# Chapter 7

# TRANSFER LEARNING FOR TEXT MINING

Weike Pan

*Hong Kong University of Science and Technology*
*Clearwater Bay, Kowloon, Hong Kong*

weikep@cse.ust.hk

Erheng Zhong

*Hong Kong University of Science and Technology*
*Clearwater Bay, Kowloon, Hong Kong*

ezhong@cse.ust.hk

Qiang Yang

*Hong Kong University of Science and Technology*
*Clearwater Bay, Kowloon, Hong Kong*

qyang@cse.ust.hk

**Abstract** Over the years, transfer learning has received much attention in machine learning research and practice. Researchers have found that a major bottleneck associated with machine learning and text mining is the lack of high-quality annotated examples to help train a model. In response, transfer learning offers an attractive solution for this problem. Various transfer learning methods are designed to extract the useful knowledge from different but related auxiliary domains. In its connection to text mining, transfer learning has found novel and useful applications. In this chapter, we will review some most recent developments in transfer learning for text mining, explain related algorithms in detail, and project future developments of this field. We focus on two important topics: cross-domain text document classification and heterogeneous transfer learning that uses labeled text documents to help classify images.

**Keywords:** Transfer learning, text mining, classification, clustering, learning-to-rank.

C.C. Aggarwal and C.X. Zhai (eds.), *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_7,

# 1. Introduction

*Transfer learning* refers to the machine learning framework in which one extracts knowledge from some auxiliary domains to help boost the learning performance in a target domain. Transfer learning as a new paradigm of machine learning has achieved great success in various areas over the last two decades [17, 67], e.g. text mining [8, 26, 23], speech recognition [95, 52], computer vision (e.g. image [75] and video [100] analysis), and ubiquitous computing [108, 93].

For text mining, transfer learning can be found in many application scenarios, e.g., knowledge transfer from Wikipedia documents (auxiliary) to Twitter text (target), from WWW webpages to Flick images, from English documents to Chinese documents in search engine, etc. One fundamental motivation of transfer learning in text mining is the so-called *data sparsity* problem in a target domain, where data sparsity can be defined by a lack of useful labels or sufficient data in the training set. For example, Twitter messages are short documents that are generated by users. These documents are often unlabeled, which are difficult to classify. Thus, it would be useful for us to transfer the supervised knowledge from another fully labeled text corpus to help classify Twitter messages. When data sparsity happens, overfitting can easily happen when we train a model. In the past, many traditional machine learning methods have been proposed for addressing the *data sparsity* problem, including semi-supervised learning [111, 18], co-training [9] and active learning [91]. However, in many practical situations, we still have to look elsewhere for additional knowledge for learning in our domain of interest.

We can take the following two views on knowledge transfer,

1 *In theory*, transfer learning can be considered as a new *learning paradigm*, where most non-transfer learning methods are considered as a special case when learning happens within a single target domain only, e.g., text classification in Twitter, and

2 *In applications*, transfer learning can be considered as a new cross-domain *learning technique*, since it explicitly addresses the various aspects of domain differences, e.g. data distribution, feature and label space, noise in the auxiliary data, relevance of auxiliary and target domains, etc. For example, we have to address most of the above issues when we transfer knowledge from Wikipedia documents to Twitter text.

Machine learning algorithms such as classification and regression (e.g. discriminative learning, ensemble learning) have been widely adopted in various text mining applications, e.g. text classification [42], sentiment analysis [68], named entity recognition (NER) [106], part-of-speech (POS) tagging [77], relation extraction (RE) [104], etc. In this chapter, we will survey some recent transfer learning extensions in aforementioned machine learning and data mining techniques and their applications for text mining. The organization of the chapter is as follows. We first give an overview of the scope of text-mining problems that we consider, and motivate the need for transfer learning in text classification. We then describe some typical approaches in transfer learning, such that we can subsequently categorize various text-mining approaches under these transfer-learning categories. This is followed by an overview of transfer learning approaches that extracts knowledge from labeled text data for the benefit of image classification and processing. This latter approach is known as heterogeneous transfer learning (HTL). Finally, we conclude the chapter with a summary and discussion of future work.

# 2. Transfer Learning in Text Classification

We first review the problem formulation in cross-domain text classification problems. In the next section, we first look at some typical benchmark data examples where the cross domain classification methods are needed. We then consider the nature of these problems, their differences from a traditional text classification problem, as well as how to formulate these problems into a machine learning problem.

## 2.1 Cross Domain Text Classification

### 2.1.1 Support Vector Machines for Text Classification.

Text classification [42] aims to categorize a document to some predefined categories $\mathcal{Y}$, where a document is usually represented in the form of bag of words $\mathcal{X}$, denoted as a vector $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$ with $d$ unique words. The entries in the feature vector $\boldsymbol{x}$ can be 1/0 indicating whether the corresponding word appears or not or TF-IDF (term frequency inverse-document frequency).

There are enormous user-generated contents in online products and services on social media forums, blogs and microblogs, social networks, etc. It is very important to be able to summarize consumers' opinions on existing products and services. Sentiment analysis (or opinion mining) [68] addresses this problem, by classifying the reviews or sentiments into positive and negative categories. Similar to text classification, reviews or sentiments can be represented as a feature vector $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$, and the label space is $\mathcal{Y} = \{\pm 1\}$.

Extension of text classification has also been done in sequence classification areas. For example, POS tagging [77] aims to assign a tag to each word in a text, or equivalently classify each word in a text to some specified categories such as norm, verb, adjective, etc. POS tagging is very important for language pre-processing, speech synthesis, word sense disambiguation, information retrieval, etc. POS tagging can be considered as a structure prediction problem, and can be reduced to multiple binary-classification problems.

As support vector machines (SVM) [42] have been recognized as a state-of-the-art model in text mining, below, we will use SVM as a representative base model among various discriminative models to illustrate how the labeled data in auxiliary domains can be used to achieve knowledge transfer from auxiliary domains to the target domain. We first consider the text data representation.

In text mining, we assume that the data are represented as a bag-of-words $\mathcal{X} = \mathbb{R}^{d\times 1}$ with the same feature space for both auxiliary and target learning domains. For notational simplicity, we consider binary classification problems, $\mathcal{Y} = \{\pm 1\}$, which can be extended to multi-class classification via common tricks of one-vs-one or one-vs-rest pre-processing. We generally assume the same feature space and label space in both auxiliary and target domains, but in Section 3, we mention some recent works on heterogeneous feature space and/or heterogeneous label space learning. We use $X$ and $Y$ to denote variables for feature space and label space, respectively, and we use $\boldsymbol{x}$, $y$, $\tilde{\boldsymbol{x}}$, $\tilde{y}$ to denote the corresponding instantiations of variables in target and auxiliary domains, respectively.

For each word in a text, we can extract a feature vector based on the context information that is represented as $\boldsymbol{x} \in \mathbb{R}^{d\times 1}$. Many text mining problems can be modeled this way. In POS tagging, for example, the learning problem is basically a classification problem by assigning a label $y$ to $\boldsymbol{x}$. In Named Entity Recognition (NER) problems [106], the aim is to classify each word in a text to some pre-defined categories, e.g. location, time, organization, etc. Another interesting problem is relation extraction [104], where each pair of entities in a sentence is represented as a feature vector $\boldsymbol{x}$, which is assigned to a certain type of relation, e.g. family, user-owner, employer-executive, etc.

Text classification can be addressed by discriminative learning methods, which explicitly model the conditional distribution $P_r(Y|X)$. We can find many text mining formulations as variants of this formulation, e.g., maximum entropy (MaxEnt) [5], logistic regression (LR) [36], conditional random field (CRF) [47]. With this in mind, we consider the following basic SVM algorithm for text classification.

**Basic SVM for Text Classification** Given $\ell$ labeled data points $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ with $\boldsymbol{x}_i \in \mathbb{R}^{d\times 1}$ and $y_i \in \{\pm 1\}$ in the target domain, we have the following optimization problem for the linear SVM with soft margin [82],

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi}} \quad & \frac{1}{2}||\boldsymbol{w}||_2^2 + \lambda \sum_{i=1}^{\ell} \xi_i \\ \text{s.t.} \quad & y_i\, \boldsymbol{w}^T \boldsymbol{x}_i \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots, \ell \end{aligned} \tag{7.1}$$

where $\boldsymbol{w} \in \mathbb{R}^{d\times 1}$ is the model parameter, $\boldsymbol{\xi} \in \mathbb{R}^{\ell\times 1}$ are the slack variables, and $\lambda > 0$ is the tradeoff parameter to balance the model complexity $||\boldsymbol{w}||_2^2$ and loss function $\sum_{i=1}^{\ell} \xi_i$. Solving the convex optimization problem in Eq.(7.1), we have a decision function

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} = \sum_{k=1}^{d} w_k x_k. \tag{7.2}$$

In this section, we will consider how to extend this formulation to include transfer learning capabilities.

**2.1.2 Cross Domain Text Classification Problems.** With the above baseline algorithm in mind, we now consider several problem domains where we show examples of cross domain text classification. These examples illustrates some of the benchmark data often used in transfer learning experiments. They also help demonstrate why transfer learning is needed when the domain difference is large between the auxiliary and target learning domains.

**20 Newsgroups** First, we consider the well-known 20-newsgroup data. The 20-newsgroup [48] is a text collection of approximately 20,000 newsgroup documents, which are partitioned across 20 different newsgroups nearly evenly. This data collection provides an ideal benchmark for evaluating and comparing different transfer learning algorithms for text classification. A typical method is to generate six different data sets from the 20-newsgroup data for evaluating cross-domain classification algorithms. For each data set, two top categories[1] are chosen, one as positive and the other as negative. Then, we can split the data based on

[1] Three top categories, `misc`, `soc` and `alt` are removed, because they are too small.

| Data Set | $\tilde{D}$ | $D$ |
| --- | --- | --- |
| comp vs sci | comp.graphics<br>comp.os.ms-windows.misc<br>sci.crypt<br>sci.electronics | comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>comp.windows.x<br>sci.med<br>sci.space |
| rec vs talk | rec.autos<br>rec.motorcycles<br>talk.politics.guns<br>talk.politics.misc | rec.sport.baseball<br>rec.sport.hockey<br>talk.politics.mideast<br>talk.religion.misc |
| rec vs sci | rec.autos<br>rec.sport.baseball<br>sci.med<br>sci.space | rec.motorcycles<br>rec.sport.hockey<br>sci.crypt<br>sci.electronics |
| sci vs talk | sci.electronics<br>sci.med<br>talk.politics.misc<br>talk.religion.misc | sci.crypt<br>sci.space<br>talk.politics.guns<br>talk.politics.mideast |
| comp vs rec | comp.graphics<br>comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>rec.motorcycles<br>rec.sport.hockey | comp.os.ms-windows.misc<br>comp.windows.x<br>rec.autos<br>rec.sport.baseball |
| comp vs talk | comp.graphics<br>comp.sys.mac.hardware<br>comp.windows.x<br>talk.politics.mideast<br>talk.religion.misc | comp.os.ms-windows.misc<br>comp.sys.ibm.pc.hardware<br>talk.politics.guns<br>talk.politics.misc |

*Table 7.1.* A description of 20-newsgroup data sets for cross-domain classification.

sub-categories. Different sub-categories can be considered as different domains, while the task is defined as top category classification. The splitting strategy ensures the domains of labeled and unlabeled data related, since they are under the same top categories. Table 7.1 shows details of this data.

**SRAA** SRAA [61] is a UseNet data set for document classification that describes documents in Simulated/Real/Aviation/Auto classes. 73,218 UseNet articles are collected from four discussion groups about simulated autos (`sim-auto`), simulated aviation (`sim-aviation`), real autos (`real-auto`) and real aviation (`real-aviation`).

For a task to predict labels of instances between *real* and *simulated*, we can use the documents in `real-auto` and `sim-auto` as auxiliary domain data, while `real-aviation` and `sim-aviation` as target domain data.

| Data Set | $\tilde{D}$ | $D$ |
|---|---|---|
| auto vs aviation | sim-auto & sim-aviation | real-auto & real-aviation |
| real vs simulated | real-aviation & sim-aviation | real-auto & sim-auto |

*Table 7.2.* The description of SRAA data sets for cross-domain classification.

| Data Set | KL($\tilde{D}\|\|D$) | Documents | | SVM | |
|---|---|---|---|---|---|
| | | $\|\tilde{X}\|$ | $\|X\|$ | $\tilde{D} \to D$ | $D$+CV |
| real vs simulated | 1.161 | 8,000 | 8,000 | 0.266 | 0.032 |
| auto vs aviation | 1.126 | 8,000 | 8,000 | 0.228 | 0.033 |
| rec vs talk | 1.102 | 3,669 | 3,561 | 0.233 | 0.003 |
| rec vs sci | 1.021 | 3,961 | 3,965 | 0.212 | 0.007 |
| comp vs talk | 0.967 | 4,482 | 3,652 | 0.103 | 0.005 |
| comp vs sci | 0.874 | 3,930 | 4,900 | 0.317 | 0.012 |
| comp vs rec | 0.866 | 4,904 | 3,949 | 0.165 | 0.008 |
| sci vs talk | 0.854 | 3,374 | 3,828 | 0.226 | 0.009 |
| orgs vs places | 0.329 | 1,079 | 1,080 | 0.454 | 0.085 |
| people vs places | 0.307 | 1,239 | 1,210 | 0.266 | 0.113 |
| orgs vs people | 0.303 | 1,016 | 1,046 | 0.297 | 0.106 |

*Table 7.3.* Description of the data sets for cross-domain text classification, including errors given by SVM. "$\tilde{D} \to D$" means training on the auxiliary domain $\tilde{D}$ and testing on the target domain $D$; "$D$+CV" means 10-fold cross-validation using target domain data only. The performances are in test error rate. The table is quoted from [22].

Then, the data set `real vs simulated` is generated as shown in Table 7.2. As a result, all the data in the auxiliary domain data set are about autos, while all the data in the target domain set are about aviation. The `auto vs aviation` data set is generated in the similar way as shown in Table 7.2.

**Reuters-21578** Reuters-21578 [49] is a well known test collections for evaluating text classification techniques. This dataset contains 5 top categories, among which `orgs`, `people` and `places` are three large ones. There is also a hierarchical structure which allows us to generate different data sets such as `orgs vs people`, `orgs vs places`, and `people vs places` for cross-domain classification in a similar way as what we have done on the 20-newsgroup and SRAA corpora.

**Properties of the Data Sets** Table 7.3 gives an overview of applying the basic SVM algorithm to the above data sets. The first three columns of the table show the statistical properties of the data sets. The first two data sets are from SRAA corpus. The next six are generated using

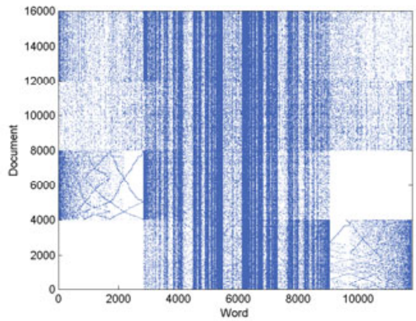


*Figure 7.1.* Document-word co-occurrence distribution on the `auto vs aviation` data set (quoted from [22]).

20-newsgroup data set. The last three are from Reuters-21578 test collection. To show the distribution differences between the training and testing data, KL-divergence values are calculated by $\text{KL}(\tilde{D}||D)$ on all the data set and are presented in the second column in the table, sorted in decreasing order from top down. Note that the Kullback-Leibler (KL) divergence [45] of two distributions of $\{p_i\}_{i=1}^{\ell}$ and $\{q_i\}_{i=1}^{\ell}$ is defined as

$$\text{KL}(\{p_i\}_{i=1}^{\ell}||\{q_i\}_{i=1}^{\ell}) = \sum_{i=1}^{\ell} p_i \ln(p_i/q_i)+(1-p_i)\ln((1-p_i)/(1-q_i)) \quad (7.3)$$

Here $\tilde{D}$ is the auxiliary domain data and $D$ is the target domain data. It can be seen that the KL-divergence values for all the data sets are much larger than the identical-distribution case which has a KL value of nearly zero. The next column titled "Documents" shows the size of the data sets used.

Under the column titled "SVM", we show two groups of classification results in two sub-columns. First, "$\tilde{D} \to D$" denotes the test error rate obtained when a classifier trained based on the auxiliary domain data set $\tilde{D}$ is applied to the target domain data set $D$. The column titled "$D$+CV" denotes the best-case obtained by the corresponding classifier, where the best case is to conduct a 10-fold cross-validation on the target domain data set $D$ using that classifier. Note that in obtaining the best case for each classifier, the training part is labeled data from $D$ and the test part is also $D$, according to different folds, which gives the best result for that classifier. It can be found that the test error rates, given by SVM, in the case of "$\tilde{D} \to D$" is much worse than those in the case of "$D$+CV". This indicates that for these data sets, it is not suitable to apply traditional supervised classification algorithms.

Figure 7.1 shows the document-word co-occurrence distribution on the `auto vs aviation` data set. In this figure, documents 1 to 8000 are from target domain $D$, while documents 8001 to 16000 are from auxiliary domain $\tilde{D}$. The documents are order first by their domains ($\tilde{D}$ or $D$), and second by their categories (positive or negative). The words are sorted by $n_+(w)/n_-(w)$, where $n_+(w)$ and $n_-(w)$ represent the number of word positions $w$ appears in positive and negative document, respectively. From Figure 7.1, it can be found that the distributions of auxiliary domain and target domain data are somewhat different, but almost consistent. That is, in general, the probabilities of a word belongs to a category in two domains do not differ very much.

## 2.2 Instance-based Transfer

One of the most intuitive methods is to transfer the knowledge between the domains by identifying a subset of source instances and insert them into the training set of the target domain data. We can observe that some instances in auxiliary domains are helpful for training the target domain model, while others may do harm to the target learning task. Thus, we need to select those that are useful and kick out those that are not. One effective way to achieve this is to perform instance weighting on the source domain data according to their importance to learning in the target domain. Taking SVM as an example, suppose that we have $\tilde{\ell}$ labeled data in the auxiliary domain, $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^{\tilde{\ell}}$ with $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{d\times 1}$ and $\tilde{y}_i \in \{\pm 1\}$, which can be incorporated into the standard SVM in Eq.(7.1) as follows [96, 54],

$$\begin{aligned} \min_{\boldsymbol{w},\boldsymbol{\xi},\tilde{\boldsymbol{\xi}}} \quad & \frac{1}{2}||\boldsymbol{w}||_2^2 + \lambda\sum_{i=1}^{\ell}\xi_i + \lambda\sum_{i=1}^{\tilde{\ell}}\tilde{\rho}_i\tilde{\xi}_i \\ \text{s.t.} \quad & y_i\,\boldsymbol{w}^T\boldsymbol{x}_i \geq 1-\xi_i,\ \xi_i \geq 0,\ i=1,\ldots,\ell \\ & \tilde{y}_i\,\boldsymbol{w}^T\tilde{\boldsymbol{x}}_i \geq 1-\tilde{\xi}_i,\ \tilde{\xi}_i \geq 0,\ i=1,\ldots,\tilde{\ell} \end{aligned} \tag{7.4}$$

where $\tilde{\rho}_i \in \mathbb{R}$ is the weight on the data point $(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$ in the auxiliary domain, which can be estimated via some heuristics [54, 40] or optimization techniques [55]. We can see that the only difference between the standard SVM in Eq.(7.1) and SVM with instance-based transfer in Eq.(7.4) is from the loss function $\lambda\sum_{i=1}^{\tilde{\ell}}\tilde{\rho}_i\tilde{\xi}_i$ and its corresponding constraints defined on the labeled data in the auxiliary domain. The auxiliary data $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^{\tilde{\ell}}$ can be the support vectors of a trained SVM in the auxiliary domain [54, 40] or the whole auxiliary data set [96, 55]. Note that the approach in [96] uses a slightly different base model of linear programming SVM (LP-SVM) [59] instead of the standard SVM in Eq.(7.1). Similar techniques are also developed in the context of *incremental learning* [80], where support vectors of a learned SVM in the auxiliary domain are combined with labeled data in the target domain with different weight.

Research works have also been done in sample selection bias [35, 103] with $\tilde{P}_r(X) \neq P_r(X)$, $\tilde{P}_r(Y|X) \neq P_r(Y|X)$, and covariate shift [88] with $\tilde{P}_r(X) \neq P_r(X)$, $\tilde{P}_r(Y|X) = P_r(Y|X)$. For example, Bickel et al. [6] explicitly consider the difference of conditional distributions, $\tilde{P}_r(Y|X) \neq P_r(Y|X)$, and propose an alternative gradient descent algorithm to automatically learn the weight of the instances besides the model parameter of Logistic regression. Jiang and Zhai [39] propose a general instance weighting framework from a distribution view considering differences from both marginal distributions, $\tilde{P}_r(X) \neq P_r(X)$, and conditional distributions, $\tilde{P}_r(Y|X) \neq P_r(Y|X)$.

Xiang et al. proposed an algorithm known as BIG (Bridging Information Gap) [97], which is a framework to make use of a wolrdwide knowledge base (e.g. Wikipedia) as a bridge to achieve knowledge transfer from an auxiliary domain with labeled data to a target domain with test data. Specifically, Xiang et al. [97] study the *information gap* between the target domain and auxiliary domain, and propose a margin related criteria to sample unlabeled data from Wikipedia to fill the *information gap*, which enables more effective knowledge transfer. Transductive SVM [41] is then trained using the improved data pool of labeled data in the auxiliary domain, unlabeled data from Wikipedia, and test data in the target domain. The proposed framework is studied in cross-domain text classification, sentiment analysis and query classification [97].

## 2.3 Cross-Domain Ensemble Learning

It is well known in text mining that ensemble methods are very effective in gaining top performance. AdaBoost [31] and Bagging [11] are two of the most popular ensemble learning algorithms in machine learning. In this section, we show how to use AdaBoost [31] as a representative base algorithm to be extended for transfer learning.

The AdaBoost [31] algorithm, as shown in Figure 7.2, starts with a uniform distribution of instance weights. It then gradually *increases* the weights of misclassified instances and *decreases* the weights of correctly classified instances, in order to concentrate more on "hard-to-learn" instances to improve overall classification performance. AdaBoost [31] finally generates a set of weighted weak learners $\{(\alpha^t, \boldsymbol{w}^t)\}_{t=1}^{\Gamma}$, which

**Input:** labeled data in the target domain $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$
**Initialization:** initialize instance weight $\{\rho_i^1\}_{i=1}^{\ell}$
**For** $t = 1 \ldots \Gamma$ Step 1. Train a model $\boldsymbol{w}^t$ using $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$
Step 2. Calculate the error $\epsilon^t$ of $\boldsymbol{w}^t$ on $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$
Step 3. Calculate the weight $\alpha^t$ from $\epsilon^t$
Step 4. Update instance weight $\{\rho_i^{t+1}\}_{i=1}^{\ell}$ using $\alpha^t$: *decrease* $\rho_i^{t+1}$ for correct predictions in the target domain *increase* $\rho_i^{t+1}$ for incorrect predictions in the target domain
**Output:** learned weight and weak models $\{(\alpha^t, \boldsymbol{w}^t)\}_{t=1}^{\Gamma}$.

*Figure 7.2.* The AdaBoost algorithm [31].

**Input:** labeled data in the target domain $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$, labeled data in the auxiliary domain $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^{\tilde{\ell}}$
**Initialization:** initialize instance weight $\{\rho_i^1\}_{i=1}^{\ell}$, $\{\tilde{\rho}_i^1\}_{i=1}^{\tilde{\ell}}$
**For** $t = 1 \ldots \Gamma$ Step 1. Train a model $\boldsymbol{w}^t$ using $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$ and $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i, \tilde{\rho}_i^t)\}_{i=1}^{\tilde{\ell}}$, which minimizes the weighted error only on labeled target data.
Step 2. Calculate the error $\epsilon^t$ of $\boldsymbol{w}^t$ on $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$
Step 3. Calculate the weight $\alpha^t$ from $\epsilon^t$
Step 4. Update instance weight $\{\rho_i^{t+1}\}_{i=1}^{\ell}$ and $\{\tilde{\rho}_i^{t+1}\}_{i=1}^{\tilde{\ell}}$ using $\alpha^t$: *decrease* $\tilde{\rho}_i^{t+1}$ for incorrect predictions in the auxiliary domain *increase* $\rho_i^{t+1}$ for incorrect predictions in the target domain
**Output:** learned weight and weak models $\{(\alpha^t, \boldsymbol{w}^t)\}_{t=\lceil\Gamma/2\rceil}^{\Gamma}$.

*Figure 7.3.* The TrAdaBoost algorithm [23].

can be used to predict the label of an incoming instance $\boldsymbol{x}$,

$$f(\boldsymbol{x}) = \sum_{t=1}^{\Gamma} \alpha^t \boldsymbol{w}^{tT} \boldsymbol{x}. \tag{7.5}$$

**TrAdaBoost** In order to leverage auxiliary instances, various ensemble learning based transfer learning algorithms are proposed. TrAdaBoost [23] is a well-known instance-based transfer learning algorithm, which is shown in Figure 7.3. The idea behind this algorithm is to pick those auxilary instances which are similar to the target domain and ignore others. One observation is that we can integrate some unlabeled data from the target domain, if there are any [23]. Although the detailed implementations of "Steps 1, 2, 3" in TrAdaBoost [23] are all different from that of AdaBoost [31], an interesting part of TrAdaBoost [23] is in "Step 4", which has a different instance weight update strategy. TrAdaBoost [23] aims at transferring the most useful instances from the auxiliary domain. Thus it *decreases* the weight of misclassified instances in the auxiliary domain. Furthermore, as in transfer learning, we care more about the prediction performance on labeled data in the target domain, thus, TrAdaBoost [23] *increases* the weights of misclassified instances in the target domain.

TransferBoost [28] extends TrAdaBoost [23] by considering both an instance level and set-of-instances level weights of an auxiliary data. By doing so it allows the model to be more robust.

TrAdaBoost.R2 [69] studies the regression problem based on TrAdaBoost [23] and AdaBoost.R2 [27]. It achieves knowledge transfer from weighted instances from the auxiliary domain. An additional feature is that TrAdaBoost.R2 [69] proposes a two-stage instance weight update strategy in order to avoid model overfitting.

MultiSourceTrAdaBoost [102] extends TrAdaBoost [23] for *multiple* auxiliary data sources, aiming at alleviating negative transfer that may happen if we only have a single auxiliary data source. MultiSourceTrAdaBoost [102] replaces "Step 1" in the TrAdaBoost algorithm in Figure 7.3 as follows,

" Step 1. Train a model using $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$ and labeled data from one of the $n_a$ auxiliary data sources. Select one model from those $n_a$ trained models that minimizes the weighted error on labeled data in the target domain. The selected model is denoted as $\boldsymbol{w}^t$. "

MultiSourceTrAdaBoost [102] combines the instance update strategy of TrAdaBoost [23] for auxiliary data and that of AdaBoost [31] for the target data.

TrAdaBoost [23] is further extended in [94] by adding an additional feature selection step. In [94], the authors replace "Step 1" of TrAdaBoost in Figure 7.3 with the following step, in order to select the most *discriminative* feature in each iteration:

"Step 1. Select a *single-feature* and train a *single-feature* model $\boldsymbol{w}^t$ using $\{(\boldsymbol{x}_i, y_i, \rho_i^t)\}_{i=1}^{\ell}$ and $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i, \tilde{\rho}_i^t)\}_{i=1}^{\tilde{\ell}}$, which minimizes the weighted error on the labeled data in the target domain."

This feature selection approach based on transfer learning models achieves very promising results in lunar crater discovery applications, as reported in [94], which is quite general and can be adapted for text classification and ranking.

## 2.4 Feature-based Transfer Learning for Document Classification

Feature-based transfer is another main transfer learning paradigm, where algorithms are designed from the perspective of feature space transformation. Examples include feature replication [37, 46], feature projection [8, 7, 64], dimensionality reduction [63, 65, 66, 89, 21], feature correlation [76, 44, 107], feature subsetting [81], feature weighting [2], etc.

**Feature Replication** The feature replication or feature augmentation approach [37] is basically a pre-processing step on the labeled data $\{(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)\}_{i=1}^{\tilde{\ell}}$ in the auxiliary domain and labeled data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ in the target domain ,

$$(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) \to ([\tilde{\boldsymbol{x}}_i^T \; \tilde{\boldsymbol{x}}_i^T \; \mathbf{0}^T]^T, \tilde{y}_i), \; i = 1, \ldots, \tilde{\ell}$$
$$(\boldsymbol{x}_i, y_i) \to ([\boldsymbol{x}_i^T \; \mathbf{0}^T \; \boldsymbol{x}_i^T]^T, y_i), \; i = 1, \ldots, \ell$$

where the feature dimensionality is expanded from $\mathbb{R}^{d\times 1}$ to $\mathbb{R}^{3d\times 1}$, and standard supervised learning methods can then be used, e.g. SVM in Eq.(7.1).

As a follow-up work, Kumar et al. [46] further generalize the idea of *feature replication* via incorporating unlabeled data $\{\boldsymbol{x}_i\}_{i=\ell+1}^{n}$ in the target domain,

$$\boldsymbol{x}_i \to ([\mathbf{0}^T \; \boldsymbol{x}^T \; -\boldsymbol{x}^T]^T, +1), \; i = \ell+1, \ldots, n$$
$$\boldsymbol{x}_i \to ([\mathbf{0}^T \; \boldsymbol{x}^T \; -\boldsymbol{x}^T]^T, -1), \; i = \ell+1, \ldots, n$$

where the processed data points are all with labels now.

The relationship of the feature replication method and the model-based transfer is discussed in [37] and some theoretical results of generalization bound are given in [46]. Feature replication approach have been successfully applied in cross-domain named entity recognition [37], part-of-speech tagging [37] and sentiment analysis [46].

**Feature Projection** Structured correspondence learning (SCL) [8] introduces the concept of *pivot features*, which possess high frequency and similar meaning in both auxiliary and target domains. Non-pivot features can be mapped to each other via the pivot features from the unlabeled data of both auxiliary and target domains. Learning in SCL [8] is based on the alternating structure optimization (ASO) algorithm [1]. Typically, SCL [8] goes through the following steps. First, it selects $n_p$ pivot features. Then, for each pivot feature, SCL trains an SVM model in Eq.(7.1) using unlabeled data instances from both domains with labels indicating whether the pivot feature appears in the data instance. In this step it obtains $n_p$ models such that $\mathbf{W} = [\boldsymbol{w}_j]_{j=1}^{n_p} \in \mathbb{R}^{d\times n_p}$. Third, SCL applies Singular Value Decomposition (SVD) to the model parameters $\mathbf{W}$, $[\mathbf{U}\,\Sigma\,\mathbf{V}^T] = \text{svd}(\mathbf{W})$, and it takes the top $k$ columns of $\mathbf{U}$ as the projection matrix $\boldsymbol{\theta} \in \mathbb{R}^{d\times k}$. Finally, it obtains the following transformation for each labeled data point in the auxiliary domain,

$$(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) \rightarrow ([\tilde{\boldsymbol{x}}_i^T\,\lambda\,(\boldsymbol{\theta}^T\tilde{\boldsymbol{x}}_i)^T]^T, \tilde{y}_i),\ i = 1, \ldots, \tilde{\ell} \tag{7.6}$$

In the above equation, $\lambda > 0$ is a tradeoff parameter. The transformed data points is augmented with $k$ additional features encoded with *structural correspondence* information between the features from auxiliary and target domains. With the transformed labeled data in the auxiliary domain, SCL can train a discriminative model, e.g. SVM in Eq.(7.1). For any future data instance $\boldsymbol{x}$, it is transformed via $\boldsymbol{x} \rightarrow [\boldsymbol{x}^T\,\lambda\,(\boldsymbol{\theta}^T\boldsymbol{x})^T]^T$ before $\boldsymbol{x}$ is classified by the learned model according to Eq.(7.2).

Blitzer et al. [7] successfully apply SCL [8] to cross-domain sentiment classification, and Prettenhofer and Stein [70, 71] extend SCL [8] with an additional cross-language translator to achieve knowledge transfer from English to German, French and Japanese for text classification and sentiment analysis. Pan et al. [64] propose a spectral learning algorithm for cross-domain sentiment classification using co-occurrence information from auxiliary-domain-specific, target-domain-specific and domain-independent features. They then align domain-specific features from both domains in a latent space via a learned projection matrix $\boldsymbol{\theta} \in \mathbb{R}^{k\times d}$. In some practical cases, the cross-domain sentiment and review classification performance of [64] is empirically shown to be superior to SCL [8] and other baselines.

**Dimensionality Reduction** In order to bridge two domains to enable knowledge transfer, Pan et al. [63] introduce *maximum mean discrepancy* (MMD) [10] as a distribution measurement of unlabeled data from auxiliary and target domains,

$$||\frac{1}{\tilde{\ell}}\sum_{i=1}^{\tilde{\ell}}\phi(\tilde{\boldsymbol{x}}_i) - \frac{1}{n-\ell}\sum_{i=\ell+1}^{n}\phi(\boldsymbol{x})_i||_2^2 \tag{7.7}$$

which is used to minimize the distribution distance in a latent space. The MMD measurement is formulated as a kernel learning problem [63], which can be solved by SDP (semi-definite programming) by learning a kernel matrix $\mathbf{K} \in \mathbb{R}^{(\tilde{\ell}+n-\ell)\times(\tilde{\ell}+n-\ell)}$. Principal Component Analysis (PCA) is then applied on the learned kernel matrix $\mathbf{K}$ to obtain a low-dimensional representation,

$$[\mathbf{U}\,\Sigma\,\mathbf{U}^T] = \mathrm{PCA}(\mathbf{K}),\ \mathbf{U} \in \mathbb{R}^{(\tilde{\ell}+n-\ell)\times k} \tag{7.8}$$

As a result of the transformation, the original data can now be represented with a reduced dimensionality of $\mathbb{R}^{k\times 1}$ in the corresponding rows of $\mathbf{U}$. Standard supervised discriminative method such as SVM in Eq.(7.1) can be used to train a model using the transformed labeled data in the auxiliary domain.

Note that as a transductive learning method, the algorithm in [63] cannot be directly used to classify out-of-sample test data, which problem is addressed in [65, 66] by learning a projection matrix to minimize the MMD [10] criteria. Si et al. [89] introduce the Bregman divergence measurement as an additional regularization term in traditional dimensionality reduction techniques to bring two domains together in the latent space.

The EigenTransfer framework [21] introduces a novel approach to integrate co-occurrence information of instance-feature, instance-label from both auxiliary and target domains in a single graph. Normalized cut [85] is then adopted to learn a low-dimensional representation from the graph to replace original data in both target and auxiliary domains. Finally, standard supervised discriminative model, e.g. SVM in Eq.(7.1) is trained using the transformed labeled data in the auxiliary domain. An advantage of EigenTrasnfer is its ability to unify almost all available information in auxiliary and target domains, allowing the consideration of heterogenous feature and label space.

**Feature Correlation** Transferring *feature correlation* from auxiliary domains to a target domain is introduced in [76, 44, 107], where a feature-feature covariance matrix $\Sigma_0 \in \mathbb{R}^{d\times d}$ estimated from some auxiliary data is taken as an additional regularization term,

$$\lambda\, \boldsymbol{w}^T \Sigma_0^{-1} \boldsymbol{w} \tag{7.9}$$

In this equation, the feature-feature correlation information is encoded in the covariance matrix $\Sigma_0$, which can be estimated from labeled or unlabeled data in auxiliary domains. $\Sigma_0$ will constrain the model parameters $w_i$ and $w_j$ of two high-correlated features $i$ and $j$ to be similar, and constrain the low-correlated features to be dissimilar. Such a regularization term is quite general and can be considered in various regularization based learning frameworks to incorporate the feature-feature correlation knowledge. Feature correlation is quite intuitive, and thus it has attracted several practical applications. For example, Raina et al. [76] transfer the feature-feature correlation knowledge from a newsgroups domain to a webpage domain for text classification, and Zhang et al. [107] study text classification with different time periods.

**Feature Subsetting** Feature selection via feature subsetting has been proposed for named entity recognition in CRF [81], which makes use of labeled data in auxiliary domains and the unlabeled data in the target domain. To illustrate the idea more clearly, we consider a simplified case of binary classification, where $y \in \{\pm 1\}$, instead of sequence labeling [81]. We re-write the optimization problem as follows,

$$\begin{aligned} \min_{\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\xi}}} \quad & \frac{1}{2}||\tilde{\boldsymbol{w}}||_2^2 + \lambda \sum_{i=1}^{\tilde{\ell}} \tilde{\xi}_i \\ \text{s.t.} \quad & \tilde{\boldsymbol{w}}^T \phi(\tilde{\boldsymbol{x}}_i, \tilde{y}_i) \geq 1 - \tilde{\xi}_i,\ \tilde{\xi}_i \geq 0,\ i = 1, \ldots, \tilde{\ell} \\ & \sum_{k=1}^{d} |\tilde{w}_k|^\gamma dist(\tilde{E}_k, E_k) \leq \epsilon \end{aligned} \tag{7.10}$$

Here we have:

$$E_k = \frac{1}{n-\ell} \sum_{i=\ell+1}^{n} (\phi_k(\boldsymbol{x}_i, +1) P_r(+1|\boldsymbol{x}_i, \tilde{\boldsymbol{w}}) + \phi_k(\boldsymbol{x}_i, -1) P_r(-1|\boldsymbol{x}_i, \tilde{\boldsymbol{w}}))$$

Furthermore, $\tilde{E}_k = \frac{1}{\tilde{\ell}} \sum_{i=1}^{\tilde{\ell}} \phi_k(\tilde{\boldsymbol{x}}_i, \tilde{y}_i)$ are expected values of the $k$th feature of the joint feature mapping function $\phi(X, Y)$ in the target and auxiliary data, respectively, and $P_r(+1|\boldsymbol{x}_i, \tilde{\boldsymbol{w}}))$ and $P_r(-1|\boldsymbol{x}_i, \tilde{\boldsymbol{w}}))$ are the posterior probabilities of instance $\boldsymbol{x}_i$ belonging to classes $+1$ and $-1$, respectively. The parameter $\gamma$ is used to control the sparsity of the model parameter $\tilde{\boldsymbol{w}}$, which produces a subset of non-zeros; this is why it is called feature subsetting. The distance $dist(\tilde{E}_k, E_k)$ can be square distance $(\tilde{E}_k - E_k)^2$ for optimization simplicity [81], which is used to punish *highly distorted features* in order to bring two domains closer. The trained model $\tilde{\boldsymbol{w}}$ will have better prediction performance in the target domain, especially when some features distort seriously in two domains.

**Feature Weighting** Arnold et al. [2] propose a feature weighting (or rescaling) approach to bridge two domains with labeled data in the auxiliary domain and test data in the target domain. Specifically, the $k$th feature of instance $\tilde{\boldsymbol{x}}_j$ in the auxiliary domain is weighted as follows,

$$\tilde{x}_{j,k} \rightarrow \tilde{x}_{j,k} \frac{E_k(\tilde{y}_j|\mathbf{X}_U, \tilde{\boldsymbol{w}})}{\tilde{E}_k(\tilde{y}_j|\tilde{D}_L)} \tag{7.11}$$

where $E_k(\tilde{y}_j|\mathbf{X}_U, \tilde{\boldsymbol{w}}) = \frac{1}{n-\ell}\sum_{i=\ell+1}^{n} x_{i,k}\, P_r(\tilde{y}_j|\boldsymbol{x}_i, \tilde{\boldsymbol{w}})$ is the expected value of $k$th feature (belonging to class $\tilde{y}_j$) in the target domain using the trained MaxEnt model $\tilde{\boldsymbol{w}}$ from auxiliary domain. The value $\tilde{E}_k(\tilde{y}_j|\tilde{D}_L) = \frac{1}{\tilde{\ell}}\sum_{i=1}^{\tilde{\ell}} \tilde{x}_{i,k}\, \delta(\tilde{y}_j, \tilde{y}_i)$ represents the expected value of $k$th feature (belonging to class $\tilde{y}_j$) in the auxiliary domain. The weighted data (feature) in the auxiliary domain then have the same expected values of joint distribution about $k$th feature and class label $y$, $\tilde{E}_k(y|\tilde{D}_L) = E_k(y|\mathbf{X}_U, \tilde{\boldsymbol{w}}), y \in \mathcal{Y}$. As a result, the two domains are brought closer together. Note that the learning procedure can be iterated with (a) learning $\tilde{\boldsymbol{w}}$ and (b) weighting the feature, and that is the reason the model is called IFT (iterative feature transformation) [2]. Since $E_k(\tilde{y}_j|\mathbf{X}_U, \tilde{\boldsymbol{w}})$ is only an estimated value, [2] adopts a common trick to preserve the original feature, which works quite well in NER problems. In particular,

$$\tilde{x}_{j,k} \rightarrow \lambda\, \tilde{x}_{j,k} + (1-\lambda)\, \tilde{x}_{j,k} \frac{E_k(\tilde{y}_j|\mathbf{X}_U, \tilde{\boldsymbol{w}})}{\tilde{E}_k(\tilde{y}_j|\tilde{D}_L)} \tag{7.12}$$

where $0 \leq \lambda \leq 1$ is a tradeoff parameter.

In the same spirit, other feature-based transfer methods have also been proposed, such as distance minimization [4], feature clustering [22, 57], kernel mapping [109], etc.

## 3. Heterogeneous Transfer Learning

Above we have surveyed transfer learning tasks where both the source and target domains are text documents in English. Recently, researchers in transfer learning area have started to consider transfer learning across heterogeneous feature and/or label space, namely heterogeneous transfer learning (HTL) [101]. HTL can be roughly categorized into two branches, (1) heterogeneous feature space, e.g. text and image space [20, 101, 87, 112, 72], English and Chinese vocabulary space [56, 105], and (2) heterogeneous label space, e.g. label space of Open Directory Project (ODP) [2] and query categories in KDDCUP 2005 [3] [84, 83], label space

[2] http://dmoz.org/

[3] http://www.sigkdd.org/kdd2005/kddcup.html

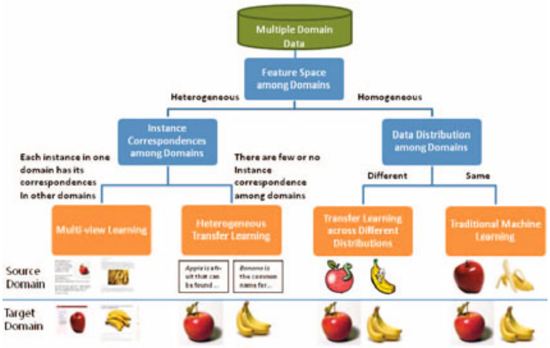


*Figure 7.4.* An intuitive illustration of heterogeneous transfer learning via classification of the images of `apple` and `banana` (quoted from [101]).

in Yahoo! Directory [4] and ODP [62], "head" (frequent) categories and "tail" (infrequent) categories in label-frequency distribution, and document categories in Newsgroup and categories in Wikipedia [98].

In Figure 7.4, we show different kinds of transfer learning and their relations to heterogeneous transfer learning. When features (or labels) are different between different domains, as shown on the left side of the figure, we have heterogeneous transfer learning when the instances in different domains lack a direct correspondence.

In general, recent works of heterogeneous transfer learning (HTL) can be classified into the following categories:

**HTL for Image Classification** An example is heterogeneous transfer learning for image classification [112]). In this work Zhu et al. consider how to use unlabeled text documents that we find on the Web to help boost the performance of image classification, by exploiting their semantic level similarity when the labeled images are in short supply.

**HTL for Image Clustering** An example of this direction is heterogeneous transfer learning for image clustering, where Yang et al. proposed a heterogenous transfer learning algorithm for image clustering by levering auxiliary annotated images ([101]).

**HTL Across Different label Space** An example is the cross-category learning in [73]. In this work, it adapts Adaboost with learning a feature correlation matrix to transfer knowledge from frequent categories to infrequent categories.

[4] `http://dir.yahoo.com/`

## 3.1 Heterogeneous Feature Space

Dai et al. [20] propose a novel approach named translated learning via risk minimization (TLRisk) to achieve knowledge transfer from text to image for image classification. The key idea is to bridge heterogeneous feature space in two domains via the co-occurrence information of image-feature and text-feature (or feature-level translator [20]) contained in the annotated auxiliary images, e.g. annotated images in Flickr. The knowledge in an auxiliary domain is then transferred along the path,

auxiliary-label $\rightarrow$ auxiliary-feature $\rightarrow$ target-feature $\rightarrow$ target-label

The TLRisk model is formulated in the risk minimization framework combining the feature translator and nearest neighbor learning, and is empirically studied for both image classification and cross-lingual (from English to German) text classification.

Yang et al. [101] proposed a probabilistic approach named annotation-based probabilistic latent semantic analysis (aPLSA) to achieve knowledge transfer from text to image for image clustering. Some multi-view auxiliary data of images and text is first transformed to a new representation of correlations between image-feature and text-feature. The aPLSA model [101] then discovers latent topics of image features of both multi-view data and target image data, which are shared as a bridge to bring two domains together.

Zhu et al. [112] propose a matrix-factorization based approach named heterogeneous transfer learning for image classification (HTLIC), in order to achieve knowledge transfer from text to image for image classification. To enable classification for out-of-sample images, HTLIC adopts collective matrix factorization [90] to learn an image-feature projection matrix from the auxiliary data of documents and the multi-view data, which is then used to obtain a new representation of the target images. Finally, a classifier (e.g. support vector machine) is trained using the newly projected target images.

Given a set of images to classify, we often need to have high-quality labeled images to train a classification model. However, obtaining the labeled image data is difficult and costly. In ([112]), the following question is addressed: is it possible for us to make use of some auxiliary labeled images and large quantities of unlabeled text to help us build a classifier? Suppose that we are given a few labeled image instances $\mathbf{X} = \{\mathbf{x}_i, y_i\}_{i=1}^{n}$ where $\mathbf{x}_i \in \mathbb{R}^d$ is an input vector of image features and $y_i$ is the corresponding label of image $i$. We assume that the labeled images are not sufficient to build a high quality image classifier. In addition, we are also given a set of auxiliary annotated images $\mathbf{I} = \{\mathbf{z}_i, \mathbf{t}_i\}_{i=1}^{l}$ and a set of text documents $\mathbf{D} = \{\mathbf{d}_i\}_{i=1}^{k}$, where $\mathbf{z}_i \in \mathbb{R}^d$ is an image represented by a feature vector as $\mathbf{x}_i$, $\mathbf{t}_i \in \mathbb{R}^h$ is its corresponding vector of tags, and $h$ is the number of tags. $\mathbf{d}_i \in \mathbb{R}^m$ is a document represented by a vector of bag-of-words, and $l$ and $k$ are the numbers of auxiliary images and documents respectively. The goal is to learn an accurate image classifier $f(\cdot)$ from $\mathbf{X}$, $\mathbf{I}$ and $\mathbf{D}$ to make predictions on $\mathbf{X}^*$, $f(\mathbf{X}^*)$.

We can make use of a set of auxiliary images $\mathbf{Z} \in \mathbb{R}^{l\times d}$ with their corresponding tags $\mathbf{T} \in \mathbb{R}^{l\times h}$ from Web resources such as Flickr. We can also easily obtain a set of unlabeled text documents $\mathbf{D} \in \mathbb{R}^{k\times m}$ via a search engine. To help build an image classifier, we need to first build some connection between image features and text features. To do this, we construct a two-layer bipartite graph based on images, tags and text documents. The top layer of the bipartite graph is used to represent the relationship between images and tags. Each image can be annotated by some tags, and some images may share one or multiple tags. If two images are annotated by some common tags, they tend to be related to each other semantically. Similarly, if two tags co-occur in annotations of shared images, they tend to be related to each other. This image-tag bipartite graph is represented by a tag matrix $\mathbf{T}$. The bottom layer bipartite graph is used to represent the relationship between tags and documents. If a tag occurs in a text document, there is an edge connecting the tag and the document.

Based on the bipartite graph, we can then learn semantic features for images by exploiting the relationship between images and text from the auxiliary sources. We first define a new matrix $\mathbf{G} = \mathbf{Z}^\top\mathbf{T} \in \mathbb{R}^{d\times h}$ to denote the correlation between low-level image features and annotations which can be referred to as high-level concepts. We then apply the Latent Semantic Analysis (LSA) as described in ([25]). Finally, we apply matrix factorization to decompose $\mathbf{G}$ into latent factor matrices as $\mathbf{G} = \mathbf{U}\mathbf{V}_1^\top$, where $\mathbf{U} \in \mathbb{R}^{d\times g}$, $\mathbf{V}_1 \in \mathbb{R}^{h\times g}$, and $g$ is the number of latent factors. Then $\mathbf{u}_i$ can be treated as a latent semantic representation of the $i^{th}$ image low-level feature, and $\mathbf{v}_{1j}$ can be treated as a latent semantic representation of $j^{th}$ tag.

Zhu et al. [112] describe a method to learn the best decomposition via collective matrix factorization, as follows.

$$\min_{\mathbf{U},\mathbf{V},\mathbf{W}} \lambda \left\|\mathbf{G} - \mathbf{U}\mathbf{V}^\top\right\|_F^2 + (1-\lambda)\left\|\mathbf{F} - \mathbf{W}\mathbf{V}^\top\right\|_F^2 + R(\mathbf{U},\mathbf{V},\mathbf{W}), \tag{7.13}$$

where $0 \leq \lambda \leq 1$ is a tradeoff parameter to control the decomposition error between the two matrix factorizations, $||\cdot||_F$ denotes the Frobenius norm of matrix, and $R(\mathbf{U}, \mathbf{V}, \mathbf{W})$ is the regularization function to control the complexity of the latent matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$. The optimization problem is an unconstrained non-convex optimization problem with three matrix variables **U**, **V** and **W**, thus it only has local optimal solutions. However, (7.13) is convex with respect to any one of the three matrices while fixing the other two. Thus a common technique to solve this kind of optimization problem is to fix two matrices and optimize the left one iteratively until the results converge.

Qi et al. [72] adopt Singular value thresholding (SVT) [14] and support vector machine to learn a low-rank feature-level correlation matrix (or translator) using multi-view data (text and images), and then the labels of text can be propagated (or transferred) to images through the feature-level translator. Note that both text and images are from the multi-view data, e.g. annotated images in Flickr. The problem setting of [72] is different from that of [112], where in [112] the multi-view data is considered as a bridge to transfer knowledge from auxiliary documents to target images, while in [72] the multi-view data is considered as a two-domain data sources in which knowledge is transferred from text to image.

## 3.2 Heterogeneous Label Space

Heterogeneous transfer learning may be needed when there is label mismatch between the auxiliary and target learning domains. The problem has attracted increasing attention in transfer learning, both in text mining and image understanding. One of the earliest works in matching labels across different classification domains is on the KDDCUP 2005 dataset, which task is to classify short, ambiguous and unlabeled search queries from a search engine log into a set of predefined categories. In [84, 83], Shen et al. considered the problem of quickly adapting a query categorization classifier when the target domain label taxonomy changes in the target learning domain. Their approach was to make the use of a large intermediate taxonomy to compile a collection of classifiers, and then adapt these classifiers to the new target label taxonomy in real time.

Shi et al. presented an approach to solving the label mismatch problem by a risk-sensitive spectral partition (RSP) algorithm [86]. A multi-task learning with mutual information (MTL-MI) is developed in [74] for learning the label correspondence.

Qi et al. [73] use quadratic programming (QP) to learn a diagonal feature-level correlation matrix on single-view data (e.g. image or video), and then use the AdaBoost framework to transfer knowledge from "head" (frequent) categories to "tail" (infrequent) categories, e.g.

from mountain images to castle images. In both [72] and [73], the decision function for a target instance is defined as a weighted linear combination of labels of auxiliary instances, where the *weight* is represented as the *similarity* of the target instance and any auxiliary instance estimated via learning a feature-level correlation matrix. The difference between [72] and [73] is that the former works on heterogeneous feature space (e.g. text and images) but same label space, while the latter focus on same feature space (e.g. images) but heterogeneous label space (e.g. semantically related categories of mountain and castle).

Rohrbach et al. [79] propose to automatically mine semantic relationships between class labels (or equivalently class attributes) from linguistic data (e.g. wikepedia, WordNet, Yahoo image, Flickr), which can be considered as a label-level translator. The trained classifiers of auxiliary classes can then be reused by target domain (different) classes through the label-level translator and Bayesian rules. The proposed approach allows different label space but assuming same feature space, and is empirically verified for image classification. A follow-up work [78] conducts extensive and in-depth study of transfer learning for image classification.

Xiang et al. [98] propose a novel approach named source-selection-free transfer learning (SSFTL) to achieve knowledge transfer from some large-scale auxiliary data set, e.g. Wikipedia, which does not require practitioners to manually select some particular part of auxiliary data to transfer from. The main idea is to bridge large-scale auxiliary label space and target label space via social tagging data, e.g. Flick. Specifically, each label (*scalar*) is represented as a *vector* in a latent space, where two vectors are similar if the corresponding labels are semantically correlated. An additional advantage of SSFTL is that the training procedure of auxiliary classifiers can be implemented offline, which makes the whole learning approach very efficient.

There are also some other heterogeneous transfer learning settings in different data domains and scenarios e.g. target domains with few instances [50], transfer from text to video [51], etc.

## 3.3 Summary

Heterogeneous transfer learning is mainly based on feature-level translator and label-level translator, which bridges heterogeneous feature space and heterogeneous label space of two domains. The techniques of heterogeneous transfer learning and transfer learning methods in previous sections are complementary, which enables knowledge transfer in a much wider application scope with very little limitation.

*Table 7.4.* Learning paradigms and techniques. The notation "req." means that the test data are *required* during model training, and "$\surd$" means the corresponding data are available to the learner. $\tilde{D}_L$ and $\tilde{D}_U$ are labeled and unlabeled data in an auxiliary domain. $D_L$, $D_U$ and $D_T$ are labeled, unlabeled and test data in the target domain. Unsupervised and supervised transfer learning are categorized by the availability of labeled data in the target domain.

| **Learning Paradigm** | | **Auxiliary** | | **Target** | | | **Learning Technique** |
|---|---|---|---|---|---|---|---|
| | | $\tilde{D}_L$ | $\tilde{D}_U$ | $D_L$ | $D_U$ | $D_T$ | |
| ML | Unsupervised | N/A | | | | req. | Spectral clustering [58], etc. |
| | Transductive | | | $\surd$ | | req. | TSVM [41], etc. |
| | Supervised | | | $\surd$ | | | AdaBoost [31], etc. |
| | Semi-supervised | | | $\surd$ | $\surd$ | | SSL [111], etc. |
| TL | *Unsupervised* | | $\surd$ | | | req. | STC [24], etc. |
| | | $\surd$ | | | | req. | LWE [32], etc. |
| | | $\surd$ | | | $\surd$ | | SCL [8], etc. |
| | *Supervised* | $\surd$ | | $\surd$ | | | MTL [30], etc. |
| | | $\surd$ | | $\surd$ | $\surd$ | | TrAdaBoost [23], etc. |
| | | $\surd$ | $\surd$ | $\surd$ | $\surd$ | req. | EigenTransfer [21], etc. |
| | | | $\surd$ | $\surd$ | | | STL [75], etc. |
| | *Heterogeneous* | across different feature space | | | | | Translated learning [20]<br>aPLSA [101]<br>TTI [72]<br>HTLIC [112], etc. |
| | | across different label space | | | | | RSP [86]<br>CCTL [73]<br>Semantic relatedness [79]<br>SSFTL [98], etc. |

# 4. Discussion

Above we have seen that there are several important applications of transfer learning. What insights can be gained from these applications and extensions on transfer learning? Below, we consider a few such issues.

**What, How and When to Transfer** As pointed out by Pan and Yang [67], there are three fundamental questions in transfer learning, namely "what to transfer", "how to transfer" and "when to transfer". We have answered the "what to transfer" question from two perspectives, (1) instance-based transfer and (2) feature-based transfer, where the corresponding knowledge are selected and weighted instances and

*Table 7.5.* Applications in text mining.

| Application | Transfer learning work |
|---|---|
| Text classification | [29, 76, 22, 107, 63, 65, 66, 89, 21, 97, 70, 71, 57], etc. |
| Sentiment analysis | [46, 7, 71, 97, 64], etc. |
| Named entity recognition | [39, 2, 37, 81], etc. |
| Part-of-speech tagging | [8, 39, 4, 37], etc. |
| Relation extraction | [38], etc. |

learned or transformed features. The "how to transfer" question [67] is quite related to "what to transfer", and we have surveyed *instance weighting*, *feature projection* and other various techniques adopted in different works to achieve knowledge transfer. The "when to transfer" question [67] is related to negative transfer, cross-domain validation and transfer bounds, where some works focus on empirical study to avoid negative transfer [28, 102, 16]. Some research works also focus on theoretical developments of transfer learning, such as [4, 53, 23, 60, 109, 46]. In addition, researchers have also proposed cross-domain cross-validation strategies [110, 12] for text mining and other learning tasks.

**Learning Paradigms and Techniques** Transfer learning can be considered as a new *learning paradigm*. One perspective is to consider transfer learning as an over-arching framework that includes the traditional learning as a special case, as shown in Table 7.4. Here we can see that traditional machine learning (ML) methods do not consider data from auxiliary domains; instead and they study the learning problems under the same data distribution $P_r(X, Y)$. In contrast, transfer learning goes beyond the learning paradigm via transferring knowledge from auxiliary domains with different distribution $\tilde{P}_r(X, Y) \neq P_r(X, Y)$.

**Text Mining Applications** As we surveyed so far, transfer learning have been wildly adopted in various text mining applications; a summary can be found in Table 7.5. Note that many transfer learning methods surveyed in previous sections have been applied to non-text mining applications as well; e.g. in speech recognition, in image and video analysis, etc.

## 5. Conclusions

In this chapter, we have focused on transfer learning approaches for text mining. Specifically, we have reviewed transfer learning techniques in text related classification tasks, including discriminative learning and ensemble learning, and heterogeneous transfer. We have considered these learning approaches from two perspectives, namely, (1) instance-based transfer and (2) feature-based transfer. Most of the surveyed transfer learning methods are proposed or can be applied in text mining applications, e.g. text classification, sentiment analysis, POS tagging, NER and relation extraction. In addition, the introduced heterogeneous transfer techniques can explore the knowledge in text to help the learning task in other domain, such as image classification.

A current research issue is how to apply transfer learning to the learning-to-rank framework [43, 13, 99], where the ranking model in the target domain may benefit from knowledge transferred from auxiliary domains. In this area, works include model-based transfer [34], instance-based transfer [19, 33, 15] and feature-based transfer [92, 19, 3], which extend the pairwise ranking algorithms of RankSVM [43], RankNet [13], or list-wise ranking model of AdaRank [99]. We expect to see much research progress in this new direction, e.g. generalizations of learning to rank to heterogeneous settings [101].

In the future, we expect to see more extensive applications of transfer learning in text mining, where the concept of "text" can be more general. For example, we expect to see transfer learning methods to be applied to analyzing microblogging contents and structure, in association with social network mining. We also expect to see more cross-domain transfer learning approaches, for knowledge transfer between very different domains, e.g., text and videos, etc.

## References

[1] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *J. Mach. Learn. Res.*, 6:1817–1853, December 2005.

[2] Andrew Arnold, Ramesh Nallapati, and William W. Cohen. A comparative study of methods for transductive transfer learning. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops*, ICDMW '07, pages 77–82, Washington, DC, USA, 2007. IEEE Computer Society.

[3] Jing Bai, Ke Zhou, Guirong Xue, Hongyuan Zha, Gordon Sun, Belle Tseng, Zhaohui Zheng, and Yi Chang. Multi-task learning for learning to rank in web search. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1549–1552, New York, NY, USA, 2009. ACM.

[4] Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, and Artur Dubrawski. Analysis of representations for domain adaptation. In *NIPS*, 2006.

[5] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22:39–71, March 1996.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 81–88, New York, NY, USA, 2007. ACM.

[7] John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Association for Computational Linguistics*, Prague, Czech Republic.

[8] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 120–128, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[9] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM.

[10] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings of the 14th International Conference on Intelligent Systems for Molecular Biology*, pages 49–57, Fortaleza, Brazil, August 2006.

[11] Leo Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, August 1996.

[12] Lorenzo Bruzzone and Mattia Marconcini. Domain adaptation problems: A dasvm classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(5):770–787, 2010.

[13] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 89–96, New York, NY, USA, 2005. ACM.

[14] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20:1956–1982, March 2010.

[15] Peng Cai and Aoying Zhou. A novel framework for ranking model adaptation. *Web Information Systems and Applications Conference*, 0:149–154, 2010.

[16] Bin Cao, Sinno Jialin Pan, Yu Zhang, Dit-Yan Yeung, and Qiang Yang. Adaptive transfer learning. In *AAAI*, 2010.

[17] Rich Caruana. Multitask learning: A knowledge-based source of inductive bias. In *ICML*, pages 41–48, 1993.

[18] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2006.

[19] Depin Chen, Yan Xiong, Jun Yan, Gui-Rong Xue, Gang Wang, and Zheng Chen. Knowledge transfer for cross domain learning to rank. *Inf. Retr.*, 13:236–253, June 2010.

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

[21] Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu. Eigentransfer: a unified framework for transfer learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 193–200, New York, NY, USA, 2009. ACM.

[22] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 210–219, New York, NY, USA, 2007. ACM.

[23] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 193–200, New York, NY, USA, 2007. ACM.

[24] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Self-taught clustering. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, volume 307, pages 200–207. ACM, 2008.

[25] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent

semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[26] Chuong B. Do and Andrew Y. Ng. Transfer learning for text classification. In *NIPS*, 2005.

[27] Harris Drucker. Improving regressors using boosting techniques. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, pages 107–115, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[28] Eric Eaton and Marie desJardins. Set-based boosting for instance-level transfer. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 422–428, Washington, DC, USA, 2009. IEEE Computer Society.

[29] Eric Eaton, Marie Desjardins, and Terran Lane. Modeling transfer relationships between learning tasks for improved inductive transfer. In *Proceedings of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases - Part I*, ECML PKDD '08, pages 317–332, Berlin, Heidelberg, 2008. Springer-Verlag.

[30] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 109–117, New York, NY, USA, 2004. ACM.

[31] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55(1):119–139, 1997.

[32] Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 283–291, New York, NY, USA, 2008. ACM.

[33] Wei Gao, Peng Cai, Kam-Fai Wong, and Aoying Zhou. Learning to rank only using training data from related domain. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 162–169, New York, NY, USA, 2010. ACM.

[34] Bo Geng, Linjun Yang, Chao Xu, and Xian-Sheng Hua. Ranking model adaptation for domain-specific search. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 197–206, New York, NY, USA, 2009. ACM.

[35] James J Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–61, January 1979.

[36] David W. Hosmer and Stanley Lemeshow. *Applied logistic regression.* Wiley-Interscience, 2 edition, 2000.

[37] Hal Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[38] Jing Jiang. Multi-task transfer learning for weakly-supervised relation extraction. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1012–1020, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[39] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *ACL*, 2007.

[40] Wei Jiang, Eric Zavesky, Shih-Fu Chang, and Alexander C. Loui. Cross-domain learning methods for high-level visual concept classification. In *ICIP*, pages 161–164, 2008.

[41] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pages 200–209, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

[42] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[43] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 133–142, New York, NY, USA, 2002. ACM.

[44] Eyal Krupka and Naftali Tishby. Incorporating prior knowledge on features into learning. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, 2007.

[45] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[46] Abhishek Kumar, Avishek Saha, and Hal Daumé III. A co-regularization based semi-supervised domain adaptation. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2010.

[47] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[48] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

[49] David Dolan Lewis. Reuters-21578 test collection. http://www.daviddlewis.com/.

[50] Fei-Fei Li, Fergus Rob, and Perona Pietro. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28:594–, April 2006.

[51] Liangda Li, Ke Zhou, Gui-Rong Xue, Hongyuan Zha, and Yong Yu. Video summarization via transferrable structured learning. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 287–296, New York, NY, USA, 2011. ACM.

[52] Xiao Li and Jeff Bilmes. Regularized adaptation of discriminative classifiers. In *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.

[53] Xiao Li and Jeff Bilmes. A bayesian divergence prior for classifier adaptation. In *Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-2007)*, March 2007.

[54] Xiao Li, Jeff Bilmes, and Joh Malkin. Maximum margin learning and adaptation of MLP classifers. September 2005.

[55] Xuejun Liao, Ya Xue, and Lawrence Carin. Logistic regression with an auxiliary data source. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 505–512, New York, NY, USA, 2005. ACM.

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

[57] Mingsheng Long, Wei Cheng, Xiaoming Jin, Jianmin Wang, and Dou Shen. Transfer learning via cluster correspondence inference. In *ICDM*, pages 917–922, 2010.

[58] Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, December 2007.

[59] Olvi L. Mangasarian. Generalized support vector machines. Technical report, Computer Sciences Department, University of Wisconsin, 1998.

[60] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2008.

[61] Andrew Kachites McCallum. Simulated/real/aviation/auto usenet data. http://www.cs.umass.edu/~mccallum/code-data.html.

[62] Tiberio Caetano S. V. N. Vishwanathan Novi Quadrianto, Alex Smola and James Petterson. Multitask learning without label correspondences. In *NIPS*, 2010.

[63] Sinno Jialin Pan, James T. Kwok, and Qiang Yang. Transfer learning via dimensionality reduction. In *AAAI*, pages 677–682, 2008.

[64] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. In *WWW*, pages 751–760, 2010.

[65] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. In *IJCAI*, pages 1187–1192, 2009.

[66] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011.

[67] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[68] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2:1–135, January 2008.

[69] David Pardoe and Peter Stone. Boosting for regression transfer. In *ICML*, pages 863–870, 2010.

[70] Peter Prettenhofer and Benno Stein. Cross-language text classification using structural correspondence learning. In *ACL*, pages 1118–1127, 2010.

[71] Peter Prettenhofer and Benno Stein. Cross-lingual adaptation using structural correspondence learning. *ACM TIST*, 3(1), 2012.

[72] Guo-Jun Qi, Charu Aggarwal, and Thomas Huang. Towards semantic knowledge propagation from text corpus to web images. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 297–306, New York, NY, USA, 2011. ACM.

[73] Guo-Jun Qi, Charu Aggarwal, Yong Rui, Qi Tian, Shiyu Chang, and Thomas Huang. Towards cross-category knowledge propagation for learning visual concepts. In *CVPR*, 2011.

[74] Novi Quadrianto, Alex J. Smola, Tiberio S. Caetano, S.V.N. Vishwanathan, and James Petterson. Multitask learning without label correspondences. In *NIPS 23*, 2010.

[75] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 759–766, New York, NY, USA, 2007. ACM.

[76] Rajat Raina, Andrew Y. Ng, and Daphne Koller. Constructing informative priors using transfer learning. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 713–720, New York, NY, USA, 2006. ACM.

[77] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, April 1996.

[78] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[79] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where - and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917, 2010.

[80] Stefan Rüping. Incremental learning with support vector machines. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, pages 641–642, Washington, DC, USA, 2001. IEEE Computer Society.

[81] Sandeepkumar Satpal and Sunita Sarawagi. Domain adaptation of conditional probability models via feature subsetting. In *Proceedings of the 11th European conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD 2007, pages 224–235, Berlin, Heidelberg, 2007. Springer-Verlag.

[82] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[83] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey Junfeng Pan, Kangheng Wu, Jie Yin, and Qiang Yang. Query enrichment for web-query classification. *ACM Trans. Inf. Syst.*, 24:320–352, July 2006.

[84] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. In *Proceedings of the 29th*

*annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 131–138, New York, NY, USA, 2006. ACM.

[85] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:888–905, August 2000.

[86] Xiaoxiao Shi, Wei Fan, Qiang Yang, and Jiangtao Ren. Relaxed transfer of different classes via spectral partition. In *ECML/PKDD*, 2009.

[87] Xiaoxiao Shi, Qi Liu, Wei Fan, Philip S. Yu, and Ruixin Zhu. Transfer learning on heterogenous feature spaces via spectral transformation. In *ICDM*, pages 1049–1054, 2010.

[88] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[89] Si Si, Dacheng Tao, and Bo Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Trans. Knowl. Data Eng.*, 22(7):929–942, 2010.

[90] Ajit P. Singh and Geoffrey J. Gordon. Relational learning via collective matrix factorization. In *KDD*, pages 650–658, 2008.

[91] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, 2:45–66, March 2002.

[92] Bo Wang, Jie Tang, Wei Fan, Songcan Chen, Zi Yang, and Yanzhu Liu. Heterogeneous cross domain ranking in latent space. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 987–996, New York, NY, USA, 2009. ACM.

[93] Hua-Yan Wang, Vincent Wenchen Zheng, Junhui Zhao, and Qiang Yang. Indoor localization in multi-floor environments with reduced effort. In *PerCom*, pages 244–252, 2010.

[94] Yang Mu Lourenco Bandeira Ricardo Ricardo Youxi Wu Zhenyu Lu Tianyu Cao Xindong Wu Wei Ding, Tomasz F. Stepinski. Subkilometer crater discovery with boosting and transfer learning. *ACM TIST*, x(x), 201x.

[95] Philip C. Woodland. Speaker adaptation for continuous density hmms: a review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, ITRW' 01, pages 29–30, August 2001.

[96] Pengcheng Wu and Thomas G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 110–, New York, NY, USA, 2004. ACM.

[97] Evan Wei Xiang, Bin Cao, Derek Hao Hu, and Qiang Yang. Bridging domains using world wide knowledge for transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(6):770–783, 2010.

[98] Evan Wei Xiang, Sinno Jialin Pan, Weik Pan, Qiang Yang, and Jian Su. Source-free transfer learning. In *IJCAI*, 2011.

[99] Jun Xu and Hang Li. Adarank: a boosting algorithm for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 391–398, New York, NY, USA, 2007. ACM.

[100] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 188–197, New York, NY, USA, 2007. ACM.

[101] Qiang Yang, Yuqiang Chen, Gui-Rong Xue, Wenyuan Dai, and Yong Yu. Heterogeneous transfer learning for image clustering via the social web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 1–9, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[102] Yi Yao and Gianfranco Doretto. Boosting for transfer learning with multiple sources. In *CVPR*, pages 1855–1862, 2010.

[103] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 114–, New York, NY, USA, 2004. ACM.

[104] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *J. Mach. Learn. Res.*, 3:1083–1106, March 2003.

[105] Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. Cross-lingual latent topic extraction. In *ACL*, pages 1128–1137, 2010.

[106] Tong Zhang and David Johnson. A robust risk minimization based named entity recognition system. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -*

*Volume 4*, CONLL '03, pages 204–207, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[107] Yi Zhang, Jeff Schneider, and Artur Dubrawski. Learning the semantic correlation: An alternative way to gain from unlabeled text. In *NIPS*, pages 1945–1952, 2008.

[108] Vincent Wenchen Zheng, Derek Hao Hu, and Qiang Yang. Cross-domain activity recognition. In *Proceedings of the 11th international conference on Ubiquitous computing*, Ubicomp '09, pages 61–70, New York, NY, USA, 2009. ACM.

[109] Erheng Zhong, Wei Fan, Jing Peng, Kun Zhang, Jiangtao Ren, Deepak Turaga, and Olivier Verscheure. Cross domain distribution adaptation via kernel mapping. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1027–1036, New York, NY, USA, 2009. ACM.

[110] Erheng Zhong, Wei Fan, Qiang Yang, Olivier Verscheure, and Jiangtao Ren. Cross validation framework to choose amongst models and datasets for transfer learning. In *ECML/PKDD (3)*, pages 547–562, 2010.

[111] Xiaojin Zhu, Andrew B. Goldberg, Ronald Brachman, and Thomas Dietterich. *Introduction to Semi-Supervised Learning.* Morgan and Claypool Publishers, 2009.

[112] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.