A Low Rank Structural Large Margin Method for Cross-Modal Ranking

Xinyan Lu
College of Computer Science
Zhejiang University, China
xinyanlu@zju.edu.cn

Fei Wu
College of Computer Science
Zhejiang University, China
wufei@cs.zju.edu.cn

Siliang Tang
College of Computer Science
Zhejiang University, China
siliang@zju.edu.cn

Zhongfei Zhang
Department of Information Science & Electronic Engineering
Zhejiang University, China
zhongfei@zju.edu.cn

Xiaofei He
College of Computer Science
Zhejiang University, China
xiaofeihe@cad.zju.edu.cn

Yueting Zhuang
College of Computer Science
Zhejiang University, China
yzhuang@zju.edu.cn

ABSTRACT

Cross-modal retrieval is a classic research topic in multimedia information retrieval. The traditional approaches study the problem as a pairwise similarity function problem. In this paper, we consider this problem from a new perspective as a listwise ranking problem and propose a general cross-modal ranking algorithm to optimize the listwise ranking loss with a low rank embedding, which we call Latent Semantic Cross-Modal Ranking (LSCMR). The latent low-rank embedding space is discriminatively learned by structural large margin learning to optimize for certain ranking criteria directly. We evaluate LSCMR on the Wikipedia and NUS-WIDE dataset. Experimental results show that this method obtains significant improvements over the state-of-the-art methods.

Categories and Subject Descriptors

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

General Terms

Algorithms, Theory, Experimentation

Keywords

Cross-Modal Retrieval; Ranking; Low Rank Embedding

1. INTRODUCTION

Nowadays many real-world applications involve multi-modal data. The ranking of cross-modal retrieval is imperative to many applications of the practical interest, such as finding relevant textual documents of a tourist spot that best match a given image of the spot or finding a set of images that visually best illustrate a given text description. Therefore, it is desirable to support the ranking of multi-modal data, e.g., identifying the most relevant textual documents in response to a query image or vice versa. It is obvious that a good performance of ranking multi-modal data hinges upon an appropriate learning of the similarity across different modalities. The *heterogeneity-gap* between multi-modal data has been widely understood as a fundamental barrier for the cross-modal metric learning.

In recent years, there has been a great deal of research devoted to the development of algorithms for learning an optimal direct similarity metric between different modalities. Among existing research of cross-modal metric learning, one kind of popular approaches is to map the data with multiple modalities into a common (or shared) space such that the distance between two similar objects is minimized, while the distance between two dissimilar objects is maximized. These approaches (e.g., Canonical Correlation Analysis (CCA) [12] and [33]) usually assume a training set of strictly paired data and exploit the symbiosis of multiple-modality data which is common to describe the rich literal and visual semantics, such as a web image with loosely related narrative text descriptions, and a news report with collateral text and images. Another kind of approaches for multi-modal metric learning is the extensions of Latent Dirichlet Allocation (LDA) [5] which are conducted on two or more collections of multiple-modality data in an unsupervised manner. The LDA-based approaches tend to model correlations among multi-modal documents at a latent semantic (*topic*) level across different modalities, e.g. correspondence LDA [4].

We are particularly interested in ranking the multi-modal data (i.e. *cross-modal ranking*) in this paper. Different from the two aforementioned categories of approaches which do not maximize a criterion related to the final ranking performance, recent years have witnessed the efforts in learning an optimal similarity (ranking) function between different modalities by using learning to rank techniques. These approaches (e.g. [11, 23, 2]) are supervised but do not enforce a strict assumption that the trained multi-modal data must be paired (e.g., one image is in pair-correspondence with its

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

collateral text). They in fact need some lists of ranked data related to the queries for training where the training examples can be easily obtained from the abundance of users' clickthrough data with little overhead [18, 10]. In this way, the learned metric for ranking multi-modal data is generally optimized for a ranking-based loss function (evaluation criterion) to *preserve* the orders of the relevance instead of the purely absolute values of similarity (dis-similarity) between multi-modal data.

A discriminative kernel-based method PAMIR proposed in [11] solves the problem of cross-modal ranking by adapting the Passive-Aggressive algorithm [8]. However, PAMIR is inherently a pairwise cross-modal ranking approach; its ranking performance is limited by the distribution of the pairs of items and skewed data may even deteriorates the ranking result. These are also the same problems for another pairwise approach SSI [2] proposed by Bai et al. Moreover in the real world, a long search query (e.g., a whole document) is beneficial because users' intents can be described in more detail [31]. Therefore, it is appropriate to support long textual documents as queries in cross-modal ranking. It is obvious that PAMIR cannot easily expedite the long queries and restrains itself from a more flexible application of cross-modal ranking in the setting of uncontrolled multi-modal data involved.

This paper aims to bridge the gap between learning a latent space and ranking for multi-modal data. We consider the problem of cross-modal retrieval from a new perspective as a listwise ranking problem in this paper, and propose a general cross-modal ranking algorithm to optimize the listwise ranking loss meanwhile considering a low rank embedding, called Latent Semantic Cross-Modal Ranking (LSCMR). LSCMR employs the structural SVM [29] to support the optimizations of various ranking evaluation measures (e.g, MAP [32] and NDCG [7]) under a unified algorithmic framework. LSCMR also incorporates a low rank embedding in the learning procedure in which the latent aspect space is induced to address the curse of dimensionality and discover the correlations between different modalities.

It is worthwhile to highlight the main motivations of the proposed method. We would like the method to benefit both from the low rank embedding and the most recent advances in learning to rank techniques:

- Similar to the ideas of the latent model SSI [2] which discovers the "latent concept" from high-dimensional space, LSCMR employs the low rank embedding to eliminate textual/visual synonymy and polysemy of multi-modal data and discovers the correlations between different modalities. This is particularly appropriate to the settings where the distributions of query data and the retrieved data are intrinsically different due to the aforementioned multi-modal heterogeneity-gap. Furthermore, from a flexible retrieval point of view, LSCMR naturally supports long queries to discern users' search intentions.

- By the introduction of structural large margin learning into the optimization of a cross-modal ranking function in a *listwise* manner, LSCMR explicitly minimizes the ranking loss of a whole permutation listwise, not individuals or pairs of items. The ranking function by the maximum margin is not biased toward the queries with more data pairs, leading to a strong generalization due to its empirical risk minimization.

We show experimental results on the ranking of cross-modal data obtained from two real-world datasets. The proposed LSCMR outperforms other cross-modal ranking approaches. LSCMR is particularly appropriate for cross-modal ranking due to its structural large margin and low-rank listwise ranking pursuing.

The rest of this paper is organized as follows. In Section 2, we describe the method in detail and show its feasibility. Section 3 discusses the existing work. We compare the proposed LSCMR with other cross-modal ranking approaches on two real-world datasets in Section 4. Conclusions are given finally.

2. THE ALGORITHM OF LSCMR

The proposed method LSCMR is a general cross-modal retrieval framework in the sense that it can be applied in both directions of image-query-text retrieval and text-query-image retrieval. Consequently, a query here may be either an image or a text document. Similarly, a retrieved document can either be an image or a text document. For a clear articulation in the rest of this section, the algorithm is only derived in the case of text-query-image retrieval. We report the experiments in both scenarios.

2.1 Notation

All vectors are assumed to be column vectors and a superscript T denotes the transpose of a matrix or vector. Denote the text query as $q \in \mathbb{R}^m$ and the retrieved image as $d \in \mathbb{R}^n$, where m is the dimension of the text space (e.g., vocabulary size of bag-of-words (BoW)) and n is the dimension of the image space (e.g., vocabulary size of bag-of-visual-words (BoVW) which is quantized by clustering from low-level visual features such as SIFT [22]). We are given a training set of N samples, in which each contains a text query q_i $(i = 1, \ldots, N)$ as well as a set of corresponding retrieved images $\mathbf{d}_i$ with their true rankings $\mathbf{y}_i^* \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of all possible permutations (rankings). We formulate a ranking as a matrix of pair orderings as [32] does, $\mathcal{Y} \subset \{-1, 0, +1\}^{|\mathbf{d}| \times |\mathbf{d}|}$ where the operator $|\cdot|$ denotes the number of elements in a set. For any $\mathbf{y} \in \mathcal{Y}$, $y_{ij} = +1$ if document d_i is ranked ahead of document d_j, and $y_{ij} = -1$ if d_j is ranked ahead d_i, and $y_{ij} = 0$ if d_i and d_j have equal rank. We consider only matrices with correspond to valid rankings (i.e., obeying antisymmetry and transitivity). Moreover, assume that the true ranking is a weak ranking with two rank values (*relevant* and *irrelevant*). For any textual query q_i, let $\mathbf{d}_i^+$ and $\mathbf{d}_i^-$ denote the set of relevant and irrelevant images in $\mathbf{d}_i$, respectively. For simplicity, we omit the subscript i of q_i and $\mathbf{d}_i$ when it is clear from the context.

2.2 The Low-Rank Learning Function

We consider that learning scores for ranking from a supervised manner, in which the ranking of images corresponding to a given textual query is available for training. Unlike the uni-modal data ranking, cross-modal ranking attempts to learn a similarity function $f(q, d)$ between a text query q and an image d according to a pre-defined ranking loss. The learned function f maps each text-image pair to a ranking score based on their semantic relevance.

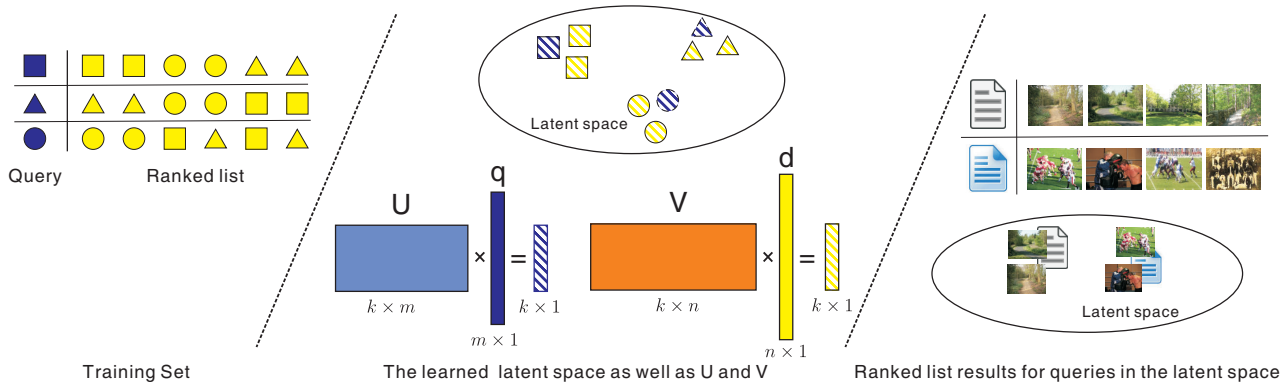


Figure 1: The algorithmic illustration of the proposed LSCMR. Shapes represent semantics (e.g., classes or categories); the same shape indicates the same (or quite relevant) semantic content. Colors represents modalities (i.e., images and text documents); the same color indicates that the data objects come from the same modality space. For simplicity in this illustration, we assume the two modalities to be text queries and image targets. Ideally, given a text query, we would like those images having the same shape (semantic content) with the query are ranked before those having different shapes in the ranking list. LSCMR tends to learn a low rank ranking function from the training set (seen queries and associated ranking lists) in a supervised manner. All the seen m-dimensional queries q and n-dimensional target documents d are mapped to a k-dimensional latent space by U and V respectively, in which those data objects with the same shape are grouped to minimize certain listwise ranking loss (e.g., MAP) directly. For unseen queries and documents, they are first mapped to the learned latent space by U and V, respectively, and then the scoring is conducted in the latent space to produce ranking results. (Figure best viewed in color)

Given a text query $q \in \mathbb{R}^m$ and an image $d \in \mathbb{R}^n$, we tend to learn a linear scoring function to measure the relevance of d given q :

$$f(q,d) = q^T W d = \sum_{i=1}^{m}\sum_{j=1}^{n} q_i W_{ij} d_j \tag{1}$$

where $f(q,d)$ is the score (or similarity) between the query q and the image d, and the only parameter $W \in \mathbb{R}^{m\times n}$ in the linear model captures the correspondence of the two different modalities of data as a weighting matrix: W_{ij} weights the correlation between the ith dimension of the text space and the jth dimension of the image space. Note that both positive values and negative values in W_{ij} are allowed in which a negative value represents a negative correlation.

Motivated by the idea of low rank embedding, a low rank prior is introduced into matrix W with $W = U^T V$ in equation (1), which results in a low rank ranking model [2]:

$$f(q,d) = q^T U^T V d = (Uq)^T V d \tag{2}$$

where $U \in \mathbb{R}^{k\times m}$ and $V \in \mathbb{R}^{k\times n}$. U refers to map the query text q from the m-dimensional text space to the k-dimensional latent space by a liner mapping, and V refers to map the retrieved image d from the n-dimensional image space to the k-dimensional latent space. Therefore, the text query and the retrieved image are mapped to a common k-dimensional latent *aspect* space, and then their similarity is measured by a dot product of the two vectors in the k-dimensional space, which is commonly used to measure the matching between textual vectors [1].

Intuitively, the low rank model in equation (2) helps us deal with the problem of textual/visual *synonymy* and *polysemy* which particularly occur in cross-modal retrieval. Note that Latent Semantic Indexing (LSI) [9] takes into account of the correlations between textual words (synonym and polysemy) in a single modality in an unsupervised manner, while the low rank model in equation (2) attempts to capture the correlation across two different modalities from a supervised manner. By constraining the form of W with a low rank form, the benefits are similar to LSI: U and V not only induce a k-dimensional latent aspect space but are also faster to compute and lead to much smaller storage requirements by representing the image documents in k dimensions instead of the original n dimensions (k is chosen much smaller than m or n). Besides, from the viewpoint of statistical learning theory, fewer parameters ($k(m+n) \ll mn$) lead to a better stability and generalization in performance.

Similar to [2], here U and V are different and there is no assumption that the query texts and the target images should be embedded to the latent space in the same way. This is appealing to cross-modal ranking since the distributions of the query texts and the target images are inherently different due to the heterogeneity-gap.

We can obtain a prediction $\mathbf{y}$ for each input query q and its corresponding ranked target images $\mathbf{d}$ by sorting the $f(q,d)$ in a descending order. The rest is to learn U and V. We aim to obtain the values of U and V by the minimization of the following empirical risk,

$$R^{\Delta}(f) = \frac{1}{N}\sum_{i=1}^{N} \Delta(\mathbf{y}_i^*, \mathbf{y}_i)\,, \tag{3}$$

where the non-negative loss function $\Delta : \mathcal{Y}\times\mathcal{Y} \to \mathbb{R}$ quantifies the penalty for making prediction $\mathbf{y}_i$ if the correct output is $\mathbf{y}_i^*$, which is typically bounded in $[0,1]$. For example, we define the loss function Δ with the average precision (AP, detailed definition in Equation (14)) loss as follows:

$$\Delta_{ap}(\mathbf{y}^*, \mathbf{y}) = 1 - \mathrm{AP}(\mathrm{rank}(\mathbf{y}^*), \mathrm{rank}(\mathbf{y}))$$

and then to minimize the empirical risk is to maximize the Mean Average Precision (MAP). Note that one can construct different ranking objective problems by considering different ranking loss function Δ.

2.3 The Formulation of LSCMR

In this section, we present the formulation of LSCMR in details. The proposed LSCMR is inspired by the structural SVM framework [29], especially SVM^{map} [32] for optimizing the average precision. The algorithmic illustration of LSCMR is presented in Figure 1.

The motivation of LSCMR is to learn a cross-modal ranking function $h : \mathcal{X} \to \mathcal{Y}$ between an input space $\mathcal{X}$ (a text query q as well as all possible target images) and output space $\mathcal{Y}$ (rankings over the image set). Similar to SVM, we can derive a prediction by finding the ranking $\mathbf{y}$ that maximizes the following discriminant function h:

$$h(q, \mathbf{d}) = \text{argmax}_{\mathbf{y}\in\mathcal{Y}} F(q, \mathbf{d}, \mathbf{y}) \tag{4}$$

where F is considered as a compatibility function parameterized by U, V that measures how compatible the triple $(q, \mathbf{d}, \mathbf{y})$ are.

By adapting the most commonly used *partial order* combined feature representation in [19] to the cross-modal ranking, we define F as:

$$F(q, \mathbf{d}, \mathbf{y}) = \sum_{i\in\mathbf{d}^+} \sum_{j\in\mathbf{d}^-} y_{ij} \frac{(Uq)^T V(d_i - d_j)}{|\mathbf{d}^+| \cdot |\mathbf{d}^-|} \tag{5}$$

where $y_{ij} = +1$ if image d_i is more preferred (more relevant to query q) than image d_j, and $y_{ij} = -1$ otherwise since we assume that the predicted rankings are complete.

One attractive property of F is that for the fixed U and V, the ranking $\mathbf{y}$ which maximizes function F (then the predicted ranking) is simply sorted by descending $f(q, d) = (Uq)^T Vd$. To see this, we note that F is a summation over the differences of all relevant/irrelevant document pairs since we assume weak rankings with two rank values. Since F decomposes linearly over the pairwise representation, we can maximize F by optimizing each y_{ij} individually: if $(Uq)^T Vd_i > (Uq)^T Vd_j$, y_{ij} is set to be 1, and $y_{ij} = -1$ otherwise. This is the same procedure as sorting documents by descending $f(q, d)$. More details can be obtained from [19]. We note that this simple prediction rule establishes a connection between the compatibility function F and the aforementioned low rank ranking model.

Since U and V are independent to the summation in equation (5), we rewrite F as a linear function of $U^T V$:

$$F(q, \mathbf{d}, \mathbf{y}) =< U^T V, \Psi(q, \mathbf{d}, \mathbf{y}) > \tag{6}$$

where

$$\Psi(q, \mathbf{d}, \mathbf{y}) = q \sum_{i\in\mathbf{d}^+} \sum_{j\in\mathbf{d}^-} y_{ij} \frac{d_i^T - d_j^T}{|\mathbf{d}^+| \cdot |\mathbf{d}^-|}\,. \tag{7}$$

Here the combined feature function $\Psi(q, \mathbf{d}, \mathbf{y})$ is a summation over the vector differences of all the relevant/irrelevant image pairs. By representing the scoring F as a Frobenius inner product of $U^T V$ and Ψ, we see that it is straightforward to extend the idea of the structural SVM to learn the cross-modal ranking function F.

For the purpose of learning to rank, the structural SVM takes a set of vector-valued features which characterize the relationship between the input query and a set of target documents as the input, and returns a ranking list $\mathbf{y} \in \mathcal{Y}$ of the target documents. The structural SVM is applied to maximize the margins between the true ranking list $\mathbf{y}^*$ and all the other possible lists $\mathbf{y}$. In this paper, LSCMR takes cross-modal ranking into consideration, for $i = 1, \ldots, N$:

$$\forall \mathbf{y} \in \mathcal{Y} : \delta F(q_i, \mathbf{d}_i, \mathbf{y}) \geq \Delta(\mathbf{y}_i^*, \mathbf{y}) - \xi_i \tag{8}$$

where for compactness, we define

$$\delta F(q_i, \mathbf{d}_i, \mathbf{y}) = F(q_i, \mathbf{d}_i, \mathbf{y}_i^*) - F(q_i, \mathbf{d}_i, \mathbf{y})\,.$$

Since we assume that the query texts and the target images are embedded into a common latent space, respectively, LSCMR adapts the original structural SVM to learn the optimal U^* and V^* which maximize the margins between the true ranking and all the other possible rankings of the target images for each text query. Hence, we replace the standard quadratic regularization $\frac{\lambda}{2}\|w\|_2^2$ with $\frac{\lambda}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2$ where $\|\cdot\|_F$ denotes the Frobenius norm. Intuitively, this extension simplifies the model complexity, thereby promoting a better generalization performance.

The optimization problem is then presented as follows:

Optimization Problem 1.

$$\min_{U,V,\xi} \quad \frac{\lambda}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 + \frac{1}{N}\sum_{i=1}^{N} \xi_i \tag{9}$$

$$s.t. \quad \forall i \in \{1, \ldots, N\}, \forall \mathbf{y} \in \mathcal{Y} : \delta F(q_i, \mathbf{d}_i, \mathbf{y}) \geq \Delta(\mathbf{y}_i^*, \mathbf{y}) - \xi_i\,. \tag{10}$$

For each triple $(q_i, \mathbf{d}_i, \mathbf{y}_i)$ in the training set, a set of constraints (10) are added to the optimization problem. To see how these constraints indeed work, note that during the prediction the model chooses the ranking $\bar{\mathbf{y}}_i$ which maximizes $F(q_i, \mathbf{d}_i, \mathbf{y})$ given the fixed U and V. If the predicted ranking is an incorrect ranking $\bar{\mathbf{y}}$, i.e., $F(q_i, \mathbf{d}_i, \bar{\mathbf{y}}_i) > F(q_i, \mathbf{d}_i, \mathbf{y}_i^*)$ where $\mathbf{y}_i^*$ is the true ranking, the corresponding slack variable ξ_i must be at least $\Delta(\mathbf{y}_i^*, \bar{\mathbf{y}}_i)$ to satisfy the constraint. Considering all the triples $(q_i, \mathbf{d}_i, \mathbf{y}_i), i = 1, \ldots, N$, the sum of slacks (i.e., $\frac{1}{N}\sum_{i=1}^{N} \xi_i$) upper bounds the empirical risk $R^{\Delta}(f)$ defined in Equation (3). This is stated formally in Proposition 1.

Proposition 1. *Denote by $\xi^*(U, V)$ the optimal solution of the slack variables in Optimization Problem 1 for the given parameters U and V. Then $\frac{1}{N}\sum_{i=1}^{N} \xi_i^*$ is an upper bound on the empirical risk $R^{\Delta}(f)$.*

Similar to SVM, to avoid overfitting, the objective function (9) to be minimized is a tradeoff between the model complexity, and a hinge loss relaxation of Δ loss. A prechosen value of parameter λ controls this tradeoff and can be tuned to achieve a good performance via the cross validation procedure.

Note that by exploring the low rank property, the optimization problem is not convex. The well-known *kernel trick* is difficult to be applied to (9), while kernel trick is considered as one of the main benefits of the traditional support vector machine. Fortunately, a linear-SVM without using kernels has been shown to give competitive performances for textual documents classification [13]. On the other hand, according to the cross-modal retrieval approach PAMIR [11], a linear mapping of BoVW yeilds the highest performance of the other kernel mapping methods. As a result, with

the multi-modal data under a certain feature representation, we argue that the model can indeed capture the linear structures of the multi-modal data to learn a cross-media semantic representation.

2.4 Algorithm and Implementation

Since $|\mathcal{Y}|$ is super-exponential in the size of the training set, our algorithm for learning U and V is adapted from the 1-slack margin-rescaling cutting-plane algorithm of Joachims et al [20]. The algorithm alternates between two steps, one optimizing the model parameters (U and V in our case) and the other updating the constraints set with a new batch of rankings $(\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N)$ ($\hat{\mathbf{y}}_i$ is one ranking for one query sample, $i = 1, \ldots, N$) which most violate the current constraints. Once reaching a stopping criterion based on the accuracy of the empirical risk (the new constraint batch's empirical risk is no more than that of the current set of constraints within a tolerance $\epsilon > 0$), the algorithm terminates.

The general optimization procedure of LSCMR is listed in Algorithm 1. The code is implemented in MATLAB. The proof of the correctness can be easily extended from [20].

Algorithm 1 Latent Semantic Cross-Modal Ranking (LSCMR).

Input: ranking examples $(q_i, \mathbf{d}_i, \mathbf{y}_i^*), i = 1, \ldots, N$, trade-off control parameter $\lambda > 0$, accuracy tolerance threshold $\epsilon > 0$

Output: mapping parameters U and V, slack variable $\xi \geq 0$

1: $\mathcal{W} \leftarrow \emptyset$
2: **repeat**
3: Solve for the optimal U, V and slack ξ:

$$\min_{U,V,\xi} \quad \frac{\lambda}{2}\|U\|_F^2 + \frac{\lambda}{2}\|V\|_F^2 + \xi$$

$$\text{s.t.} \quad \forall(\mathbf{y}_1, \ldots, \mathbf{y}_N) \in \mathcal{W}:$$

$$\frac{1}{N}\sum_{i=1}^{N} \delta F(q_i, \mathbf{d}_i, \mathbf{y}_i) \geq \frac{1}{N}\sum_{i=1}^{N} \Delta(\mathbf{y}_i^*, \mathbf{y}_i) - \xi$$

4: **for** $i = 1$ **to** N **do**
5: $\hat{\mathbf{y}}_i \leftarrow \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} \Delta(\mathbf{y}_i^*, \mathbf{y}) + F(q_i, \mathbf{d}_i, \mathbf{y}_i)$
6: **end for**
7: $\mathcal{W} \leftarrow \mathcal{W} \cup (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_N)$
8: **until**

$$\frac{1}{N}\sum_{i=1}^{N} \Delta(\hat{\mathbf{y}}_i^*, \hat{\mathbf{y}}_i) - \frac{1}{N}\sum_{i=1}^{N} \delta F(q_i, \mathbf{d}_i, \hat{\mathbf{y}}_i) \leq \xi + \epsilon$$

9: **return** U, V, ξ;

To solve the optimization problem in Algorithm 1, there are two key issues to be resolved. One is searching for the most violated constraints, the so-called *separation oracle*, in Step 5. For different loss functions $\Delta(\mathbf{y}^*, \mathbf{y})$, different methods are proposed to address this issue, for example, [19] for AUC loss (defined as $1 - \text{AUC}(\mathbf{y}^*, \mathbf{y})$) and [32] for MAP loss. Recalling that F is the Frobenius inner product of $U^T V$ and Ψ, their work [19, 32] can be easily applied to this algorithm with a minor modification in implementation to reduce the computational complexity.

The other key issue is how to solve the optimization problem in Step 3. Since the problem is not a convex problem, the parameters U and V are initialized with their previous (local) optimal values while in the beginning they are randomly initialized using a normal distribution with mean zero and standard deviation one. We have implemented a subgradient descent solver adapted from Pegasos algorithm [27] originally proposed for solving a traditional support vector machine. The Pegasos algorithm is a simple iterative algorithm which alternates between stochastic subgradient descent and projection steps, and is shown to be effective to solve the primal problem of SVM. In the problem, the subgradient descent is performed by iteratively picking the most violated ranking tuple $(\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_N)$ from the set $\mathcal{W}$ to minimize the slack variable.

On iteration t, the update for U is given by:

$$U_{t+\frac{1}{2}} \leftarrow (1 - \eta_t \lambda) U_t + \frac{\eta_t}{N}\sum_{i=1}^{N} V_t (\delta\Psi(q_i, \mathbf{d}_i, \hat{\mathbf{y}}_i))^T \qquad (11)$$

where η_t is the learning rate on iteration t which is adjustable and $\delta\Psi(q_i, \mathbf{d}_i, \hat{\mathbf{y}}_i) \triangleq \Psi(q_i, \mathbf{d}_i, \mathbf{y}_i^*) - \Psi(q_i, \mathbf{d}_i, \hat{\mathbf{y}}_i)$. U_{t+1} is obtained by projecting $U_{t+\frac{1}{2}}$ onto the set for acceleration (see [27]):

$$B = \{U : \|U\|_F \leq 1/\sqrt{\lambda}\} \qquad (12)$$

The update for V can be derived similarly except for the most violated ranking tuple which is computed using the updated U_{t+1}. The update is calculated exactly as given by:

$$V_{t+\frac{1}{2}} \leftarrow (1 - \eta_t \lambda) V_t + \frac{\eta_t}{N}\sum_{i=1}^{N} U_{t+1} \delta\Psi(q_i, \mathbf{d}_i, \hat{\mathbf{y}}_i) \qquad (13)$$

followed by the projection step (12).

Moreover, our problem is a bit different from Pegasos [27]: the objective function is penalized by $\frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)$ to control the model perplexity. It should be noted that the optimal U and V must satisfy the condition $\|U\|_F = \|V\|_F$ since the prediction rule uses the product $U^T V$ only. Thus after each subgradient descent, the updated U and V are forced to be multiplied with a constant respectively to ensure $\|U\|_F = \|V\|_F$ while keeping $\|U^T V\|_F$ fixed. Let $\alpha = \sqrt{\|U\|_F \|V\|_F}$,

$$U \leftarrow \alpha U / \|U\|_F$$
$$V \leftarrow \alpha V / \|V\|_F$$

The experiments show that this strategy yields a much faster convergence rate. For fixing tolerance $\epsilon = 0.01$, the loop in Algorithm 1 usually terminates within 200 iterations.

3. PRIOR WORK

There has been a great deal of research devoted to the development of algorithms for learning the similarity between the data with different modalities in order to perform cross-modal retrieval. Most of cross-modal metric learning approaches tend to project multi-modal data into a common (or shared) subspace so that the correlation between multi-modal data is preserved or maximized. As one of the most popular approaches to finding a pair of linear transformations to maximize the correlations between two variables, Canonical Correlation Analysis (CCA) [15] and its extensions are applied in cross-modal similarity learning. For

example, after the maximally correlated subspaces of text and image features are obtained by CCA, logistic regression is employed to cross-modal retrieval in [26]. As a supervised kernelizable extension of CCA, Generalized Multiview Analysis [28] is conducted to map data in different modality spaces to a single (non)linear subspace. Motivated by the fact that dictionary learning (DL) methods have the intrinsic power of capturing the heterogeneous features by generating different dictionaries for multi-modal data, multi-modal dictionary learning is recently applied to cross-modal metric learning [16, 24]. Following the the seminal work of Blei et al. [5], LDA has been extended to learn the joint distribution of multi-modal data (e.g., text and imagery) such as correspondence LDA [4], topic-regression multi-modal LDA [25], Multi-modal Document Random Field [17] and hierarchical Dirichlet process (HDP)-based LDA [30].

The aforementioned approaches, either optimizing the similarity (distance) between pairs of samples or optimizing the likelihood of the topic models, do not optimize for the final ranking performance directly. While bearing a resemblance to multi-modal metric learning which aims at learning the similarity or the distance measure from multi-modal data, the multi-modal ranking function is generally optimized by an evaluation criterion or a loss function defined over the *permutation* space induced by the scoring function over the target documents.

Traditionally, algorithms of learning to rank can be categorized into the pointwise approaches, the pairwise approaches, and the listwise approaches. The main differences among these three categories of approaches actually lie in the input representations and the loss functions employed in training. It is observed that the listwise and pairwise approaches usually outperform the pointwise approaches [21].

In [18] Joachims et al. trained a Ranking Support Vector Machine (RankSVM) to learn the weights of the hand-designed features in which the training set was a set of documents preference pairs obtained through the clickthrough data from the query-log of a search engine. The retrieval function is automatically learned by taking a support vector machine. The goal of RankSVM is to minimize the average number of the inversions in ranking; thus the method is considered as a pairwise preference satisfaction approach.

Unlike the pairwise approaches, Cao et al. [6] first noticed the fact that ranking was a prediction task on a list of documents and took the ranking lists as training instances. They trained two probabilistic models, respectively referred to as permutation probability and top k probability, to define a listwise loss function for learning.

Yue et al. [32] proposed another listwise approach SVM^{map} to solve the problem of learning to rank in a discriminative way. The method uses the structural SVM framework [29] with the loss function defined as MAP (mean average precision) loss that globally optimizes a hinge-loss relaxation of MAP loss. This method simplifies the process of obtaining ranking functions with a high MAP performance by avoiding the additional intermediate steps and heuristics.

Chakrabarti et al. [7] proposed almost-linear-time algorithms to optimize MRR (mean reciprocal rank) and NDCG (normalized discounted cumulative gain). Further, they folded multiple ranking loss functions into a multi-criteria max-margin optimization problem to develop a single, robust ranking model with close to the best accuracy of the learners trained on individual criterions.

Different from the aforementioned *uni*-modal learning to rank techniques, to the best of our knowledge, Passive Aggressive Model for Image Retrieval (PAMIR) is the first attempt to address the problem of ranking images by text query directly [11]. PAMIR formulates the cross-modal retrieval problem similar to RankSVM and derives an efficient training procedure by adapting the Passive-Aggressive algorithm.

The authors of [23] studied metric learning as a problem of learning to rank. They presented a general metric learning algorithm based on the structural SVM, to learn a metric such that the ranking of data induced by the distance from a query can be optimized again various ranking measures. Different from LSCMR, they focused on learning a intra-modality metric which restricts a positive semi-define matrix W (see Eq. (1)) to learn a valid metric and does not introduce a low rank embedding.

The text and imagery are usually represented as BoW and BoVW in a high-dimensional vector space. However, the high-dimensional vector space representation suffers from its inability to cope with two classic problems, i.e., synonymy and polysemy. To capture the latent semantic associations of data and to address these problems, embedding words in a low-dimensional latent space to capture the semantics is a classic approach in text retrieval such as Latent Semantic Indexing (LSI) [9] and pLSA [14]. The idea of low rank embedding is introduced into Supervised Semantic Indexing (SSI) for cross-lingual retrieval [2]. SSI defines a set of linear low rank models to take account of correlations between words (synonymy and polysemy). Related to SSI, Polynomial Semantic Indexing (PSI) [3] generalizes and extends the SSI approach to general polynomial models which could be used to capture the higher order relationships among words.

4. EXPERIMENTS AND RESULTS

The main goal of the experiments is to evaluate the effectiveness of the proposed LSCMR approach. To show its competitive performance, LSCMR is compared with other three state-of-the-art approaches (CCA, PAMIR and SSI) for cross-modal ranking.

These comparative methods are elaborately chosen for the fair comparisons. Comparing with the classical CCA method aims to test LSCMR's ability to learn a useful latent space; PAMIR has been shown to outperform pLSA and SVM [11]; however it does not consider a latent space; SSI introduces the low rank parameterizations while it minimizes a pairwise ranking loss and lacks of parameter regularization. Since LSCMR can generate low rank matrices U and V (see Eq. (2)), in the experiments, we demonstrate that the learned model also discovers the latent correlations between textual words and topics.

4.1 Experimental Setup

Two public real-world datasets are used in the comparative experiments. They are the largest available multimodal datasets that are fully paired and labeled (tagged), to the best of our knowledge. Both datasets are bi-modal with the image and the associated text modalities. The statistics of the two datasets are summarized in Table 1.

The first dataset, Wikipedia feature articles[1], consists of 2,866 images, each with a short paragraph describing the

[1] http://www.svcl.ucsd.edu/projects/crossmodal/

Table 1: The statistics of the datasets used.

	Wikipedia	NUS-WIDE
BoVW vocabulary size	1000	500
BoW vocabulary size	5000	1000
Avg. # of words/image	**117.5**	**7.73**
Documents	1,500/500	2,664/23,977
Partition[a]	866	106,567
Queries	1,500/500	2,664/2,000
Partition[a]	866	2,000

[a] Partitions are ordered by training/validation/test.

image. The images are labeled with exactly one of the 10 different semantic classes, such as art and geography. In the originally provided dataset, the text comes with a 10 dimensional feature vector representing the probabilistic distributions over the 10 topics, which are derived from a Latent Dirichlet Allocation (LDA) model [5]. We note that LSCMR and the comparative methods all resort to the raw low-level features rather than the high-level semantic features. For the training text, we extract 5,000-dimensional feature vectors using the bag of words (BoW) representation with the TF-IDF weighting scheme. For images, we first extract SIFT points from each images in the dataset. The randomly selected SIFT points are clustered by k-means to generate 1,000 centers as the visual dictionary. Then each image is quantized into a 1,000 dimensional histogram feature vector using the bag-of-visual-words (BoVW) model.

The second dataset, NUS-WIDE[2], contains 133,208 images with 1,000 tags and 81 concepts, which are pruned from the NUS dataset by keeping the images that have at least one tag and one concept. For the feature representation, we use the publicly available 1,000 dimensional text feature vector (namely tags) and 500 dimensional image feature vector based on SIFT BoVW kindly provided by the authors.

Another reason why we choose these two datasets is due to the large difference on the average number of the textual words per image and the dimensionality of the text space. In Wiki dataset the textual descriptions are based on Wikipedia surrounding paragraphs which yield a 5,000 dimension text space and in average there are 117.5 surrounding words per image. The NUS dataset, on the other hand, is based on Flickr user-provided tags which yield a 1,000 dimension text space and in average there are 7.73 words (tags) per image. A manual examination reveals that the synonymy and polysemy problem may occur more frequently in the Wiki dataset than in the NUS dataset. For their difference, first we want to examine our algorithm's ability to learn a latent space for the Wiki dataset and second we want to see whether our algorithm decays rapidly with the NUS dataset.

Note that the two datasets are both presented by pairs of text and imagery where CCA can be trained by this setting. For the other three methods (PAMIR, SSI and LSCMR), the restriction of paired correspondence of a text document and an image is not needed. On the contrary, the queries and the corresponding ranking lists are needed as training examples of the three methods. These training examples originate from both direction of the text-query-image retrieval and the image-query-text retrieval. For this purpose, we first define the relevance assessment. For the Wiki dataset, we define that a target document d is relevant to a query q if d and q belong to the same semantic class. Similarly, for the NUS dataset, a target document d is relevant if it shares at least one concept with query q. The ranking examples are generated as follows: for each text (image) query, we randomly selected 40 images (text documents) in the other modality in the training set as candidates and then the selected target documents are automatically labeled as relevant or irrelevant to form a ranking example.

For all the 2,866 generated ranking examples in the Wiki dataset, we randomly sample 1,500 examples to form the training set, of which 500 examples form the validation set. The rest are used to form the testing set. For NUS, 2,664 ranking examples are randomly selected to be the training samples and 2,000 to be validation samples (see Table 1). To be fair, all the comparative approaches are trained and tested on the same training set and testing set respectively.

For both datasets, performance evaluations are conducted using standard information retrieval metrics. We use *Mean Average Precision* (MAP) as the performance measures. Let $p^* = \text{rank}(\mathbf{y}^*)$ (true ranking with two rank value $+1$ and -1) and $p = \text{rank}(\mathbf{y})$ (predicted ranking with a total order). Given a query and a set of R retrieved target documents, the *Average Precision* (AP) is defined as

$$\text{AP}(p^*, p) = \frac{1}{L} \sum_{j=1}^{R} Prec(j) \cdot Rel(j) \tag{14}$$

where L is the number of the relevant documents in the retrieved set, $Prec(j)$ is the percentage of the relevant documents in the top j documents in predicted ranking p and $Rel(j)$ is an indicator function equaling 1 if the item at rank j in predicted ranking p is a relevant document, zero otherwise. We then average the AP values from all the queries in the query set to obtain the MAP score. The larger the MAP, the better the performance. In the experiments, R is the number of the retrieved documents to be examined, where we set $R = 50$ or $R = all$ for all the retrieved documents. Recalling that our model can be optimized for various ranking measures, we implement the greedy algorithm for optimizing the average precision proposed in [32].

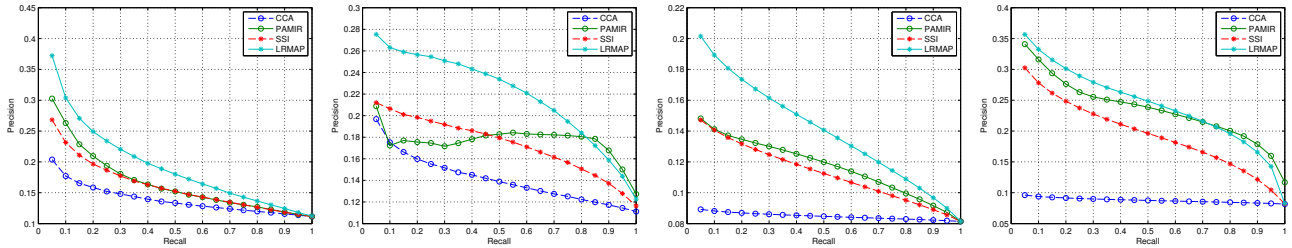
We report the performance results on both directions of ranking images from text queries (*text-query-image*) and ranking text documents from image queries (*image-query-text*). Besides, to give an pictorial demonstration of an algorithm's performance, the *Precision-Recall* curves are also reported on all the approaches.

4.2 Results on the Wiki Dataset

Table 2 reports the performance of LSCMR and the other comparative models on the testing set of the Wiki dataset, showing that LSCMR outperforms all the comparative methods on both directions of the retrieval tasks. Compared to the best comparative methods, the minimum relative improvement is 9.6 percent gained by LSCMR for the image-query-text retrieval with $R = 50$ and the maximum is 25.3 percent also for the image-query-text retrieval with $R = all$.

This improvement is due to the latent semantic space. To verify this, we note that the low-rank based SSI also outperforms PAMIR in the image-query-text retrieval while PAMIR even controls the model complexity by optimizing an adapted cross-modal RankSVM model. Further, LSCMR outperforms SSI due to the structural large margin that reg-

[2] `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm`



(a) Wiki Text Query (b) Wiki Image Query (c) NUS Text Query (d) NUS Image Query

Figure 2: Precision-Recall curves on the two datasets (LRMAP is short for LSCMR optimizing MAP).

Table 2: The performance comparison in terms of MAP@R scores on the Wiki dataset. Each text document is represented as 5000-D BoW and each image is presented as 1000-D BoVW. Both directions of ranking tasks are reported. The results shown in boldface are the best results.

		CCA	PAMIR	SSI	LSCMR
Text Query	$R = 50$	0.2343	0.3093	0.2821	**0.3663**
	$R = all$	0.1433	0.1734	0.1664	**0.2021**
Image Query	$R = 50$	0.2208	0.1797	0.2344	**0.2570**
	$R = all$	0.1451	0.1779	0.1759	**0.2229**

Table 3: The performance comparison in terms of MAP@R scores on the NUS dataset. Each text document is represented as 1000-D BoW and each image is represented as 500-D BovW. Both directions of ranking tasks are reported. The results shown in boldface are the best results.

		CCA	PAMIR	SSI	LSCMR
Text Query	$R = 50$	0.1497	0.2046	0.2156	**0.2781**
	$R = all$	0.0851	0.1184	0.1140	**0.1424**
Image Query	$R = 50$	0.1523	**0.5003**	0.4101	0.4997
	$R = all$	0.0883	0.2410	0.1992	**0.2491**

ularizes the model and optimizes for MAP ranking loss directly. The Precision-Recall curves on both directions are reported in Figure 2(a) and 2(b). The Precision-Recall curves further validate the superiority of LSCMR for the cross-modal ranking.

Recall that the CCA model is trained by the pairs of data objects with different modalities and learns a unified model for the retrieval tasks of both directions. Hence, CCA has achieved nearly the same performance in both image-query-text retrieval and text-query-image retrieval. On the other hand, the performances of other three approaches (including LSCMR) are noticeably different in the corresponding both directions of the retrieval.

4.3 Results on the NUS Dataset

The improvement of LSCMR on the NUS dataset is not as significant as that on the Wiki dataset. The MAP scores of all the methods are shown in Table 3 and the Precision-Recall curves are reported in Figure 2(c) and Figure 2(d). For text-query-image retrieval, LSCMR outperforms the other comparative methods again while for image-query-text retrieval LSCMR outperforms all the comparative methods in all the cases except for the case of $R = 50$ where PAMIR has a slightly better overall performance than LSCMR.

Recall that in the NUS-WIDE dataset, one image is associated with about seven annotated words in average. The low rank embedding does not help much for querying short texts in the task of image-query-text retrieval. PAMIR and LSCMR both train a regularized model, and therefore the performances of PAMIR and LSCMR is undoubtedly super to SSI in the image-query-text retrieval.

Once again, it is observed that CCA has a very similar performance in both directions of the retrieval.

4.4 The Performance Discussions

It is noted that LSCMR has a better overall performance for text-query-image retrieval than for image-query-text retrieval in the Wiki dataset and a better overall performance for image-query-text retrieval than for text-query-image retrieval in the NUS dataset. The reason exactly lies in the low-rank embedding of LSCMR that is capable of discerning the latent aspect space and consequently supports rich-semantic queries. For the Wiki dataset, each document has over one hundred words in average, resulting in a long text query which is presumably much richer in semantics than the case in the NUS dataset where each image has only about seven text words in given as the annotation that is much shorter when posed as a query. On the other hand, the overall image quality in the NUS dataset is much richer and more diverse in semantics than that in the Wiki dataset, and thus resulting in a better overall performance for image-query-text-retrieval.

As we have stated that for both datasets CCA achieves very similar performances in both directions of the retrieval tasks. The reason is that CCA learns a unified model from paired multi-modal data in which the pair-correspondence of images and text documents ensure an equal contribution to the learned metric between both modalities. However, a unified model like CCA does not give a good performance which is verified in both datasets. Our explanation is as follows: it is not the problem of unified models but the problem of taking strictly paired data as training instances; in such settings CCA can not capture the complete ranking information (e.g. dissimilarity between data in different modalities). Nevertheless, learning a unified model with ranking lists is still our interests for future work.

Overall, the results on the Wiki dataset and NUS dataset demonstrate that the use of LSCMR is advantageous for rich-semantic queries and has a superior performance in both

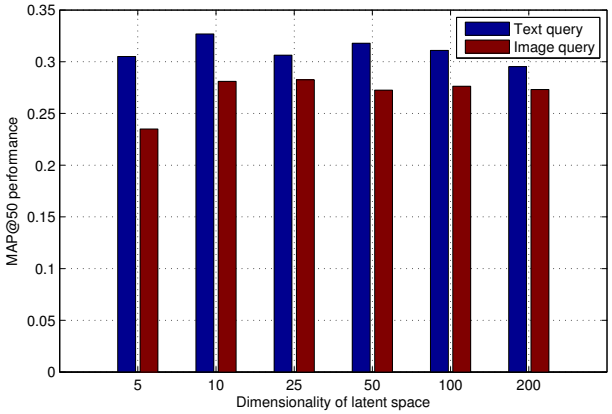


Figure 3: The performance (MAP@50) comparisons for the Wiki dataset when the dimensions of latent space are set to different values. The dimensionality of 10 is reported to be the best for the overall performance of text query and image query.

Table 4: Exemplar words along with their top 9 neighboring words from the Wiki dataset.

Fiction	Stories Manuscript Tales Language Theatre Poetry Editor Publisher Powell
Football	Teams Player NHL Goal Conference Compete Olympics Matches Yards
DVD	Moves Brand Broadcasting Movies Victims Commented Sequel Direction Producer
Airport	Passenger Oregon Democratic Communities Bring Lands Telephone Sale Coverage

directions of the cross-modal retrieval than that of the peer methods from the state-of-the-art literature.

4.5 The Embedding Latent Space

The results over the Wiki dataset outline the advantage of LSCMR over the comparative solutions, especially PAMIR. This observation is certainly due to the low rank embedding used in LSCMR. Now we look into the learned embedding latent aspect space for the Wiki dataset. Since the BoVW features used in our experiments originate from the SIFT points which are difficult to illustrate, we only demonstrate the latent space by textual words.

First, the dimensionality of the subspace (also the size of U and V) is a parameter to be determined before solving Optimization Problem 1. We tune this parameter via a validation set and the MAP@50 performance over a candidate set of the chosen dimensionalities is reported in Figure 3. It is observed that LSCMR performs the best with the optimal subspace dimensionality of 10. When the chosen dimensionality of the latent subspace is larger (even much larger) than 10, the performances do not decay rapidly at the same time it is observed that the topics overlap. But we find something interesting that though some topics overlap, some smaller but more precise topics are discovered (see the two "Biology" topics in Table 5).

Second, consider the mapping of textual words into the latent space in LSCMR. For a text query q, Uq maps q into the latent space. Note that $U \in \mathbb{R}^{k \times m}$ where k represents the dimensionality of the latent space. For each row in U, the row weights the contribution of all words to the corresponding "topic" in the latent space. The larger the number U_{ij}, the more positive correlation between topic i and word j. We then sort every row by which the most relevant words are ranked ahead. We present some topic examples in Table 5. The columns in U acts a similar way like the rows except for that each column represents the relevance between the corresponding word and all topics. We define two words are neighbors if their relevance with all topics are similar. Some examples on the neighboring words are shown in Table 4.

5. CONCLUSIONS

In this work, we have presented a new approach to solving the problem of cross-modal retrieval by casting the problem as a problem of learning to rank in a supervised manner with the idea of low rank embedding. We have demonstrated the effectiveness of our proposed method LSCMR and have shown significant improvements over the comparative methods especially on two datasets. We have also investigated the interpretability of the leaned low rank model by showing some examples on the textual topics and the neighboring words.

6. ACKNOWLEDGEMENTS

This work is supported by 973 Program (No. 2012CB316400), NSFC (61070068, 90920303), 863 program (2012AA012505), Chinese Knowledge Center of Engineering Science and Technology (CKCEST) and China Academic Digital Associative Library (CADAL). Zhongfei Zhang is also supported by US NSF (IIS-0812114, CCF-1017828).

7. REFERENCES

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314, 2010.

[3] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, C. Cortes, and M. Mohri. Polynomial semantic indexing. *Advances in Neural Information Processing Systems*, 22:64–72, 2009.

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 127–134, 2003.

[5] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *the Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, pages 129–136, 2007.

[7] S. Chakrabarti, R. Khanna, U. Sawant, and C. Bhattacharyya. Structured learning for non-smooth ranking losses. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 88–96, 2008.

[8] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer. Online passive-aggressive algorithms. *The Journal of Machine Learning Research*, 7:551–585, 2006.

Table 5: Exemplar topics from the Wiki dataset. We assign each learned topic to its most probable category. Topic words are sorted by their importance values in the descending order.

Category	Topic Words
Geography & Places	Guadalcanal Corps Sale Airport Battleships Iowa Aircraft Carriers Chicago Puerto
Biology	Subspecies Breeding Dinosaurs Fossils Wallace Genus Marsh Females Wings Virginia
Warface	Puerto Battalion Guadalcanal Oklahoma Soviet Soldier Tank Bishop Bradman Infantry
Literature & Theatre	Hamlet Theatre Gilbert Uncle Stories Shakespeare Refugees Manuscript Fiction Punk
Sports & Recreation	Tech Jordan Players Championship Tournament Olympic NHL Cricket EP Coach
Biology	Puerto Zoo Skull Nest Augustus Specimen Tail Teeth Organisms Darwin

[9] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[10] J. Gao, W. Yuan, X. Li, K. Deng, and J. Nie. Smoothing clickthrough data for web search ranking. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 355–362, 2009.

[11] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, 2008.

[12] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

[13] C. Ho and C. Lin. Large-scale linear support vector regression. Technical report, National Taiwan University, 2012.

[14] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[15] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[16] Y. Jia, M. Salzmann, and T. Darrell. Factorized latent spaces with structured sparsity. *Advances in Neural Information Processing Systems*, 23:982–990, 2010.

[17] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *IEEE International Conference on Computer Vision*, pages 2407–2414, 2011.

[18] T. Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[19] T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 377–384, 2005.

[20] T. Joachims, T. Finley, and C. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[21] H. Li. Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1):1–113, 2011.

[22] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[23] B. McFee and G. Lanckriet. Metric learning to rank. In *Proceedings of the 27th International Conference on Machine Learning*. Citeseer, 2010.

[24] G. Monaci, P. Jost, P. Vandergheynst, B. Mailhe, S. Lesage, and R. Gribonval. Learning multimodal dictionaries. *IEEE Transactions on Image Processing*, 16(9):2272–2283, 2007.

[25] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3415, 2010.

[26] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*, pages 251–260, 2010.

[27] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, pages 807–814, 2007.

[28] A. Sharma, A. Kumar, H. Daume, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2160–2167, 2012.

[29] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6(2):1453–1484, 2006.

[30] S. Virtanen, Y. Jia, A. Klami, and T. Darrell. Factorized multi-modal topic model. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 843–851, 2012.

[31] Y. Yang, N. Bansal, W. Dakka, P. Ipeirotis, N. Koudas, and D. Papadias. Query by document. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 34–43, 2009.

[32] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 271–278, 2007.

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