Heterogeneous Transfer Learning for Image Clustering via the Social Web

Qiang Yang
Hong Kong University of Science and Technology, Clearway Bay, Kowloon, Hong Kong
qyang@cs.ust.hk

Yuqiang Chen **Gui-Rong Xue** **Wenyuan Dai** **Yong Yu**
Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
{yuqiangchen,grxue,dwyak,yyu}@apex.sjtu.edu.cn

Abstract

In this paper, we present a new learning scenario, *heterogeneous transfer learning*, which improves learning performance when the data can be in different feature spaces and where no correspondence between data instances in these spaces is provided. In the past, we have classified Chinese text documents using English training data under the heterogeneous transfer learning framework. In this paper, we present image clustering as an example to illustrate how unsupervised learning can be improved by transferring knowledge from auxiliary heterogeneous data obtained from the social Web. Image clustering is useful for image sense disambiguation in query-based image search, but its quality is often low due to image-data sparsity problem. We extend PLSA to help transfer the knowledge from social Web data, which have mixed feature representations. Experiments on image-object clustering and scene clustering tasks show that our approach in heterogeneous transfer learning based on the auxiliary data is indeed effective and promising.

1 Introduction

Traditional machine learning relies on the availability of a large amount of data to train a model, which is then applied to test data in the same feature space. However, labeled data are often scarce and expensive to obtain. Various machine learning strategies have been proposed to address this problem, including semi-supervised learning (Zhu, 2007), domain adaptation (Wu and Dietterich, 2004; Blitzer et al., 2006; Blitzer et al., 2007; Arnold et al., 2007; Chan and Ng, 2007; Daume, 2007; Jiang and Zhai, 2007; Reichart and Rappoport, 2007; Andreevskaia and Bergler, 2008), multi-task learning (Caruana, 1997; Reichart et al., 2008; Arnold et al., 2008), self-taught learning (Raina et al., 2007), etc. A commonality among these methods is that they all require the training data and test data to be in the same feature space. In addition, most of them are designed for supervised learning. However, in practice, we often face the problem where the labeled data are scarce in their own feature space, whereas there may be a large amount of labeled heterogeneous data in another feature space. In such situations, it would be desirable to transfer the knowledge from heterogeneous data to domains where we have relatively little training data available.

To learn from heterogeneous data, researchers have previously proposed multi-view learning (Blum and Mitchell, 1998; Nigam and Ghani, 2000) in which each instance has multiple views in different feature spaces. Different from previous works, we focus on the problem of *heterogeneous transfer learning*, which is designed for situation when the training data are in one feature space (such as text), and the test data are in another (such as images), and there may be no correspondence between instances in these spaces. The type of heterogeneous data can be very different, as in the case of text and image. To consider how heterogeneous transfer learning relates to other types of learning, Figure 1 presents an intuitive illustration of four learning strategies, including traditional machine learning, transfer learning across different distributions, multi-view learning and heterogeneous transfer learning. As we can see, an important distinguishing feature of heterogeneous transfer learning, as compared to other types of learning, is that more constraints on the problem are relaxed, such that data instances do not need to correspond anymore. This allows, for example, a collection of Chinese text documents to be classified using another collection of English text as the

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 1–9,
Suntec, Singapore, 2-7 August 2009.

training data (c.f. (Ling et al., 2008) and Section 2.1).

In this paper, we will give an illustrative example of heterogeneous transfer learning to demonstrate how the task of image clustering can benefit from learning from the heterogeneous social Web data. A major motivation of our work is Web-based image search, where users submit textual queries and browse through the returned result pages. One problem is that the user queries are often ambiguous. An ambiguous keyword such as "Apple" might retrieve images of Apple computers and mobile phones, or images of fruits. Image clustering is an effective method for improving the accessibility of image search result. Loeff et al. (2006) addressed the image clustering problem with a focus on image sense discrimination. In their approach, images associated with textual features are used for clustering, so that the text and images are clustered at the same time. Specifically, spectral clustering is applied to the distance matrix built from a multimodal feature set associated with the images to get a better feature representation. This new representation contains both image and text information, with which the performance of image clustering is shown to be improved. A problem with this approach is that when images contained in the Web search results are very scarce and when the textual data associated with the images are very few, clustering on the images and their associated text may not be very effective.

Different from these previous works, in this paper, we address the image clustering problem as a *heterogeneous transfer learning* problem. We aim to leverage heterogeneous auxiliary data, social annotations, etc. to enhance image clustering performance. We observe that the World Wide Web has many annotated images in Web sites such as Flickr (`http://www.flickr.com`), which can be used as auxiliary information source for our clustering task. In this work, our objective is to cluster a small collection of images that we are interested in, where these images are not sufficient for traditional clustering algorithms to perform well due to data sparsity and the low level of image features. We investigate how to utilize the readily available socially annotated image data on the Web to improve image clustering. Although these auxiliary data may be irrelevant to the images to be clustered and cannot be directly used to solve the data sparsity problem, we show that they can still be used to estimate a good *latent feature representation*, which can be used to improve image clustering.

2 Related Works

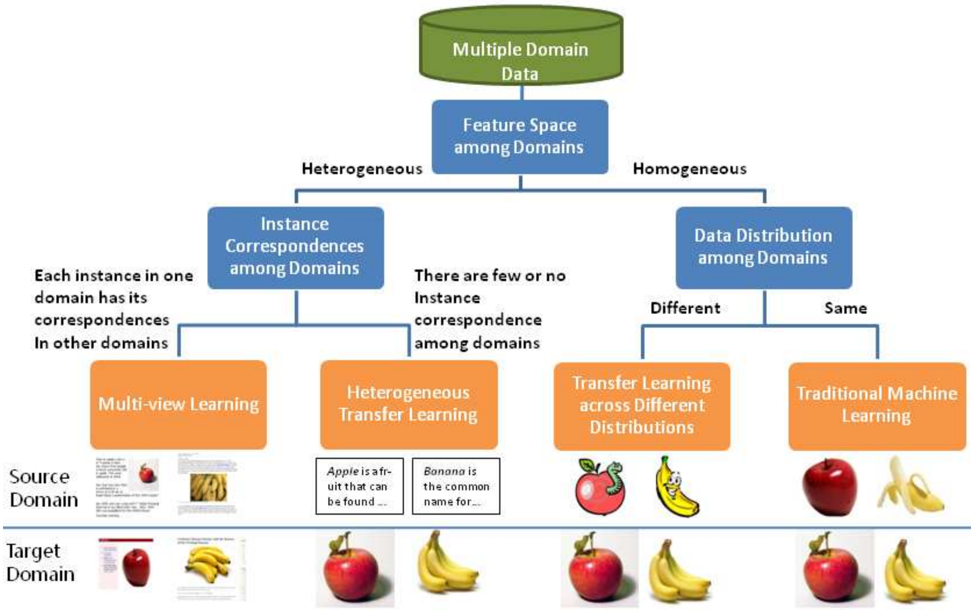
2.1 Heterogeneous Transfer Learning Between Languages

In this section, we summarize our previous work on cross-language classification as an example of heterogeneous transfer learning. This example is related to our image clustering problem because they both rely on data from different feature spaces.

As the World Wide Web in China grows rapidly, it has become an increasingly important problem to be able to accurately classify Chinese Web pages. However, because the labeled Chinese Web pages are still not sufficient, we often find it difficult to achieve high accuracy by applying traditional machine learning algorithms to the Chinese Web pages directly. Would it be possible to make the best use of the relatively abundant labeled English Web pages for classifying the Chinese Web pages?

To answer this question, in (Ling et al., 2008), we developed a novel approach for classifying the Web pages in Chinese using the training documents in English. In this subsection, we give a brief summary of this work. The problem to be solved is: we are given a collection of labeled English documents and a large number of unlabeled Chinese documents. The English and Chinese texts are not aligned. Our objective is to classify the Chinese documents into the same label space as the English data.

Our key observation is that even though the data use different text features, they may still share many of the same semantic information. What we need to do is to uncover this latent semantic information by finding out what is common among them. We did this in (Ling et al., 2008) by using the information bottleneck theory (Tishby et al., 1999). In our work, we first translated the Chinese document into English automatically using some available translation software, such as Google translate. Then, we encoded the training text as well as the translated target text together, in terms of the information theory. We allowed all the information to be put through a 'bottleneck' and be represented by a limited number of *code-*



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 1: An intuitive illustration of different kinds learning strategies using classification/clustering of image `apple` and `banana` as the example.

words (i.e. labels in the classification problem). Finally, information bottleneck was used to maintain most of the common information between the two data sources, and discard the remaining irrelevant information. In this way, we can approximate the ideal situation where similar training and translated test pages shared in the common part are encoded into the same codewords, and are thus assigned the correct labels. In (Ling et al., 2008), we experimentally showed that heterogeneous transfer learning can indeed improve the performance of cross-language text classification as compared to directly training learning models (e.g., Naive Bayes or SVM) and testing on the translated texts.

2.2 Other Works in Transfer Learning

In the past, several other works made use of transfer learning for cross-feature-space learning. Wu and Oard (2008) proposed to handle the cross-language learning problem by translating the data into a same language and applying kNN on the latent topic space for classification. Most learning algorithms for dealing with cross-language heterogeneous data require a *translator* to convert the data to the same feature space. For those data that are in different feature spaces where no translator is available, Davis and Domingos (2008) proposed a Markov-logic-based transfer learning algorithm, which is called *deep transfer*, for transferring knowledge between biological domains and Web domains. Dai et al. (2008a) proposed a novel learning paradigm, known as translated learning, to deal with the problem of learning heterogeneous data that belong to quite different feature spaces by using a risk minimization framework.

2.3 Relation to PLSA

Our work makes use of PLSA. *Probabilistic latent semantic analysis* (PLSA) is a widely used probabilistic model (Hofmann, 1999), and could be considered as a probabilistic implementation of *latent semantic analysis* (LSA) (Deerwester et al., 1990). An extension to PLSA was proposed in (Cohn and Hofmann, 2000), which incorporated the hyperlink connectivity in the PLSA model by using a joint probabilistic model for connectivity and content. Moreover, PLSA has shown a lot of applications ranging from text clustering (Hofmann, 2001) to image analysis (Sivic et al., 2005).

2.4 Relation to Clustering

Compared to many previous works on image clustering, we note that traditional image clustering is generally based on techniques such as K-means (MacQueen, 1967) and hierarchical clustering (Kaufman and Rousseeuw, 1990). However, when the data are sparse, traditional clustering algorithms may have difficulties in obtaining high-quality image clusters. Recently, several researchers have investigated how to leverage the auxiliary information to improve target clustering performance, such as supervised clustering (Finley and Joachims, 2005), semi-supervised clustering (Basu et al., 2004), self-taught clustering (Dai et al., 2008b), etc.

3 Image Clustering with Annotated Auxiliary Data

In this section, we present our *annotation-based probabilistic latent semantic analysis* algorithm (aPLSA), which extends the traditional PLSA model by incorporating annotated auxiliary image data. Intuitively, our algorithm aPLSA performs PLSA analysis on the target images, which are converted to an image instance-to-feature co-occurrence matrix. At the same time, PLSA is also 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. Through these common latent variables, which are now constrained by both target image data and auxiliary annotation data, a better clustering result is expected for the target data.

3.1 Probabilistic Latent Semantic Analysis

Let $\mathcal{F} = \{f_i\}_{i=1}^{|\mathcal{F}|}$ be an image feature space, and $\mathcal{V} = \{v_i\}_{i=1}^{|\mathcal{V}|}$ be the image data set. Each image $v_i \in \mathcal{V}$ is represented by a *bag-of-features* $\{f | f \in v_i \wedge f \in \mathcal{F}\}$.

Based on the image data set $\mathcal{V}$, we can estimate an image instance-to-feature co-occurrence matrix $A^{|\mathcal{V}|\times|\mathcal{F}|} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{F}|}$, where each element A_{ij} ($1 \leq i \leq |\mathcal{V}|$ and $1 \leq j \leq |\mathcal{F}|$) in the matrix A is the frequency of the feature f_j appearing in the instance v_i.

Let $\mathcal{W} = \{w_i\}_{i=1}^{|\mathcal{W}|}$ be a text feature space. The annotated image data allow us to obtain the co-occurrence information between images v and text features $w \in \mathcal{W}$. An example of annotated image data is the Flickr (`http://www.flickr.com`), which is a social Web site containing a large number of annotated images.

By extracting image features from the annotated images v, we can estimate a text-to-image feature co-occurrence matrix $B^{|\mathcal{W}|\times|\mathcal{F}|} \in \mathbb{R}^{|\mathcal{W}|\times|\mathcal{F}|}$, where each element B_{ij} ($1 \leq i \leq |\mathcal{W}|$ and $1 \leq j \leq |\mathcal{F}|$) in the matrix B is the frequency of the text feature w_i and the image feature f_j occurring together in the annotated image data set.

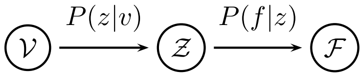


Figure 2: Graphical model representation of PLSA model.

Let $\mathcal{Z} = \{z_i\}_{i=1}^{|\mathcal{Z}|}$ be the latent variable set in our aPLSA model. In clustering, each latent variable $z_i \in \mathcal{Z}$ corresponds to a certain cluster.

Our objective is to estimate a clustering function $g : \mathcal{V} \mapsto \mathcal{Z}$ with the help of the two co-occurrence matrices A and B as defined above.

To formally introduce the aPLSA model, we start from the *probabilistic latent semantic analysis* (PLSA) (Hofmann, 1999) model. PLSA is a probabilistic implementation of *latent semantic analysis* (LSA) (Deerwester et al., 1990). In our image clustering task, PLSA decomposes the instance-feature co-occurrence matrix A under the assumption of conditional independence of image instances $\mathcal{V}$ and image features $\mathcal{F}$, given the latent variables $\mathcal{Z}$.

$$P(f|v) = \sum_{z \in \mathcal{Z}} P(f|z)P(z|v). \qquad (1)$$

The graphical model representation of PLSA is shown in Figure 2.

Based on the PLSA model, the log-likelihood can be defined as:

$$\mathcal{L} = \sum_i \sum_j \frac{A_{ij}}{\sum_{j'} A_{ij'}} \log P(f_j|v_i) \qquad (2)$$

where $A^{|\mathcal{V}|\times|\mathcal{F}|} \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{F}|}$ is the image instance-feature co-occurrence matrix. The term $\frac{A_{ij}}{\sum_{j'} A_{ij'}}$ in Equation (2) is a normalization term ensuring each image is giving the same weight in the log-likelihood.

Using EM algorithm (Dempster et al., 1977), which locally maximizes the log-likelihood of the PLSA model (Equation (2)), the probabilities $P(f|z)$ and $P(z|v)$ can be estimated. Then, the clustering function is derived as

$$g(v) = \operatorname*{argmax}_{z \in \mathcal{Z}} P(z|v). \qquad (3)$$

Due to space limitation, we omit the details for the PLSA model, which can be found in (Hofmann, 1999).

3.2 aPLSA: Annotation-based PLSA

In this section, we consider how to incorporate a large number of socially annotated images in a

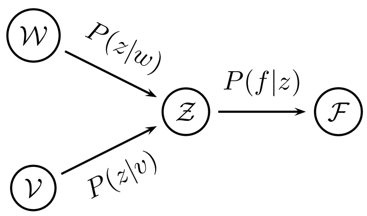


Figure 3: Graphical model representation of aPLSA model.

unified PLSA model for the purpose of utilizing the correlation between text features and image features. In the auxiliary data, each image has certain textual tags that are attached by users. The correlation between text features and image features can be formulated as follows.

$$P(f|w) = \sum_{z \in \mathcal{Z}} P(f|z)P(z|w). \tag{4}$$

It is clear that Equations (1) and (4) share a same term $P(f|z)$. So we design a new PLSA model by joining the probabilistic model in Equation (1) and the probabilistic model in Equation (4) into a unified model, as shown in Figure 3. In Figure 3, the latent variables $\mathcal{Z}$ depend not only on the correlation between image instances $\mathcal{V}$ and image features $\mathcal{F}$, but also the correlation between text features $\mathcal{W}$ and image features $\mathcal{F}$. Therefore, the auxiliary socially-annotated image data can be used to help the target image clustering performance by estimating good set of latent variables $\mathcal{Z}$.

Based on the graphical model representation in Figure 3, we derive the log-likelihood objective function, in a similar way as in (Cohn and Hofmann, 2000), as follows

$$\mathcal{L} = \sum_j \left[\lambda \sum_i \frac{A_{ij}}{\sum_{j'} A_{ij'}} \log P(f_j|v_i) \right. \\ \left. +(1-\lambda) \sum_l \frac{B_{lj}}{\sum_{j'} B_{lj'}} \log P(f_j|w_l) \right], \tag{5}$$

where $A^{|\mathcal{V}| \times |\mathcal{F}|} \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{F}|}$ is the image instance-feature co-occurrence matrix, and $B^{|\mathcal{W}| \times |\mathcal{F}|} \in \mathbb{R}^{|\mathcal{W}| \times |\mathcal{F}|}$ is the text-to-image feature-level co-occurrence matrix. Similar to Equation (2), $\frac{A_{ij}}{\sum_{j'} A_{ij'}}$ and $\frac{B_{lj}}{\sum_{j'} B_{lj'}}$ in Equation (5) are the normalization terms to prevent imbalanced cases.

Furthermore, λ acts as a trade-off parameter between the co-occurrence matrices A and B. In the extreme case when $\lambda = 1$, the log-likelihood objective function ignores all the biases from the text-to-image occurrence matrix B. In this case, the aPLSA model degenerates to the traditional PLSA model. Therefore, aPLSA is an extension to the PLSA model.

Now, the objective is to maximize the log-likelihood $\mathcal{L}$ of the aPLSA model in Equation (5). Then we apply the EM algorithm (Dempster et al., 1977) to estimate the conditional probabilities $P(f|z)$, $P(z|w)$ and $P(z|v)$ with respect to each dependence in Figure 3 as follows.

- E-Step: calculate the posterior probability of each latent variable z given the observation of image features f, image instances v and text features w based on the old estimate of $P(f|z)$, $P(z|w)$ and $P(z|v)$:

$$P(z_k|v_i, f_j) = \frac{P(f_j|z_k)P(z_k|v_i)}{\sum_{k'} P(f_j|z_{k'})P(z_{k'}|v_i)} \tag{6}$$

$$P(z_k|w_l, f_j) = \frac{P(f_j|z_k)P(z_k|w_l)}{\sum_{k'} P(f_j|z_{k'})P(z_{k'}|w_l)} \tag{7}$$

- M-Step: re-estimates conditional probabilities $P(z_k|v_i)$ and $P(z_k|w_l)$:

$$P(z_k|v_i) = \sum_j \frac{A_{ij}}{\sum_{j'} A_{ij'}} P(z_k|v_i, f_j) \tag{8}$$

$$P(z_k|w_l) = \sum_j \frac{B_{lj}}{\sum_{j'} B_{lj'}} P(z_k|w_l, f_j) \tag{9}$$

  and conditional probability $P(f_j|z_k)$, which is a mixture portion of posterior probability of latent variables

$$P(f_j|z_k) \propto \lambda \sum_i \frac{A_{ij}}{\sum_{j'} A_{ij'}} P(z_k|v_i, f_j) \\ + (1-\lambda) \sum_l \frac{B_{lj}}{\sum_{j'} B_{lj'}} P(z_k|w_l, f_j) \tag{10}$$

Finally, the clustering function for a certain image v is

$$g(v) = \underset{z \in \mathcal{Z}}{\operatorname{argmax}}\, P(z|v). \tag{11}$$

From the above equations, we can derive our annotation-based probabilistic latent semantic analysis (aPLSA) algorithm. As shown in Algorithm 1, aPLSA iteratively performs the E-Step and the M-Step in order to seek local optimal points based on the objective function $\mathcal{L}$ in Equation (5).

Algorithm 1 Annotation-based PLSA Algorithm (aPLSA)

Input: The $\mathcal{V}$-$\mathcal{F}$ co-occurrence matrix A and $\mathcal{W}$-$\mathcal{F}$ co-occurrence matrix B.

Output: A clustering (partition) function $g : \mathcal{V} \mapsto \mathcal{Z}$, which maps an image instance $v \in \mathcal{V}$ to a latent variable $z \in \mathcal{Z}$.

```
Initial $\mathcal{Z}$ so that $|\mathcal{Z}|$ equals the number clusters desired.
Initialize $P(z|v)$, $P(z|w)$, $P(f|z)$ randomly.
while the change of $\mathcal{L}$ in Eq. (5) between two sequential iterations is greater than a predefined threshold do
    E-Step: Update $P(z|v, f)$ and $P(z|w, f)$ based on Eq. (6) and (7) respectively.
    M-Step: Update $P(z|v)$, $P(z|w)$ and $P(f|z)$ based on Eq. (8), (9) and (10) respectively.
end while
for all $v$ in $\mathcal{V}$ do
    $g(v) \leftarrow \underset{z}{\mathrm{argmax}} P(z|v)$.
end for
Return $g$.
```

4 Experiments

In this section, we empirically evaluate the aPLSA algorithm together with some state-of-art baseline methods on two widely used image corpora, to demonstrate the effectiveness of our algorithm aPLSA.

4.1 Data Sets

In order to evaluate the effectiveness of our algorithm aPLSA, we conducted experiments on several data sets generated from two image corpora, Caltech-256 (Griffin et al., 2007) and the fifteen-scene (Lazebnik et al., 2006). The Caltech-256 data set has 256 image objective categories, ranging from animals to buildings, from plants to automobiles, etc. The fifteen-scene data set contains 15 scenes such as `store` and `forest`. From these two corpora, we randomly generated eleven image clustering tasks, including seven 2-way clustering tasks, two 4-way clustering task, one 5-way clustering task and one 8-way clustering task. The detailed descriptions for these clustering tasks are given in Table 1. In these tasks, `bi7` and `oct1` were generated from fifteen-scene data set, and the rest were from Caltech-256 data set.

DATA SET	INVOLVED CLASSES	DATA SIZE
bi1	skateboard, airplanes	102, 800
bi2	billiards, mars	278, 155
bi3	cd, greyhound	102, 94
bi4	electric-guitar, snake	122, 112
bi5	calculator, dolphin	100, 106
bi6	mushroom, teddy-bear	202, 99
bi7	MIThighway, livingroom	260, 289
quad1	calculator, diamond-ring, dolphin, microscope	100, 118, 106, 116
quad2	bonsai, comet, frog, saddle	122, 120, 115, 110
quint1	frog, kayak, bear, jesus-christ, watch	115, 102, 101, 87, 201
oct1	MIThighway, MITmountain, kitchen, MITcoast, PARoffice, MIT-tallbuilding, livingroom, bedroom	260, 374, 210, 360, 215, 356, 289, 216
tune1	coin, horse	123, 270
tune2	socks, spider	111, 106
tune3	galaxy, snowmobile	80, 112
tune4	dice, fern	98, 110
tune5	backpack, lightning, mandolin, swan	151, 136, 93, 114

Table 1: The descriptions of all the image clustering tasks used in our experiment. Among these data sets, `bi7` and `oct1` were generated from *fifteen-scene* data set, and the rest were from *Caltech-256* data set.

To empirically investigate the parameter λ and the convergence of our algorithm aPLSA, we generated five more date sets as the development sets. The detailed description of these five development sets, namely `tune1` to `tune5` is listed in Table 1 as well.

The auxiliary data were crawled from the Flickr (`http://www.flickr.com/`) web site during August 2007. Flickr is an internet community where people share photos online and express their opinions as social tags (annotations) attached to each image. From Flicker, we collected $19,959$ images and $91,719$ related annotations, among which $2,600$ words are distinct. Based on the method described in Section 3, we estimated the co-occurrence matrix B between text features and image features. This co-occurrence matrix B was used by all the clustering tasks in our experiments.

For data preprocessing, we adopted the *bag-of-features* representation of images (Li and Perona, 2005) in our experiments. Interesting points were found in the images and described via the *SIFT descriptors* (Lowe, 2004). Then, the interesting points were clustered to generate a codebook to form an image feature space. The size of codebook was set to $2,000$ in our experiments. Based on the codebook, which serves as the image feature space, each image can be represented as a corresponding feature vector to be used in the next step.

To set our evaluation criterion, we used the

Data Set	KMeans		PLSA		STC	aPLSA
	separate	combined	separate	combined		
bi1	0.645±0.064	0.548±0.031	0.544±0.074	0.537±0.033	0.586±0.139	**0.482±0.062**
bi2	0.687±0.003	0.662±0.014	0.464±0.074	0.692±0.001	0.577±0.016	**0.455±0.096**
bi3	1.294±0.060	1.300±0.015	1.085±0.073	1.126±0.036	1.103±0.108	**1.029±0.074**
bi4	1.227±0.080	1.164±0.053	0.976±0.051	1.038±0.068	1.024±0.089	**0.919±0.065**
bi5	1.450±0.058	1.417±0.045	1.426±0.025	1.405±0.040	1.411±0.043	**1.377±0.040**
bi6	1.969±0.078	1.852±0.051	1.514±0.039	1.709±0.028	1.589±0.121	**1.503±0.030**
bi7	0.686±0.006	0.683±0.004	0.643±0.058	0.632±0.037	0.651±0.012	**0.624±0.066**
quad1	0.591±0.094	0.675±0.017	0.488±0.071	0.662±0.013	0.580±0.115	**0.432±0.085**
quad2	0.648±0.036	0.646±0.045	0.614±0.062	0.626±0.026	0.591±0.087	**0.515±0.098**
quint1	0.557±0.021	0.508±0.104	0.547±0.060	0.539±0.051	0.538±0.100	**0.502±0.067**
oct1	0.659±0.031	0.680±0.012	0.340±0.147	0.691±0.002	0.411±0.089	**0.306±0.101**
average	0.947±0.029	0.922±0.017	0.786±0.009	0.878±0.006	0.824±0.036	**0.741±0.018**

Table 2: Experimental result in term of entropy for all data sets and evaluation methods.

entropy to measure the quality of our clustering results. In information theory, entropy (Shannon, 1948) is a measure of the uncertainty associated with a random variable. In our problem, entropy serves as a measure of randomness of clustering result. The entropy of g on a single latent variable z is defined to be $H(g, z) \triangleq -\sum_{c\in\mathcal{C}} P(c|z)\log_2 P(c|z)$, where $\mathcal{C}$ is the class label set of $\mathcal{V}$ and $P(c|z) = \frac{|\{v|g(v)=z\wedge t(v)=c\}|}{|\{v|g(v)=z\}|}$, in which $t(v)$ is the *true* class label of image v. Lower entropy $H(g, \mathcal{Z})$ indicates less randomness and thus better clustering result.

4.2 Empirical Analysis

We now empirically analyze the effectiveness of our aPLSA algorithm. Because, to our best of knowledge, few existing methods addressed the problem of image clustering with the help of social annotation image data, we can only compare our aPLSA with several state-of-the-art clustering algorithms that are not directly designed for our problem. The first baseline is the well-known KMeans algorithm (MacQueen, 1967). Since our algorithm is designed based on PLSA (Hofmann, 1999), we also included PLSA for clustering as a baseline method in our experiments.

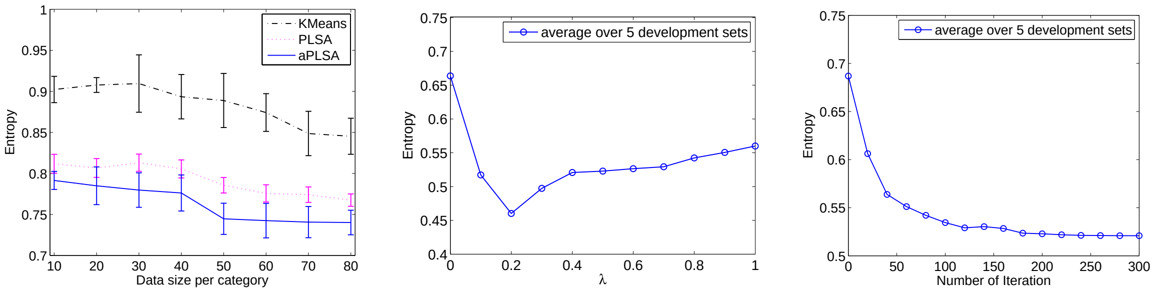
For each of the above two baselines, we have two strategies: (1) separated: the baseline method was applied on the target image data only; (2) combined: the baseline method was applied to cluster the combined data consisting of both target image data and the annotated image data. Clustering results on target image data were used for evaluation. Note that, in the combined data, all the annotations were thrown away since baseline methods evaluated in this paper do not leverage annotation information.

In addition, we compared our algorithm aPLSA to a state-of-the-art transfer clustering strategy, known as *self-taught clustering* (STC) (Dai et al., 2008b). STC makes use of auxiliary data to estimate a better feature representation to benefit the target clustering. In these experiments, the annotated image data were used as auxiliary data in STC, which does not use the annotation text.

In our experiments, the performance is in the form of the average entropy and variance of five repeats by randomly selecting 50 images from each of the categories. We selected only 50 images per category, since this paper is focused on clustering sparse data. Table 2 shows the performance with respect to all comparison methods on each of the image clustering tasks measured by the entropy criterion. From the tables, we can see that our algorithm aPLSA outperforms the baseline methods in all the data sets. We believe that is because aPLSA can effectively utilize the knowledge from the socially annotated image data. On average, aPLSA gives rise to 21.8% of entropy reduction and as compared to KMeans, 5.7% of entropy reduction as compared to PLSA, and 10.1% of entropy reduction as compared to STC.

4.2.1 Varying Data Size

We now show how the data size affects aPLSA, with two baseline methods KMeans and PLSA as reference. The experiments were conducted on different amounts of target image data, varying from 10 to 80. The corresponding experimental results in average entropy over all the 11 clustering tasks are shown in Figure 4(a). From this figure, we observe that aPLSA always yields a significant reduction in entropy as compared with two baseline methods KMeans and PLSA, regardless of the size of target image data that we used.



(a) (b) (c)

Figure 4: (a) The entropy curve as a function of different amounts of data per category. (b) The entropy curve as a function of different number of iterations. (c) The entropy curve as a function of different trade-off parameter λ.

4.2.2 Parameter Sensitivity

In aPLSA, there is a trade-off parameter λ that affects how the algorithm relies on auxiliary data. When $\lambda = 0$, the aPLSA relies only on annotated image data B. When $\lambda = 1$, aPLSA relies only on target image data A, in which case aPLSA degenerates to PLSA. Smaller λ indicates heavier reliance on the annotated image data. We have done some experiments on the development sets to investigate how different λ affect the performance of aPLSA. We set the number of images per category to 50, and tested the performance of aPLSA. The result in average entropy over all development sets is shown in Figure 4(b). In the experiments described in this paper, we set λ to 0.2, which is the best point in Figure 4(b).

4.2.3 Convergence

In our experiments, we tested the convergence property of our algorithm aPLSA as well. Figure 4(c) shows the average entropy curve given by aPLSA over all development sets. From this figure, we see that the entropy decreases very fast during the first 100 iterations and becomes stable after 150 iterations. We believe that 200 iterations is sufficient for aPLSA to converge.

5 Conclusions

In this paper, we proposed a new learning scenario called heterogeneous transfer learning and illustrated its application to image clustering. Image clustering, a vital component in organizing search results for query-based image search, was shown to be improved by transferring knowledge from unrelated images with annotations in a social Web. This is done by first learning the high-quality latent variables in the auxiliary data, and then transferring this knowledge to help improve the clustering of the target image data. We conducted experiments on two image data sets, using the Flickr data as the annotated auxiliary image data, and showed that our aPLSA algorithm can greatly outperform several state-of-the-art clustering algorithms.

In natural language processing, there are many future opportunities to apply heterogeneous transfer learning. In (Ling et al., 2008) we have shown how to classify the Chinese text using English text as the training data. We may also consider clustering, topic modeling, question answering, etc., to be done using data in different feature spaces. We can consider data in different modalities, such as video, image and audio, as the training data. Finally, we will explore the theoretical foundations and limitations of heterogeneous transfer learning as well.

Acknowledgement Qiang Yang thanks Hong Kong CERG grant 621307 for supporting the research.

References

Alina Andreevskaia and Sabine Bergler. 2008. When specialists and generalists work together: Overcoming domain dependence in sentiment tagging. In *ACL-08: HLT*, pages 290–298, Columbus, Ohio, June.

Andrew Arnold, Ramesh Nallapati, and William W. Cohen. 2007. A comparative study of methods for transductive transfer learning. In ICDM 2007 Workshop on Mining and Management of Biological Data, pages 77-82.

Andrew Arnold, Ramesh Nallapati, and William W. Cohen. 2008. Exploiting feature hierarchy for transfer learning in named entity recognition. In *ACL-08: HLT*.

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. 2004. A probabilistic framework for semi-supervised clustering. In *ACM SIGKDD 2004*, pages 59–68.

John Blitzer, Ryan Mcdonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *EMNLP 2006*, pages 120–128, Sydney, Australia.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL 2007*, pages 440–447, Prague, Czech Republic.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT 1998*, pages 92–100, New York, NY, USA. ACM.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *ACL 2007*, Prague, Czech Republic.

David A. Cohn and Thomas Hofmann. 2000. The missing link - a probabilistic model of document content and hypertext connectivity. In *NIPS 2000*, pages 430–436.

Wenyuan Dai, Yuqiang Chen, Gui-Rong Xue, Qiang Yang, and Yong Yu. 2008a. Translated learning: Transfer learning across different feature spaces. In *NIPS 2008*, pages 353–360.

Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. 2008b. Self-taught clustering. In *ICML 2008*, pages 200–207. Omnipress.

Hal Daume, III. 2007. Frustratingly easy domain adaptation. In *ACL 2007*, pages 256–263, Prague, Czech Republic.

Jesse Davis and Pedro Domingos. 2008. Deep transfer via second-order markov logic. In *AAAI 2008 Workshop on Transfer Learning*, Chicago, USA.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. L, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, pages 391–407.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *J. of the Royal Statistical Society*, 39:1–38.

Thomas Finley and Thorsten Joachims. 2005. Supervised clustering with support vector machines. In *ICML 2005*, pages 217–224, New York, NY, USA. ACM.

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

Thomas Hofmann. 1999 Probabilistic latent semantic analysis. In Proc. of Uncertainty in Artificial Intelligence, UAI99. Pages 289–296

Thomas Hofmann. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*. volume 42, number 1-2, pages 177–196. Kluwer Academic Publishers.

Jing Jiang and Chengxiang Zhai. 2007. Instance weighting for domain adaptation in NLP. In *ACL 2007*, pages 264–271, Prague, Czech Republic, June.

Leonard Kaufman and Peter J. Rousseeuw. 1990. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR 2006*, pages 2169–2178, Washington, DC, USA.

Fei-Fei Li and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *CVPR 2005*, pages 524–531, Washington, DC, USA.

Xiao Ling, Gui-Rong Xue, Wenyuan Dai, Yun Jiang, Qiang Yang, and Yong Yu. 2008. Can chinese web pages be classified with english data source? In *WWW 2008*, pages 969–978, New York, NY, USA. ACM.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *COLING/ACL 2006 Main conference poster sessions*, pages 547–554.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV) 2004*, volume 60, number 2, pages 91–110.

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 1:281–297, Berkeley, CA, USA.

Kamal Nigam and Rayid Ghani. 2000. Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management*, pages 86–93, New York, USA.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *ICML 2007*, pages 759–766, New York, NY, USA. ACM.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptation of statistical parsers trained on small datasets. In *ACL 2007*.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *ACL-08: HLT*, pages 861–869.

C. E. Shannon. 1948. A mathematical theory of communication. *Bell system technical journal*, 27.

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. 2005. Discovering object categories in image collections. In *ICCV 2005*.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. 1999. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377.

Pengcheng Wu and Thomas G. Dietterich. 2004. Improving svm accuracy by training on auxiliary data sources. In *ICML 2004*, pages 110–117, New York, NY, USA.

Yejun Wu and Douglas W. Oard. 2008. Bilingual topic aspect classification with a few training examples. In *ACM SIGIR 2008*, pages 203–210, New York, NY, USA.

Xiaojin Zhu. 2007. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.