Discriminative Factor Alignment across Heterogeneous Feature Space

Fangwei Hu[1], Tianqi Chen[1], Nathan N. Liu[2], Qiang Yang[2], and Yong Yu[1]

[1] Shanghai Jiao Tong University, Shanghai, China
{hufangwei,tqchen,yyu}@apex.sjtu.edu.cn
[2] Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong*
{nliu,qyang}@cse.ust.hk

Abstract. Transfer learning as a new machine learning paradigm has gained increasing attention lately. In situations where the training data in a target domain are not sufficient to learn predictive models effectively, transfer learning leverages auxiliary source data from related domains for learning. While most of the existing works in this area are only focused on using the source data with the same representational structure as the target data, in this paper, we push this boundary further by extending transfer between text and images.

We integrate documents , tags and images to build a heterogeneous transfer learning factor alignment model and apply it to improve the performance of tag recommendation. Many algorithms for tag recommendation have been proposed, but many of them have problem; the algorithm may not perform well under cold start conditions or for items from the long tail of the tag frequency distribution. However, with the help of documents, our algorithm handles these problems and generally outperforms other tag recommendation methods, especially the non-transfer factor alignment model.

1 Introduction

Tag recommendation has found many applications ranging from personal photo albums to multimedia information delivery. In the past, tag recommendation has met two major difficulties. First, the annotated images for training are often in short supply, and annotating new images involves much human labor. Hence, annotated training data is often sparse, and further tags included in training data may be from the long tail of the frequency distribution. Second, words usually have synonym; e.g. two words may have same or similar meaning. We would like to find the latent links between these words. How to effectively overcome these difficulties and build a good tag recommendation system therefore becomes a challenging research problem. While annotated images are expensive, abundant text data are easier to obtain. This motivates us to find way to use the readily available text data to help improve the tag recommendation performance.

In the past, several approaches have been proposed to solve the 'lack of data' problem. Recently, transfer learning methods [1] have been proposed to use knowledge from

* Part of this work was done while the first author was a visiting student in Hong Kong University of Science Technology.

P. Flach et al. (Eds.): ECML PKDD 2012, Part II, LNCS 7524, pp. 757–772, 2012.

auxiliary data in a different but related domain to help learn the target tasks. However, a commonality among most transfer learning methods so far is that the data from different domains have the same feature space.

In some scenarios, given a target task, one may easily collect much auxiliary data that are represented in a different feature space. For example, our task is to recommend tags for an image with a tiger in it. We have only a few annotated images for training. And we can easily collect a large number of text documents from the Web, e.g. Wikipedia. In this case, we can model the tag recommendation task as the target task, where we have some annotated data and some auxiliary data. In the target domain, the data are represented in pixels. Also in our case, the auxiliary domain, or the source domain, is the text domain, which contains text documents. Now, what we care about is whether it is possible to use the cheap auxiliary data to help improve the performance of the tag recommendation task. This is an interesting and difficult question, since the relationship between text and images is not explicitly given. This problem has been referred to as a *Heterogeneous Transfer Learning* problem [2]. In this paper, we focus on heterogeneous transfer learning for tag recommendation by exploring knowledge transfer from auxiliary text data.

In tag recommendation, a key issue for us to address is to discover a new and improved common representation for both images and tags to boost the recommendation performance. In this paper, we investigate how to obtain a reasonable common feature space from both annotated images and auxiliary text data. Although images and text are represented in different feature spaces, we can transfer knowledge from text to images via tags which are related both to images and text. We propose a factor alignment model to discover the common feature space and modify it into a heterogeneous transfer learning factor alignment model, which can handle the auxiliary data effectively. A common space is then learned to better calculate the metric between an image and a tag. We illustrate the overall framework in Figure 1. Compared to self-taught learning [3], our approach can use a different feature representation (i.e., text) for transfer learning. Compared to translated learning [4] and Zhu et al. [5], their tasks are different from ours and our approach does not need to compute the total correlation between image feature and word features.

2 Related Work

2.1 Image Annotation

A closely related area is image annotation. Duygulu et al. [6] regarded image annotation as a machine translating process. Some other researchers model the joint probability of images regions and annotations. Barnard et al. [7] investigated image annotation under probabilistic framework and put forward a number of models for the joint distribution of image blobs and words. Blei et al. [8] developed *correspondence latent Dirichlet allocation* to model the joint distribution. In Lavrenko et al. [9], the *continuous-space relevance model* was proposed to better handle continuous feature and be free from the influence of image blob clustering. In Carneiro et al. [10], image annotation is posed

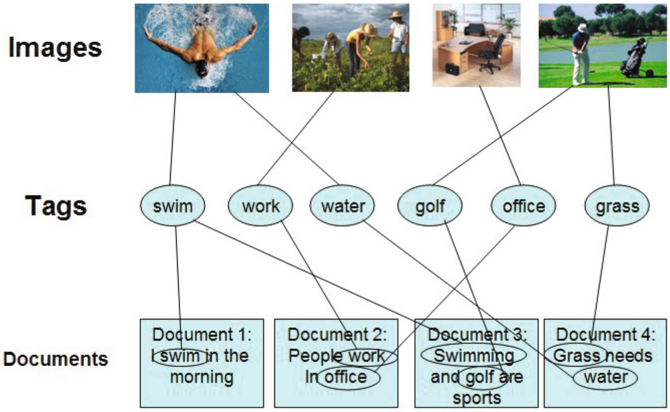

Images
Tags
Documents
swim
work
water
golf
office
grass
Document 1: I swim in the morning
Document 2: People work In office
Document 3: Swimming and golf are sports
Document 4: Grass needs water

Fig. 1. Source data used for our transfer learning algorithms. Our proposed *heterogeneous transfer learning for tag recommendation* takes all three information, i.e. images, tags and documents as inputs.

as classification problems where each class is defined by images sharing a common semantic label. In this paper, we use an image annotation algorithm proposed in Makadia et al. [11] to solve the problem of tag recommendation as a baseline.

2.2 Image Retagging

There are some efforts on improving unreliable descriptive keywords of images. They focus on imprecise annotation refinement, i.e., identifying and eliminating the imprecise annotation keywords produced by the automatic image annotation algorithms. As a pioneering work, Jin et al. [12] used WordNet to estimate the semantic correlation among the annotated keywords and remove the weakly-correlated ones. However, this method can only achieve limited success since it totally ignores the visual content of the images. To address this problem, Wang et al. [13] proposed a content-based approach to re-rank the automatically annotated keywords of an image and only reserve the top ones as the refined results and Liu et al. [14] proposed to refine the tags based on the visual and semantic consistency residing in the social images, which assigns similar tags to visually similar images. Later, Liu et al. [15] formulated this retagging process as a multiple graph-based multi-label learning problem, which simultaneously explores the visual content of the images, semantic correlation of the tags and the prior information provided by users.

2.3 Visual Contextual Advertising

Another closely related area is visual contextual advertising, which aims to recommend advertisements for Web images without the help of any textual context, such as surrounding text for the images. In Chen et al. [16], they exploit the annotated image data

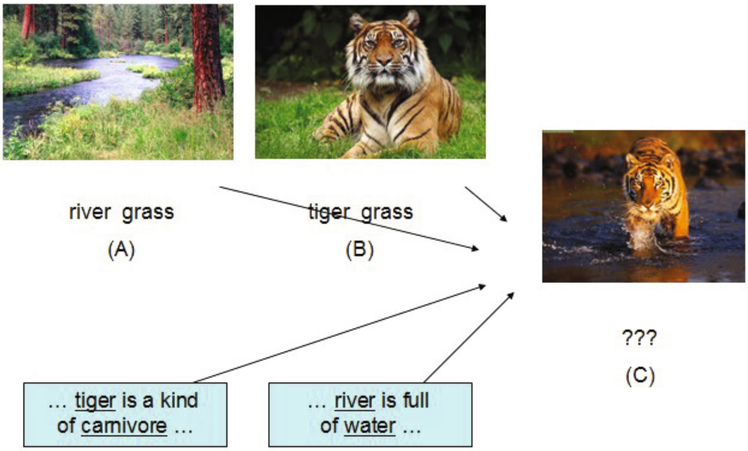

river grass
(A)
tiger grass
(B)
???
(C)
... tiger is a kind of carnivore ...
... river is full of water ...

Fig. 2. A case in which document may help the tag recommendation

from social Web sites such as Flickr to link the visual feature space and the word space. To be specific, they present a unified generative model, ViCAD, to deal with visual contextual advertising. ViCAD runs in several steps. First, they model the visual contextual advertising problem with a Markov chain which utilizes annotated images to transform images from the image feature space to the word space. With the representations of images in word space, a language model for information retrieval is then applied to find the most relevant advertisements. If we regard tag as one-word advertisement, ViCAD can handle the problem of tag recommendation. Hence we use ViCAD as one of our baselines.

3 Motivation

Before describing our proposed method in detail we first illustrate a motivating example showing how text data help improve the tag recommendation performance for images. As shown in Figure 2, we may have image A and image B as training data and image C as testing data. As mentioned before, we can't recommend other tags except 'river' and 'tiger' for image C. However, by using some additional auxiliary text documents where tags co-occur frequently, we may establish a strong similarity between tags. If we have a document including '*... a tiger is a kind of carnivore ...*', we will build a similarity between 'tiger' and 'carnivore', which may cause the recommendation system to recommend 'carnivore' for image C. And if we have a document including '*... a river is full of water ...*', 'river' and 'water' will also be regarded as being related. Consequently, 'water' may be also recommended; this can reduce the data sparsity in the image domain and word domain.

4 Algorithm

4.1 Problem Formulation

First we define the problem of our tag recommendation task formally. Suppose we are given annotated data instances $X = \{z_i, t_i\}_{i=1}^{n}$ and some test images $X^* = \{z_i^*, t_i^*\}_{i=n+1}^{n+m}$, where $z_i \in \mathbb{R}^d$ is an input vector of image features and $t_i \in \mathbb{R}^h$ is the corresponding tags of image i, where h is the number of tags. For example, if an image z_i is annotated by tags α and β with $\alpha, \beta \in \{1, \ldots, h\}$, then $t_i = [0, \ldots, 1, \ldots, 1, \ldots, 0]$ is a vector of dimensionality h with all zeros but one's in the α and β positions. Using "bag-of-words" [17] to represent image features, we can assume that the feature values are nonnegative. n and m are the numbers of training and testing instances, respectively. In addition, we also have a set of auxiliary text documents $F = \{f_i\}_{i=1}^{k}$, $f_i \in \mathbb{R}^s$ is a document represented by a vector of bag-of-words, and k is the number of auxiliary documents. Here, we notice that s doesn't have to be equal to h. Actually set of tags for words is a subset of the set for documents, which means $h \leq s$. Our goal is to learn a function $g(\cdot, \cdot)$ from X and F, that can estimate the correlation between a given image and a tag as accurately as possible. Applying $g(\cdot, \cdot)$ on X^*, we can rank the correlation to obtain the recommended tag list for test images. We summarize the problem definition in Table 1. For convenience, we denote $Z = \{z_i\}_{i=1}^{l} \in \mathbb{R}^{l \times d}$ and $T = \{t_i\}_{i=1}^{l} \in \mathbb{R}^{l \times h}$ the image features and text tags of images separately. Furthermore, we abuse the notation X, X^*, Z, and T to represent the data matrices with instances x_i, x_i^*, z_i, and t_i being row vectors in them.

Table 1. Problem formulation

Learning objective	Make predictions on target test images
Target tag recommendation	Training images: $X = \{z_i, t_i\}_{i=1}^{n}$ Testing images: $X^* = \{z_i^*, t_i^*\}_{i=n+1}^{n+m}$
Auxiliary source data	Text documents: $F = \{f_i\}_{i=1}^{k}$

4.2 Algorithm Description

Given a set of images $Z \in \mathbb{R}^{l \times d}$ with their corresponding tags $T \in \mathbb{R}^{l \times h}$, and a set of documents $F \in \mathbb{R}^{k \times s}$, we wish build a connection between images and text documents. Each image can be annotated by tags, and some images may share one or multiple tags. If two images are annotated by shared 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 via the tag matrix T. If a tag, more precisely, the text word of the tag, occurs in a document, then there is an edge connecting the tag and the document. We use a matrix $F \in \mathbb{R}^{k \times s}$ to represent the document-tag bipartite graph, where $F_{ij} = n$ if the j^{th} tag appears n times in the i^{th} document.

Image-Tag Ranking Model. In this section, we would like to first give the ranking model only leveraging the information of annotated images. Our intuitive idea is to project the image feature space and tag feature space to a common latent space, and then we can use dot product to calculate the correlation between an image and a tag. Hence we define $W \in \mathbb{R}^{d \times g}$ as a projection matrix to project the image feature space to the latent space and we also define $P \in \mathbb{R}^{h \times g}$as the latent space matrix of tags (i.e. $\mathbf{p}_i \in \mathbb{R}^g$ is the latent feature vector of tag i). Now for any image i and any tag j, we first project $\mathbf{z}_i \in \mathbb{R}^d$ to the common space as

$$\mathbf{c}_i = \mathbf{z}_i W \in \mathbb{R}^g \tag{1}$$

We can calculate the correlation between image i and tag j as

$$\hat{T}_{ij} = \mathbf{c}_i \mathbf{p}_j^\top = \mathbf{z}_i W \mathbf{p}_j^\top \tag{2}$$

Using matrix format to represent the formula, we obtain the equation

$$\hat{T} = ZWP^\top \tag{3}$$

Now we would like to define a loss function to measure the rank distance of the predicting matrix and real training matrix. Inspired by the metric of area under the ROC curve (AUC) [18], we define the loss function as

$$L(T, \hat{T}) = -\sum_{u,i,j} r_{ij}^{(u)} \ln \hat{r}_{ij}^{(u)} + (1 - r_{ij}^{(u)}) \ln(1 - \hat{r}_{ij}^{(u)}) \tag{4}$$

where

$$r_{ij}^{(u)} = \begin{cases} 1 \text{ }, T_{ui} - T_{uj} > 0 \\ 0 \text{ }, otherwise \end{cases} \tag{5}$$

$$\hat{r}_{ij}^{(u)} = \frac{1}{1 + e^{-(\hat{T}_{ui} - \hat{T}_{uj})}} \tag{6}$$

We notice that

$$r_{ij}^{(u)} \ln \hat{r}_{ij}^{(u)} + (1 - r_{ij}^{(u)}) \ln(1 - \hat{r}_{ij}^{(u)}) \tag{7}$$

is comfortingly symmetric (swapping i and j should leave the result invariant). In order to simplify the formula, we define

$$D(T) = \{(u, i, j) | T_{ui} - T_{uj} > 0\} \tag{8}$$

Now we can rewrite the loss function as

$$\begin{aligned} L(T, \hat{T}) &= -\sum_{u,i,j \in D(T)} r_{ij}^{(u)} \ln \hat{r}_{ij}^{(u)} + (1 - r_{ij}^{(u)}) \ln(1 - \hat{r}_{ij}^{(u)}) & (9) \\ &= -\sum_{u,i,j \in D(T)} 1 \cdot \ln \hat{r}_{ij}^{(u)} + (1 - 1) \ln(1 - \hat{r}_{ij}^{(u)}) & (10) \\ &= -\sum_{u,i,j \in D(T)} \ln \hat{r}_{ij}^{(u)} & (11) \end{aligned}$$

So far, our corresponding optimization problem becomes

$$\min_{W,P} L(T, \hat{T}) + \lambda_1 \|W\|_F^2 + \lambda_2 \|P\|_F^2 \tag{12}$$

where λ_1, λ_2 are nonnegative parameters to control the responding regularization terms, and $\| \cdot \|_F$ is the Frobenius norm. In next section, we will apply this model on text documents.

Doc-Tag Ranking Model. Now let us consider the information included by text documents, and we would like to apply the ranking model mentioned in the last section to text documents. Hence the problem in this section is to recommend tags for each text document. Similarly, we define $Q \in \mathbb{R}^{h \times g}$ as the latent space matrix of tags. However, unlike images, we directly define $B \in \mathbb{R}^{s \times g}$ as the latent space matrix of documents. Now for any document i and any tag j, we can obtain the correlation between document i and tag j as

$$\hat{F}_{ij} = \mathbf{b}_i \mathbf{q}_j^\top \tag{13}$$

The matrix format of the calculation is

$$\hat{F} = BQ^\top \tag{14}$$

Therefore, the corresponding optimization of this problem is

$$\min_{B,Q} Loss(F, \hat{F}) + \lambda_3 \|B\|_F^2 + \lambda_4 \|Q\|_F^2 \tag{15}$$

where λ_3, λ_4 are small nonnegative numbers, $\lambda_3\|B\|_F^2$ and $\lambda_4\|Q\|_F^2$ serve as a regularization term to improve the robustness.

Joint Ranking Model. There are many ways to transfer knowledge of text data to image data. Our idea in this paper is not to calculate the correlation between image features and text features, but rather to transfer via the latent space of tags. By forcing P and Q to be approximate, the ranking model only for images can incorporate the information of text data. Compared to forcing two latent matrices (i.e. P and Q) to be absolutely the same, our model uses soft constraints leveraging λ_0 to control the similarity of these two matrices, which is more flexible. $\lambda_0 = 0$ implies that P and Q are uncorrelated and documents do nothing to help learning. Then $\lambda_0 = \infty$ denotes that P should be equal to Q, which becomes a hard constraint. To achieve this goal, we solve the following problem,

$$\min_{W,B,P,Q} L(T, \hat{T}) + L(F, \hat{F}) + \lambda_0 \|P - Q\|_F^2 + R(W, B, P, Q) \tag{16}$$

where $R(W, B, P, Q)$ is the regularization function to control the complexity of the latent matrices W, B, P and Q, and λ_0 controls the strength to constrain the similarity of P and Q. As mentioned in previous sections, we define the regularization function as

$$R(W, B, P, Q) = \lambda_1 \|W\|_F^2 + \lambda_2 \|P\|_F^2 + \lambda_3 \|B\|_F^2 + \lambda_4 \|Q\|_F^2 \tag{17}$$

where λ_1, λ_2, λ_3 and λ_4 are nonnegative parameters to control the responding regularization terms. In this paper, we set $\lambda_1 = \lambda_3 = 0.004$ and $\lambda_2 = \lambda_4 = 0.006$ using cross validation. Equation 16 is the joined ranking model of T and F with regularization. The bridge of transfer P and Q are ensured to capture both the structures of image matrix T and the document matrix F. Once we find the optimal W and P and project test images to latent space, we can apply the dot product to estimate the rating for tags given by images.

Update Rule. Equation 16 can be solved using gradient methods, such as the stochastic gradient descent and quasi-Newton methods. In this paper, we use stochastic gradient descent. Then the main computation of the gradient method gives the update rule for all variables. For $(u, i, j) \in D(T)$,

$$\mathbf{p}_k = \mathbf{p}_k + \eta(\hat{e}_T(u,i,j)Z_{uk}(\mathbf{p}_i - \mathbf{p}_j) - \lambda_1 \mathbf{p}_k) \tag{18}$$

$$\mathbf{p}_k = \mathbf{p}_k + \eta(\hat{e}_T(u,i,j)\gamma_{ij}(k)(\mathbf{z}_u W) - \lambda_2 \mathbf{p}_k - \lambda_0(\mathbf{p}_k - \mathbf{q}_k)) \tag{19}$$

where

$$\hat{e}_T(u,i,j) = 1 - \frac{1}{1 + e^{-(\hat{T}_{ui} - \hat{T}_{uj})}} \tag{20}$$

and

$$\gamma_{ij}(k) = \begin{cases} 1 & , k = i \\ -1 & , k = j \\ 0 & , otherwise \end{cases} \tag{21}$$

Then for $(u, i, j) \in D(F)$,

$$\mathbf{b}_k = \mathbf{b}_k + \eta(\hat{e}_F(u,i,j)\sigma_u(k)(\mathbf{q}_i - \mathbf{q}_j) - \lambda_3 \mathbf{b}_k) \tag{22}$$

$$\mathbf{q}_k = \mathbf{q}_k + \eta(\hat{e}_F(u,i,j)\gamma_{ij}(k)\mathbf{b}_u - \lambda_4 \mathbf{q}_k - \lambda_0(\mathbf{p}_k - \mathbf{q}_k)) \tag{23}$$

Here

$$\hat{e}_F(u,i,j) = 1 - \frac{1}{1 + e^{-(\hat{F}_{ui} - \hat{F}_{uj})}} \tag{24}$$

and

$$\sigma_u(k) = \begin{cases} 1 & , k = u \\ 0 & , otherwise \end{cases} \tag{25}$$

Prediction. Now we will present the process of prediction of test image set. For any image $\mathbf{x}_i^*$ in test set X^*, we first project $\mathbf{z}_i^*$ to the common space as

$$\mathbf{d}_i = \mathbf{z}_i^* W \in \mathbb{R}^g \tag{26}$$

Therefore, the definition of $f(x_i^*, tag_j)$ is as

$$g(x_i^*, tag_j) = \mathbf{d}_i \mathbf{p}_j^\top = \mathbf{z}_i^* W \mathbf{p}_j^\top \tag{27}$$

Using matrix format to represent the prediction matrix, we obtain the equation

$$\hat{R} = Z^* W P^\top \tag{28}$$

Notice that $\hat{R} \in \mathbb{R}^{m \times h}$ can help us rank the tags for testing images. For any testing image u, if $\hat{R}_{ui} > \hat{R}_{uj}$, we can consider tag i is more related to the image than tag j. Thus we will get the recommended tag list in order by ranking the rating. Putting it together, our overall heterogeneous transfer learning algorithm is referred to as **HTLFA**, which stands for Heterogeneous Transfer Learning for Factor Alignment.

5 Experiments

Our experiments are designed to demonstrate the effectiveness of exploiting text in our heterogeneous learning algorithm.

5.1 Dataset and Processing

We use annotated images from Flickr crawled during December 2009. We collected 528,587 images and 20,000 distinct related tags. Each of these tags is a single word. We crawled text from Wikipedia as our auxiliary data, from which we picked out 539,460 documents. Each of the documents includes more than 100 tag words mentioned above.

Data preprocessing is applied to the raw data. We use the "bag-of-words" model [17] to represent each image. First interesting points were detected and described by SIFT descriptors [19]. Then we cluster a random subset of all interesting points to obtain a codebook. Similar to Sivic et al. [20], we set the number of clusters to be 1,000. Using this codebook, each image is converted into a vector for further tag-recommendation uses. One image at most has 208 tags and at least has 1 tag. On average each image has about 9 tags. And for documents, we also converted them into 30,000-dim vectors using 30,000 tag words. The extra 10000 tags are selected from popular tags included in documents which do not appear in origin 20000 tags.

Actually, we don't regard all tags not related to images (or documents) as negative samples. For an image (or document), we just sample n' tags from unrelated tags as negative sample and the other tags are regarded as neutral items. Here n' is proportional to the size of the positive set. For example, if image A has 10 annotated tags, then we randomly select $n' = 10k'$ tags from unrelated ones as negative sample. Here k' is a proportion parameter. At last, we set latent space dimension $g = 100$.

5.2 Evaluation and Baseline Methods

As for evaluation metrics, we choose the precision at n (or **P**@n), which is the portion of related tags in the topmost n recommendations to evaluate the experimental results. In order to make the results more solid, we also use **MAP** to measure the precision. To obtain ground truth results, we reserve about a quarter of all images from the Flickr image set as test set and regard the annotated tags as ground truth. Before we formulate these two evaluation metrics, we first define $M(i)$ as the number of annotated tags belonging to test image i and $loc_i(j)$ as the position of tag j in the recommendation list for test image i. Then for any test image i, we rank the loc_i by ascending order. Now we can present the expression of **P**@n,

$$\mathbf{P}@\mathrm{n} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{M(i)} \frac{\psi(loc_i(j)<n)}{n}}{m} \tag{29}$$

where m is the number of test images and $\psi(Con)$ is a condition function,

$$\psi(Con) = \begin{cases} 1 \text{ }, Con \text{ holds} \\ 0 \text{ }, otherwise \end{cases} \tag{30}$$

Then the expression of **MAP** is formulated as:

$$\mathbf{MAP} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{M(i)} \frac{j}{loc_i(j)}}{m} \tag{31}$$

We compare our proposed method with three baselines for tag recommendation. The three baselines and our proposed method are summarized as follows,

- **NN** This baseline is reported in Makadia et al. [11]. This baseline annotation methods are comprised of an image distance measure for nearest neighbor ranking, combined with a label transfer measure. In this paper, we measure the distance using SIFT feature vector of images.
- **ViCAD** We implemented the method proposed in Chen et al. [16] as another baseline, which directly transforms images from a image feature space to a word space utilizing the knowledge from images with annotations from Flickr. Then each dimension of the word space means the correlation between the image and a word. Since tags can be regarded as single word advertisements, we can directly get the recommendation list of tags by ranking the correlation.
- **FA** The name of this baseline is short for factor alignment, which is a non-transfer algorithm corresponding to the loss function 12. In this baseline, we do not use the documents to help learning the projection matrix. Without the transformation of knowledge from documents to images, we simply train a ranking model only on the image-tag matrix. As mentioned above, we apply an image-tag ranking model on training data to discover the common representation.
- **HTLFA** This denoted our proposed method, which uses all the auxiliary data i.e. documents. The parameter settings are discussed in the following section.

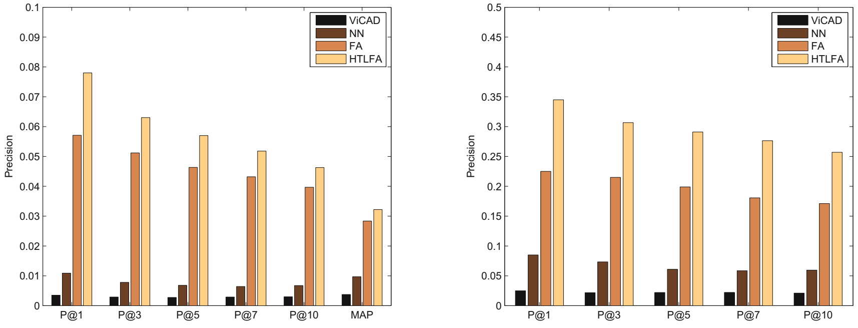
Table 2. Comparison with baselines

	ViCAD	**NN**	**FA**	**HTLFA**
P@1	0.003500	0.010900	0.057100	0.077990
P@2	0.003050	0.009200	0.054250	0.069550
P@3	0.002900	0.007800	0.051200	0.063033
P@4	0.002750	0.006975	0.048975	0.059275
P@5	0.002760	0.006820	0.046340	0.057000
P@6	0.002900	0.006550	0.044850	0.053933
P@7	0.002886	0.006400	0.043200	0.051800
P@8	0.002925	0.006513	0.041712	0.049675
P@9	0.002922	0.006511	0.040778	0.048000
P@10	0.002970	0.006740	0.039670	0.046270
MAP	0.003744	0.009711	0.028368	0.032200

5.3 Comparison with Baselines

In the first experiment, we compare our method with three baselines on the same tag recommendation task. We randomly select 10,000 images from the test set as test data. The **P**@n results with respect to **NN**, **FA**, **ViCAD** and our model are given in Figure 3(a) and Table 2. In this experiment, for **HTLFA**, we set the parameter λ_0 in Equation 16 to 0.05. As we can see from Figure 3(a), our proposed **HTLFA**, which use documents to help learn projection matrices for rating, outperforms other baselines, especially it is better than **FA**. **NN** and **ViCAD** which are generative models without learning process perform pretty poorly in this task. This implies that with the help of documents our proposed method is powerful for tag recommendation.

However, the value of precision is a bit low since the groundtruth of testdata does not include all relevant tags. In fact, we crawl the image data from Flickr which is a social image hosting website. Hence the data is not unified and the groundtruth of test data may not be correct. For example, an image with a cat has a tag "people" but has does not have a tag "pet". Consequently, it's necessary for us to do a user study to check the result, in which we randomly select 200 test images and check the top 20 tags. Since this requires much human labor, we are not able to check all 10,000 test images. Six people check the results, and the final precision is the average of these six people. Figure 3(b) displays the results of the user study, which indicates our model can obtain high precision.



(a) Comparison with baselines

(b) The result of user study

Fig. 3. Evaluation of HTLFA

5.4 Impact of Constraint Parameter

In the second experiment, we study the parameter sensitivity of λ_0 on the overall performance of **HTLFA** in tag recommendation. As mentioned before, λ_0 controls the strength of the constraint between P and Q. In this experiment we tune the value of λ_0 to obtain a set of results. Figure 4 shows the recommendation accuracy **P**@10 of **HTLFA** under varying values of λ_0. We find that **HTLFA** performs best and steadily when λ_0 falls at about 0.05 , which implies the document-tag matrix can indeed help learning a more precise latent factor matrix and a soft constraint is better than a hard constraint. Because if λ_0 approaches zero, it means we don't leverage the help of documents, and if λ_0 approaches infinity, it means we are using a hard constraint to force $P \approx Q$ instead of using a soft constraint.

Furthermore, a comparison is made between **HTLFA** and **CMF**(i.e. Collective Matrix Factorization) [21]. In **CMF**, they simultaneously factor several matrices, sharing parameters among factors when an entity participates in multiple relations. That is **CMF** transfers knowledge from a matrix to another by forcing two latent matrices, which are P and Q in our model, to be equal. Actually, equality is a special case of our model. If we let $\lambda_0 = \infty$, P and Q will be forced to be equal, which makes our model more flexible. Since it's not feasible to set $\lambda_0 = \infty$, we directly use two equal latent matrices(i.e. $P = Q$) abandoning the term $\lambda_0 \|P - Q\|$, which is called hard constrain. Figure 5 shows that a soft constraint outperforms a hard constraint.

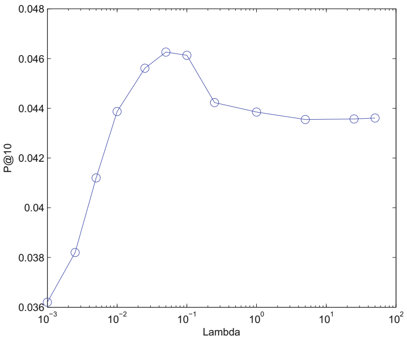


Fig. 4. Varying values of λ_0

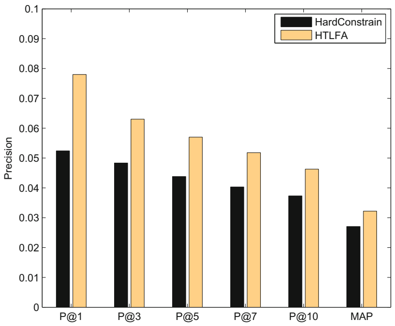


Fig. 5. Comparison between **CMF** and **HTLFA**

5.5 Performance under Longtail Condition

In the third experiment, we also analyze the impact of the amount of training data on the performance of **FA** and **HTLFA** in tag recommendation. We randomly select 30,000, 50,000, 80,000, 120,000 images from all training data to train four models, and evaluate the results. The experimental results are shown in Figure 6. As we can see, compared with the first experiment, the advantage of **HTLFA** becomes larger when training data is sparse. The reason is that when the amount of training data is smaller, the documents can make a larger contribution to supplement the training data. In other words, documents increase the number of heterogeneous training data instances. It properly proves the advantage of transfer learning, that is to remedy the sparsity of the source data. However, **HTLFA** cannot perform well either when training data is extremely sparse. There is no doubt that the distribution of image-tag features is different from the distribution of document-tag features. Hence when the amount of training data approach zero and documents becomes dominant, the model naturally does not work on the task of recommending tags for images.

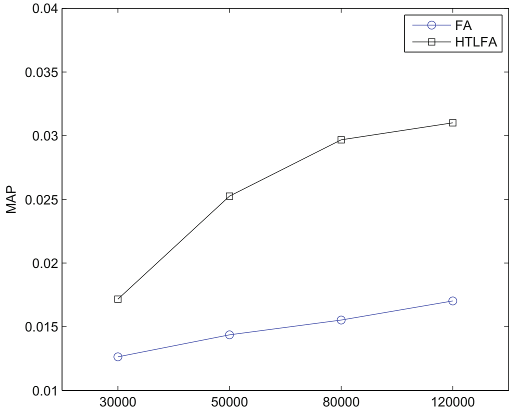
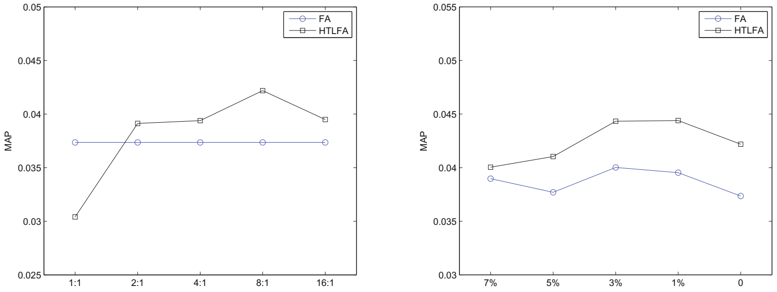


Fig. 6. Varying number of training data

We also compare **HTLFA** and **FA** under the long tail condition. Since words which are in the tail of the frequency distribution for images are almost always in the tail of the distribution for documents, we randomly select 2000 tags and use them to form a long tail tag set. We down sample these tags in images to simulate a long tail condition. We construct several test settings, corresponding to different down sample rates. When evaluating the results, we ignore the groundtruth not included in long tail tag set and focus on recommending long tail groundtruth from the other tags. In Figure 7(a), we vary the iteration number ratio between images and documents fixing the down sample rate at 0. The result shows that $8 : 1$ (i.e., iterate through 8 images before iterating a document) obtains the best performance. Fixing the iteration number ratio at $8 : 1$, Figure 7(b) shows that our model outperforms **FA** under long tail condition. However, as mentioned above, when the down sample rate is extremely low (e.g. zero percent), the shortcoming that the two heterogeneous data sets have different distributions becomes obvious, which finally leads to a decrease in the precision.

5.6 Case Study

Table 3 shows the tag recommendation result of our algorithm **HTLFA** for Flickr data sets. In this figure, six images are from the Flickr data set. Three recommended tags are given at the bottom of each image. From the table we can see that our algorithm can indeed find related tags based on visual contextual information of an image. Here we demonstrate a case that makes a difference among the compared algorithms. Table 4 provides some tags recommended by the four algorithms for a target image about a baby. The top three recommended tags are shown on the right of each algorithm name. As the table shows, **ViCAD** and **NN** have really poor performance on this case. Non-transfer **FA** recommends only one related tag while the top three tags recommended by



(a) Varying the iteration number ratio (b) Varying down sample rate

Fig. 7. Evaluation under long tail condition

Table 3. The **HTLFA** results

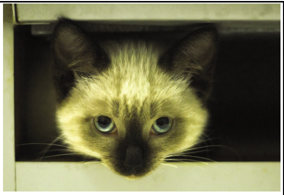
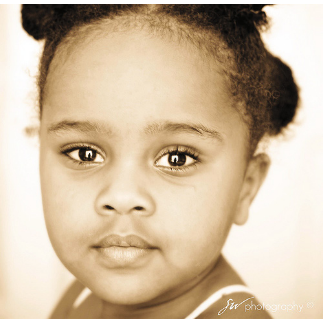
		
cat, grey, pets	clouds, seascape, beach	wheels, voiture, ford
concert, music, guitar	nature, wildlife, plant	flight, aviation, plane

Table 4. Tags recommended by the compared algorithms on one case

	**ViCAD**	farriery laminitis arthrosis
	NN	cool surgery rachel
	FA	lips freckles jackson
	HTLFA	lips eyelashes toddler

HTLFA are all related. And we also discover that the tag "toddler" does not appear in the top ten of the recommendation list of **FA**. Hence it is reasonable to believe that it is documents that help discover the related tag "toddler".

6 Conclusion

In this paper, we explore heterogeneous transfer learning for factor alignment integrating documents, tags and images and apply it on the task of tag recommendation by leveraging the help of long text from Wikipedia. We expect to make use of auxiliary text documents to supplement the target domain training data since auxiliary data may reduce the data sparsity in the image domain and word domain. Then we propose a factor alignment ranking model and modify it to make it able to handle the auxiliary text data. Using soft constraints to control the similarity between two latent matrices, we manage to incorporate the information of auxiliary text data in the model. We compare the method with three baselines, and prove that our method outperforms all three. We vary the value of parameter λ_0 to illustrate that a soft constraint is necessary. Comparing our model to one without transfer, we also show that auxiliary text data also improves handling of items from the long tail of the tag frequency distribution. So overall, we have shown that the performance of tag recommendation can be improved by utilizing textual information.

In the future, we will consider other types of auxiliary data as well as more than one data source.

Acknowledgement. We thank Weinan Zhang and Erheng Zhong for discussions. We thank the support of grants from NSFC-RGC joint research project HKUST 624/09 and 60931160445.

References

1. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359 (2010)
2. Yang, Q., Chen, Y., Xue, G.-R., Dai, W., Yu, Y.: 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: vol. 1, ACL 2009, pp. 1–9. Association for Computational Linguistics, Stroudsburg (2009)
3. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: Proceedings of the 24th International Conference on Machine Learning, ICML 2007, pp. 759–766. ACM, New York (2007)
4. Dai, W., Chen, Y., Xue, G.R., Yang, Q., Yu, Y.: Translated learning: Transfer learning across different feature spaces (2008)
5. Zhu, Y., Chen, Y., Lu, Z., Pan, J.S., Xue, G.R., Yu, Y., Yang, Q.: Heterogeneous transfer learning for image classification (2011)
6. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)

7. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M., Jordan, M.I.: Matching words and pictures. J. Mach. Learn. Res. 3, 1107–1135 (2003)
8. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR 2003, pp. 127–134. ACM, New York (2003)
9. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures (2003)
10. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(3), 394–410 (2007)
11. Makadia, A., Pavlovic, V., Kumar, S.: A New Baseline for Image Annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
12. Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, MULTIMEDIA 2005, pp. 706–715. ACM, New York (2005)
13. Wang, C., Jing, F., Zhang, L., Zhang, H.-J.: Content-based image annotation refinement. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007 (2007)
14. Liu, D., Hua, X.S., Wang, M., Zhang, H.J.: Image retagging. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 491–500. ACM, New York (2010)
15. Liu, D., Yan, S., Hua, X.-S., Zhang, H.-J.: Image retagging using collaborative tag propagation. IEEE Transactions on Multimedia 13(4), 702–712 (2011)
16. Chen, Y., Jin, O., Xue, G.R., Chen, J., Yang, Q.: Visual contextual advertising: Bringing textual advertisements to images (2010)
17. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 524–531 (2005)
18. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2009, Arlington, Virginia, United States. AUAI Press (2009)
19. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
20. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections (2005)
21. Singh, A.P., Gordon, G.J.: Relational learning via collective matrix factorization. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2008, pp. 650–658. ACM, New York (2008)