Chapter 11

TEXT MINING IN MULTIMEDIA

Zheng-Jun Zha
School of Computing, National University of Singapore
zhazj@comp.nus.edu.sg

Meng Wang
School of Computing, National University of Singapore
wangm@comp.nus.edu.sg

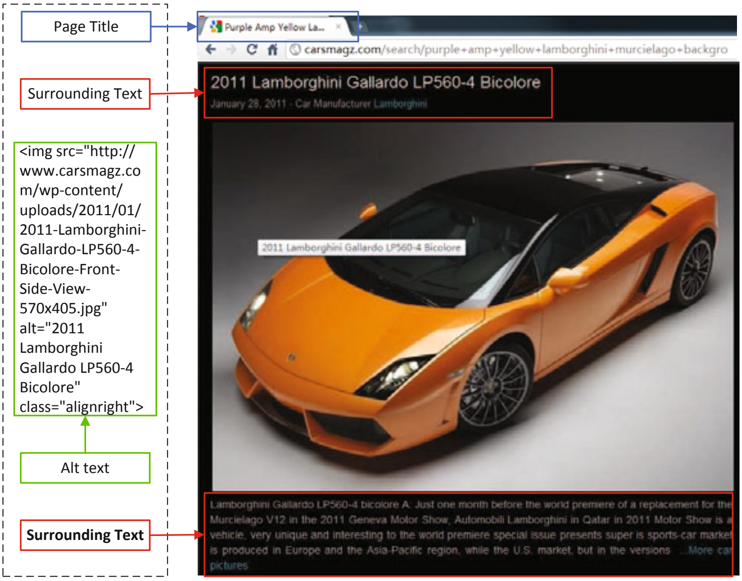
Jialie Shen
Singapore Management University
jlshen@smu.edu.sg

Tat-Seng Chua
School of Computing, National University of Singapore
chuats@comp.nus.edu.sg

Abstract A large amount of multimedia data (e.g., image and video) is now available on the Web. A multimedia entity does not appear in isolation, but is accompanied by various forms of metadata, such as surrounding text, user tags, ratings, and comments etc. Mining these textual metadata has been found to be effective in facilitating multimedia information processing and management. A wealth of research efforts has been dedicated to text mining in multimedia. This chapter provides a comprehensive survey of recent research efforts. Specifically, the survey focuses on four aspects: (a) surrounding text mining; (b) tag mining; (c) joint text and visual content mining; and (d) cross text and visual content mining. Furthermore, open research issues are identified based on the current research efforts.

Keywords: Text Mining, Multimedia, Surrounding Text, Tagging, Social Network

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_11,


Page Title
Surrounding Text
Alt text
Surrounding Text
Purple Amp Yellow La...
carsmagz.com/search/purple+amp+yellow+lamborghini+murcielago+backgro
2011 Lamborghini Gallardo LP560-4 Bicolore

```
<img src="http://www.carsmagz.com/wp-content/uploads/2011/01/2011-Lamborghini-Gallardo-LP560-4-Bicolore-Front-Side-View-570x405.jpg" alt="2011 Lamborghini Gallardo LP560-4 Bicolore" class="alignright">
```

2011 Lamborghini Gallardo LP560-4 Bicolore

January 28, 2011 - Car Manufacturer Lamborghini

Lamborghini Gallardo LP560-4 bicolore A: Just one month before the world premiere of a replacement for the Murcielago V12 in the 2011 Geneva Motor Show, Automobili Lamborghini in Qatar in 2011 Motor Show is a vehicle, very unique and interesting to the world premiere special issue presents super is sports-car market is produced in Europe and the Asia-Pacific region, while the U.S. market, but in the versions ...More car pictures

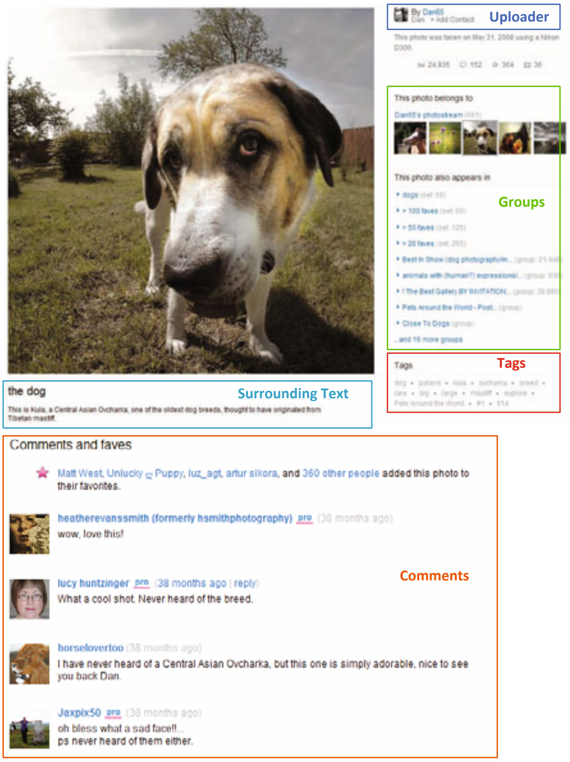
Figure 11.1. Illustration of textual metadata of an embedded image in a Web page.

1. Introduction

Lower cost hardware and growing communications infrastructure (e.g. Web, cell Phones, etc.) have led to an explosion in the availability of ubiquitous devices to produce, store, view and exchange multimedia entities (images, videos). A large amount of image and video data are now available. Take one of the most popular photo sharing services Flickr [1] as example, it has accumulated several billions of images. Another example is Youtube [2], which is a video sharing Web site that is hosting billions of videos. As the largest photo sharing site, Facebook [3] currently stores hundreds of hundreds of billions of photos.

On the other hand, a multimedia entity does not appear in isolation but is accompanied by various forms of textual metadata. One of the most typical examples is the surrounding text appearing around the embedded images or videos in the Web page (See Figure 11.1). With recent proliferation of social media sharing services, the newly emerging textual meatadata include user tags, ratings, comments, as well as

[1] http://www.flickr.com/
[2] http://www.youtube.com/
[3] http://www.facebook.com/



By Dan65
Dan • Add Contact

This photo was taken on May 31, 2008 using a Nikon D300.

24,935 152 364 36

This photo belongs to

Dan65's photostream (551)

This photo also appears in

- dogs (set 56)
- > 100 faves (set 63)
- > 50 faves (set 125)
- > 20 faves (set 265)
- Best in Show (dog photographyin... (group: 21,448)
- animals with (human?) expressions!... (group: 836)
- ! The Best Gallery BY INVITATION... (group: 28,980)
- Pets Around the World - Post... (group)
- Close To Dogs (group)

...and 16 more groups

Tags

dog • poland • kula • ovcharka • breed • cara • big • large • mastiff • explore • Pets Around the World • #1 • 514

the dog

This is Kula, a Central Asian Ovcharka, one of the oldest dog breeds, thought to have originated from Tibetan mastiff.

Comments and faves

Matt West, Unlucky Puppy, luz_agt, artur sikora, and 360 other people added this photo to their favorites.

heatherevanssmith (formerly hsmithphotography) pro (38 months ago)
wow, love this!

lucy huntzinger pro (38 months ago | reply)
What a cool shot. Never heard of the breed.

horselovertoo (38 months ago)
I have never heard of a Central Asian Ovcharka, but this one is simply adorable, nice to see you back Dan.

Jaxpix50 pro (38 months ago)
oh bless what a sad face!!...
ps never heard of them either.

Figure 11.2. Illustration of textual metadata of an image on a photo sharing Web site.

the information about the uploaders and their social network (See Figure 11.2). These metadata, in particular the tags, have been found to be an important resource for facilitating multimedia information processing and management. Given the wealth of research efforts that has been done, there have been various studies in multimedia community on the mining of textual metadata. In this chapter, a multimedia entity refers to an image or a video. For the sake of simplicity and without lost of generality, we use the term image to refer to multimedia entity for the rest of this chapter.

In this chapter, we first review the related works on mining surrounding text for image retrieval as well as the recent research efforts that explore surrounding text for image annotation and clustering in Section 2. In Section 3, we provide a literature review on tag mining and show that the main focus of existing tag mining works includes three aspects: tag ranking, tag refinement, and tag information enrichment. In

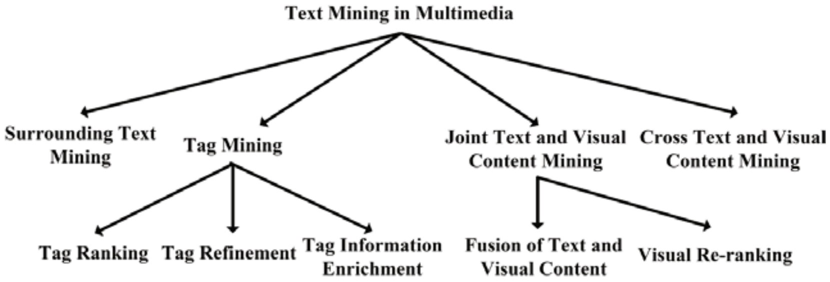

Text Mining in Multimedia
Surrounding Text Mining
Tag Mining
Joint Text and Visual Content Mining
Cross Text and Visual Content Mining
Tag Ranking
Tag Refinement
Tag Information Enrichment
Fusion of Text and Visual Content
Visual Re-ranking

Figure 11.3. A taxonomy consisting of the research works reviewed in this chapter.

Section 4, we survey the recent progress in integrating textual metadata and visual content. We categorize the exiting works into two categories: the fusion of text and visual content as well as visual re-ranking. In Section 5, we provide a detailed discussion on recent research on cross text and visual content mining. We organize all the works reviewed in this chapter into a taxonomy as shown in Figure 11.3. The taxonomy provides an overview of state-of-the-art research and helps us to identify open research issues to be presented in Section 6.

2. Surrounding Text Mining

In order to enhance the content quality and improve user experience, many hosting Web pages include different kinds of multimedia entities, like image or video. These multimedia entities are frequently embedded as part of the text descriptions which we called the surrounding text. While there is no standard definition, surrounding text generally refers to the text consisting of words, phrases or sentences that surrounds or close to the embedded images, such as those that appear at the top, below, left or right region of images or connected via Web links. The effective use of surrounding texts is becoming increasingly important for multimedia retrieval. However, developing effective extraction algorithm for the comprehensive analysis of surrounding text has been a very challenging task. In many cases, automatically determining which page region is more relevant to the image than the others could be difficult. Moreover, how large the region nearby should be considered is still an open question. Further, the quality of surrounding texts could be low and inconsistent. These problems make it very hard to directly apply the surrounding text information to facilitate accurate retrieval. Thus, refinement process or combining it with other cues is essential.

The earliest efforts on modeling and analyzing surrounding texts to facilitate multimedia retrieval occurred in the 1990s. AltaVista's A/V

Photo Finder applies textual and visual cues to index image collections [1]. The indexing terms are precomputed based on the HTML documents containing the Web images. With a similar approach, the WebSeer system harvests the information for indexing Web images from two different sources: the related HTML text and the embedded image itself [12]. It extracts keywords from page title, file name, caption, alternative text, image hyperlinks, and body text titles. A weight is calculated for each keyword based on its location inside a page. In PICITION system [40], an interesting approach is developed to exploit both textual and visual information to index a pictorial database. Image captions are used as an important cue to identify faces appearing in a related newspaper photograph. The empirical study based on a data set containing 50 pictures and captions obtained from the $Buffalo\ News$ and the $New\ York\ Times$ is used to demonstrate the effectiveness of the PICITION system. While the system can be successfully adopted for accessing photographs in newspaper or magazine, it is not straightforward to apply it for Web image retrieval.

In [39], Smith and Chang proposed the WebSeek framework designed to search images from the Web. The key idea is to analyze and classify the Web multimedia objects into a predefined taxonomy of categories. Thus, an initial search can be performed to explore a catalog associated with the query terms. The image attribute (e.g., color histogram for images) is then computed for similarity matching within the category.

Besides its efficacy in image retrieval, surrounding text has been explored for image annotation recently. Feng et al. presented a bootstrapping framework to label and search Web images based on a set of predefined semantic concepts [9]. To achieve better annotation effectiveness, a co-training scheme is designed to explore the association between the text features computed using corresponding HTML documents and visual features extracted from image content. Observing that the links between the visual content and the surrounding texts can be modeled via Web page analysis, a novel method called Iterative Similarity Propagation is proposed to refine the closeness between the Web images and their annotations [50]. On the other hand, it is not hard to find that images from the same cluster may share many similar characteristics or patterns with respect to relevance to information needs. Consequently, accurate clustering is a very crucial technique to facilitate Web multimedia search and many algorithms have recently been proposed based on the analysis of surrounding texts and low level visual features [3][13][34]. For example, Cai et al. [3] proposed a hierarchical clustering method that exploits visual, textual, and link analysis. A webpage is partitioned into blocks, and the textual and link information of an image are extracted from the block containing that image. By using block-level link analysis techniques, an image graph is constructed. They then applied spectral techniques to find a Euclidean embedding of the images. As a result, each image has three types of representations: visual feature, textual feature, and graph-based representation. Spectral clustering techniques are employed to cluster search results into various clusters. Gao et al. [13] and Rege et al. [34] used a tripartite graph to model the relations among visual features, images and their surrounding text. The clustering is performed by partitioning this tripartite graph.

3. Tag Mining

In newly emerging social media sharing services, such as the Flickr and Youtube, users are encouraged to share multimedia data on the Web and annotate content with tags. Here a tag is referred to as a descriptive keyword that describes the multimedia content at semantic or syntactic level. These tags have been found to be an important resource for multimedia management and have triggered many innovative research topics [61][51][38][36]. For example, with accurate tags, the retrieval of multimedia content can be easily accomplished. The tags can be used to index multimedia data and support efficient tag-based search. Nowadays, many online media repositories, such as Flickr and Youtube, support tag-based multimedia search. However, since the tags are provided by grassroots Internet users, they are often noisy and incomplete and there is still a gap between these tags and the actual content of the images[20][26][48]. This deficiency has limited the effectiveness of tag-based applications.

Recently, a wealth of research has been proposed to enhance the quality of human-provided tags. The existing works mainly focus on the following three aspects: (a) tag ranking, which aims to differentiate the tags associated with the images with various levels of relevance; (b) tag refinement with the purpose to refine the unreliable human-provided tags; and (c) tag information enrichment, which aims to supplement tags with additional information [26]. In this section, we present a comprehensive review of existing tag ranking, tag refinement, and tag information enrichment methods.

3.1 Tag Ranking

As shown in [25], the relevance level of the tags cannot be distinguished from the tag list of an image. The lack of relevance information in the tag list has limited the application of tags. Recently, tag ranking has been studied to infer the relevance levels of tags associated with an

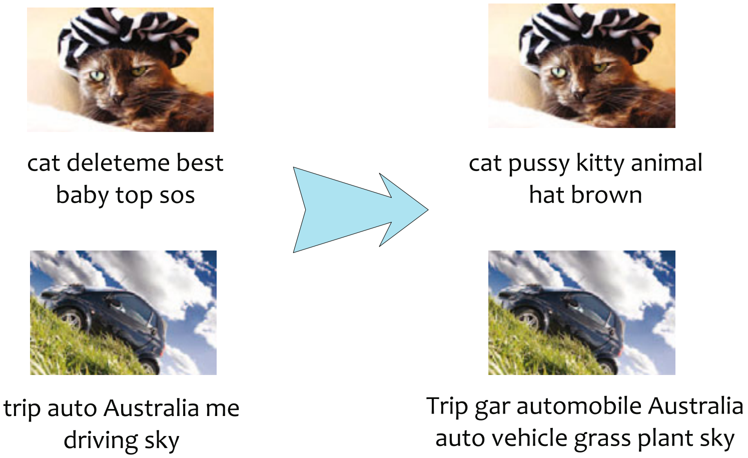

cat deleteme best baby top sos
cat pussy kitty animal hat brown
trip auto Australia me driving sky
Trip gar automobile Australia auto vehicle grass plant sky

Figure 11.4. Examples of of tag refinement. The left side of the figure shows the original tags while the right side shows the refined tags. The technique is able to remove irrelevant tags and add relevant tags to obtain better description of multimedia contents.

image. As a pioneering work, Liu et al. [25] proposed to estimate tag relevance scores using kernel density estimation, and then employ random walk to boost this primary estimation. Li et al. [22] proposed a data driven method for tag ranking. They learned the relevance scores of tags by a neighborhood voting approach. Given an image and one of its associated tag, the relevance score is learned by accumulating the votes from the visual neighbors of the image. They then extended the work to multiple visual spaces [23]. They learned the relevance scores of tags and ranked them by neighborhood voting in different feature spaces, and the results are aggregated with a score fusion or rank fusion method. Different aggregation methods have been investigated, such as the average score fusion, Borda count and RankBoost. The results show that a simple average fusion of scores is already able to perform closed to supervised fusion methods like RankBoost.

3.2 Tag Refinement

User-provided tags are often noisy and incomplete. The study in [20] shows that when a tag appears in a Flickr image, there is only about a 50% chance that the tag is really relevant, and the study in [38] shows that more than half of Flickr images are associated with less than three tags. Tag refinement technologies are proposed aiming at obtaining more accurate and complete tags for multimedia description, as shown in Figure 11.4.

A lot of tag refinement approaches have been developed based on various statistical learning techniques. Most of them are based on the following three assumptions.

- The refined tags should not change too much from those provided by the users. This assumption is usually used to regularize the tag refinement.

- The tags of visually similar images should be closely related. This is a natural assumption that most automatic tagging methods are also built upon.

- Semantically close or correlative tags should appear with high correlation. For example, when a tag "sea" exists for an image, the tags "beach" and "water" should be assigned with higher confidence while the tag "street" should have low confidence.

For example, Chen et al. [6] first trained a SVM classifier for each tag with the loosely labeled positive and negative samples. The classifiers are used to estimate the initial relevance scores of tags. They then refined the scores with a graph-based method that simultaneously considers the similarity between images and semantic correlation among tags. Xu et al. [52] proposed a tag refinement algorithm from topic modeling point of view. A new graphical model named regularized latent Dirichlet allocation (rLDA) is presented to jointly model the tag similarity and tag relevance. Zhu et al. [64] proposed a matrix decomposition method. They used a matrix to represent the image-tag relationship: the (i, j)-th element is 1 if the i-th image is associated with the j-th tag, and 0 otherwise. The matrix is then decomposed into a refined matrix plus an error matrix. They enforced the error matrix to be sparse and the refined matrix to follow three principles: (a) let the matrix be low-rank; (b) if two images are visually similar, the corresponding rows are with high correlation; and (c) if two tags are semantically close, the corresponding vectors are with high correlation. Fan et al. [8] grouped images with a target tag into clusters. Each cluster is regarded as a unit. The initial relevance scores of the clusters are estimated and then refined by a random walk process. Liu et al. [24] adopted a three-step approach. The first step filters out tags that are intrinsically content-unrelated based on the ontology in WordNet. The second step refines the tags based on the consistency of visual similarity and semantic similarity of images. The last step performs tag enrichment, which expands the tags with their appropriate synonyms and hypericum.

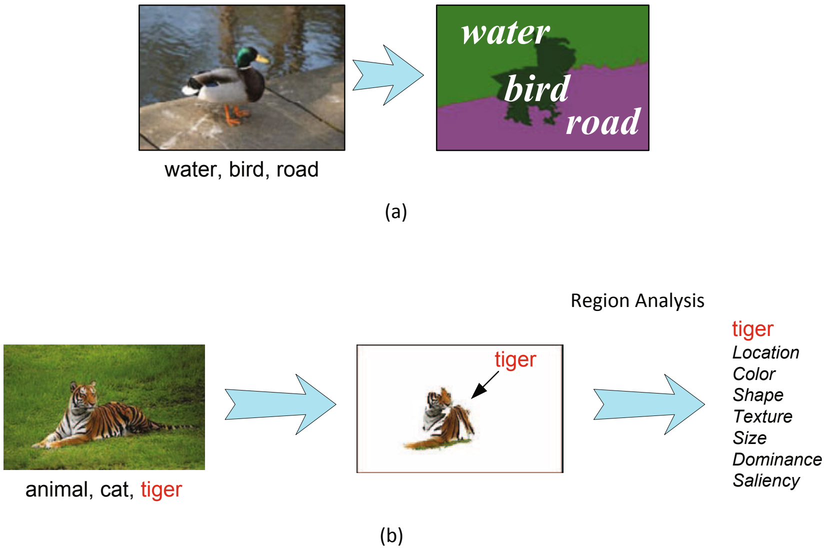


Figure 11.5. (a) An example of tag localization, which finds the regions that the tags describe. (b) An illustration of tag information enrichment. It first finds the corresponding region of the target tag and then analyze the properties of the region.

3.3 Tag Information Enrichment

In the manual tagging process, generally human labelers will only assign appropriate tags to multimedia entities without any additional information, such as the image regions depicted by the corresponding tags. But by employing computer vision and machine learning technologies, certain information of the tags, such as the descriptive regions and saliency, can be automatically obtained. We refer to these as tag information enrichment.

Most existing works employ the following two steps for tag information enrichment. First, tags are localized into regions of images or sub-clips of videos. Second, the characteristics of the regions or sub-clips are analyzed, and the information about the tags is enriched accordingly. Figure 11.5 (a) illustrates the examples of tag localization for image and video data. Liu et al. [28] proposed a method to locate image tags to corresponding regions. They first performed over-segmentation to decompose each image into patches and then discovered the relationship between patches and tags via sparse coding. The over-segmented regions are then merged to accomplish the tag-to-region process. Liu et al. extended the approach based on image search [29]. For a tag of the target image, they collected a set of images by using the tag as query with an image search engine. They then learned the relationship between the tag and the patches in this image set. The selected patches are used to reconstruct each candidate region, and the candidate regions are ranked based on the reconstruction error. Liu et al. [27] accomplished the tag-to-region task by regarding an image as a bag of regions and then performed tag propagation on a graph, in which vertices are images and edges are constructed based on the visual link of regions. Feng et al. [10] proposed a tag saliency learning scheme, which is able to rank tags according to their saliency levels to an image's content. They first located tags to images' regions with a multi-instance learning approach. In multi-instance learning, an image is regarded as a bag of multiple instances, i.e., regions [58]. They then analyzed the saliency values of these regions. It can provide more comprehensive information when an image is relevant to multiple tags, such as those describing different objects in the image. Yang et al. [55] proposed a method to associate a tag with a set of properties, including location, color, texture, shape, size and dominance. They employed a multi-instance learning method to establish the region that each tag is corresponding to, and the region is then analyzed to establish the properties, as shown in Figure 11.5 (b). Sun and Bhowmick [41] defined a tag's visual representativeness based on a large image set and the subset that is associated with the tag. They employed two distance metrics, cohesion and separation, to estimate the visual representativeness measure.

Ulges et al. [43] proposed an approach to localize video-level tags to keyframes. Given a tag, it regards whether a keyframe is relevant as a latent random variable. An EM-style process is then adopted to estimate the variables. Li et al. [21] employed a multi-instance learning approach to accomplish the video tag localization, in which video and shot are regarded as bag and shot, respectively.

By supplementing tags with additional information, a lot of tag-based applications can be facilitated, such as tag-based image/video retrieval and intelligent video browsing etc.

4. Joint Text and Visual Content Mining

Beyond mining pure textual metadata, researchers in multimedia community have started making progress in integrating text and content for multimedia retrieval via joint text and content mining. The integration of text and visual content has been found to be more effective than exploiting purely text or visual content separately. The joint text and content mining in multimedia retrieval often comes down to finding effective mechanisms for fusing multi-modality information from textual metadata and visual content. Existing research efforts can generally be categorized into four paradigms: (a) linear fusion; (b) latent-space-based fusion; (c) graph-based fusion; and (d) visual re-ranking that exploits visual information to refine text-based retrieval results. In this section, we first briefly review linear, latent space based, and graph based fusion methods and then provide comprehensive literature review on visual re-ranking technology.

Linear fusion combines the retrieval results from various modalities linearly [18][4][31]. In [18], visual content and text are combined in both online learning stage with relevance feedback and offline keyword propagation. In [31], linear, max, and average fusion strategies are employed to aggregate the search results from visual and textual modalities. Chang et al. [4] adopted a query-class-dependent fusion approach. The critical task in linear fusion is the estimation of fusion weights of different modalities. A certain amount of training data is usually required for estimating these weights. The latent space based fusion assumes that there is a latent space shared by different modalities and thus unify different modalities by transferring the features of these modalities into the shared latent space [63][62]. For example, Zhao et al. [63] adopted the Latent Semantic Indexing (LSI) method to fuse text and visual content. Zhang et al. [62] proposed a probabilistic context model to explicitly exploit the synergy between text and visual content. The synergy is represented as a hidden layer between the image and text modalities. This hidden layer constitutes the semantic concepts to be annotated through a probabilistic framework. An Expectation-Maximization (EM) based iterative learning procedure is developed to determine the conditional probabilities of the visual features and the words given a hidden concept class. Latent space based methods usually require a large amount of training samples for learning the feature mapping from each modality into the unified latent space. Graph based approach [49] first builds the relations between different modalities, such as relations between images and text using the Web page structure. The relations are then utilized to iteratively update the similarity graphs computed from different modalities. The difficulty of creating similarity graphs for billions of images on the Web makes this approach insufficiently scalable.

4.1 Visual Re-ranking

Visual re-ranking is emerging as one of the promising technique for automated boosting of retrieval precision [42] [30] [55]. The basic functionality is to reorder the retrieved multimedia entities to achieve the optimal rank list by exploiting visual content in a second step. In particular, given a textual query, an initial list of multimedia entities is returned using the text-based retrieval scheme. Subsequently, the most relevant results are moved to the top of the result list while the less relevant ones are reordered to the lower ranks. As such, the overall search precision at the top ranks can be enhanced dramatically. According to the statistical analysis model used, the existing re-ranking approaches can roughly be categorized into three categories including the clustering based, classification based and graph based methods.

Cluster analysis is very useful to estimate the inter-entity similarity. The clustering based re-ranking methods stem from the key observation that a lot of visual characteristics can be shared by relevant images or video clips. With intelligent clustering algorithms (e.g., mean-shift, K-means, and K-medoids), initial search results from text-based retrieval can be grouped by visual closeness. One good example of clustering based re-ranking algorithms is an Information Bottle based scheme developed by Hsu et al. [16]. Its main objective is to identify optimal clusters of images that can minimize the loss of mutual information. The cluster number is manually configurated to ensure the each cluster contains the same number of multimedia entities (about 25). This method was evaluated using the TRECVID 2003-2005 data and significant improvements were observed in terms of MAP measures. In [19], a fast and accurate scheme is proposed for grouping Web image search results into semantic clusters. For a given query, a few related semantic clusters are identified in the first step. Then, the cluster names relating to query are derived and used as text keywords for querying image search engine. The empirical results from a set of user studies demonstrate an improvement in performance over Google image search results. It is not hard to show that the clustering based re-ranking methods can work well when the initial search results contain many near-duplicate media documents. However, for queries that return highly diverse results or without clear visual patterns, the performance of the clustering-based methods is not guaranteed. Furthermore, the number of clusters has large impact on the final effectiveness of the algorithms. However, determining the optimal cluster number automatically is still an open research problem.

In the classification based methods, visual re-ranking is formulated as a binary classification problem aiming to identify whether each search result is relevant or not. The major process for result list reordering consists of three major steps: (a) the selection of pseudo-positive and pseudo-negative samples; (b) use the samples obtained in step (a) to train a classification scheme; and (c) reorder the samples according to their relevance scores given by the trained classifier. For existing classification methods, pseudo relevance feedback (PRF) is applied to select the training examples. It assumes that: (a) a limited number of top-ranked entities in the initial retrieval results are highly relevant to the search queries; and (b) automatic local analysis over the entities can be very helpful to refine query representation. In [54], the query images or video clip examples are used as the pseudo-positive samples. The pseudo-negative samples are selected from either the least relevant samples in the initial result list or the databases that contain less samples related to the query. The second step of the classification based methods aim to train classifiers and a wide range of statistical classifiers can be adopted. They include the Support Vector Machine (SVM) [54], Boosting [53] and ListNet [57]. The main weakness for the classification based methods is that the number and quality of training data required play a very important role in constructing effective classifiers. However, in many real scenarios, the training examples obtained via PRF are very noisy and might not be adequate for training effective classifier. To address this issue, Fergus et al. [11] used RANSAC to sample a training subset with a high percentage of relevant images. A generative constellation model is learned for the query category while a background model is learned from the query "things". Images are re-ranked based on their likelihood ratio. Observing that discriminative learning can lead to superior results, Schroff et al. [35] first learned a query independent text based re-ranker. The top ranked results from the text based re-ranking are then selected as positive training examples. Negative training examples are picked randomly from the other queries. A binary SVM classifier is then used to re-rank the results on the basis of visual features. This classifier is found to be robust to label noise in the positive training set as long as the non-relevant images are not visually consistent. Better training data can be obtained from online knowledge resources if the set of queries restricted. For instance, Wang et al. [44] learned a generative text model from the query's Wikipedia [4] page and a discriminative image model from the Caltech [15] and Flickr data sets. Search results are then re-ranked on the basis of these learned probability models. Some user interactions are required to disambiguate the query.

Graphs provide a natural and comprehensive way to explore complex relations between data at different levels and have been applied to a wide range of applications [59][46][47][60]. With the graph based re-ranking methods, the multimedia entities in top ranks and their associations/dependencies can be represented as a collection of nodes (vertices) and edges. The local patterns or salient features discover using graph

[4] http://www.wikipedia.org/

analysis are very helpful to improve effectiveness of rank lists. In [16], Hsu et al. modeled the re-ranking process as a random walk over the context graph. In order to effectively leverage the retrieved results from text search, each sample corresponds to a "dongle" node containing ranking score based on text. For the framework, edges between "dongle" nodes are weighted with multi-modal similarities. In many cases, the structure of large scale graphs can be very complex and this easily makes related analysis process very expensive in terms of computational cost. Thus, Jing and Baluja proposed a VisualRank framework to efficiently model similarity of Google image search results with graph [17]. The framework casts the re-ranking problem as random walk on an affinity graph and reorders images according to the visual similarities. The final result list is generated via sorting the images based on graph nodes' weights. In [42], Tian et al., presented a Bayesian video search re-ranking framework formulating the re-ranking process as an energy minimization problem. The main design goal is to optimize the consistency of ranking scores over visually similar videos and minimize the disagreement between the optimal list and the initial list. The method achieves a consistently better performance over several earlier proposed schemes on the TRECVID 2006 and 2007 data sets. The graph based re-ranking algorithms mentioned above generally do not consider any initial supervision information. Thus, the performance is significantly dependent on the statistical properties of top ranked search results. Motivated by this observation, Wang et al, proposed a semi-supervised framework to refine the text based image retrieval results via leveraging the data distribution and the partial supervision information obtained from the top ranked images [45]. Indeed, graph analysis has been shown to be a very powerful tool for analyzing and identifying salient structure and useful patterns inside the visual search results. With recent progresses in graph mining, this research stream is expected to continue to make important contributions to improve visual re-ranking from different perspectives.

5. Cross Text and Visual Content Mining

Although the joint text and visual content mining approaches described above facilitate image retrieval, they require that the test images have associated text modality. However, in some real world applications, images may not always have associated text. For example, most surveillance images/videos in in-house repository are not accompanied with any text. Even on social media Website such as the Flickr, there exist a substantial number of images without any tags. In such cases, joint

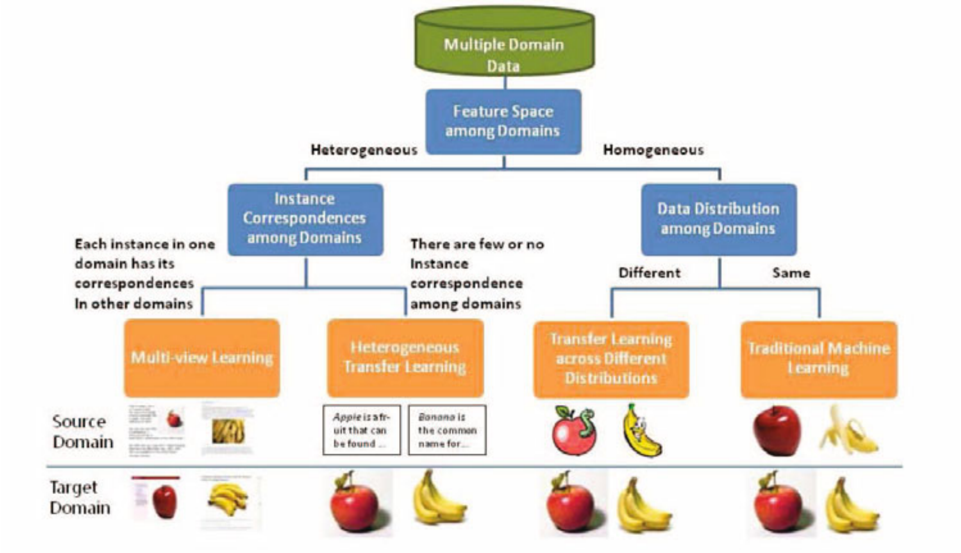

Multiple Domain Data
Feature Space among Domains
Heterogeneous
Homogeneous
Instance Correspondences among Domains
Data Distribution among Domains
Each instance in one domain has its correspondences In other domains
There are few or no Instance correspondence among domains
Different
Same
Multi-view Learning
Heterogeneous Transfer Learning
Transfer Learning across Different Distributions
Traditional Machine Learning
Source Domain
Target Domain
Apple is a fr-uit that can be found ...
Banana is the common name for...

Figure 11.6. An illustration of different types of learning paradigms using image classification/clustering in the domains of apple and banana. Adapted from [56].

text and visual content mining cannot be applied due to missing text modality.

Recently, cross text and visual content mining has been studied in the context of transfer learning techniques. This class of techniques emphasizes the transferring of knowledge across different domains or tasks [32]. Cross text and visual content mining does not require that a test image has an associated text modality, and is thus beneficial to dealing with the images without any text by propagating the semantic knowledge from text to images [5]. It is also motivated by two observations. First, visual content of images is much more complicated than the text feature. While the textual words are easier to interpret, there exist a tremendous semantic gap between visual content and high-level semantics. Second, image understanding becomes particularly challenging when only a few labeled images are available for training. This is a common challenge, since it is expensive and time-consuming to obtain labeled images. On the contrary, labeled/unlabeled text data are relatively easier to collect. For example, millions of categorized text articles are freely available in Web

[5] Cross text and visual content can also facilitate text understanding in special cases by propagating knowledge from images to text.

text collections, such as Wikipedia, covering a wide range of topics from culture and arts, geography and places, history and events, to natural and physical science. A large number of Wikipedia articles are indexed by thousands of categories in these topics [33]. This provides abundant labeled text data. Thus, it is desirable to propagate semantic knowledge from text to images to facilitate image understanding. However, it is not trivial to transfer knowledge between various domains/tasks due to the following challenges:

- The target data may be drawn from a distribution different from the source data.
- The target and source data may be in different feature spaces (e.g., image and text) and there may be no correspondence between instances in these spaces.
- The target and source tasks may have different output spaces.

While the traditional transfer learning techniques focus on the distribution variance problem, the recent proposed heterogenous transfer learning approaches aim to tackle both the distribution variance and heterogenous feature space problems [56][7][65][33], or all the three challenges listed above [37]. Figure 11.6 from [56] presents an intuitive illustration of four learning paradigms, including traditional machine learning, transfer learning across different distributions, multi-view learning and heterogenous transfer learning. As we can see, heterogenous transfer learning is usually much more challenging due to the unknown correspondence across the distinct feature spaces. In order to learn the underlying correspondence for knowledge transformation, a "semantic bridge" is required. The "semantic bridge" can be obtained from the co-occurrence information between text and images or the linkage information in social media networks. For example, while the traditional webpages provide the co-occurrence information between text and images, the social media sites contain a large number of linked information between different types of entities, such as the text articles, tags, posts, images and videos. This linkage information provide a "semantic bridge" to learn the underlying correspondence [2].

Most existing works exploit the tag information that provide text-to-image linking information. As a pioneering work, Dai et al. [7] showed that such information can be effectively leveraged for transferring knowledge between text and images. The key idea of [7] is to construct a correspondence between the images and the auxiliary text data with the use of tags. Probabilistic latent semantic analysis (PLSA) model is employed to construct a latent semantic space which can be used for transferring knowledge. Chen et al. [56] proposed the concept of heterogeneous transfer learning and applied it to improve image clustering by leveraging auxiliary text data. They collected annotated images from the social web, and used them to construct a text to image mapping. The algorithm is referred to as aPLSA (Annotated Probabilistic Latent Semantic Analysis). The key idea is to unify two different kinds of latent semantic analysis in order to create a bridge between the text and images. The first kind of technique performs PLSA analysis on the target images, which are converted to an image instance-to-feature co-occurrence matrix. The second kind of PLSA is applied to the annotated image data from social Web, which is converted into a text-to-image feature co-occurrence matrix. In order to unify those two separate PLSA models, these two steps are done simultaneously with common latent variables used as a bridge linking them. It has been shown in [5] that such a bridging approach leads to much better clustering results. Zhu et al. [65] discussed how to create the connections between images and text with the use of tag data. They showed how such links can be used more effectively for image classification. An advantage of [65] is that it exploits unlabeled text data instead of labeled text as in [7].

In contrast to these methods that exploit tag information to link images and auxiliary text articles, Qi et al. [33] proposed to learn a "translator" which can directly establish the semantic correspondence between text and images even if they are new instances of the image data with unknown correspondence to the text articles. This capability increase the flexibility of the approach and makes it more widely applicable. Specifically, they created a new topic space into which both the text and images are mapped. A translator is then learned to link the instances across heterogeneous text and image spaces. With the resultant translator, the semantic labels can be propagated from any labeled text corpus to any new image by a process of cross-domain label propagation. They showed that the learned translator can effectively convert the semantics from text to images.

6. Summary and Open Issues

In this chapter, we have reviewed the active research on text mining in multimedia community, including surrounding text mining, tag mining, joint text and visual content mining, and cross text and visual content mining. Although research efforts in this filed have made great progress in various aspects, there are still many open research issues that need to be explored. Some examples are listed and discussed as follows.

Joint text and visual content multimedia ranking

Despite the success of visual re-ranking in multimedia retrieval, visual re-ranking only employs the visual content to refine text-based retrieval results; visual content has not been used to assist in learning the ranking model of search engine, and sometimes it is only able to bring in limited performance improvements. In particular, if text-based ranking model is biased or over-fitted, re-ranking step will suffer from the error that is propagated from the initial results, and thus the performance improvement will be negatively impacted. Therefore, it is worthwhile to simultaneously exploit textual metadata and visual content to learn a unified ranking model. A preliminary work has been done in [14], where a content-aware ranking model is developed to incorporate visual content into text-based ranking model learning. It shows that the incorporation of visual content into ranking model learning can result in a more robust and accurate ranking model since noise in textual features can be suppressed by visual information.

Scalable text mining for large-scale multimedia management

Despite of the success of existing text mining in multimedia, most existing techniques suffer from difficulties in handling large-scale multimedia data. Huge amount of training data or high computation powers are usually required by existing methods to achieve acceptable performance. However, it is too difficult, or even impossible, to meet this requirement in real-world applications. Thus there is a compelling need to develop scalable text mining techniques to facilitate large-scale multimedia management.

Multimedia social network mining

In recent years, we have witnessed the emergence of multimedia social network communities like Napster [6], Facebook [7], and Youtube, where millions of users and billions of multimedia entities form a large-scale multimedia social network. Multimedia social networking is becoming an important part of media consumption for Internet users. It brings in new and rich metadata, such as user preferences, interests, behaviors, social relationships, and social network structure etc. These information present new potential for advancing current multimedia analysis

[6] http://music.napster.com/

[7] http://www.facebook.com/

techniques and also trigger diverse multimedia applications. Numerous research topics can be explored, including (a) the combination of conventional techniques with information derived from social network communities; (b) fusion analysis of content, text, and social network data; and (c) personalized multimedia analysis in social networking environments.

Acknowledgements

This work was in part supported by A*Star Research Grant R-252-000-437-305 and NRF (National Research Foundation of Singapore) Research Grant R-252-300-001-490 under the NExT Center.

References

[1] Altavista's a/v photo finder. http://www.altavista.com/sites/search/simage.

[2] C. C. Aggarwal, H. Wang. *Text Mining in Social Networks.* Social Network Data Analytics, Springer, 2011.

[3] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *Proceedings of the ACM Conference on Multimedia*, 2004.

[4] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia university trecvid-2006 video search and high-level feature extraction. In *Proceedings of NIST TRECVID workshop*, 2006.

[5] L. Chen and A. Roy. Event detection from Flickr data through wavelet-based spatial analysis. In *Proceedings of the ACM conference on Information and knowledge management*, pages 523–532. ACM, 2009.

[6] L. Chen, D. Xu, I. W. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[7] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu. Translated learning: Transfer learning across difference feature spaces. In *NIPS*, pages 353–360, 2008.

[8] J. Fan, Y. Shen, N. Zhou, and Y. Gao. Harvesting large-scaleweakly-tagged image databases from the web. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[9] H. Feng, R. Shi, and T.-S. Chua. A bootstrapping framework for annotating and retrieving www images. In *Proceedings of the ACM Conference on Multimedia*, 2004.

[10] S. Feng, C. Lang, and D. Xu. Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking. In *Proceedings of International Conference on Image and Video Retrieval*, 2010.

[11] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *Proceedings of the European Conference on Computer Vision*, 2004.

[12] C. Frankel, M. J. Swain, and V. Athitsos. Webseer: An image search engine for the world wide web. Technical report, University of Chicago, Computer Science Department, 1996.

[13] B. Gao, T.-Y. Liu, Q. Tao, X. Zheng, Q. Cheng, and W.-Y. Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the ACM Conference on Multimedia*, 2005.

[14] B. Geng, L. Yang, C. Xu, and X.-S. Hua. Content-aware ranking for visual search. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 2010.

[15] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[16] W. Hsu, L. Kennedy, , and S.-F. Chang. Reranking methods for visual search. *IEEE Multimedia*, 14:14–22, 2007.

[17] F. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1877–1890, 2008.

[18] F. Jing, M. Li, H.-J. Zhang, and B. Zhang. A unified framework for image retrieval using keyword and visual features. *IEEE Transactions on Image Processing*, 2005.

[19] F. Jing, C. Wang, Y. Yao, K. Deng, L. Zhang, and W.-Y. Ma. Igroup: Web image search results clustering. In *Proceedings of the ACM Conference on Multimedia*, pages 377–384, 2006.

[20] L. S. Kennedy, S. F. Chang, and I. V. Kozintsev. To search or to label? predicting the performance of search-based automatic image classifiers. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*, 2006.

[21] G. Li, M. Wang, Y. T. Zheng, Z.-J. Zha, H. Li, and T.-S. Chua. Shottagger: Tag location for internet videos. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2011.

[22] X. Li, C. G. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *Pattern Recognition Letters*, 11(7), 2009.

[23] X. Li, C. G. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *Proceedings of the International Conference on Image and Video Retrieval*, 2010.

[24] D. Liu, X. C. Hua, M. Wang, and H. Zhang. Image retagging. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[25] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the International Conference on World Wide Web*, 2009.

[26] D. Liu, X.-S. Hua, and H.-J. Zhang. Content-based tag processing for internet social images. *Multimedia Tools and Application*, 51:723–738, 2010.

[27] D. Liu, S. Yan, Y. Rui, and H. J. Zhang. Unified tag analysis with multi-edge graph. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[28] X. Liu, B. Cheng, S. Yan, J. Tang, T. C. Chua, and H. Jin. Label to region by bi-layer sparsify priors. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[29] X. Liu, S. Yan, J. Luo, J. Tang, Z. Huang, and H. Jin. Nonparametric label-to-region by search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2010.

[30] Y. Liu, T. Mei, and X.-S. Hua. Crowdreranking: Exploring multiple search engines for visual search reranking. In *Proceedings of the ACM SIGIR Conference*, 2009.

[31] T. Mei, Z.-J. Zha, Y. Liu, M. Wang, and et al. Msra at trecvid 2008: High-level feature extraction and automatic search. In *Proceedings of NIST TRECVID workshop*, 2008.

[32] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 2010.

[33] G.-J. Qi, C. C. Aggarwal, and T. Huang. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the International Conference on World Wide Web*, 2011.

[34] M. Rege, M. Dong, and J. Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient

web image clustering. In *Proceedings of the International Conference on World Wide Web*, 2008.

[35] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting images databases from the web. In *Proceedings of the International Conference on Computer Vision*, 2007.

[36] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what i watch: using community activeity to understand content. In *Proceedings of the ACM Workshop on Multimedia Information Retrieval*, 2007.

[37] X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu. Transfer learning on heterogenous feature spaces via spectral tranformation. In *Proceedings of the International Conference on Data Mining*, 2010.

[38] B. Sigurbjörnsson and R. V. Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of International Conference on World Wide Web*, 2008.

[39] J. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4:12–20, 1995.

[40] R. Srihari. Automatic indexing and content-based retrieval of captioned images. *IEEE Computer*, 28:49–56, 1995.

[41] A. Sun and S. S. Bhowmick. Quantifying tag representativeness of visual content of social images. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[42] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[43] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *Proceedings of the International Conference on Image and Video Retrieval*, 2008.

[44] G. Wang and D. A. Forsyth. Object image retrieval by exploiting online knowledge resources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[45] J. Wang, Y.-G. Jiang, and S.-F. Chang. Label diagnosis through self tuning for web image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[46] M. Wang, X. S. Hua, R. Hong, J. Tang, G. J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(5), 2009.

[47] M. Wang, X. S. Hua, J. Tang, and R. Hong. Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Transactions on Multimedia*, 11(3), 2009.

[48] M. Wang, B. Ni, X.-S. Hua, and T.-S. Chua. Assistive multimedia tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Survey*, 2011.

[49] X.-J. Wang, W.-Y. Ma, G.-R. Xue, and X. Li. Multi-model similarity propagation and its application for web image retrieval. In *Proceedings of the ACM Conference on Multimedia*, pages 944–951, 2004.

[50] X.-J. Wang, W.-Y. Ma, L. Zhang, and X. Li. Iteratively clustering web images based on link and attribute reinforcements. In *Proceedings of the ACM Conference on Multimedia*, 2005.

[51] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li. Flickr distance. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[52] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag refinement by regularized LDA. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[53] R. Yan and A. G. Hauptmann. Co-retrieval: A boosted reranking approach for video retrieval. In *Proceedings of the ACM Conference on Image and Video Retrieval*, 2004.

[54] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *Proceedings of the ACM Conference on Image and Video Retrieval*, 2003.

[55] K. Yang, X.-S. Hua, M. Wang, and H. C. Zhang. Tagging tags. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[56] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu. Heterogeneous transfer learning from image clustering via the social web. In *Proceedings of the Joint Conference of the Annual Meeting of the ACL*, 2009.

[57] Y.-H. Yang, P. Wu, C. W. Lee, K. H. Lin, W. Hsu, and H. H. Chen. Contextseer: Context search and recommendation at query time for shared consumer photos. In *Proceedings of the ACM Conference on Multimedia*, 2008.

[58] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[59] Z.-J. Zha, T. Mei, J. Wang, X.-S. Hua, and Z. Wang. Graph-based semi-supervised learning with multiple labels. *Journal of Visual Communication and Image Representation*, 2009.

[60] Z.-J. Zha, M. Wang, Y.-T. Zheng, Y. Yang, R. Hong, and T.-S. Chua. Interactive video indexing with statistical active learning. *IEEE Transactions on Multimedia*, 2011.

[61] Z.-J. Zha, L. Yang, T. Mei, M. Wang, and Z. Wang. Viusal query suggestion. In *Proceedings of the ACM Conference on Multimedia*, 2009.

[62] R. Zhang, Z. M. Zhang, M. Li, W.-Y. Ma, and H.-J. Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *Proceedings of the International Conference on Computer Vision*, pages 846–851, 2005.

[63] R. Zhao and W. I. Grosky. Narrowing the semantic gap - improved text-based web document retireval using visual fetures. *IEEE Transactions on Multimedia*, 4, 2002.

[64] G. Zhu, S. Yan, and Y. Ma. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *Proceedings of the ACM Conference on Multimedia*, 2010.

[65] Y. Zhu, Y. Chen, Z. Lu, S. J. Pan, G.-R. Xue, Y. Yu, and Q. Yang. Heterogeneous transfer learning for image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2011.