Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 14, no. 4, April 2018, pp. 795-804
DOI: 10.23940/ijpe.18.04.p21.795804

Two-Stage Semantic Matching for Cross-Media Retrieval

Gongwen Xu[a,*], Lina Xu[a], Meijia Zhang[a], and Xiaomei Li[b]

[a]*School of Information Science and Engineering, Shandong Normal University, Jinan, 250358, China*
[b]*The Second Hospital of Shandong University, Jinan, 250033, China*

Abstract

With the development of information technology, there exists a large amount of multi-media data in our lives; the data is heterogeneous with low-level features while consistent with semantic information. Traditional mono-media retrieval can't cross the heterogeneous gap of multi-media data, and cross-media retrieval is arousing many researchers' interests. In this paper, we propose a two-stage semantic matching for cross-media retrieval based on support vector machines (called TSMCR). Our approach uses a combination of testing images' predictive labels and testing texts' predictive labels as the next training labels. It makes full use of semantic information of both training samples and testing samples, and the experimental results on four state-of-the-art datasets show that the TSMCR algorithm is effective.

Keywords: cross-media retrieval; two-stage; semantic matching; support vector machine

(Submitted on December 29, 2017; Revised on February 2, 2018; Accepted on March 20, 2018)

1. Introduction

With the development of science and technology, the network information has been explosively growing, and the new media forms emerge in an endless stream. In general, multimodal data is similar to expression but in the form of different modalities, different sources, and different backgrounds. Multimodal data usually have the characteristics of cross-modal, cross-source, and cross-space [5]. Cross-modal refers to the common expression of a concept or event by various types of organization of data, such as texts and images. For example, the symbiotic texts and images that exist on the same page carry similar semantic information in cross-media retrieval. Cross-source refers to the source of data is different, but the meaning of the expression is similar. For example, news reports that express a same event can come from two different portals from Sina and Sohu. Cross-space refers to the leap between the information space and the physical world. The picture of a tiger can be described with two different feature sets to form two isomeric feature spaces [11].

There are a motley variety of approaches to achieve information, not only from the original newspapers, radio, television and other traditional media, but also from WeChat, micro-blog, blogs, news websites and other Internet media. In addition, the entity of information content has also changed a lot; it has gradually changed from single media to cross-media. When users browse the Internet web pages, they often notice that a news text is equipped with corresponding news photos or videos. That is to say, a variety of media forms complement each other to describe a common event or a goal. However, traditional media processing methods are no longer suitable for dealing with such complex information entities. New methods and new means need to be studied to cater to people's information retrieval needs. As a new way of media retrieval, cross-media retrieval [22] has attracted more and more researchers' attention.

Currently, the correlation modeling among multi-modal media data still faces some challenges. The research objects of cross-media retrieval are various types of organization of multimedia data. In low-level features, the data is heterogeneous. But in semantic, the date is related. The multimedia data have complex structures, so there are most changeful organizational forms and association structures between distinguished types of data. Multimedia data such as images and videos are semi-

* Corresponding author.
E-mail address: xugongwen@163.com.

structured or unstructured. It is difficult for computer to understand high-level semantics accurately based on the underlying visual or auditory features.

In this paper, we propose a two-stage semantic matching for cross-media retrieval. This method is supported by support vector machines (called TSMCR), which integrates Semi-supervised [12,23], support vector machines (SVM) [2,16] and semantic matching [17,28]. Firstly, we use the training samples of texts and images with their labels to train two independent support vector machines. Then, we input the testing samples of texts and images to the two trained SVMs and obtain their predict labels and use the jointly representation of them as the labels of testing sample. Secondly, we use the testing samples of texts and images with their labels to train two new SVMs, and regard the outputs of the SVMs as the common semantic space of texts and images to conduct cross-media retrieval. What's more, our method greatly uses semantic information, and compared with other methods, it's effective.

The main works of our article are as follows: (1) We propose an efficient and concise cross-media retrieval algorithm called TSMCR, which utilizes a combination of testing images' predictive labels and testing texts' predictive labels as the next training labels, and greatly improves the accuracy of cross-media retrieval. (2) Our two-stage method: once the first training stage has trained the model, when a testing sample is arrived, it is only necessary to retrain the testing samples. The complexity of time is greatly reduced. (3) We carried out the experiments of TSMCR compared with some well-known methods on four frequently used datasets: Wikipedia dataset, Pascal-CNN dataset, NUS-WIDE dataset, Wikipedia-CNN dataset. The experimental results verify its effectiveness.

2. Related Work

Recently, cross-media retrieval has universally appeared in pattern recognition, image retrieval, and biological information and so on. Nowadays, a series of cross-media learning methods have been put forward by researchers at home and abroad to explore potential relationships between multimodal data, in order to increase the accuracy rate of cross media learning. According to whether training data have class labels, we divide existing cross-media methods into supervised, unsupervised and semi-supervised learning.

Supervised learning is the most classic method of cross-media retrieval and is widely used. It can effectively use the label information of training samples, thus greatly improving the accuracy of cross-media retrieval. However, most of the images and texts in real life do not have labeled information. The cost of manually tagging the semantics is too high. Supervised learning method cannot deal with these problems effectively. Rasiwasia et al. [13] proposed Semantic Correlation Matching (SCM), which combines Canonical Correlation Analysis (CCA) [20,3,18] and Semantic Matching(SM). It projects texts and images to a semantic space learned using CCA representation from their own natural spaces. Sharma et al. [14] propose Generalized Multiview Linear Discriminant Analysis (GMLDA), which is the extension of Linear Discriminant Analysis (LDA) [15] and Generalized Multiview Marginal Fisher Analysis (GMMFA), which is the development of Marginal Fisher Analysis (MFA) [21].

Unsupervised learning uses unlabeled data and relatively has low accuracy. Hardoon et al. [9] presented a general method of using Kernel Canonical Correlation Analysis (KCCA) [27] to learn semantic description between the web pages and related text. Besides KCCA, Partial Least Squares (PLS) [10,30] and Bilinear Model (BLM) [25,8], there are some other unsupervised methods for the cross-media problem.

In traditional machine learning area, there are two commonly used learning algorithms: supervised learning and unsupervised learning. Neither of them is suitable for dealing with the situation where there are few labeled data and a great number of unlabeled data. Semi-supervised learning algorithms, which consider both labeled and unlabeled data, can improve learning effectiveness significantly. Liang Zhang et al. [29] proposed Generalized Semi-supervised Structured Subspace Learning (GSS-SL), which firstly adds some relevant labels for unlabeled data, then takes advantage of the semantic information that the class labels reflect. Xiaohua Zhai et al. [26] proposed a semi-supervised learning algorithm called joint representation learning (JRL), which can consider not only the paired samples but also the semantic information.

3. A Two-Stage Semantic Matching for Cross-Media Retrieval

3.1. Support Vector Machines

Nowadays, the support vector machines method has made breakthroughs in their theoretical research and algorithm implementation, and are essentially different from the traditional methods. The performance of traditional statistical methods

is guaranteed theoretically only when the number of samples tends to infinity. Support vector machine (SVM) is a small method of sample learning, which can achieve good generalization ability when training samples are small, especially for dealing with nonlinear classification and nonlinear regression problems. In addition, the amount of calculation of support vector method is almost independent of the dimension of the sample vector, which avoids the "dimension disaster" in some sense. For the classification of a small number of samples, the support vector machine has the advantages of less adjustment parameters and faster operation speed.

Given a set of training samples $\{(x_i, y_i)\}_{i=1}^{l}, i = 1,...,l.$ where $x_i \in R^m$ is a sample point, and its components are called sample features, $y_i \in \{1,-1\}$ is a sample's class label, m is the feature dimension of the samples, l is the number of the training ones.

If the training samples set is linearly separable, for any $i = 1,...,l$, there are $w \in R^m, b \in R$ in SVM that

$$\begin{aligned} &w^{\mathrm{T}} x_i + b > 0, y_i = 1, \\ &w^{\mathrm{T}} x_i + b < 0, y_i = -1. \end{aligned} \tag{1}$$

where $w^{\mathrm{T}} x + b = 0$ is the classification hyperplane.

It is not hard to see that the bigger the interval between the two supporting planes, the better the separation effect of the corresponding classification hyperplane. Based on the above idea, we can get the following linear separable SVM optimization model:

$$\begin{aligned} &\min_{w,b} \frac{1}{2}\|w\|^2 \\ &s.t. y_i(w^{\mathrm{T}} x_i + b) \geq 1, i = 1,2,...,l,\ w \in R^m, b \in R. \end{aligned} \tag{2}$$

It is a convex quadratic optimization problem, if (w^*, b^*) is the optimum solution of above problem, we can get $(w^*)x + b^* = 0$ and the classification decision function is obtained:

$$f(x) = \mathrm{sgn}((w^*)^{\mathrm{T}} x + b^*) \tag{3}$$

If the training samples set are nonlinearly separable, introduce nonnegative slack variables $\xi = (\xi_1, \xi_2, ..., \xi_l)$ The component of ξ corresponds to the degree of error of each sample point. By changing the constraint condition to $y_i(w^{\mathrm{T}} x_i + b) \geq 1 - \xi_i, i = 1,2,...,l$ and limiting the error degree of each sample point as far as possible, we get the following optimization problem:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \xi_i \\ &s.t. y_i(w^{\mathrm{T}} x_i + b) \geq 1 - \xi_i, i = 1,2,...,l, \\ &\boldsymbol{\xi} = (\xi_1, \xi_2, ..., \xi_l)^{\mathrm{T}} \geq \mathbf{0}_l, \qquad w \in R^m, b \in R \end{aligned} \tag{4}$$

where *C* is a penalty parameter for balancing two items in the objective function. Similarly, if (w^*, b^*) is the optimum solution of above problem, we can construct the classification hyperplane $(w^*)^{\mathrm{T}} \phi(x) + b^* = 0$ in high dimensional feature space. The classification decision function is obtained:

$$f(x) = \mathrm{sgn}((w^*)^{\mathrm{T}} \phi(x) + b^*) \tag{5}$$

Support vector machine can be divided into two algorithms: classification algorithms and regression algorithms. Traditional support vector machines also expose some shortcomings in practical applications such as large amount of computation, slow speed, strong parameter selection and experience, and cannot well solve multiple classification problems.

The speed problem restricts the application of SVM to a large extent, and becomes the bottleneck of the support vector machine approach into the large-scale practical stage. The main reason for the slow training of support vector machines is that a lot of quadratic programming calculations have been put forward during the training process. These calculations have high complexity and long time consuming. The main reason for the slow classification is that there is a lot of support vectors involved in the calculation in the classification process. In order to solve the problems of support vector machine, many experts and scholars have made a lot of improvements and creative works on support vector machine classification algorithms and regression algorithms through years of efforts. ν-Support Vector Regression (ν-SVR) is a more effective support vector regression machine method.

3.2. ε-Support Vector Regression (ε-SVR) and ν-Support Vector Regression (ν-SVR)

Let $\{(x_1, z_1), ..., (x_l, z_l)\}$ is a set of data points, where $x_i \in R^n$ as the input and $z_i \in R^l$ as the target output, the general formula of support vector regression [24, 1] is:

$$\begin{aligned} \min_{w,b,\xi,\xi^*} \quad & \frac{1}{2} w^{\mathrm{T}} w + C\sum_{i=1}^{l} \xi_i + C\sum_{i=1}^{l} \xi_i^* \\ & w^{\mathrm{T}} \phi(x_i) + b - z_i \le \varepsilon + \xi_i, \\ & z_i - w^{\mathrm{T}} \phi(x_i) - b \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0, i = 1, ..., l. \end{aligned} \tag{6}$$

The dual is:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\alpha - \alpha^*)^{\mathrm{T}} Q(\alpha - \alpha^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} z_i(\alpha_i - \alpha_i^*) \\ & \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, 0 \le \alpha_i, \alpha_i^* \le C, i = 1, ..., l, \end{aligned} \tag{7}$$

where $Q_{ij} = K(x_i, x_j) \equiv \phi(x_i)^{\mathrm{T}} \phi(x_j)$

The approximate function is:

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*) K(x_i, x) + b \tag{8}$$

In this paper, we use a more advanced support vector regression algorithm: ν-Support Vector Regression (ν-SVR).

The primal form is:

$$\begin{aligned} \min_{w,b,\xi,\xi^*,\varepsilon} \quad & \frac{1}{2} w^{\mathrm{T}} w + C(v\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)) \\ & (w^{\mathrm{T}} \phi(x_i) + b) - z_i \le \varepsilon + \xi_i, \\ & z_i - (w^{\mathrm{T}} \phi(x_i) - b) \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0, i = 1, ..., l, \varepsilon \ge 0 \end{aligned} \tag{9}$$

and the dual is

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\alpha - \alpha^*)^{\mathrm{T}} Q(\alpha - \alpha^*) + z^T(\alpha_i - \alpha_i^*) \\ & e^T(\alpha - \alpha^*) = 0, e^T(\alpha + \alpha^*) \le Cv \\ & 0 \le \alpha_i, \alpha_i^* \le C/l, i = 1, ..., l \end{aligned} \tag{10}$$

In this paper, we consider $C \leftarrow C/l$; we can solve the dual problem, as shown below:

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \quad & \frac{1}{2}(\alpha - \alpha^*)^{\mathrm{T}} Q(\alpha - \alpha^*) + z^T(\alpha_i - \alpha_i^*) \\ & e^T(\alpha - \alpha^*) = 0, e^T(\alpha + \alpha^*) \le Clv \\ & 0 \le \alpha_i, \alpha_i^* \le C, i = 1, ..., l \end{aligned} \tag{11}$$

The decision function can be gained:

$$\sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(x_i, x) + b \tag{12}$$

3.3. A Two-Stage Cross-media Semantic Matching

Our algorithm is divided into 2 stages. first, train two independent support vector machines (SVM) with training texts, training images and their corresponding labels. Then, we input the testing samples of texts and images to the two trained SVMs and obtain their predict labels and use the jointly representation of them as the labels of testing sample. Second, train two new SVMs with testing texts, testing images and their corresponding labels, and regard the outputs of the SVMs as the common semantic space of texts and images to conduct cross-media retrieval. Our method gets the highest MAP scores for all the cross-media retrieval tasks. The reason is that the integration of the predictive labels for the testing images and testing texts can promote each other and thus effectively improve the retrieval accuracy. It makes full use of semantic information of both training samples and testing samples, it's simple but effective. The detailed process of the algorithm is shown as follows: A Two-stage Cross-media Semantic Matching (TSMCR)

Algorithm 1 A Two-stage Cross-media Semantic Matching (TSMCR)

Input: training samples of images $I \in R^{p\times n}$ and texts $T \in R^{q\times n}$, their labels $label \in R^{k\times n}$ testing samples of images $Iy \in R^{p\times te}$ and texts $Ty \in R^{q\times te}$.

Output: S_I and S_T.

1. Use *I* and *label* to train SVM of images.
   Input *Iy* to trained SVM and obtain its predictive labels $label_I$.
2. Use *T* and *label* to train SVM of texts.
   Input *Ty* to trained SVM and obtain its predictive labels $label_T$.
3. $label_{te} = label_T - label_I$
4. Use *Iy* and $label_{te}$ to retrain the SVM of images.
   Input *Iy* to trained SVM and obtain its predictive labels S_I.
5. Use *Ty* and $label_{te}$ to retrain the SVM of texts.
   Input *Ty* to trained SVM and obtain its predictive labels S_T.

4. Experiments

We brought out a series of experiments of TSMCR and the compared methods to testify the availability of the TSMCR for the tasks of cross-media retrieval, including text query image task (T2I) and image query text task (I2T). We test all the methods on four datasets: Wikipedia dataset [13], NUS-WIDE dataset [4], Wikipedia dataset with CNN features [6], and Pascal dataset with CNN features [7].

4.1. Experimental Datasets

Wikipedia dataset: This dataset contains 2,173 training image-text pairs and 693 testing image-text image-text pairs belonging to 10 categories. In this dataset, Latent Dirichlet Allocation (LDA) method is used to extracted low-level features of texts while Scale Invariant Feature Transformation (SIFT) is used to extracted 128-dimensional features for images [15].

NUS-WIDE dataset: This dataset consists of 40834/27159 (training/testing) image-tag pairs. We conduct our experiments in 10 categories with maximum number of samples from all the 81 semantic categories. In this dataset, the low-level features of texts are 1000-dimensional tag feature vectors while images are 500-dimensional SIFT features.

Wikipedia-CNN dataset: Similar to the first dataset, it has 2866 image-text pairs, and they are divided into 10 semantic classes. CNN extracts 4096 dimensional visual features and LDA extracts 100 dimensional textual features [15].

Pascal-CNN dataset: It has 600/400 (training/testing) image-text pairs and they are divided into 20 semantic classes. CNN extracts 4096 dimensional visual features and LDA extracts 100 dimensional textual features too [14].

4.2. Compared Methods and Evaluation Metric

Our TSMCR algorithm is compared with six different well-known methods.

CCA (Canonical Correlation Analysis): CCA with the help of the concept of principal component analysis (PCA). It uses each set of variables as a whole research object rather than internal analysis of each group of variables. It is the most classic cross-media retrieval method.

SCM (Semantic correlation matching): It is a simple combination of the CCA algorithm and the SM algorithm. First, SCM uses CCA algorithm to project the texts and images to the related subspace. Then, the related subspace is mapped to the same semantic subspace with the SM algorithm, and the subsequent processing is the same as the SM algorithm [13].

CCA-3V (Three-view CCA) [7]: By introducing the third kind of high-level semantic information, texts and images with the same semantics have good aggregation in the isomorphic space.

GMLDA (Generalized Multiview Linear Discriminant Analysis) which is the extension of LDA (Linear Discriminant Analysis) and GMMFA (Generalized Multiview Marginal Fisher Analysis) which is the development of MFA (Marginal Fisher Analysis) [15].

MDCR (Modality-dependent cross-media retrieval) [19]: It belongs to Task-specific Cross-modal Retrieval (TSCR). In other words, it uses different mapping mechanisms for different cross-media retrieval tasks.

There are multiple assessing criteria to measure the efficiency of cross-media retrieval. In this paper, the Mean Average Precision (MAP) and Precision-Recall (PR) are used to assess the efficiency of our method and other methods.

4.3. Experimental Results

4.3.1. Results on the Wikipedia dataset and NUS-WIDE dataset

Table 1. MAP scores on Wikipedia dataset and NUS-WIDE dataset

methods	Wikipedia			NUS-WIDE		
	I2T	T2I	Avg	I2T	T2I	Avg
CCA	0.241	0.184	0.212	0.287	0.284	0.286
SCM	0.277	0.226	0.252	0.351	0.347	0.349
CCA-3V	0.275	0.224	0.250	0.351	0.326	0.339
GMMFA	0.275	0.214	0.245	0.298	0.294	0.296
GMLDA	0.275	0.210	0.243	0.324	0.308	0.316
MDCR	0.271	0.225	0.248	0.287	0.242	0.264
TSMCR	0.333	0.241	0.287	0.435	0.388	0.412

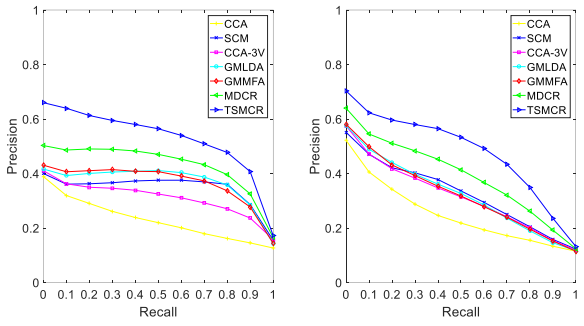
The experimental results of I2T, T2I and their average retrieval scores on Wikipedia dataset and NUS-WIDE dataset are expressed in Table 1. We can see clearly that TSMCR achieves superior performance. The reason why our method has the best results is that it utilizes a combination of testing images' predictive labels and testing texts' predictive labels as the next training labels, which makes the semantic information more accurate. Because CCA is an unsupervised learning method, it achieves the worst performance on Wikipedia dataset. GMLDA and GMMFA focus on the semantic information, so they get better results than CCA on Wikipedia dataset [15, 21]. The MAP value of our method is 6%, 1.6%, 3.9% higher than that of MDCR for text query image task, image query text task and average scores on Wikipedia respectively.

On NUS-WIDE, the MAP value of our method is far higher compared with the previous and well-known methods. It is about 14.8%, 14.6%, 14.8% higher than that of MDCR for text query image task [19], image query text task and average scores. The reason is that the integration of the predictive labels for the testing images and testing texts can promote each other. Our approach considers both labeled and unlabeled data, and can improve learning effectiveness significantly.

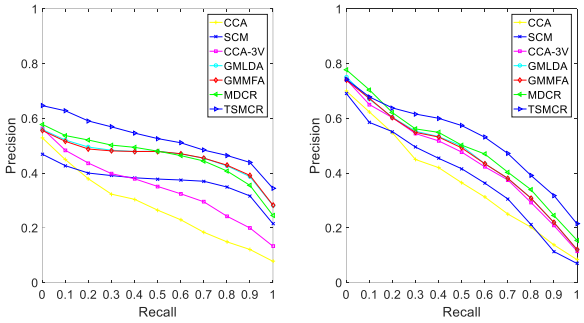
In addition, we can see that great majority methods get higher MAP value on the NUS-WIDE dataset than that of Wikipedia dataset. The reason is that the feature extraction method of the former is better than the latter, and it can effectively enhance the MAP value of cross-media retrieval. NUS-WIDE dataset is one of large-scale datasets that is used to measure the efficiency of cross-media retrieval; many other works also used this dataset to assess the efficiency of the methods.

4.3.2. Results on the Wikipedia-CNN dataset and pascal-CNN dataset

The PR curves for all of retrieval tasks on Wikipedia-CNN dataset and Pascal-CNN dataset can be found in Figure 1. TSMCR method has the best results for I2T and T2I task on Wikipedia-CNN dataset. It also gets better results on Pascal-CNN dataset.

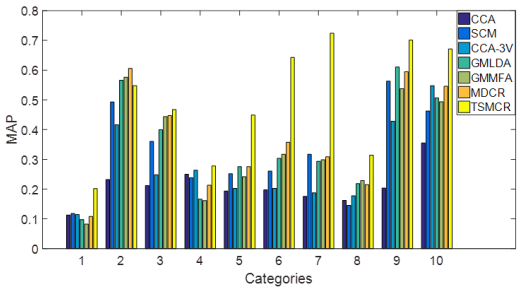


(a)

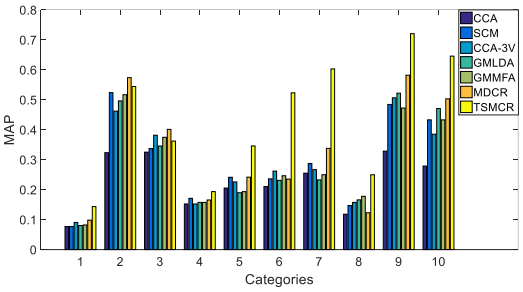


(b)

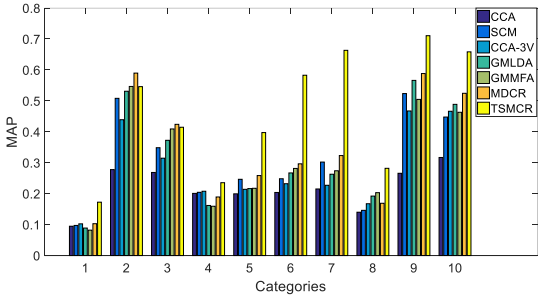
Figure 1. PR curves for image query (left) and text query (right) on (a) Wikipedia-CNN dataset (b) and Pascal-CNN dataset



(a)

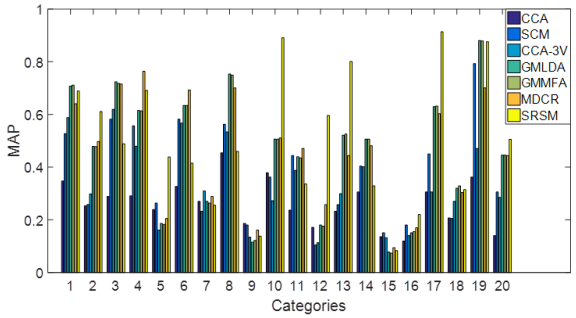


(b)

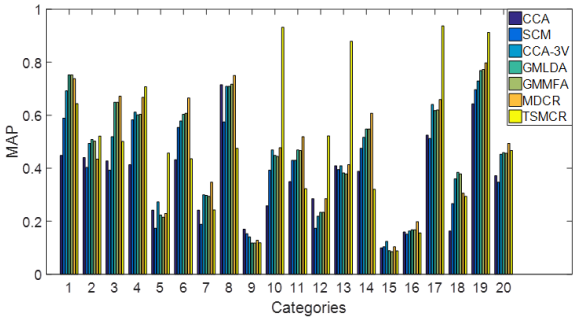


(c)

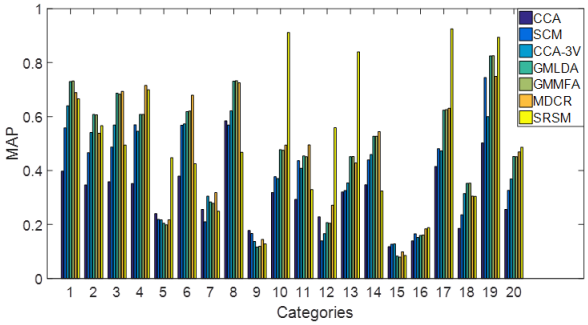
Figure 2. MAP scores per class of TSMCR and compared methods for image query (a), text query (b) and average performance (c) on Wikipedia-CNN dataset



(a)



(b)



(c)

Figure 3. MAP scores per class of TSMCR and compared methods for image query (a), text query (b) and average performance (c) on Pascal-CNN dataset

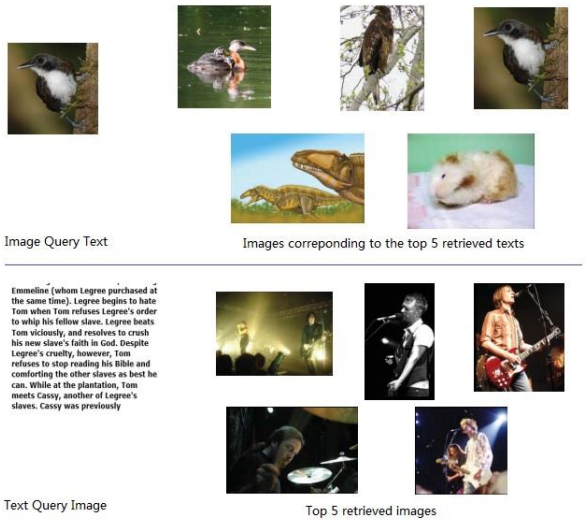


Figure 4. The retrieval results on Wikipedia-CNN dataset.

Furthermore, the MAP for I2T, T2I and the average retrieving efficiency for every class on Wikipedia-CNN dataset is shown in Figure 2 and Pascal-CNN dataset is shown in Figure 3. It can be found that our method has better performance

compared with other methods in nearly the whole classes. In particular, it has obvious advantages over the fifth to seventh classes. Our method also achieves better results on Pascal-CNN dataset; it gains obvious advantages in the tenth, twelfth, thirteenth, and seventeenth class.

It shows the results of image retrieving text, and the top five images corresponding to texts are shown in the top of Figure 4. For convenience, we use the images corresponding to the texts to represent the texts. The samples are selected from "biology" category. For the example of text query image, the query text is about singer from "music". All of retrieved images by TSMCR come from "music" category. It also shows that TSMCR has good results in all the cross-media retrieval tasks.

5. **Conclusions**

In this paper, we proposes a two-stage semantic matching algorithm called TSMCR, which utilizes a combination of testing images' predictive labels and testing texts' predictive labels as the next training labels, and greatly improves the accuracy of cross-media retrieval. It cleverly uses label information of semantic information of both training samples and testing samples. We measure the performance of our method on Wikipedia dataset, NUS-WIDE dataset, Wikipedia dataset with CNN features and Pascal dataset with CNN features. The experimental results validate its effectiveness compared with some well-known algorithms on the Mean Average Precision.

Acknowledgements

This work was partly financially by the Key Research and Development Foundation of Shandong Province (2016GGX101035), the Development Projects of Science and Technology of Jinan (201602151), Shandong Housing and Urban Rural Construction Science Planning Project (2017-R1-001), and Shandong Soft Science Research Planning Project (2017RKB01077).

References

1. A. Abdiansah, and R. Wardoyo, "Time complexity analysis of support vector machines (svm) in libsvm," *International Journal of Computer Applications*, vol. 128, no.3, pp. 975-8887, 2015
2. M. M. Adankon, and M. Cheriet, "Support Vector Machine", *Computer Science*, vol. 1, no.4, pp.1-28, 2002
3. G. Andrew, R. Arora, and J. Bilmes, "Deep canonical correlation analysis," *International Conference on International Conference on Machine Learning. JMLR.org*, pp. III-1247, 2013
4. T. S. Chua, J. Tang, and R. Hong, "NUS-WIDE: a real-world web image database from National University of Singapore," *ACM International Conference on Image and Video Retrieval. ACM*, pp. 48, 2009
5. P. J. Costa, E. Coviello, and G. Doyle, "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 36, no. 3, pp.521-35, 2014
6. J. Deng, L. Du, and Y. D. Shen, "Heterogeneous Metric Learning for Cross-Modal Multimedia Retrieval," *International Conference on Web Information Systems Engineering, Springer, Berlin, Heidelberg.* Vol. 8180, pp. 43-56, 2013
7. Y. Gong, Q. Ke, and M. Isard, "A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210-233, 2014
8. A. Halimi, Y. Altmann, and N. Dobigeon, "Nonlinear unmixing of hyperspectral images using a generalized bilinear model," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 49, no. 11, pp. 4153-4162, 2014
9. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods," *Neural Computation*, vol. 16, no. 12, pp. 2639-2664, 2014
10. J. He, B. Ma, and S. Wang, "Cross-modal Retrieval by Real Label Partial Least Squares," *ACM on Multimedia Conference. ACM*, pp. 227-231, 2016
11. C. Kang, S. Xiang, and S. Liao, "Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp.370-381, 2015
12. Y. Peng, X. Zhai, and Y. Zhao, "Semi-Supervised Cross-Media Feature Learning With Unified Patch Graph Regularization," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 26, no.3, pp.583-596, 2016
13. N. Rasiwasia, J. C. Pereira, and E. Coviello, "A new approach to cross-modal multimedia retrieval," *International Conference on Multimedia. ACM*, pp.251-260, 2010
14. A. Sharma, "Generalized Multiview Analysis: A discriminative latent space," *IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society*, pp. 2160-2167, 2012
15. A. Sharma, K. K. Paliwal, "A deterministic approach to regularized linear discriminant analysis," *Neurocomputing*, vol. 151, no. 1, pp. 207-214, 2015
16. Y. Verma, and C. V. Jawahar, "A Support Vector Approach for Cross-Modal Search of Images and Texts," *Computer Vision & Image Understanding*, pp.154, 2016.
17. D. Wang, X. Gao, and X. Wang, "Semantic topic multimodal hashing for cross-media retrieval," *International Conference on Artificial Intelligence. AAAI Press*,pp.3890-3896, 2015

18. W. Wang, R. Arora, and K. Livescu, "Unsupervised learning of acoustic features via deep canonical correlation analysis," *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE*, pp. 4590-4594, 2015
19. Y. Wei, Y. Zhao, and Z. Zhu, "Modality-Dependent Cross-Media Retrieval," *ACM Transactions on Intelligent Systems & Technology*, vol. 7, no.4, pp. 5, 2016
20. F. Wu, H. Zhang, and Y. Zhuang, "Learning Semantic Correlations for Cross-Media Retrieval," *IEEE International Conference on Image Processing. IEEE*, pp.1465-1468, 2007
21. D. Xu, S. Yan, and D. Tao, "Marginal Fisher analysis and its variants for human gait recognition and content- based image retrieval," *IEEE Transactions on Image Processing*, vo. 16, no. 11, pp. 2811-21, 2007
22. X. Xu, Y. Yang, A. Shimada, R. I. Taniguchi, and L. He, "Semi-supervised Coupled Dictionary Learning for Cross-modal Retrieval in Internet Images and Texts," *ACM International Conference on Multimedia*, *ACM* , pp.847-850, 2015
23. Z. Xue, G. Li, and W. Zhang, "Topic detection in cross-media: a semi-supervised co-clustering approach," *International Journal of Multimedia Information Retrieval*, vol. 3, no. 3, pp.193-205, 2014
24. K. Yanai, "Tools on Support Vector Machines : SVMLight, LIBSVM, SHOGUN," *Journal of the Institute of Image Information & Television Engineers*, vol. 63, pp. 1778-1781, 2009
25. Z. Yu, J. Yu, J. Fan, D. Tao, "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering," *IEEE International Conference on Computer Vision* , I*EEE Computer Society*, pp. 1839-1848, 2017
26. X. Zhai, Y. Peng, and J. Xiao, "Learning Cross-Media Joint Representation With Sparse and Semisupervised Regularization," *IEEE Transactions on Circuits & Systems for Video Technology*, vol. 24, no.6, pp. 965-978, 2014
27. H. Zhang, and L. Chen, "Learning optimal data representation for cross-media retrieval," *IEEE International Conference on Image Processing. IEEE*, pp. 1925-1928, 2013
28. H. Zhang, X. Gao, and P. Wu, "A cross-media distance metric learning framework based on multi-view correlation mining and matching," *World Wide Web-internet & Web Information Systems*, vol. 19, no.2, pp.181-197, 2016
29. L. Zhang, B. Ma, and G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, no. 99, pp.1-1. 2017
30. Y. Zhou, X. Cui, Q. Hu, and Y. Jia, "Improved multi-kernel SVM for multi-modal and imbalanced dialogue act classification," *International Joint Conference on Neural Networks, IEEE*, pp. 1-8, 2015

Gongwen Xu is a Ph. D. student from the School of Information Science and Engineering, Shandong Normal University. His research interest is cross-media retrieval.

Lina Xu is a Ph. D. student from the School of Information Science and Engineering, Shandong Normal University. Her research interest is statistical sparse learning.

Meijia Zhang is a Ph. D. student from the School of Information Science and Engineering, Shandong Normal University. Her research interest is machine learning.

Xiaomei Li received her Ph. D. degree in 2014 from Shandong University. She is a member of Cancer Center of the Second Hospital, Shandong University. Her research interest is medical image processing.