Learning Semantic Correlation of Web Images and Text with Mixture of Local Linear Mappings

Youtian Du, Kai Yang
Ministry of Education Key Lab for Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an 710049, China
duyt@mail.xjtu.edu.cn, scfy.217@stu.xjtu.edu.cn

ABSTRACT

This paper proposes a new approach, called *mixture of local linear mappings* (*MLLM*), to the modeling of semantic correlation between web images and text. We consider that close examples generally represent a uniform concept and can be supposed to be locally transformed based on a linear mapping into the feature space of another modality. Thus, we use a mixture of local linear transformations, each local component being constrained by a neighborhood model into a finite local space, instead of a more complex nonlinear one. To handle the sparseness of data representation, we introduce the constraints of sparseness and non-negativeness into the approach. MLLM is with good interpretability due to its explicit closed form and concept-related local components, and it avoids the determination of capacity that is often considered for nonlinear transformations. Experimental results demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models

General Terms

Theory

Keywords

Semantic correlation; heterogeneous modalities; cross-media retrieval

1. INTRODUCTION

There has been a massive explosion of multimedia contents on the web such as text, images and videos, which usually co-exist in a multimedia document and describe the similar semantic concepts. For example, an image can give us a vivid imagination but incomplete about a concept. In contrast, the text could accurately describe the details of a concept, but it is not intuitive enough. Consequently, it could make the content understanding more accurate to jointly exploit the full information from different modalities, and be significant to many applications such as image tagging [1, 2, 3] and cross-media retrieval [4, 5, 6]. However, the different representation and different ways of stimulating to human being's brains derived from heterogenous modalities lead to challenges of semantic correlation mining among the heterogenous modalities of medias [7].

The main related work can be categorized into three classes. 1) *Linear/nonlinear mapping*: The class of methods build a closed-form transformation between heterogeneous spaces [8, 9]. In general, complex models lead to high capacities but low generalization. A challenge in these methods is to choose the nonlinear model of a suitable capacity. 2) *Probabilistic models* such as probabilistic latent semantic analysis (PLSA) [10, 11]: These methods focus on the dominant global semantics of medias but ignore the direct relationship among the local components of documents, i.e., the textual words and local visual regions. 3) *Graph-based correlation propagation*: Graph-based methods, with documents as vertices and correlation as the weight of edges [4, 12], focus on the local structure of data distributions and are effective to the data with complicated distributions. The computational cost generally increases rapidly for a large scale of data.

This paper proposes a new approach to the analysis of semantic correlation between text and images based on *mixture of local linear mappings* (*MLLM*). We consider that data in a local region generally represent a uniform concept and can be supposed to be transformed based on a linear mapping into the feature space of another modality. Therefore, this work uses a mixture of local linear mapping, each local component being constrained by a neighborhood model into a finite local space, to substitute for a more complex nonlinear mapping between the feature spaces of textual and visual modalities. The MLLM model also considers the sparseness and non-negativeness and leads to the sparse output given a query, which is consistent with the true sparse representation in feature spaces of both text and images.

2. THE PROPOSED METHOD

2.1 Mixture of Local Linear Mapping

Since different representations tend to be adopted for images and text, there is typically no explicit correspondence between the representations, and it is difficult to construct a uniform correlated model over the whole distributions of

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806331.

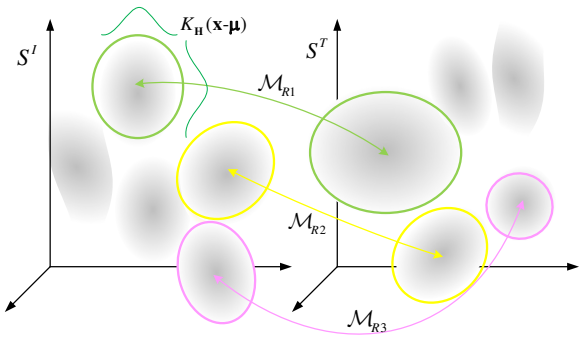


Figure 1: Illustration of the mixture of local linear mappings. The scale of darkness represents the distribution of data in feature space S^I and S^T, and the colored ellipses denote the local regions over which local linear mapping models are built.

images and text. Hence, in this work we present a mixture of local linear mappings (MLLM).

We assume that over a local region there exists a linear mapping model from textual (visual) space to visual (textual) space, and characterize this model by the concatenation of two matrices as follows:

$$\mathbf{y}_i = U \cdot W\mathbf{x}_i + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\mathbf{x}_i$ denotes a query of d_q-dimensional vector, $\mathbf{y}_i$ is the d_o-dimensional example associated to query $\mathbf{x}_i$, $W \in \mathbf{R}^{K\times d_q}$ denotes the transformation from input space to a K-dimensional latent semantic space and $U \in \mathbf{R}^{d_o\times K}$ means that from the semantic space to output space, and $\boldsymbol{\varepsilon}_i$ is the fitness error. For simplicity, we let $\boldsymbol{\varepsilon}_i$ be a vector of independent and normally distributed elements with zero mean, i.e. $\Sigma_{\boldsymbol{\varepsilon}} = diag(\sigma_1^2, \sigma_2^2, \cdots, \sigma_{d_o}^2)$. The linear mapping in Eq.1 can be rewritten as the following probabilistic form:

$$\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathcal{M}) = \frac{1}{(2\pi)^{\frac{d_o}{2}}|\Sigma_{\varepsilon_i}|^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{\varepsilon}_i^T\Sigma_{\varepsilon}^{-1}\boldsymbol{\varepsilon}_i} \tag{2}$$

where $\mathcal{M}$ denotes the model in Eq.1 that $\{\mathbf{x}_i, \mathbf{y}_i\}$ follows.

Intuitively, a set of close examples, denoted by $\mathcal{R}$, tend to have similar semantics and approximately follow the same cross-media mapping model. The data near the centroid of $\mathcal{R}$ follow the model with high confidence. In contrast, the data far away from the centroid follow the model with low confidence. Thus, we characterize the difference of confidence resulted from the location by introducing a neighborhood model $K_{\mathbf{H}}(\mathbf{x}-\boldsymbol{\mu})$ and a symmetric positive definite $d_q \times d_q$ bandwidth matrix $\mathbf{H}$. We choose Gaussian function as the neighborhood model:

$$K_{\mathbf{H}}(\mathbf{x}_i - \boldsymbol{\mu}) = \frac{1}{(2\pi)^{\frac{d_q}{2}}|\mathbf{H}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\mathbf{H}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})} \tag{3}$$

where $\boldsymbol{\mu}$ denotes the centroid of set $\mathcal{R}$ that $\mathbf{x}_i$ belongs to. For simplicity, we assume that $\mathbf{H}$ is a diagonal matrix and $\mathbf{H} = diag(h_1, h_2, \cdots, h_{dq})$. $K_{\mathbf{H}}(\mathbf{x}_i - \boldsymbol{\mu})$ can be used to describe the probability (or confidence) of data that follow the mapping model, i.e. $\Pr(\mathbf{x}_i|\mathcal{M}_R)$.

Now, we define the following likelihood function for the pair $(\mathbf{x}_i, \mathbf{y}_i)$ given a specific local linear mapping $\mathcal{M}_R$:

$$\begin{aligned}\mathcal{L}_{F,i,m} &= \Pr(\mathbf{x}_i|\mathcal{M}_{Rm})\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathcal{M}_{Rm}) \\ &= K_{\mathbf{H}}(\mathbf{x}_i - \boldsymbol{\mu}_m)\Pr(\mathbf{y}_i|\mathbf{x}_i, \mathcal{M}_{Rm})\end{aligned} \tag{4}$$

Actually, due to the complicated data distribution and nonlinear mappings between textual and visual spaces, a unique linear model is of insufficient capacity in modeling the correlation between them. Consequently, we develop a mixture of local linear mappings to characterize the cross-media correlation as follows:

$$\begin{aligned}\mathcal{L}_F &= \prod_{i=1}^{N}\Pr(\mathbf{x}_i, \mathbf{y}_i) \\ &= \prod_{i=1}^{N}\sum_{m=1}^{M}\omega_m c_m e^{\sum_{j=1}^{d_q}\frac{-(\mathbf{x}_{i,j}-\boldsymbol{\mu}_{m,j})^2}{2h_{j,m}^2}} \cdot e^{\sum_{j=1}^{d_o}\frac{-\varepsilon_{i,m,j}^2}{2\sigma_{j,m}^2}}\end{aligned} \tag{5}$$

where M is the number of components of the mixture model, ω_m denotes the positive weight of the m-th component and $\sum_{m=1}^{K}\omega_m = 1$, c_m is a factor to normalize $\mathcal{L}_F$ to be a probability density function, $\mathbf{x}_{i,j}$ and $\boldsymbol{\mu}_{m,j}$ denote the j-th entry of $\mathbf{x}_i$ and $\boldsymbol{\mu}_m$, $\varepsilon_{i,m,j}$, $h_{j,m}$ and $\sigma_{j,m}$ are the j-th entry of $\boldsymbol{\varepsilon}_i$, $\mathbf{H}$ and $\Sigma_{\boldsymbol{\varepsilon}}$ corresponding to the model $\mathcal{M}_{Rm}$, respectively. In our mixture model in Eq.5, the first exponential term plays the role of making the dense points tend to share one model, and the second one focuses on modeling the correlation between two heterogeneous modalities.

2.2 Constraints to MLLP

In this section, we consider some extra constraints to the mixture of local linear mappings.

1) *Smoothness.* As mentioned above, close examples, with proper representation, tend to indicate the similar concept in both textual space and visual space. Hence, it is necessary to return the similar retrieved outputs for two close queries. We formulate the smoothness constraint as follows:

$$J_W = \sum\nolimits_{i\sim j}\|e_{ij}W(\mathbf{x}_i - \mathbf{x}_j)\|_2 \le \|D_X\|_F\|W\|_F \tag{6}$$

$$J_U = \sum\nolimits_{i\sim j}\|e_{ij}UW(\mathbf{x}_i - \mathbf{x}_j)\|_2 \le \|WD_X\|_F\|U\|_F \tag{7}$$

where $i \sim j$ means that $\mathbf{x}_i$ is the neighbor of $\mathbf{x}_j$, e_{ij} denotes the weight of the pair $\mathbf{x}_i$ and $\mathbf{x}_j$, and D_X is the matrix whose columns are vectors $\{\mathbf{x}_i - \mathbf{x}_j\}$ in a certain order for all $i \sim j$. We let $e_{ij} = 1$ for all i and j in Eqs.6 and 7 since we have no clear prior information to show which pair is more important. The neighbors of $\mathbf{x}_j$ is determined by choosing C-nearest points. Thus, we measure the smoothness of cross-media mappings with the following expression:

$$\begin{aligned}J_{sm}(U, W) &= J_W + J_U \\ &\triangleq \lambda_W\|W\|_F + \lambda_U\|U\|_F\end{aligned} \tag{8}$$

where λ_W and λ_U, which replace the coefficients on the right-hand side of Eqs.6 and 7 respectively, are used to control the importance of the two terms. The first term on the right-hand side of Eq.8 measures the distance in the concept space, and the second one measures the distance of two retrieved outputs. To obtain a smooth mapping model, $J_{sm}(U, W)$ needs to be constrained to a small value.

2) *Sparseness and Non-negativeness.* With the bag-of-word representation, both text and images are generally represented as the feature vectors of non-negative and sparse components. Take the mapping from textual space to concept space as example. When a document is mapped into the concept space, we expect that it corresponds to a few concepts that can describe its semantics. That is, the other components in the concept vector are zeros. Thus, the matrix W needs to be sparse. Further, we think every unit

appearing in a document supplies positive contribution to the corresponding concept. In other words, a whole concept is formed by combining a set of units and can be considered as the linear combination of elements. The non-negative constraint makes only additive combination be allowed.

Finally, we have the following optimization problem:

$$\min_{U_m \in \mathbf{R}^{d_o \times K}, W_m \in \mathbf{R}^{K \times d_q}} \mathcal{L}(U_m, W_m) \tag{9}$$

where

$$\mathcal{L}(U_m, W_m) = -\ln \mathcal{L}_F + \lambda_1 \sum_{m=1}^{M} J_{sm}(U_m, W_m) + \lambda_2 \sum_{m=1}^{M} J_{sp}(U_m, W_m)$$
$$S.t.\ U_m \geq 0, W_m \geq 0 \tag{10}$$

where $J_{sp}(U_m, W_m) \triangleq \|U_m\|_1 + \|W_m\|_1$ denotes the sparseness term, and λ_1 and λ_2 are used to control the balance between these terms.

2.3 Optimization Process

The optimization problem (9) is different from many regularization based learning because it contains some hidden information. More specifically, we don't know which local linear mapping that $(\mathbf{x}_i, \mathbf{y}_i)$ follows. We develop a sub-gradient descent approach under the framework of Expectation- Maximization. The parameter set is denoted by $\Theta = \{U_m, W_m, \omega_m, \mathbf{H}_m, \boldsymbol{\mu}_m, \Sigma_{\varepsilon,m}\}_{m=1}^{M}$.

1) *Initialization.* First, we use Kmeans clustering algorithm over training data $\{\mathbf{x}_i\}_{i=1}^{N}$ and achieve M clusters. Let the initial value of $\boldsymbol{\mu}_m$ and $\mathbf{H}_m$ (written as $\boldsymbol{\mu}_m^{(0)}$ and $\mathbf{H}_m^{(0)}$, respectively) equal the mean vector and covariance matrix of the m-th cluster; Initialize $U_m \in \mathbf{R}^{d_o \times K}$ and $W_m \in \mathbf{R}^{K \times d_q}$ with non-negative matrix.

2) *Determination of hidden information.* Given $\Theta^{(t-1)}$, we compute the probability that $(\mathbf{x}_i, \mathbf{y}_i)$ follows $\mathcal{M}_{Rm}$.

$$\Pr(\mathcal{M}_{Rm}|\mathbf{x}_i, \mathbf{y}_i, \Theta^{(t-1)}) = \frac{\Pr(\mathcal{M}_{Rm}, \mathbf{x}_i, \mathbf{y}_i|, \Theta^{(t-1)})}{\sum_{m=1}^{M} \Pr(\mathcal{M}_{Rm}, \mathbf{x}_i, \mathbf{y}_i|, \Theta^{(t-1)})} \tag{11}$$

where

$$\Pr(\mathcal{M}_{Rm}, \mathbf{x}_i, \mathbf{y}_i|\Theta^{(t-1)}) = \omega_m c_m e^{\sum_{j=1}^{d_q} \frac{-(\mathbf{x}_{i,j} - \boldsymbol{\mu}_{m,j})^2}{2h_{j,m}^2}} \cdot e^{\sum_{j=1}^{d_o} \frac{-\varepsilon_{i,m,j}^2}{2\sigma_{j,m}^2}} \tag{12}$$

For simplicity, we rewrite $\Pr(\mathcal{M}_{Rm}|\mathbf{x}_i, \mathbf{y}_i, \Theta^{(t-1)})$ as $P_{i,m}^{(t-1)}$.

3) *Update.* The parameters of $\mathcal{M}_{Rm}$ are updated at t time sequentially by:

$$\omega_m^{(t)} = \frac{1}{N} \sum_{i=1}^{N} P_{i,m}^{(t-1)},\ \boldsymbol{\mu}_m^{(t)} = \sum_{i=1}^{N} \mathbf{x}_i P_{i,m}^{(t-1)} / \sum_{i=1}^{N} P_{i,m}^{(t-1)} \tag{13}$$

$$h_{j,m}^{(t)\ 2} = \sum_{i=1}^{N} (\mathbf{x}_{i,j} - \boldsymbol{\mu}_{m,j}^{(t-1)})^2 P_{i,m}^{(t-1)} / \sum_{i=1}^{N} P_{i,m}^{(t-1)} \tag{14}$$

For the update of the other parameters, we define another cost function based on problem (9):

$$\mathcal{Q}(U_m, W_m) = \sum_{m=1}^{M} \Big(\sum_{i=1}^{N} (-\ln c_m - \sum_{j=1}^{d_o} \frac{\varepsilon_{i,m,j}^2}{2\sigma_{j,m}^2}) P_{i,m}^{(t-1)} + \lambda_1 J_{sm}(U_m, W_m) + \lambda_2 J_{sp}(U_m, W_m) \Big) \tag{15}$$

Table 1: MAP scores (Top 50 retrieved examples).

	Corel5K		Wikipedia	
	T→I	I→T	T→I	I→T
CCA	0.1078	0.1725	0.2485	0.1646
KCCA	0.2115	0.1501	0.2807	0.2121
MLLM	**0.2787**	**0.2832**	**0.3581**	**0.2793**

Minimizing $\mathcal{Q}(U_m, W_m)$ subjected to $U_m \geq 0, W_m \geq 0$ with sub-gradient descent approaches can obtain the solution.

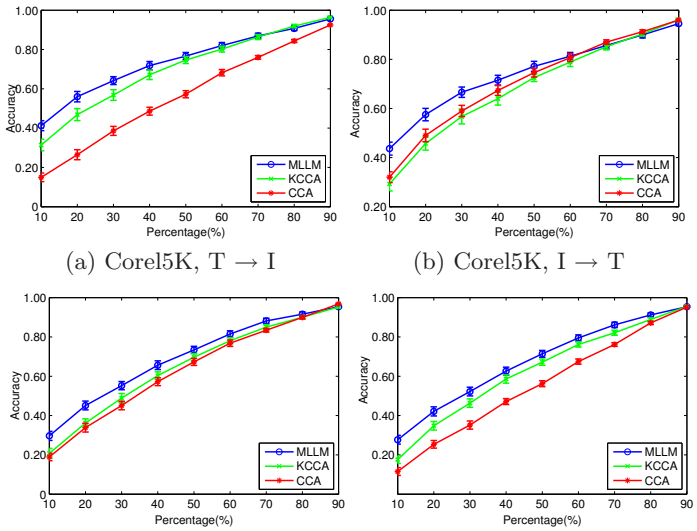
3. EXPERIMENTAL RESULTS

Two public real-world datasets are used in our experiments. 1) *Corel5K*: There are 50 categories in this dataset and each category is made up by 100 images. Each image has a caption of 1-5 keywords. We represent text and images of Corel5K by 374-dimensional and 500-dimensional bag-of-word feature vectors, respectively. We random choose 2/3 examples for training and the rest for test. 2) *Wikipedia dataset*: The set consists of 2,866 images, each with a paragraph of text to describe the image. We use the originally provided representation: 10-dimensional topic-based features for text and 128-dimensional SIFT feature for images, and the originally provided ratio 2173/693 (training/test). Each experiment is implemented 10 times independently.

We compare the proposed method to two previous representative methods, including 1) *Canonical Correlation Analysis* (*CCA*) [8]: The technique linearly transforms two kind of data from initial spaces to two new spaces respectively to achieve the largest relevance; 2) *KCCA* [8]: a non-linear version of CCA by introducing kernels. In this work, we employ the RBF kernel. We use two measures to evaluate the performance. 1) Percentage score: An image(or text) is considered correctly retrieved if it appears in the first t percent of the predicted list from its corresponding text (or image) query. 2) Mean average precision (MAP). Because users generally pay more attention to the front retrieved results, we only analyze the results of the first 50 returned samples in MAP analysis.

For parameter tuning, we split the training set into 5 folds and employ 5-fold cross-validation. For MLLM, we choose M and K as 50 and 15 for Corel5K, and 15 and 8 for Wikipedia. We search the other parameters over the grid: $\lambda_1, \lambda_2, \lambda_U, \lambda_W \in \{0.01, 0.1, 1, 10, 100\}$. In the implementation, we let $\lambda_U = \lambda_W$ for low complexity of search.

Fig.2 illustrates the average percentage score of 10 independent experiments over both datasets. For the often used percentage $t = 20\%$, MLLM achieves the precision of 0.561, 0.575, 0.452 and 0.421, which are clearly superior to the compared two methods. In some cases, we note that for large percentage t (e.g., $t > 70\%$), the precision of our method is similar with or slightly lower than CCA or KCCA. We consider that a highly possible reason is our method tends to describe the local property of data distribution over the similar data instead of the distribution over the whole space. Table 1 shows the MAP result, which measures whether the retrieved data belong to the same category as the query or not. Fig.3 shows an example of sparseness in the prediction of MLLM given a text query. Given the query including textual words "plane", "runway" and "boeing", the predicted visual words is highly correlated to the true ones of the image associated with the query, which means that MLLM method can predict the sparse output well given a query. In



(a) Corel5K, T → I (b) Corel5K, I → T

(c) Wikipedia, T → I (d) Wikipedia, I → T

Figure 2: The percentage score evaluation on two sets. $T \rightarrow I$ and $I \rightarrow T$ denote text query and image query, respectively. The error bars indicate ± 1 standard error.

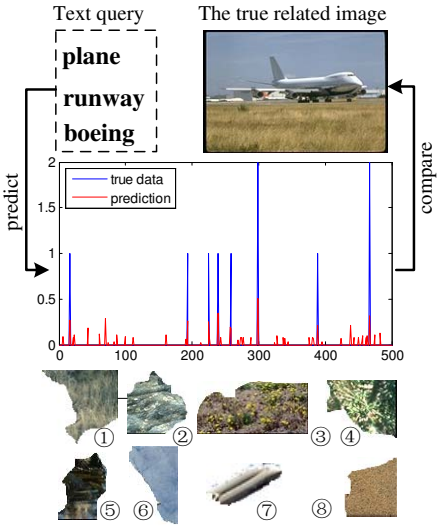


Figure 3: Illustration of sparseness. Small image patches 1-8 are those closest to the centers of the clusters corresponding orderly to the visual words indicated by blue vertical lines.

a majority of the data especially in Corel5K, we can find the results like this example.

4. CONCLUSIONS

In this paper, we proposes a new approach, named MLLM, to the modeling of semantic correlation between text and images. The close data in a local region can be supposed to be transformed based on a linear mapping into the feature space of another modality. Therefore, In the whole space, the correlation between heterogenous modalities can be modeled by a mixture of local linear mappings. We try to discover a transformation from one modality of information to the other based on mixture of local linear mapping that can replace and be superior to the more complex nonlinear transformation. Besides of the better performance shown in the experiments, our method also have some other advantages. It is with good interpretability due to its explicit closed form and concept-related local components. Moreover, it avoids the complicated analyse of the capacity that is often considered in nonlinear transformations.

5. ACKNOWLEDGMENTS

This work is supported in part by the National Natural Science Foundation (61375040, 61202392, 61175039, 61221063) of China.

6. REFERENCES

[1] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, pages 119–126. ACM, 2003.

[2] L. Wu, R. Jin, and A.K.Jain. Tag completion for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):716–727, 2013.

[3] M. Wang, B. Ni, X. Hua, and T. Chua. Assistive tagging: a survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys*, 44(4):25–25, August 2012.

[4] Y. Zhuang, Y. Yang, and F. Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transaction on Multimedia*, 10(2):221–229, 2008.

[5] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing*, 22(1):363 – 376, 2013.

[6] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, pages 1070–1076, 2013.

[7] X. Wu, Y. Qiao, X. Wang, and X. Tang. Cross matching of music and image. In *ACM Multimedia*, pages 837–840. ACM, 2012.

[8] J. Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[9] T. Jiang and A. Tan. Learning image-text associations. *IEEE Transactions on Knowledge and Data Engineering*, 21(2):161–177, 2009.

[10] F. Monay and D.G.Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007.

[11] Y.Zhou, M. Liang, and J. Du. Study of cross-media topic analysis based on visual topic model. In *CCDC*, pages 3467–3470, 2012.

[12] X. Zhai, Y. Peng, and J. Xiao. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In *MMM*, pages 312–322, 2012.