracy (%) of different tagging methods over 14 concepts with top $K = 100, 90, 80, 70$ and 60 automatically generated tags.

	SNTag	RWTag	TagRank	Our method ($\beta = 0.2$)	Our method ($\beta = 0.6$)
$K = 100$	56.34	55.76	56.27	59.64	60.13
$K = 90$	56.06	56.54	56.07	60.07	60.43
$K = 80$	57.71	57.39	57.94	60.73	61.21
$K = 70$	59.51	57.40	60.76	61.54	61.70
$K = 60$	60.47	60.11	62.06	62.40	62.59

explains the decrease in the performance improvement. Among the two β settings in our method, $\beta = 0.6$ performs slightly better than $\beta = 0.2$, which corresponds with the results in Section 5.3.2. This results indicate that relatively balanced combination ($\beta = 0.6$) is able to generate better performance than imbalanced one ($\beta = 0.2$), which further proves the effectiveness of both of our proposed visual and corpus tag relevances.

Besides, we also observe that as K decreases, there are consistently rising trends in both of comparing algorithms and our methods ($\beta = 0.2$ and $\beta = 0.6$). Note that in the results of our methods, when $K = 60$ two curves reaches 62.59 and 62.40% which improve the results of $K = 100$ by nearly 5%. This phenomenon proves that our tag denoising step with the new tag model of multiple latent semantic groups is really more capable of promoting relevant tags and removing noisy tags than existing methods.

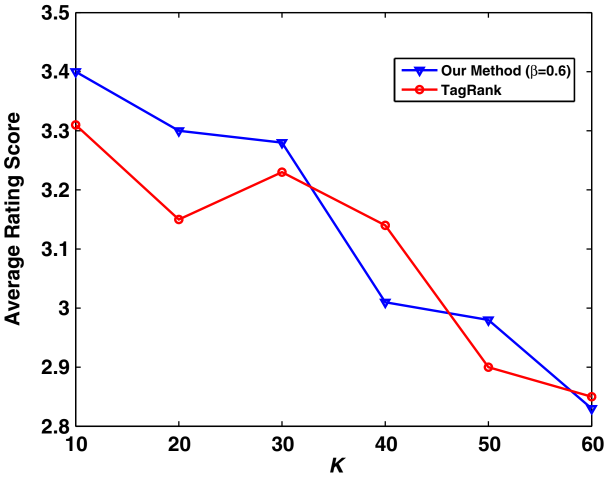
5.4 User-involved assessment

To further evaluate the tagging performance of our method, we also conduct a subjective assessment experiment. Test images and their automatically generated tags are together presented to invited users for direct and subjective assessment. In this experiment, we compare the results generated by two tagging methods—TagRank and our method with $\beta = 0.6$ as they provide better results than other methods in terms of the average classification accuracy as shown in Table 6.

For practical evaluation, assessors are provided two tag lists for each image. The tag lists are produced by the two methods and sorted in descend order according to tag relevance score. As manual evaluation is an extremely time-consuming process, we only randomly choose 50 test images and top 60 tags for evaluation. Assessor are asked to manually score each tag with a five-grade relevance scale: 1 (strongly irrelevant), 2 (irrelevant), 3 (uncertain), 4 (relevant), and 5 (strongly relevant).

We analyze the average rating scores to evaluate the effect of our tag relevance estimation strategy. To this end, average rating scores over top $K = 10, 20, 30, 40, 50$

Figure 5 User involved assessments.



and 60 tags are summarized and the results are illustrated in Figure 5. Generally speaking, as K rises, both methods show descending trends. This phenomenon reveals that the proportion of relevant tags in K results decreases as K grows. This outcome is expectedly in correspondence with our previous classification results. When comparing the two methods, we observe that the difference between their results is marginal. On average, our method slightly outperforms TagRank, especially when the number of tags is between 10 to 30.

6 Conclusions

In this paper, we have introduced an image tagging approach based on near-duplicate image content and collective multi-tag association mining. More specifically, confronted with the tag semantic incomplete issue, we first adopt a near-duplicate clustering algorithm to aggregate tags of near-duplicate images as a cluster document, which can help enhance and elicit tag correlation. Then, by regarding each cluster document as a transaction, multi-tag correlations are mined via a weighted association rule mining algorithm. Given a test image, its near-duplicates are retrieved to generate its candidate tags and the initial corpus relevance score for each candidate tag is estimated from corpus knowledge. More potentially relevant tags are subsequently expanded based on the multi-tag association rules. Meanwhile, we build a visual tag model for each tag using a KL Clustering method in order to elicit the underlying multiple latent semantics. For each tag, its visual tag relevance score can be calculated by comparing the test image and multiple clusters. Finally, visual and corpus relevance scores are combined together to obtain the overall relevance score. Experiments on a real-world Flickr image dataset shows that our proposed method outperforms existing tagging methods in terms of classification performance. We believe that our proposed tagging approach can bring direct improvement for current multimedia indexing and search.

Since our approach is composed of several parametric data mining techniques, it means our approach is probably sensitive to parameters and not easy to extend to general cases. In future, we intend to focus on exploring nonparametric techniques for developing more robust tagging framework. Also we plan to take into account multiple sources such as web shared image (Flickr), videos (YouTube), bookmarks (Delicious), etc. to extract more complete and meaningful knowledge for boosting the performance of current tagging approaches. In addition, we will further seek for more adaptive algorithm to elicit tag latent semantics. Finally, different near-duplicate detection methods will also be further investigated to see how they affect the performance of our method.

References

1. Agrawal, R., Imieliński, T. Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec. **22**(2), 207–216 (1993)
2. Ames, M., Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In: SIGCHI, pp. 971–980 (2007)
3. Amir, A., Argillander, J., Campbell, M., Haubold, A., Iyengar, G., Ebadollahi, S., Kang, F., M. Naphade, R., Natsev, A., Smith, J.R., Tei, J., Volkmer, T.: Ibm research trecvid-2005 video retrieval system. In: TREC Video Retrieval Evaluation Proceedings (2006)

4. Bailloeul, T., Zhu, C., Xu, Y.: Automatic image tagging as a random walk with priors on the canonical correlation subspace. In: MIR '08: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, pp. 75–82. ACM, New York (2008)
5. Cao, L., Yu, J., Luo, J., Huang, T.S.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In: MM '09: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 125–134, ACM, New York (2009)
6. Chang, C.-C., Lin, C.-J.: LIBSVM: A Library for Support Vector Machines (2001)
7. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of Singapore. In: CIVR, pp. 1–9 (2009)
8. Guan, Z., Bu, J., Mei, Q., Chen, C., Wang, C.: Personalized tag recommendation using graph-based ranking on multi-type interrelated objects. In: SIGIR, pp. 540–547 (2009)
9. Han, Y.: Multi-label boosting for image annotation by structural grouping sparsity. In: ACM Multimedia (2010)
10. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: SIGIR, pp. 531–538 (2008)
11. Hua, X.-S., Qi, G.-J.: Online multi-label active annotation: towards large-scale content-based video search. In: ACM Multimedia, pp. 141–150 (2008)
12. Kennedy, L.S., Chang, S.-F., Kozintsev, I.V.: To search or to label?: predicting the performance of search-based automatic image classifiers. In: MIR, pp. 249–258 (2006)
13. Krestel, R., Fankhauser, P., Nejdl, W.: Latent dirichlet allocation for tag recommendation. In: RecSys, pp. 61–68 (2009)
14. Li, X., Snoek, C., Worring, M.: Learning social tag relevance by neighbor voting. IEEE Trans. Multimedia **11**(7), 1310–1322 (2009)
15. Liu, D., Hua, X., Zhang, H.-J.: Image retagging. In: ACM Multimedia (2010)
16. Liu, D., Hua, X.-S., Yang, L., Wang, M., Zhang, H.-J.: Tag ranking. In: WWW, pp. 351–360 (2009)
17. Liu, X., Cheng, B., Yan, S., Tang, J., Chua, T.S., Jin, H., Label to region by bi-layer sparsity priors. In: MM '09: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 115–124, ACM, New York (2009)
18. Liu, Y., Wu, F., Zhuang, Y., Xiao, J.: Active post-refined multimodality video semantic concept detection with tensor representation. In: ACM Multimedia, pp. 91–100 (2008)
19. Mei, T., Wang, Y., Hua, X.-S., Gong, S., Li, S.: Coherent Image Annotation by Learning Semantic Distance (2008)
20. Moxley, E., Mei, T., Manjunath, B.: Video annotation through search and graph reinforcement mining. IEEE Trans. Multimedia **12**(3), 184–193 (2010)
21. Noh, T.-G., Park, S.-B., Yoon, H.-G., Lee, S.-J., Park, S.-Y.: An automatic translation of tags for multimedia contents using folksonomy networks. In: SIGIR, pp. 492–499 (2009)
22. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Zhang, H.J.: Correlative multi-label video annotation. In: ACM Multimedia, pp. 17–26. New York (2007)
23. Rui, X., Li, M., Li, Z., Ma, W.-Y., Yu, N.: Bipartite graph reinforcement model for web image annotation. In: ACM Multimedia, pp. 585–594 (2007)
24. Siersdorfer, S., San Pedro, J., Sanderson, M.: Automatic video tagging using content redundancy. In: SIGIR, pp. 395–402 (2009)
25. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW, pp. 327–336 (2008)
26. Sun, K., Bai, F.: Mining weighted association rules without preassigned weights. IEEE Trans. Knowl. Data Eng. **20**(4), 489–495 (2008)
27. Tang, J., Hua, X.-S., Qi, G.-J., Song, Y., Wu, X.: Video annotation based on kernel linear neighborhood propagation. IEEE Trans Multimedia **10**(4), 620–628 (2008)
28. Tang, J., Yan, S., Hong, R., Qi, G.-J., Chua, T.-S.: Inferring semantic concepts from community-contributed images and noisy tags. In: ACM Multimedia, pp. 223–232 (2009)
29. Wang, C., Jing, F., Zhang, L., Zhang, H.-J.: Image annotation refinement using random walk with restarts. In: ACM Multimedia, pp. 647–650 (2006)
30. Wang, C., Yan, S., Zhang, L., Zhang, H.-J., Multi-label sparse coding for automatic image annotation. In: Proceedings of IEEE Int. Conf. Computer Vision and Pattern Recognition, pp. 1643–1650. Florida, USA (2009)
31. Wang, F., Ding, C.H.Q., Li, T.: Integrated kl (k-means—laplacian) clustering: a new clustering approach by combining attribute data and pairwise relations. In: SDM, pp. 38–48 (2009)
32. Wang, N., Parthasarathy, S., Tan, K.-L., Tung, A.K.H.: Csv: visualizing and mining cohesive subgraphs. In: SIGMOD, pp. 445–458 (2008)

33. Wang, X.-J., Zhang, L., Li, X., Ma, W.-Y.: Annotating images by mining image search results. IEEE Trans. Pattern Anal. Mach. Intell. **30**(11), 1919–1932 (2008)
34. Weinberger, K.Q., Slaney, M., Van Zwol, R.: Resolving tag ambiguity. In: ACM Multimedia, pp. 111–120 (2008)
35. Wu, L., Hoi, S.C., Jin, R., Zhu, J., Yu, N.: Distance metric learning from uncertain side information with application to automated photo tagging. In: MM '09: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 135–144, ACM, New York (2009)
36. Wu, L., Yang, L., Yu, N., Hua, X.-S.: Learning to tag. In: WWW, pp. 361–370 (2009)
37. Xu, Z., Fu, Y., Mao, J., Su, D.: Towards the semantic web: collaborative tag suggestions. In: Collaborative Web Tagging Workshop. Edinburgh, Scotland (2006)
38. Yang, Y., Xu, D., Nie, F., Luo, J., Zhuang, Y.: Ranking with local regression and global alignment for cross media retrieval. In: ACM Multimedia (2009)
39. Yang, Y., Xu, D., Nie, F., Yan, S., Zhuang, Y.: Image clustering using local discriminant models and global IEEE Trans. Image Process. **19**(10), 2761–2773 (2010)
40. Yang, Y., Zhuang, Y.-T., Wu, F., Pan, Y.-H., Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Trans Multimedia **10**(3), 437–446 (2008)
41. Yin, Z., Li, R., Mei, Q., Han, J.: Exploring social tagging graph for web object classification. In: KDD, pp. 957–966 (2009)
42. Yuan, X., Hua, X.-S., Wang, M., Wu, X.: Manifold-ranking based video concept detection on large database and feature pool. In: ACM Multimedia, pp. 623–626 (2006)
43. Zha, Z.-J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: MM '09: Proceedings of the Seventeen ACM International Conference on Multimedia. ACM (2009)
44. Zhang, S., Huang, J., Huang, Y., Yu, Y., Li, H., Metaxas, D.N.: Automatic image annotation using group sparsity. In IEEE Conference on Computer Vision and Pattern Recognition, 2010. CVPR 2010 (2010)
45. Zhao, W., Ngo, C.-W.: Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. IEEE Trans. Image Process. **18**(2), 412–423 (2009)