Cross-Media Retrieval based on Pseudo-Label Learning and Semantic Consistency Algorithm

Gongwen Xu*, Zhiqi Sangb, and Zhijun Zhanc

*aSchool of Information Science and Engineering, Shandong Normal University, Jinan, 250358, China
bCollege of Architecture and Urban Planning, Shandong Jianzhu University, Jinan, 250101, China
cSchool of Computer Science and Technology, Shandong Jianzhu University, Jinan, 250101, China

Abstract

To retrieve heterogeneous multimodal data with the same semantics, many algorithms for retrieval over multimodal data have been suggested. The organization and analysis of heterogeneous data have become the focus of intensive research. Here, a new and efficient algorithm for cross-media retrieval is proposed based on pseudo-label learning and semantic consistency (PLSC). In this algorithm, an adaptive learning projection matrix optimization method is proposed, and in the process of learning the projection matrices, the method fully considers the semantic information of the labeled and unlabeled samples. Thus, the PLSC algorithm can utilize more useful information than other methods and can learn the more efficient projection matrices. Firstly, the class centers of labeled text are computed. We use median feature vectors as the class center vectors. Next, unlabeled images are projected onto the text space and are assigned pseudo-labels by comparing with the class center vectors of the text data. Finally, a new training dataset, which includes labeled and unlabeled data, is generated for training the projection matrix. Using the projection matrix to project image or text data onto the same feature space, the data can be compared with each other for similarity, and the distance between data points can be calculated using the Euclidean metric. Validation experiments suggest that the PLSC outperforms other state-of-the-art algorithms.

Keywords: cross-media retrieval; pseudo-label; semi-supervised; semantic analysis

(Submitted on June 13, 2018; Revised on July 24, 2018; Accepted on August 30, 2018)
© 2018 Totem Publisher, Inc. All rights reserved.

1. Introduction

With advances in information technology, the amount of multimodal data has steadily increased. Multimodal data are ubiquitous, with humans sharing personal text, audio, image, and video information using the Internet. Multimodal data refers to the data that describe the same object/concept using different modalities. The different components of the multimodal data of a certain object/concept exist in different modalities but are associated on a high semantic level [1-3]. Multimodal data are widely present and used in everyday life; thus, the detection and analysis of multimodal data is an important area of research.

Retrieval over multimodal data has recently become a topic of active research. Methods for retrieval over multimodal data differ from traditional data retrieval methods, which mine information from same-modality data. In cross-media retrieval, a media object in one modality can be used for the retrieval of data in other modalities. For example, when an image of a ‘plane’ is considered, text, images, audio recordings, and video recordings of ‘plane’ can be retrieved. These different images can be used to retrieve other images, according to our demand. Figure 1 shows the concept of cross-media retrieval schematically.

A good method for cross-media retrieval can significantly shorten the retrieval time and reduce the workload, and it can effectively improve the precision and recall of information retrieval. Because the data to be retrieved are in different modalities, low-level features of these data are heterogeneous, while high-level semantics are inter-related [4-5]. The respective data structures are complex and vary significantly. Multimedia data such as images, audio recordings, and video
recordings are semi-structured or unstructured, making it difficult to describe high-level semantics in terms of low-level features. In addition, the features of the different modality data span spaces with different dimensionalities; therefore, it is difficult to compare data in different modalities using existing methods. Thus, feature heterogeneity of multi-modal data is the strongest challenge associated with retrieval over multimodal data.

Figure 1. Under the cross-media retrieval method, the text, image, wave, and video can retrieve each other

Here, we consider retrieval and comparison of text and image information. Concretely, we consider the problem of using text to retrieve images (i.e. text illustration), and using images to retrieve text. The problem is conceptualized in Figure 2.

A new retrieval method, which we name pseudo-label learning and the semantic consistency (PLSC) algorithm, is proposed here. In this method, an adaptive learning projection matrix optimization method is proposed, and in the process of learning the projection matrix, PLSC fully considers the semantic information of the labeled and unlabeled samples. Thus, it
can utilize more useful information than other methods and can learn the more efficient projection matrices. Firstly, the class centers of labeled text are computed. We use median feature vectors as the class center vectors. Next, unlabeled images are projected onto the text space and are assigned pseudo-labels by comparing with the class center vectors of the text data. Finally, a new training dataset, which includes labeled and unlabeled data, is generated for training the projection matrix. Using the projection matrix to project image or text data onto the same feature space, the data can be compared with each other for similarity, and the distance between data points can be calculated using the Euclidean metric. Validation experiments suggest that the PLSC outperforms other state-of-the-art algorithms.

In this method, high-dimensional features of different modality data are projected onto a shared latent space, and then we can calculate the similarity between different data points using the same distance metric. First, image and text data are projected from their own feature spaces onto the same semantic space using linear regression. Next, the inner-modality and intra-modality are both considered, and the correlation between image and text data is computed to ensure their similarity in the same space. Finally, a group of maps relating image and text modality data are learned under joint optimization. In our method, unlabeled data are valued as much as labeled data during training, and they are assigned pseudo-labels, which significantly improves retrieval.

This paper firstly introduces the related work briefly in Section 2. In Section 3, the pseudo-labeling method is introduced. The proposed method, the PLSC method, is introduced in Section 4 in detail. Section 5 describes the validation experiment and its results. Finally, we provide conclusions in Section 6.

2. Related Work

Methods for retrieval over multimodal data assume that heterogeneous data are uniformly represented. When multimedia data are projected onto an isomorphic space, they can be compared using the same distance metric, e.g. using the Euclidean metric or the Hamming metric [6].

Much research has been performed recently on cross-media retrieval. The concept of a shared subspace was introduced, allowing different modality data to be represented consistently in the same subspace for mutual retrieval. Canonical correlation analysis (CCA) is a typical uniform representation and is widely used in the fields of computational biology, financial analysis, and information retrieval [2]. Methods have been developed to maximize the similarity between pairwise training data in a common space. Kernel canonical correlation analysis has been applied to learn correlations between images and corresponding texts. Clustering methods have been widely used [7]. A method that preserves local correlations among class data, named the locality correlation preserving based support vector machine (LCPSVM), has been proposed [8]. A set of linear mapping matrices were obtained and data with heterogeneous features were projected onto the same semantic space [9-12]. However, it is not sufficient to only consider correlations between data points in the training set. In cross-media retrieval, it is expected that retrieved data are semantically similar to the querying image/text data. Thus, semantically similar multimedia data are expected to form clusters in the shared subspace. For this purpose, [13-14] used supervised class information or supervised information obtained by clustering to cluster semantically similar multimedia data in the common subspace. However, these methods only learn one group projection per one text/image retrieval task; thus, their performance is not satisfactory.

Zhai et al. [15] proposed the joint representation learning (JRL) method. JRL considers the semantic information and correlations at the same time. Heterogeneous metric learning with joint graph regularization (JGRHML) [16] uses a joint graph regularization to retrieve among different media. The cross modality correlation propagation (CMCP)[17] uses the positive and negative correlation at the same time. HSNN [18] is a method that measures heterogeneous similarity with nearest neighbors. In [19], a method for modality-dependent cross-media retrieval (MDCR) was proposed, and it uses various projections according to various retrieval tasks. In this approach, for different retrieval tasks, different mapping matrices are learned, and the matrices are selected according to different retrievals, such as text or image retrieval tasks. This method demonstrated very good performance. However, the MDCR methods consider only labeled data and use supervised learning methods. The PLSC algorithm that is proposed in this paper learns the mapping matrices from labeled data as well as unlabeled data; thus, this algorithm implements semi-supervised learning.

Using unlabeled data is advantageous for the following reasons. First, labeled data are expensive and difficult to acquire, while unlabeled data can be easily acquired at a relatively low cost. Second, unlabeled data can increase the robustness of the retrieval model by increasing the accuracy of the decision boundary. The PLSC approach for cross-media retrieval uses the pseudo-labeling method [20], which is described next.
3. Pseudo-Labeling Method

Significant progress has been made recently in the field of artificial intelligence, especially in deep learning [21-22]. Existing deep learning models mainly focus on pretreatment of training and detail tuning. Pretreatment of training methods belong to the family of unsupervised learning category, while fine-tuning methods belong to the supervised learning category. On the one hand, the large number of unlabeled pre-training data affects the recognition results. On the other hand, the fine-tuning can improve the recognition accuracy with additional labeled training data [23].

Accordingly, the pseudo-labeling method was developed for training both labeled and unlabeled data. In the training processes, both convolutional neural networks and auto-encoders include classification [24]. In supervised learning, a measure of discrepancy (cross-entropy between the actual input label and model output) is calculated each time the model produces output in response to some labeled input. On the other hand, for dealing with unlabeled samples, pseudo-labels are required.

\[
y_i' = \begin{cases} 1, & \text{if } i = \arg\max_{i'} f_{i'}(x) \\ 0, & \text{otherwise} \end{cases}
\]

A pseudo-label \(y_i'\) is determined according to the above definition [20]. Before training this model, we define the loss function as below [20]:

\[
L = \sum L(f_i, y_i) + \alpha(t) \sum L(f_i, y_i')
\]

Where \(y\) and \(f\) represent the input and output of the supervised learning process, respectively. The label \(y'\) is a pseudo-label, \(\alpha\) is a weight coefficient, and \(t\) is the current iteration times. Thus, in this semi-supervised learning approach, both the unlabeled and labeled data can be used for training at the same time.

The main idea of pseudo-labeling is as follows. First and foremost, a learning model is trained on the labeled data. Then, the generating model is applied to label the untagged data, and pseudo-labels are created using the above approach. All the data are then combined in a new dataset that is used to train the model [20]. The above-described process is schematically shown in Figure 3.

\[\text{in the following section, we will implement the pseudo-labeling method in the traditional machine learning models.}\]

4. Proposed Method

In the following, the proposed method is described in detail. The PLSC algorithm uses the unlabeled data as well as the labeled data. The PLSC training process comprises three steps: 1) obtaining class centers, 2) producing the pseudo-labels, and 3) training projection matrices. The process is schematically shown in Figure 4.
Here, $V$ is the projection matrix for projecting the image space onto the text space:

$$V = \arg\max_{V} |V|_F$$

Before learning the text data class center location, we use the following equation to learn the projection matrix for projecting the text data space onto the image space:

$$V' = \arg\min_{V'} |V' - I||_2 + \lambda |V'||_2$$

In the retrieval over multimedia data, the similarity between different modality data cannot be measured directly. To correlate data in a certain modality with other modality data, we have learned projection matrices using labeled and unlabeled data. In the PLS-C data retrieval approach, the class center location of labeled text is computed first, and then unlabeled images are projected onto the space and assigned pseudo-labels by comparing the data with the class center location of the text data. This yields a new text sample data for training the projection matrix.

Before learning the projection matrix, the data are expressed using the following equation:

$$V = 1 \times V'$$

$$T_i = 1 \times V'$$

For unlabeled data, we further define $T_i = \{T_i \vdash r \in r\}$ and $T_i = \{T_i \vdash I \in r\}$, where $T_i$ and $T_i$ represent the tagged and untagged data in the image form, respectively, while $T_i$ and $T_i$ represent the tagged and untagged data in the text form, respectively.

For a given training dataset, $G = \{G_i \}_{i=1}^n$, contains $n$ pairs of data examples. The quantities $u_i \in \mathbb{R}^p$ and $l_i \in \mathbb{R}^q$ represent the low-level features of the image and text, respectively, in the training dataset. Assuming there are $c$ classes in the training dataset, the semantic feature matrix is

Figure 4: Schematic of the PLS-C algorithm

Assume that a given training dataset, $G = \{G_i \}_{i=1}^n$, contains $n$ pairs of data examples. The quantities $u_i \in \mathbb{R}^p$ and $l_i \in \mathbb{R}^q$ represent the low-level features of the image and text, respectively, in the training dataset. Assuming there are $c$ classes in the training dataset, the semantic feature matrix is $S = \{s_k \}_{k=1}^c$, where $s_k$ is the projection matrix for obtaining a better measure of similarity between semantically similar but different data points.
Considering that the data in the sample dataset may be distributed non-uniformly and there may be significant deviations from average, we use median feature vectors as the class center locations. In the experiment, the data with the same label has similar features. The features can be treated as the semantic of the same class. After data projection, the median feature vectors can be used to approximately represent the class label after the iterative operation. Thus, the class center location can be calculated as follows.

The \( j^{th} \) class center of \( T'_l \) and \( T_l \) is expressed as

\[
C_{T'_l} = \text{median}(T'_l)
\]

(5)

\[
C_{T_l} = \text{median}(T_l)
\]

(6)

Then, the class matrices \( C_{T'_l} \in \mathbb{R}^{p \times p} \) and \( C_{T_l} \in \mathbb{R}^{q \times q} \) are obtained.

Next, unlabeled image data are projected onto the text space and pseudo-labels are obtained, and then the data are expressed in the text space.

In many methods for cross-media retrieval, unlabeled data are not considered but become useful for improving the retrieval performance. Based on the above work, the unlabeled image data are assigned pseudo-labels. Now, \( T'_u \) and \( T_u \) are in the common space. \( S_u \) is the similarity of the unlabeled image data \( I_u \), namely \( T'_u \), and the center \( C_{T'_1} \), namely the labeled text data. Furthermore, \( S_{uji} \), the element of \( S_u \), is calculated as

\[
S_{uji} = e^{-\|T'_ui - C_{T'_1}\|^2_2}
\]

(7)

Where \( \gamma \) is the kernel coefficient and set to 8 in this paper. Then, we obtain a new similarity matrix \( S=[S_l; S_u] \).

Then, the unlabeled image data \( I_u \) can be replaced with the text mode \( T' \ u \). The most similar element \( C_{T'_1} \) in \( C_T \) is selected as the element \( T' \ u_i \).

After the above steps, we obtain pseudo-labels of the unlabeled data \( I_u \) and the corresponding data \( T' \ u \), which are expressed in the text mode. This provides an effective approach to bridging the semantic gap.

4.3. Optimization Objective Function

The PLSC algorithm proposed in this paper uses the training dataset \( G \) to learn the projection matrices \( V \in \mathbb{R}^{p \times p} \) and \( W \in \mathbb{R}^{q \times q} \) using the optimization objective function. \( V \) is the projection matrix for images, and \( W \) is the projection matrix for text. Image and text data are projected onto the same semantic space in which the distances between the projected data can be calculated. The optimization framework can be formally expressed as

\[
\min_{V, W} f(V, W) = C(V, W) + L(V, S) + R(V, W)
\]

(8)

Where \( f \) is used as the objective function; \( C(V, W) \) is used for the correlation analysis-related term that ensures pair-wise closeness and feature consistency in the shared projection space; \( L(V, S) \) is the linear regression term that captures clustering of semantically similar multi-modal data in the common latent space and ensures semantic consistency; and \( R(V, W) \) is the regularization term that controls projection matrices \( V \) and \( W \), which is turn helps to avoid overfitting the model to the training set data.

Furthermore, the optimization function can be expressed as

\[
\min_{V, W} f(V, W) = \lambda ||V^T - TW^T||_F^2 + (1 - \lambda)||V^T - S||_F^2 + \eta_1 ||V||_F^2 + \eta_2 ||W||_F^2
\]

(9)
Where \( \eta_1 \) and \( \eta_2 \) are nonnegative balanced parameters for the regularization term.

The solutions for \( V \) and \( W \) are obtained in unconstrained optimization. The function \( f(I, W) \) is a nonconvex function and only has a locally optimal solution. Otherwise, when fixing one of them, the solution of the other will be that of a convex problem, and then they can be calculated alternately using the gradient descent method.

The partial derivatives are calculated as

\[
\frac{\partial F_1}{\partial V} = VIT + 2[\eta_1 V - \lambda WT^T I - (1 - \lambda)ST^T I] \\
\frac{\partial F_1}{\partial W} = 2[\eta_2 W + \lambda (WT^T T - VT^T T)]
\]

(10) (11)

In general, the alternate iteration method is used to solve this problem. The alternate iteration method terminates on convergence. The following algorithm describes the alternate iteration process in detail:

**Algorithm 1 PLSC algorithm.**

**Input:**
- The matrix of image features \( I = [I_1, I_2]^T \in \mathbb{R}^{n \times p} \) and the matrix of text features \( T = [T_1, T_2]^T \in \mathbb{R}^{n \times q} \), the matrix of semantic features \( S = [S_1, S_2] \).

**Initialization:** \( \nu \in \mathbb{R}^{n \times q} \), \( w \in \mathbb{R}^q \), \( \eta_1 \), \( \eta_2 \), \( \epsilon \), \( \mu \), \( \epsilon \) is the convergence condition, \( \mu \) is the descent step size.

**Repeat**
- **Repeat**
  - Set \( fv_{1-1} = f(V^{(0)}, W^{(0)}) \);
  - Update \( V^{(t+1)} = V^{(t)} + \mu f(V^{(t)}, W^{(t)}) \);
  - Set \( V_{2-1} = f(V^{(t)}, W^{(t+1)}) \), \( \nu \leftarrow \nu + 1 \);
- **Until** \( fv_{1-1} = f(V^{(t)}, W^{(t+1)}) \) and \( fv_{2-1} = f(V^{(t)}, W^{(t+1)}) \) and the convergence or until maximal number of iterations is reached
- **Output:** Projection matrices \( V^{(t)}, W^{(t)} \)

In the above, \( V \) and \( W \) can be solved using the alternate iteration update approach. When the condition is matched, the solution process is terminated. Unlike other methods, when updating the \( V \) and \( W \), the semantic information of the unlabeled sample data is fully considered.

With the projection matrices \( V \) and \( W \), image and text data can be considered in the same space; thus, distances between different modality data points can be calculated. The most widely used measure of distance is the Euclidean distance [11]:

\[
dist(I, T) = \sqrt{\sum_{j=1}^{n}(i_j-t_j)^2}
\]

(12)

5. Validation Experiments

5.1. Datasets

In the following validation experiments, Wikipedia, Pascal Sentence, and INRIA-Websearch datasets were used to verify the performance of the PLSC method with those of conventional methods.

**Wikipedia:** It contained 2,866 text-image pairs in 10 categories. There are 2,173 samples for training and 693 for testing [25]. The image visual features were adopted from the 128-dimensional SIFT Bag-of-Visual-Words (BoW) [26], and the text features were adopted from the 10-dimensional latent Dirichlet allocation (LDA), which are described in [12, 27]. Based on the same feature data as in [12, 27], the performance of the PLSC algorithm was evaluated and compared with those of other algorithms. Another Wikipedia dataset has 4,096 dimensional convolution neural network (CNN) visual features and 100 dimensional LDA text features [19]. We refer to the former as wiki128 and the latter as wiki4096.
**Pascal Sentence:** There were 1,000 text-image pairs and 20 classes in this dataset [28]. Each class contained 50 pairs of data. In our method, 30 pairs of each class were selected as the training set, and the remainder comprised the test set. Thus, overall there were 600 training examples and 400 testing examples [29]. For the image data, CNN visual features were used. For the text features, each text for the 100 latent topics was computed using the LDA based on the BoW presentation of 300 tokens [26-27].

Semantic features were constructed according to the data class tags; thus, the dimensionalities of the Wikipedia and Pascal Sentence datasets were 10 and 20, respectively.

**INRIA-Websearch Dataset:** There were 71,743 image-text pairs in this dataset and 353 categories. The CNN visual features were 4,096 dimensional [11]. We selected the first top 100 categories to construct 14,698 pairs in the experiment.

### 5.2. Experimental Setup

The Euclidean distance was used as a measure of distance between the text and image data in the isomorphic space. The average precision (AP) and mean AP (mAP) measures were designed for evaluation of the performance of different retrieval methods [15]. The definition of AP was

\[
AP = \frac{\sum_{k=1}^{R} P(k) \text{rel}(k)}{\sum_{k=1}^{R} \text{rel}(k)}
\]

Where \( R \) is the number of querying results. If a \( k \)-th example is consistent with the querying item, \( \text{rel}(k)=1 \); otherwise, \( \text{rel}(k)=0 \). \( P(k) \) quantifies the accuracy of top \( k \) results. mAP is obtained by averaging over all AP.

### 5.3. Results of Analysis

#### 5.3.1. mAP

Firstly, in our validation experiment, the mAP performance of the PLSC algorithm is compared with four state-of-the-art methods, which are briefly described below.

JGRHML [16] retrieves among different media using a joint graph regularization. HSNN [18] is a method measuring heterogeneous similarity with nearest neighbors. CMCP [17] uses the positive and negative correlation at the same time. JRL [15] considers the semantic information and correlations simultaneously.

Table 1 lists the mAP scores for the PLSC method and other state-of-the-art methods. Compared with other methods, the PLSC algorithm performs better, especially on T2I retrieval.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>JGRHML</th>
<th>HSNN</th>
<th>CMCP</th>
<th>JRL</th>
<th>PLSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki128</td>
<td>T2I</td>
<td>11.79</td>
<td>30.92</td>
<td>32.52</td>
<td>32.1</td>
<td>32.3</td>
</tr>
<tr>
<td>Average</td>
<td>T2I</td>
<td>11.77</td>
<td>27.54</td>
<td>28.77</td>
<td>28.5</td>
<td>42.4</td>
</tr>
<tr>
<td>Wiki4096</td>
<td>T2I</td>
<td>34.75</td>
<td>44.72</td>
<td>44.76</td>
<td>47.39</td>
<td>31.44</td>
</tr>
<tr>
<td>Average</td>
<td>T2I</td>
<td>28.88</td>
<td>40.79</td>
<td>41.01</td>
<td>42.04</td>
<td>53.77</td>
</tr>
<tr>
<td>Pascal Sentence</td>
<td>T2I</td>
<td>31.8</td>
<td>42.76</td>
<td>42.89</td>
<td>44.72</td>
<td>42.61</td>
</tr>
<tr>
<td>Average</td>
<td>T2I</td>
<td>42.34</td>
<td>40.64</td>
<td>41.1</td>
<td>48.13</td>
<td>44.57</td>
</tr>
<tr>
<td>INRIA-Websearch</td>
<td>T2I</td>
<td>43.52</td>
<td>41.42</td>
<td>41.47</td>
<td>46.24</td>
<td>57.65</td>
</tr>
<tr>
<td>Average</td>
<td>T2I</td>
<td>42.93</td>
<td>41.03</td>
<td>41.29</td>
<td>47.19</td>
<td>51.11</td>
</tr>
<tr>
<td></td>
<td>I2T</td>
<td>12.87</td>
<td>35.03</td>
<td>34.56</td>
<td>50.7</td>
<td>49.17</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>17.95</td>
<td>36.54</td>
<td>36.38</td>
<td>54.43</td>
<td>57.09</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>15.41</td>
<td>35.79</td>
<td>35.47</td>
<td>52.57</td>
<td>53.13</td>
</tr>
</tbody>
</table>
5.3.2. Precision-Recall Curves

Now, we compared the performance of the PLSC algorithm with the performances of some of the most popular algorithms for cross-media retrieval, such as canonical correlation analysis (CCA) [2], semantic matching (SM) [25], semantic correlation matching (SCM) [25], three-view CCA (T-V CCA) [30], generalized multi-view marginal Fisher analysis (GMMFA) [31], generalized multi-view linear discriminant analysis (GMLDA) [31], and MDCR [19]. The MDCR method was published in 2016, and it achieved the best retrieval results in Wikipedia at the time. Among the methods above, CCA and T-V CCA belong to the unsupervised learning methods, while MDCR and other methods belong to the supervised learning methods. The algorithm proposed in this paper, PLSC, is semi-supervised.

Figure 5 show the performance of the PLSC algorithm compared with that of popular and classical algorithms. The figure shows the precision-recall curves on the Wikipedia and Pascal Sentence datasets. In particular, the precision-recall of T2I on the Wikipedia dataset is much better than that of other algorithms. The scores of the PLSC algorithm are superior to these algorithms; especially in the T2I task, the scores are obviously outstanding.

![Figure 5. Precision-recall curves](image)

5.3.3. Querying Results

Figure 6 presents some successful examples of our method on the Wikipedia dataset. The query images are selected from sport, biology, and warfare categories. The query image was used to rank the text articles in the Wikipedia dataset. The left-most image is the query images, and the screenshots of the retrieved texts are shown in the right four columns. The corresponding images are listed under the texts.
6. Conclusions

In this paper, an efficient method for cross-media retrieval was introduced and is based on pseudo-label learning and semantic consistency. In the proposed approach, unlabeled data are considered fully using the pseudo-labeling method. The unlabeled data are used to train the model, along with labeled data. Projection matrices are learned after iterative operation, and different modality data can be projected onto the same space for comparison. According to the results of our validation experiment on the Wikipedia and Pascal Sentence datasets, the proposed method is efficient and practical. The experimental results suggest that the PLSC method is better than the other algorithms. In future work, the method will be applied to the retrieval of information from different other modalities, such as audio and video, and is likely to gain widespread use.

Acknowledgements

This work was supported by the Key Research and Development Foundation of Shandong Province (2016GGX101035), Shandong Housing and Urban Rural Construction Science Planning Project (2017-R1-001), and Shandong Soft Science Research Planning Project (2017RKB01077).

References
