Integrating Visual and Semantic Contexts for Topic Network Generation and Word Sense Disambiguation

Jianping Fan
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
jfan@uncc.edu

Hangzai Luo
Software Engineering Institute
East China Normal University
Shanghai, 200062, China
hluo@sei.ecnu.edu.cn

Yi Shen, Chunlei Yang
Dept of Computer Science
UNC-Charlotte
Charlotte, NC 28223, USA
yshen9,cyang36@uncc.edu

ABSTRACT

To support more effective searches in large-scale weakly-tagged image collections, we have developed a novel algorithm to integrate both the visual similarity contexts between the images and the semantic similarity contexts between their tags for topic network generation and word sense disambiguation. First, a topic network is generated to characterize both the semantic similarity contexts and the visual similarity contexts between the image topics more sufficiently. By organizing large numbers of image topics according to their cross-modal inter-topic similarity contexts, our topic network can make the semantics behind the tag space more explicit, so that users can gain deep insights rapidly and formulate their queries more precisely. Second, our word sense disambiguation algorithm can integrate the topic network to exploit both the visual similarity contexts between the images and the semantic similarity contexts between their tags for addressing the issues of polysemes and synonyms more effectively, thus it can significantly improve the precision and recall rates for image retrieval. Our experiments on large-scale Flickr and LabelMe image collections have provided very positive results.

Keywords
Topic network, semantic and visual contexts, word sense disambiguation.

1. INTRODUCTION

Collaborative image tagging has become a very popular way for people to share and annotate images. In a collaborative image tagging system [17-19], people can tag the images according to their social or cultural backgrounds, personal expertise and perception. We call such the collaboratively-tagged images as *weakly-tagged images* because their social tags may not be strongly related to the underlying image semantics. With the exponential growth of such the weakly-tagged images, it has become increasingly important to have mechanisms that can support more effective searches in large-scale weakly-tagged image collections. Unfortunately, searching from large-scale weakly-tagged image collections is not a trivial task because of the following challenging problems: (1) there is a *vocabulary discrepancy* between the keywords for query formulation and the text terms for image tagging (i.e., image annotators and users may not think of the same tags); and (2) without controlling the word vocabulary, many text terms for image tagging may be *synonymous* or *polysemous* (e.g., which may result in low precision and recall rates for image retrieval).

To assist users on query formulation, a tag cloud [12] has been used to provide a list of the most popular tags and support tag space exploration. Inter-tag contexts can be used to support more effective tag space navigation and achieve more precise characterization of the interestingness of the tags (i.e., like page linkages for characterizing the importance of web pages [2]), but the tag cloud completely ignores such the inter-tag contexts. Concept ontology [3-5] has been recently used to exploit the hierarchical inter-concept semantic contexts (i.e., namely the parent, siblings and children concepts) for large-scale text/image organization and navigation. However, such a concept ontology may not be suitable for organizing large-scale weakly-tagged image collections because there are no explicit hierarchical inter-tag semantic contexts in a collaborative image tagging space [17-19, 22-28].

To support more effective searches in large-scale weakly-tagged image collections, there is an urgent need to develop new algorithms for addressing the following problems more effectively:

(a) **Automatic Topic Network Generation:** Because only the inter-concept semantic context is exploited for concept ontology construction [3-4], the concept ontology cannot allow users to navigate large-scale weakly-tagged image collections according to their visual properties. It is well-accepted that the visual properties of the images are very important for users to search for images [1, 13-16]. Thus it is necessary to exploit both the inter-topic semantic contexts and the inter-topic visual contexts for generating a more precise topic network.

(b) **Synonymous Tags:** Different people may use different tags, which have the same or close meanings (synonyms), to tag semantically-related or visually-related images. For example, *car*, *auto*, and *automobile* are a set of synonyms. The synonyms will result in incomplete returns of relevant images in the image search side (i.e., with a low recall rate). Tag clustering [22-28] has been used to tackle

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CIVR '09 July 8-10, 2009, Santorini, GR
Copyright 2009 ACM 978-1-60558-480-5/09/07 ...$10.00.

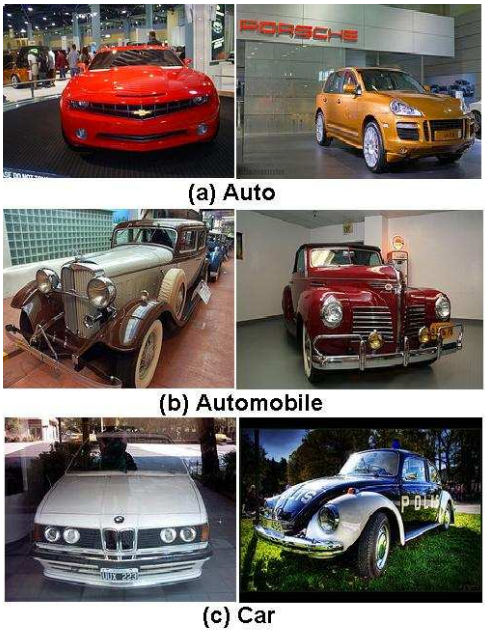

Figure 1: Visual properties of the relevant images may be helpful for tackling the synonyms.

the issue of synonyms by exploiting the inter-tag semantic contexts. However, not much work has been done by considering the visual properties of the relevant images for tag clustering [9]. As shown in Fig. 1, visual similarities between the relevant images may provide an alternative information source for achieving more accurate tag clustering in a collaborative image tagging space.

(c) **Polysemous (ambiguous) Tags:** Collaborative image tagging is an ambiguous process. People may tag an image based on: (1) an object in the image; (2) all objects in the image; (3) an overall "look" of the image; (4) an image event (i.e., human and object actions); or (5) image capture time and location [20-21]. Without controlling the vocabulary, different people may apply the same tag in different ways (i.e., the same tag may have different meanings under different contexts), which may result in imprecise and ambiguous results in the image search side. For example, the text term "bank" can be used to tag "bank office", "river bank" and "cloud bank". Forcing people to resolve the ambiguity of tagging terms at the image sharing time would be difficult and may prevent them from sharing and tagging their images. Word sense disambiguation is one potential solution for addressing this ambiguity issue [6-8], but exploiting only the semantic contexts between the nearby text terms may not be able to tackle the issue of polysemes effectively [9]. As shown in Fig. 2, the visual properties of the relevant images may offer an alternative information source for dealing with this ambiguity issue more effectively, but there is no comprehensive framework that can incorporate the visual similarity contexts between the relevant images for word sense disambiguation.

Some pioneering work has been done to combine both the visual contents of the images and the keywords for image concept modeling when the relationships between the keywords for image semantics interpretation and the visual contents of the images are explicit [14-16]. Because of the polysemes, the relationships between the social tags and the visual contents of the images are less clear in a collaborative image tagging space. Thus these existing techniques for image concept modeling cannot be extended for addressing the issue of polysemes effectively.

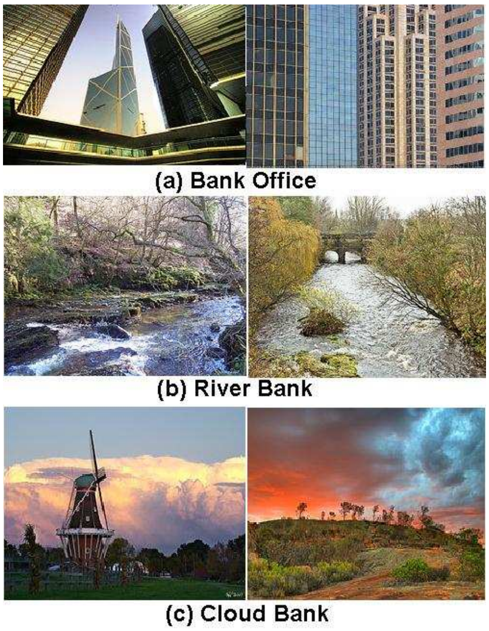

Figure 2: Visual properties of the relevant images may be helpful for tackling the polysemes.

In this paper, we have developed a novel algorithm to incorporate the topic network for addressing the issues of polysemes and synonyms in a collaborative image tagging space. The paper is organized as follows. In section 2, an interesting algorithm is introduced for automatic image topic extraction. In section 3, an automatic topic network generation algorithm is introduced, where both the semantic similarity contexts between the image topics and the visual similarity contexts between their images are exploited. In section 4, a novel cross-modal tag clustering algorithm is introduced for addressing the issue of synonyms. In section 5, a new cross-modal image clustering algorithm is developed for addressing the issue of polysemes. The algorithm evaluation results are given in section 6. We conclude this paper at section 7.

2. IMAGE TOPIC EXTRACTION

Each image in a collaborative tagging system is associated with the image holder's taggings of the underlying image contents and other users' taggings or comments. It is worth noting that entity extraction can be done more effectively in a collaborative image tagging space. In this paper, we first focus on extracting the social tags which are strongly related to the underlying image semantics. The social tags, which are related to image capture time, are also very attractive for searching in large-scale weakly-tagged image collections [21], but they are beyond the scope of this paper. Thus the image tags are first partitioned into two categories: noun phrases *versus* verb phrases. The noun phrases are further partitioned into two categories automatically: *content-relevant tags* (i.e., tags that are relevant to image contents) and *content-irrelevant tags*. The verb phrases are further partitioned into two categories automatically: *event-relevant tags* (i.e., tags that are relevant to image events) and *event-*

irrelevant tags.

The occurrence frequency for each content-relevant tag and each event-relevant tag is counted automatically by using the number of relevant images. The misspelling tags may have low frequencies (i.e., different people may make different types of tagging mistakes), thus it is easy for us to correct such the misspelling tags and their images are added into the relevant tags automatically. Two tags, which are used for tagging the same image, are considered to co-occur once without considering their order. A co-occurrence matrix is obtained by counting the frequencies of such pairwise tag co-occurrences.

The content-relevant tags and the event-relevant tags are further partitioned into two categories according to their interestingness scores: *interesting tags* and *uninteresting tags*. In this paper, multiple information sources have been exploited for determining the interesting tags more accurately. For a given tag C, its interestingness score $\omega(C)$ depends on: (1) *its occurrence frequency* $t(C)$ (e.g., higher occurrence frequency corresponds to higher interestingness score); (2) *its co-occurrence frequency* $\vartheta(C)$ with any other tag in the vocabulary (e.g., higher co-occurrence frequency corresponds to higher interestingness score); and (3) *users' query frequency* $\alpha(C)$ (e.g., higher query frequency corresponds to higher interestingness score). The occurrence frequency $t(C)$ for a given tag C is equal to the number of images that are tagged by the given tag C. The co-occurrence frequency $\vartheta(C)$ for the given tag C is equal to the number of images that are tagged jointly by the given tag C and any other tag in the vocabulary.

Thus the ***interestingness score*** $\omega(C)$ for a given tag C is defined as:

$$\omega(C) = \xi \cdot \frac{e^{t(C)} - e^{-t(C)}}{e^{t(C)} + e^{-t(C)}} + \zeta \cdot \frac{e^{\vartheta(C)} - e^{-\vartheta(C)}}{e^{\vartheta(C)} + e^{-\vartheta(C)}} + \lambda \cdot \frac{e^{\alpha(C)} - e^{-\alpha(C)}}{e^{\alpha(C)} + e^{-\alpha(C)}} \tag{1}$$

where $\xi + \zeta + \lambda = 1$, the first part is used to characterize the interestingness score of the given tag C gained from its occurrence frequency $t(C)$ in large-scale image collections, the second part is used to characterize the interestingness score gained from its co-occurrence frequency $\vartheta(C)$ with any other tag in the vocabulary, and the third part is used to characterize the interestingness score gained from users' query frequency $\alpha(C)$, ξ, ζ and λ are the relative importance factors.

In our definition, the interestingness score is normalized within the interval [0, 1]. The content-related tags and the event-related tags, which have larger values of the interestingness scores, are identified as the ***interesting tags***.

To enlarge the vocabulary of the interesting tags, a new algorithm is developed for discovering *latent interesting tags*, so that users can have more choices on query formulation. The co-occurrences of each uninteresting tag and the interesting tags are first counted. The χ^2-measure is used to characterize the degree of the bias of the co-occurrence distribution between the interesting tags and the uninteresting tags. If an uninteresting tag appears frequently with a particular subset of the interesting tags in large-scale weakly-tagged image collections, it will be likely to be very important and should be treated as the latent interesting tag for image topic extraction.

All the interesting tags and the latent interesting tags are treated as ***image topics*** of interest in a collaborative image tagging space. Rather than using only the query frequencies and the tag frequencies for image topic extraction (as done by a tag cloud), multiple alternative information sources (i.e., co-occurrence frequencies and latent interesting tags) are exploited for discovering more meaningful image topics. By mapping the images and their social tags onto a conceptual space, our image topic extraction algorithm can create a new form to support more effective representation and access of large-scale weakly-tagged image collections.

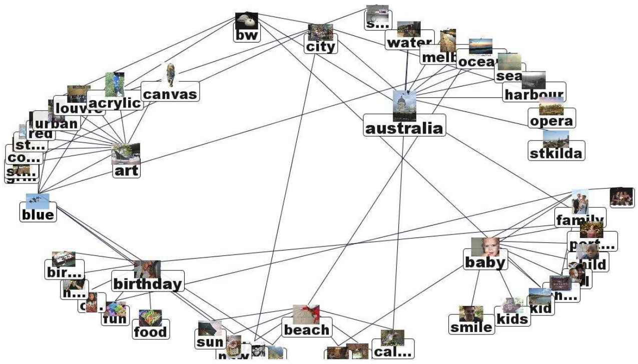

Figure 3: Topic network for Flickr image set.

3. TOPIC NETWORK GENERATION

Our topic network consists of two key components: (1) *large numbers of image topics*; and (2) *their cross-modal inter-topic similarity contexts*. The cross-modal inter-topic similarity contexts consists of both the inter-topic semantic contexts and the inter-topic visual contexts.

In this paper, multiple criteria (both flat and hierarchical semantic contexts) are considered to achieve more precise characterization of the inter-topic semantic contexts in a collaborative image tagging space. For two image topics C_i and C_j, their inter-topic semantic context $\phi(C_i, C_j)$ consists of two components: (1) the *flat inter-topic semantic context* because of their co-occurrences in large-scale weakly-tagged image collections (e.g., higher co-occurrence probability corresponds to stronger inter-topic semantic context); and (2) the *hierarchical inter-topic semantic context* because of their inherent correlation defined by WordNet (e.g., stronger inherent correlation (i.e., closer on WordNet [5]) corresponds to stronger inter-topic semantic context).

For two image topics C_i and C_j, their inter-topic semantic context $\phi(C_i, C_j)$ is defined as:

$$\phi(C_i, C_j) = -\upsilon \cdot \frac{\theta(C_i, C_j)}{\log \theta(C_i, C_j)} - \varsigma \cdot \theta(C_i, C_j) \cdot \log \frac{L(C_i, C_j)}{2 \cdot D} \tag{2}$$

where $\upsilon + \varsigma = 1$, the first part is used to characterize the flat inter-topic semantic context according to their concurrence in large-scale image collections, the second part is used to characterize the hierarchical inter-topic semantic context, υ and ς are the relative importance factors, $\theta(C_i, C_j)$ is the co-occurrence probability for the image topics C_i and C_j, $L(C_i, C_j)$ is the number of nodes between the text terms for interpreting the image topics C_i and C_j on WordNet, D is the maximum number of nodes from root node to leaf node on WordNet.

Our inter-topic semantic context measure $\phi(\cdot, \cdot)$ can simultaneously consider both the flat inter-topic semantic context and the hierarchical inter-topic semantc context in a collaborative image tagging space.

It is well-accepted that the visual properties of the images are very important for image retrieval [1, 34], thus the inter-topic visual contexts may also play an important role in generating a more precise topic network. To achieve more sufficient characterization of various visual properties of the images, both global and local visual features are extracted for image content representation. The following visual features are extracted for image content representation: (1) 36-bin RGB color histogram to characterize the global color distributions of the images; (2) 48-dimensional texture features from Gabor filter banks to characterize the global visual properties (i.e., global structures) of the images; and (3) a number of interest points and their SIFT (scale invariant feature transform) features to characterize the local visual properties of the underlying salient image components.

The high-dimensional visual features are first partitioned automatically into multiple feature subsets and each feature subset is used to characterize one certain type of the visual properties of the images. For each feature subset, a suitable base kernel is designed for image similarity characterization.

For a given image topic C_j in the vocabulary, different base image kernels may play different roles on characterizing the diverse visual similarity relationships between the images. Thus the diverse visual similarities between the images are characterized more precisely by using a mixture-of-kernels [29-32]:

$$\kappa(x, y) = \sum_{l=1}^{\tau} \alpha_l \kappa_l(x, y), \qquad \sum_{l=1}^{\tau} \alpha_l = 1 \tag{3}$$

where τ is the number of feature subsets (i.e., the number of base image kernels), $\alpha_l \geq 0$ is the importance factor for the lth base image kernel $\kappa_l(x, y)$.

The inter-topic visual contexts may also play an important role in generating a more precise concept network. The inter-topic visual context $\gamma(C_i, C_j)$ between the image topics C_i and C_j can be determined by performing canonical correlation analysis [33] on their image sets S_i and S_j:

$$\gamma(C_i, C_j) = \begin{matrix} max \\ \theta, \vartheta \end{matrix} \frac{\theta^T \kappa(S_i)\kappa(S_j)\vartheta}{\sqrt{\theta^T \kappa^2(S_i)\theta \cdot \vartheta^T \kappa^2(S_j)\vartheta}} \tag{4}$$

where θ and ϑ are the parameters for determining the optimal projection directions to maximize the correlations between two image sets S_i and S_j for the image topics C_i and C_j, $\kappa(S_i)$ and $\kappa(S_j)$ are the cumulative kernel functions for characterizing the visual correlations between the images in the same image sets S_i and S_j.

$$\kappa(S_i) = \sum_{x_l, x_m \in S_i} \kappa(x_l, x_m), \quad \kappa(S_j) = \sum_{x_h, x_k \in S_j} \kappa(x_h, x_k) \tag{5}$$

where the visual correlation between the images is defined as their kernel-based visual similarity $\kappa(\cdot, \cdot)$ in Eq.(5).

The parameters θ and ϑ for determining the optimal projection directions are obtained automatically by solving the following eigenvalue equations:

$$\kappa(S_i)\kappa(S_i)\theta - \lambda_\theta^2 \kappa(S_i)\kappa(S_i)\theta = 0$$

$$\kappa(S_j)\kappa(S_j)\vartheta - \lambda_\vartheta^2 \kappa(S_j)\kappa(S_j)\vartheta = 0 \tag{6}$$

where the eigenvalues λ_θ and λ_ϑ follow the additional constraint $\lambda_\theta = \lambda_\vartheta$.

The inter-topic visual context $\gamma(C_i, C_j)$ is first normalized into the same interval as the inter-topic semantic context $\phi(C_i, C_j)$. The inter-topic semantic context and the inter-topic visual context are further integrated to achieve more precise characterization of their cross-modal inter-topic similarity context $\varphi(C_i, C_j)$:

$$\varphi(C_i, C_j) = \epsilon \cdot \frac{e^{\phi(C_i, C_j)} - e^{-\phi(C_i, C_j)}}{e^{\phi(C_i, C_j)} + e^{-\phi(C_i, C_j)}} + \eta \cdot \frac{e^{\gamma(C_i, C_j)} - e^{-\gamma(C_i, C_j)}}{e^{\gamma(C_i, C_j)} + e^{-\gamma(C_i, C_j)}} \tag{7}$$

where $\epsilon + \eta = 1$, the first part denotes the semantic similarity context between the image topics C_j and C_i, the second part indicates their inter-topic visual context, $\gamma(C_i, C_j)$ is the visual similarity context between the image sets for the image topics C_i and C_j.

When large numbers of image topics and their cross-modal inter-topic similarity contexts are available, they are used to construct a topic network. Unlike the one-direction IS-A hierarchy [3-5], each image topic can be linked with all the other image topic on the topic network, thus the maximum number of such inter-topic associations could be $\frac{T(T-1)}{2}$, where T is the total number of image topics on the topic network. However, the strength of the associations between some image topics may be very weak, thus it is not necessary for each image topic to be linked with all the other image topics on the topic network. Eliminating the weak inter-topic links can increase the visibility of the image topics of interest dramatically, but also allow users to concentrate on the most significant cross-modal inter-topic similarity contexts. Based on this understanding, each image topic is automatically linked with the most relevant image topics with larger values of the inter-topic similarity contexts $\varphi(\cdot, \cdot)$ (i.e., their values of $\varphi(\cdot, \cdot)$ are above a threshold $\delta = 0.25$). The value of the threshold is determined by making a tradeoff between the computational complexity for interactive topic network exploration and the effectiveness for characterizing the most significant inter-topic contexts.

The topic networks for our test image sets (LabelMe and Flickr) are shown in Fig. 3 and Fig. 4, where each image topic is linked with multiple relevant image topics with larger values of $\varphi(\cdot, \cdot)$. It is worth noting that different image topic can have different numbers of the most relevant image topics on the topic network. By visualizing large numbers of image topics according to their cross-modal inter-topic similarity contexts, our topic network can make the semantics behind the tag space more explicit and can allow users to navigate large-scale weakly-tagged image collections effectively according to their semantic and visual similarity contexts. By supporting interactive context-driven topic network exploration and navigation, users can gain deep insights rapidly, build up their mental query models interactively, and exploit their background knowledge and strong pattern recognition capability to select the visible image topics on the topic network for query formulation. Thus the user's image needs can be efficiently translated into the available image topics on the topic network.

4. COMBINING SYNONYMOUS TOPICS

Some image topics on the topic network may be synonymous (i.e., multiple image topics share the same meaning), which may result in incomplete returns of images in the search side (i.e, with a low recall rate). In this paper, a cross-modal tag clustering algorithm is developed to combine the synonymous topics, which may significantly increase the recall rate for image retrieval.

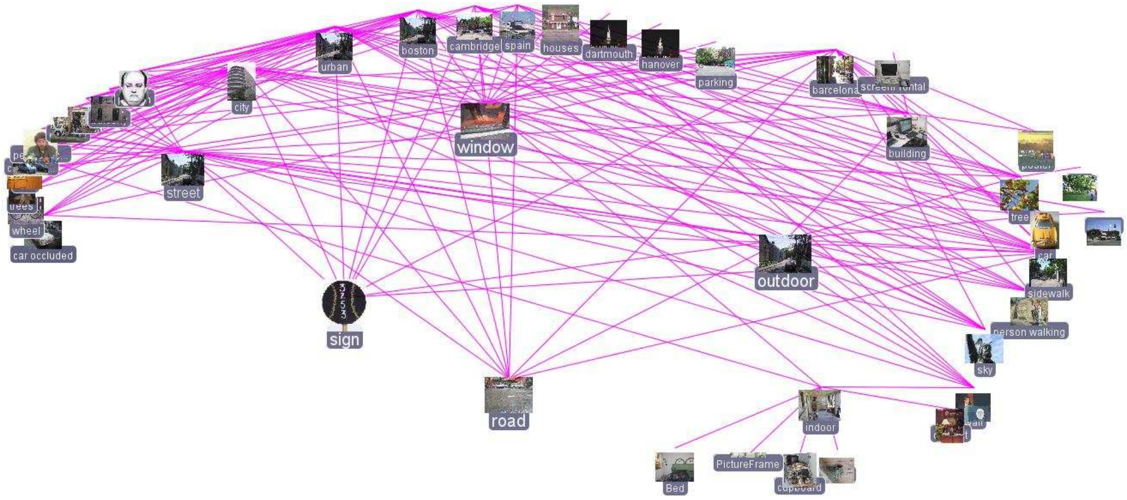

Figure 4: Topic network for LabelMe image set.

Our cross-modal tag clustering algorithm consists of two critical components: (1) distance or similarity measures for characterizing the pairwise tag similarity; and (2) optimization criteria for tag grouping according to their pairwise similarity measures. Since the pairwise similarity measure is fundamental to the definition of a tag cluster, the topic network is incroporated to define a more accurate similarity measure for tag clustering.

Because image topics and their cross-modal inter-topic similarity contexts are indexed coherently by the topic network, a constraint-driven clustering algorithm is developed to achieve more accurate cross-modal tag clustering. For two image topics C_i and C_j on the topic network, their constrained cross-modal similarity context $\psi(C_i, C_j)$ depends on two issues: (1) cross-modal inter-topic similarity context $\varphi(C_i, C_j)$ (e.g., similar image topics should have larger values of $\varphi(\cdot, \cdot)$); and (2) constraint and linkage relatedness on the topic network (e.g., similar image topics should be closer on the topic network). The constrained cross-modal similarity context $\psi(C_i, C_j)$ between two image topics C_i and C_j is defined as:

$$\psi(C_i, C_j) = \varphi(C_i, C_j) \times \begin{cases} e^{-\frac{l^2(C_i,C_j)}{\sigma^2}}, & if \quad l(C_i, C_j) \leq \Delta \\ 0, & otherwise \end{cases} \tag{8}$$

where the first part $\varphi(C_i, C_j)$ denotes the cross-modal inter-topic similarity context between C_i and C_j, the second part indicates the constraint and linkage relatedness between C_i and C_j on the topic network, $l(C_i, C_j)$ is the distance between the physical locations for the image topics C_i and C_j on the topic network, σ is the variance of their physical location distances, and Δ is a pre-defined threshold which largely depends on the size of the nearest neighbors to be considered. In this paper, the first-order nearest neighbors is considered as shown in Fig. 5, $\Delta = 1$.

After such the constrained cross-modal inter-topic similarity contexts are obtained, graph-cut algorithm is used for tag clustering [11]. Thus the synonymous topics, which have large values of the constrained cross-modal inter-topic similarity contexts, are grouped into the same cluster and can be combined as one *super-topic*. The images for these synonymous topics in the same cluster are assigned to the super-topic automatically, so that users can obtain more comprehensive returns of the relevant images (i.e., with a higher recall rate). Multiple tags for interpreting these synonymous topics are combined as one union phrase for tagging the super-topic, so that users can have more flexible choices on query formulation (i.e., any part of such a union phrase can be used as the query term).

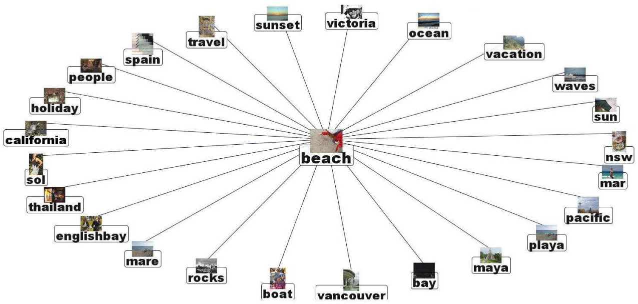

Figure 5: The first-order nearest neighbors of "beach" in Flickr image set.

When multiple synonymous image topics $\{c_1, \cdots, c_n\}$ are integrated as one super-topic C_s, the inter-topic similarity contexts between one given image topic on the topic network and the super-topic C_s largely depend on its inter-topic similarity contexts with all these synonymous image topics $\{c_1, \cdots, c_n\}$. Based on this understanding, a novel algorithm is developed for calculating the aggregated similarity contexts between the super-topic C_s and other image topics on the topic network more effectively. Thus the aggregated semantic context $\hat{\phi}(C_s, C_j)$ between the super-topic C_s and the image topic C_j on the topic network is obtained by using Eq. (2) with the cumulative probabilities $\hat{\theta}(C_s, C_j)$ and $\hat{\theta}(C_s)$ and a new inherent correlation $\hat{L}(C_s, C_j)$. The cumulative co-occurrence probability $\hat{\theta}(C_s, C_j)$, the cumulative occurrence probability $\hat{\theta}(C_s)$, and the new inherent correlation $\hat{L}(C_s, C_j)$ is defined as:

$$\hat{\theta}(C_s, C_j) = \sum_{l=1}^{n} \theta(c_l, C_j), \qquad \hat{\theta}(C_s) = \sum_{l=1}^{n} \theta(c_l)$$

$$\hat{L}(C_s, C_j) = min\{L(c_l, C_j) | l = 1, \cdots, n\} \tag{9}$$

where $\theta(c_l, C_j)$ is the co-occurrence probability for the syn-

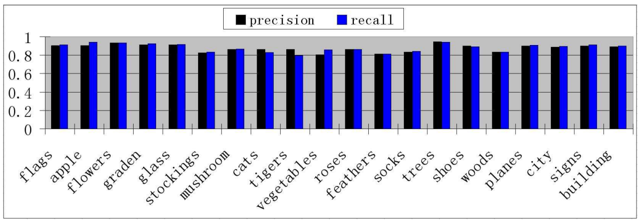

Figure 6: The precision and recall for our system to support image retrieval.

onymous image topic c_l and the image topic C_j, $\theta(c_l)$ is the individual occurrence probability of the synonymous image topic c_l, $L(c_l, C_j)$ is the inherent correlation between the synonymous image topic c_l and the image topic C_j on WordNet.

The aggregated visual context $\hat{\gamma}(C_s, C_j)$ between the super-topic C_s and the image topic C_j on the topic network is defined by using Eq. (4), where the image set S_s for the super-topic C_s is defined as an union of the image sets for the synonymous image topics $\{c_1, \cdots, c_n\}$: $S_s = S_1 \cup \cdots \cup S_n$.

By incorporating the topic network to achieve more accurate cross-modal tag clustering, our algorithm can address the issue of synonyms more effectively and may result in a higher recall rate for image retrieval. By merging the synonymous topics, our algorithm can offer more coherent representation of the topic network, thus it can reduce the users' perceptual cost significantly for interactive topic network exploration and allow them to gain the deep insights rapidly for query formulation.

5. SPLITTING POLYSEMOUS TOPICS

Some image topics on the topic network may be polysemous, which may result in large numbers of weakly-related images (i.e., with a low precision rate). To address the polysemes, automatic image clustering is performed to split the polysemous topics, so that users can obtain more precise returns of the relevant images (i.e., with a higher precision rate). Thus a new algorithm is developed for partitioning the images under the same polysemous topic into multiple groups automatically and each group may correspond to one certain sub-topic of the same polysemous topic.

Each image consists of two information sources: (1) visual properties; and (2) multiple tags for image semantics interpretation. The visual similarity context between the images is characterized by using a mixture-of-kernels, and the semantic similarity context between their tags is characterized by using a semantic kernel.

To determine the cross-modal similarity contexts between the images, both the visual similarity context (i.e., mixture-of-kernels) and the semantic similarity context (i.e., semantic kernel) are aligned onto the same kernel space. Second, the visual similarity context and the semantic similarity context are normalized into the same interval, so that they can be comparable. Finally, a linear combination of the normalized visual similarity context and the normalized semantic similarity context is used to measure the cross-modal similarity context between the images $\hat{\kappa}(\cdot, \cdot)$ [29-32].

Our kernel alignment approach can achieve more precise characterization of the cross-modal similarity contexts between the images, thus the images under the same polysemous topic can be partitioned into multiple clusters more accurately. The optimal image partition is obtained by minimizing the trace of the within-cluster scatter matrix [10], S_w^ϕ. The trace of the scatter matrix $Tr(S_w^\phi)$ is given by:

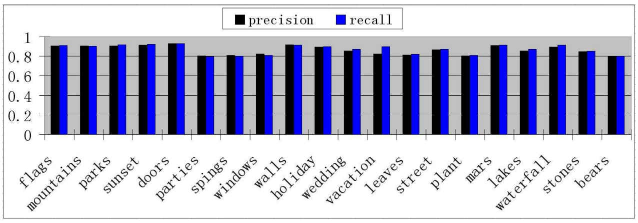

Figure 7: The precision and recall for our system to support image retrieval.

$$Tr(S_w^\phi) = \frac{1}{N} \sum_{l=1}^{\tau} \sum_{i=1}^{N} \alpha_{li} \left(\hat{\kappa}(x_i, x_j) - \frac{2}{N_l} \sum_{j=1}^{N} \alpha_{lj} \hat{\kappa}(x_i, x_j) \right.$$
$$\left. + \frac{1}{N_l^2} \sum_{j=1}^{N} \sum_{m=1}^{N} \alpha_{lj} \alpha_{lm} \hat{\kappa}(x_j, x_m) \right) \quad (10)$$

where $\hat{\kappa}(\cdot, \cdot)$ is the cross-modal similarity context between two images, N is the total number of images, τ is the number of clusters, and $N_l = \sum_{i=1}^{N} \alpha_{li}$ is the number of images for the lth cluster. Searching the optimal values of the elements α that minimizes the trace of the within-cluster scatter matrix will be achieved effectively by an iterative procedure [10]. Each image cluster may correspond to one certain sub-topic for the same polysemous topic.

Semantic image classification is performed for tagging the sub-topics (i.e., image clusters) under the same polysemous topic. First, the similarity contexts between the polysemous topic and its sub-topics is exploited to learn multiple inter-related image classifiers more effectively. The similarity contexts between the polysemous topic and its multiple sub-topics are well-defined, thus the image classifiers with higher discrimination power can be learned for tagging the sub-topics. For the polysemous topic "bank" as shown in Fig. 2, multiple inter-related image classifiers are learned for tagging its sub-topics. For example, the "building" classifier is used to tag the sub-topic "bank office", the image classifiers for "water", "sand", "grass", and "tree" are integrated to tag the sub-topic "river bank", and the image classifiers for "sky" and "cloud" are integrated to tag the sub-topic "cloud bank".

When the polysemous topic C_p is split into multiple sub-topics $\{\hat{c}_1, \cdots, \hat{c}_m\}$, a novel algorithm is developed to determine their split inter-topic similarity contexts with the residue image topics on the topic network more effectively. The split inter-topic semantic context between the sub-topic $\hat{c}_l$ and any other image topic on the topic network is defined by using Eq. (2), where the co-occurrence probability is replaced by the split co-occurrence probability. The split co-occurrence probability $\widetilde{\theta}(\hat{c}_l, C_j)$, the split occurrence probability $\widetilde{\theta}(\hat{c}_l)$, and the new inherent correlation $\widetilde{L}(\hat{c}_l, C_j)$ are refined as:

$$\widetilde{\theta}(\hat{c}_l, C_j) = \frac{|\hat{S}_l|}{|S_p|} \theta(C_p, C_j), \qquad \widetilde{\theta}(\hat{c}_l) = \frac{|\hat{S}_l|}{|S_p|} \theta(C_p)$$

$$\widetilde{L}(\hat{c}_l, C_j) = L(C_p, C_j) + 1 \quad (11)$$

where $\theta(C_p, C_j)$ is the co-occurrence probability for the polysemous image topic C_p and the image topic C_j, $\theta(C_p)$ is the occurrence probability for the polysemous image topic C_p, $|\hat{S}_l|$ is the size of the image set $\hat{S}_l$ for the sub-topic $\hat{c}_l$, $|S_p|$ is

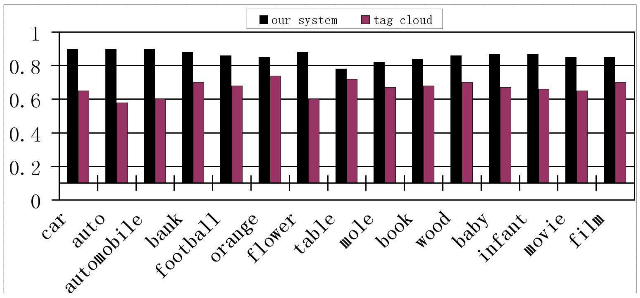

Figure 8: The comparison results between our prototype system and Flickr image search engine for some polysemous or synonymous image topics.

the size of the image set S_p for the polysemous image topic C_p, $|S_P| = \sum_{l=1}^{m} |\hat{S}_l|$, $L(C_p, C_j)$ is the inherent correlation between the polysemous image topic C_p and the image topic C_j on WordNet.

The split inter-topic visual context between the sub-topic $\hat{c}_l$ and any other image topic on the topic network is determined automatically according to the visual correlation between their images by using kernel CCA. The smaller image set $\hat{S}_l$ for the sub-topic $\hat{c}_l$ is defined as a subset of the image set S_p for the polysemous topic C_p (e.g., $\hat{S}_l$ can be obtained automatically by our image clustering technique).

By exploiting multiple cross-modal information sources for cross-modal image clustering, our algorithm can address the issue of polysemes more effectively and result in a higher precision rate for image retrieval. By splitting the polysemous topics automatically and supporting automatic subtopic tagging, our algorithm can enrich the semantics on the topic network significantly, so that users can select more specific image topics to formulate their queries more precisely and obtain more accurate returns of the relevant images.

6. ALGORITHM EVALUATION

We have carried out our experimental studies by using large-scale weakly-tagged Flickr and LabelMe images [12-13]. We have downloaded more than 1.5 billions Flickr images and 1.2 millions LabelMe images and their tagging documents. We have generated a topic network with more than 4000 most popular image topics (i.e., most popular taggings).

The image seeking process is necessarily initiated by an image need on user's side, thus the success of an image retrieval system depends largely on its ability to allow user to communicate his/her image needs effectively [1]. Therefore, traditional performance criteria (precision and recall) may not be sufficient enough for evaluating the significance of our system. Assessing the performance of image retrieval systems is partly subjective, therefore we have developed new evaluation models to assess our system. In this paper, we have incorporated a **user study** and a **new performance metric** (i.e., rating score sheet) for evaluating our system. Our user study focuses on evaluating the benefits of users from using our system in the context of integrating the topic network on query formulation and word sense disambiguation.

Our algorithm evaluation work focuses on four criteria: (1) how well the topic network summarizes and represents large-scale weakly-tagged image collections, (2) how well the topic network assists them on query formulation, (3) how well the image topics on the topic network match the diverse image needs of users, and (4) how well our techniques address the issues of polysemes and synonyms. The first three criteria are assessed through user study, the last criterion is assessed by using the precision and recall rates for image retrieval.

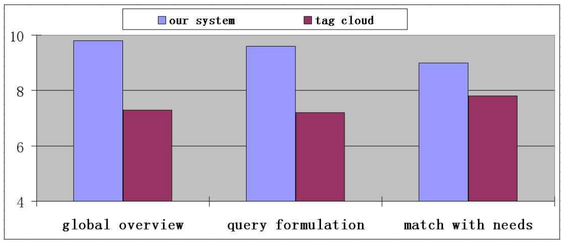

Figure 9: The comparison results on user study on three criteria between our prototype system and Flickr image search engine.

For evaluating the effectiveness of our word sense disambiguation algorithm, the *benchmark metric* includes *precision* ρ and *recall* ϱ. They are defined as:

$$\rho = \frac{\vartheta}{\vartheta + \xi}, \qquad \varrho = \frac{\vartheta}{\vartheta + \nu} \qquad (12)$$

where ϑ is the set of true positive images that are related to the corresponding image topic and are returned correctly, ξ is the set of true negative images that are irrelevant to the corresponding image topic and are returned incorrectly, and ν is the set of false positive images that are related to the corresponding image topic but are returned incorrectly. The precision is used to characterize the accuracy of our system for finding the particular images of interest, and the recall is used to characterize the efficiency of our system for finding the particular images of interest.

Fig. 6 and Fig. 7 give the precision and recall of our system for social image retrieval. From these experimental results, one can observe that our system can support social image retrieval effectively. For some polysemous or synonymous image topics, we have compared the precision and recall rates between our system and a tag cloud-based Flickr system as shown in Fig. 8. One can observe that our system can achieve higher precision and recall rates for image retrieval, thus our word sense disambiguation algorithm can address the issues of polysemes and synonyms effectively.

Image retrieval is a lucid example of user-centric computing because both the judgment of image relevance and the interpretation of image semantics are user-dependent. Based on this observation, user study has been conducted by comparing our prototype system with keyword-based image rerieval system (such as Flickr) on three key issues: 1) how well the topic network summarizes and represents large-scale weakly-tagged image collections, (2) how well the topic network assists them on query formulation, (3) how well the image topics on the topic network match the diverse image needs of users.

We have invited 18 students to participate this user study, where 10 undergraduate students from database class without any knowledge on image indexing and retrieval, 5 graduate students with some experiences on keyword-based Flickr image search engine, 3 graduate students who are familar with image retrieval systems. The students are asked to score our system and Flickr system. The scores for system evaluation are set from 10 (highest) to 1 (lowest). Each student is required to submit 500 queries which are randomly sampled from 4000 available image topics, different query sets are assigned for different students with some overlappings for cross validation. Therefore, each image topic is searched by at least two students independently. The comparison results are given in Fig. 9, one can observe that our proposed prototype system is very competitive.

7. CONCLUSIONS

In this paper, we have introduced a new framework to incorporate the topic network for word sense disambiguation. First, a topic network is generated to characterize both the semantic similarity contexts and the visual similarity contexts between the image topics more sufficiently. To address the issues of polysemes and synonyms and enhance the precision and recall rates for image retrieval, a novel algorithm is developed by integrating the topic network to exploit both the visual similarity contexts between the images and the semantic similarity contexts between their tags for word sense disambiguation. Our experiments on large-scale weakly-annotated Flickr and LabelMe image collections have provided very positive results. Our future work will focus on making our system available online, so that more Internet users can join our user study.

8. REFERENCES

[1] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, “Content-based image retrieval at the end of the early years”, *IEEE Trans. on PAMI*, 2000.

[2] S. Brin, L. Page, “The anatomy of a large-scale hypertextual web search engine”, WWW, 1998.

[3] M. Naphade, J.R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, J. Curtis, “Large-scale concept ontology for multimedia”, *IEEE Multimedia*, 2006.

[4] M. Sanderson, W. B. Croft, “Deriving concept hierarchies from text”, ACM SIGIR, pp.206-213, 1999.

[5] C. Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Boston, MA, 1998.

[6] J. Stetina, S. Kurohashi, M. Nagao, “General word sense method based on a full sentential context”, COLING-ACL Workshop, 1998.

[7] M. Sussna, “Word sense disambiguation for free-text indexing using a massive semantic network”, ACM CIKM, pp.67-74, 1993.

[8] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods”, ACL, 1995.

[9] K. Barnard, M. Johnson, ”Word sense disambiguation with pictures”, *Artificial Intelligence*, vol. 167, pp. 13-30, 2005.

[10] M. Girolami, “Mercer kernel-based clustering in feature space”, *IEEE Trans. on Neural networks*, vol.13, no.3, pp.780-784, 2002.

[11] J Shi, J Malik, “Normalized cuts and image segmentation”, *IEEE Trans. on PAMI*, 2000.

[12] Flickr, http://www.flickr.com.

[13] B. Russell, A. Torralba, W.T. Freeman, http://labelme.csail.mit.edu/

[14] R. Zhang, Z. Zhang, M. Li, W.-Y. Ma, H.-J. Zhang, “A probabilistic semantic model for image annotation and multi-modal image retrieval”, IEEE ICCV, 2005.

[15] K. Barnard and D. Forsyth, “Learning the semantics of words and pictures”, IEEE ICCV, pp.408-415, 2001.

[16] J. Jeon, V. Lavrenko, R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models”, ACM SIGIR, 2003.

[17] S. Golder, B. Huberman, “The structure of collaborative tagging systems”, HP Lab Report, 2006.

[18] M. Guy, E. Tonkin, “Folksonomies: Tidying up tags?”, *D-Lib Magazine*, vol.12, 2006.

[19] A. Mathes, “Folksonomies-Comperative classification and communication through shared metadata”, 2004.

[20] M. Naaman, Y. Song, A. Paepcke, H. Garcia-Molina, “Automatic organization for digital photographs with geographic coordinates”, ACM/IEEE JCDL, 2004.

[21] A. Graham, H. Garcia-Molina, A. Paepcke, T. Winograd, “Time as essence for photo browsing through personal digital libraries”, ACM/IEEE JCDL, pp.326Ű335, 2002.

[22] X. Wu, L. Zhang, Y. Yu, “Exploring social annotations for the semantic web”, ACM WWW, pp.417-426, 2006.

[23] C.H. Brooks, N. Montanez, “Improved annotation of the blogosphere via autotagging and hierarchical clustering”, ACM WWW, 2006.

[24] S. Bao, X. Wu, B. Fei, G. Xue, Z. Su, Y. Yu, “Optimizing web search using social annotations”, WWW, pp.501-510, 2007.

[25] G. Begelman, P. Keller, F. Smadja, “Automated tag clustering: Improving search and exploration in the tag space”, ACM WWW, 2006.

[26] J. Gemmell, A. Shepitsen, B. Mobasher, R. Burke, “Personalized navigation in folksonomies using hierarchical tag clustering”, AAAI Workshop, 2008.

[27] E. Simpson, “Clustering tags in enterprise and web folksonomies”, HPL-2007-190, 2007.

[28] M. Grineva, M. Grinev, D. Turdakov, P. Velikhov, “Harnessing Wikipedia for smart tags clustering”, AAAI, 2008.

[29] M. Varma, D. Ray, “Learning the discriminative power-invariance trade-off”, IEEE ICCV, 2007.

[30] A. Frome, Y. Singer, F. Sha, J. Malik, “Learning globally-consistent local distance functions for shape-based image retrieval and classification”, IEEE ICCV, 2007.

[31] A. Bosch, A. Zisserman, X. Munoz, “Representing shape with a spatial pyramid kernel”, ACM CIVR, 2007.

[32] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, “Local features and kernels for classification of texture and object catetories: A comprehensive study”, *Intl. Journal of Computer Vision*, vol.73, no.2, 213-238, 2007.

[33] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods”, Technical Report, CSD-TR-03-02, University of London, 2003.

[34] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, S. Li, “Flickr distance”, ACM Multimedia, 2008.