CROSS-MODALITY CORRELATION PROPAGATION FOR CROSS-MEDIA RETRIEVAL

Xiaohua Zhai, Yuxin Peng , and Jianguo Xiao*

Institute of Computer Science and Technology, Peking University, Beijing 100871, China
{zhaixiaohua, pengyuxin, xjg}@icst.pku.edu.cn

ABSTRACT

We consider the problem of cross-media retrieval, where the query and the retrieved results can be of different modalities. In this paper, we propose a novel cross-modality correlation propagation approach to simultaneously deal with positive correlation and negative correlation between media objects of different modalities, while existing works focus solely on the positive correlation. Negative correlation is very important because it provides the effective exclusive information. The correlation is modeled as must-link constraints and cannot-link constraints respectively. Furthermore, our approach is able to propagate the correlation between heterogeneous modalities. Experiments on the wikipedia dataset show the effectiveness of our cross-modality correlation propagation approach, compared with state-of-the-art methods.

Index Terms— cross-modality, correlation propagation, cross-media retrieval

1. INTRODUCTION

In recent years, there has been rapid growth of multimedia content on the web. Many real-world applications involve multiple-modality data. Although many techniques have been proposed to leverage text associated with images, they typically assume the texts contain mostly words or short image description to describe visible objects[1, 2, 3]. However, they didn't well explore the textual information.

Multiple-modality data such as texts, images and videos always co-exist in a multimedia document to describe the same semantic concept, such as the E-business web-pages, newspaper articles and multimedia cyclopedias. For instance, we can quickly draw a vivid but incomplete imagination about a concept through an image. In contrast, texts could accurately describe the details of the concept, but it is not intuitive enough. Consequently, jointly modeling and retrieving across the multi-modality media contents becomes increasingly important. By fully exploiting the rich information of the media objects of different modalities, we could understand the content of multimedia more accurately.

To exploit the correlation between different modalities, Bredin and Chollet [4] applied canonical correlation analysis (CCA) to the task of audiovisual based talking-face biometric verification. Li et al [5] introduced a cross-modal factor analysis (CFA) approach to evaluate the association between two modalities, and demonstrated superior performance than CCA. Zhuang et al. [6, 7, 8] explored the co-occurrence information in different modalities, that is, if a media object is shared by two multimedia documents, then they are of the same semantic. Rasiwasia et al. [9] proposed two kinds of supervised learning methods for cross-media retrieval. The first is to learn the subspace that maximizes the correlation between images and texts by canonical correlation analysis (CCA). The other is to represent the media objects by high level semantic vectors. Each dimension of the semantic vectors corresponding to the probabilities of the media object belonging to a semantic category. More recently, [10] proposed a new heterogeneous similarity measure by computing the probability for two media objects belonging to the same semantic category.

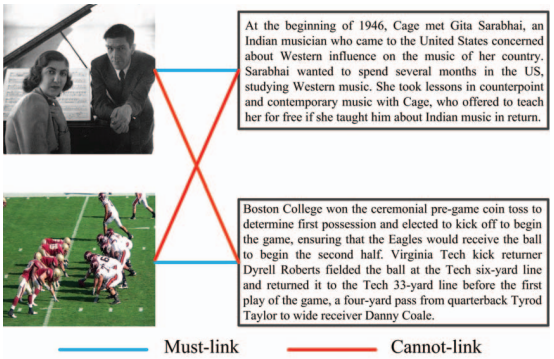


Fig. 1: Illustration of the cross-modality correlation, both images and texts come from Wikipedia article. The upper part describes the 2008 ACC Championship Game (http://en.wikipedia.org/wiki/2008_ACC_Championship_Game) and the lower part describes the Sonatas and Interludes (http://en.wikipedia.org/wiki/Sonatas_and_Interludes). The must-link and cannot-link constraints derived from the category label of articles.

Totally, the existing methods only focus on the positive correlation between media objects of the same category, while the negative correlation between media objects of different categories is omitted. In our opinion, both kinds of correlation are very important because positive correlation provides the concurrence information and negative correlation reflects the exclusive information. Figure 1 gives an example to depict such constraints. In this example, the images and texts come from "music" category and "sport" category respectively. A section of text description about the "sport" may have strong positive correlation with the image of players on the playground. And it also has negative correlation with the image about two people sit before a piano. To address above problems, in this paper we propose a cross-modality correlation propagation (CMCP) algorithm which jointly exploit the positive and negative correlation for cross-media retrieval. The correlation is modeled through must-link constraints and cannot-link constraints respectively. We obtain both kinds of pairwise constraints from the category label of the media objects. Since traditional constraints propagation techniques [11, 12, 13] have been designed basically for single modal-

*Corresponding author.

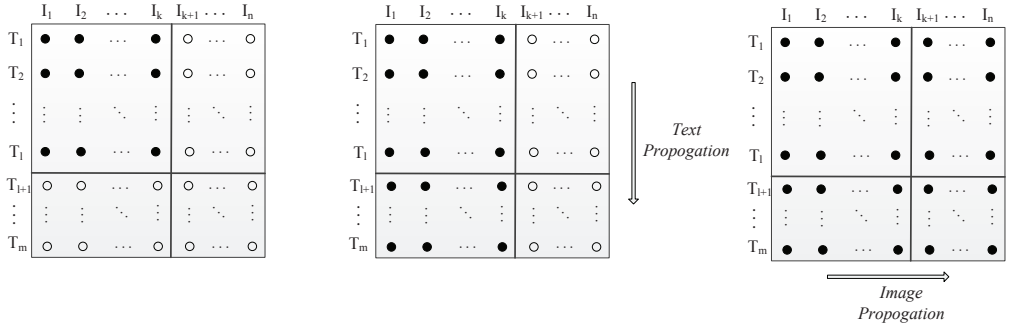


Fig. 2: Illustration of the semi-supervised semantic propagation.

ity data and cannot be readily applied for cross-media retrieval, our proposed CMCP provides an efficient solution to the cross-media retrieval problem by dividing the cross-modality correlation propagation problem into two sets of independent label propagation subproblems. Each set of subproblems propagate the initial correlation along corresponding media modality. Through combining the propagation along each modality, the semantic correlation can be propagated throughout the whole dataset. Then the achieved correlation on the target objects can naturally meet the requirement of cross-media retrieval problem, in which the retrieval results are of the same semantic and can be of different media type from the query.

In summary, the proposed approach has the following advantages: Firstly, it is able to propagate the correlation between any combination of heterogeneous data. Secondly, it can simultaneously deal with positive correlation and negative correlation between media objects of different modalities, which is omitted by existing works. Experiments on the wikipedia dataset show the effectiveness of our CMCP algorithm compared with state-of-the-art methods.

The rest of this paper will be organized as follows. In section 2, we demonstrate the proposed cross-modality correlation propagation algorithm for cross-media retrieval. Section 3 shows the experimental results. Finally, we conclude this paper in section 4.

2. CROSS-MODALITY CORRELATION PROPAGATION

In this section, we present the proposed cross-modality correlation propagation algorithm for cross-media retrieval. Although the fundamental ideas are applicable to any combination of media types, we restrict the discussion to documents containing images and texts as [9]. The goal is to support truly cross-media queries: to retrieve text articles in response to query images and vice-versa.

Given a dataset $\mathcal{D} = \{D_1, ..., D_N\}$, in which D_i denotes a multimedia document containing heterogeneous media objects such as images and texts. Images and texts are represented as feature vectors $f_{I_i} \in \mathcal{R}^I$ and $f_{T_i} \in \mathcal{R}^T$, respectively. Our goal is to exploit the semantic correlation between heterogeneous objects for cross-media query. Firstly, as shown in the left part of Figure 2, we construct the semantic correlation matrix $Y = \{Y_{ij}\}_{m*n}$, where m and n are the object number of texts and images in the dataset. The element Y_{ij} stands for the pairwise constraint between the ith text and the jth image. The definition of Y is given as following:

$$Y_{ij} = \begin{cases} +1, & C(I_i) = C(T_j), i \in \{1, 2, ..., m\}, j \in \{1, 2, ..., n\}; \\ -1, & C(I_i) \neq C(T_j), i \in \{1, 2, ..., m\}, j \in \{1, 2, ..., n\}; \\ 0, & C(I_i) \text{ or } C(T_j) \text{ is unkown.} \end{cases} \tag{1}$$

where $C(I_i)$ and $C(T_j)$ represent the category label of image or text respectively. $Y_{ij} = 1$ means the ith text and the jth image have positive correlation, which is referred to as must-link constraint. $Y_{ij} = -1$ means the ith text and the jth image have negative correlation, which is referred to as cannot-link constraint. $Y_{ij} = 0$ means we do not have any information about the category of the ith text and the jth image. In Figure 2, black filled circles indicate the already known correlation and empty circles indicate the unknown semantic correlation. We place the labeled media objects in the left-top region and the unlabeled media objects in the right-bottom region for observation. Consequently, the correlation propagation problem is reduced to how to fill the elements which are empty circles according to the the already known filled circles.

We make further observations on Y column by column. For the j-th column $Y_{\cdot j}$ which corresponds to the semantic correlation between the j-th image and all the texts, the elements $Y_{\cdot j}$ actually provides the initial configuration of a two-class semi-supervised learning problem. The "positive class"($Y_{ij} = 1$) consists of texts with the same semantic of the j-th image and the "negative class"($Y_{ij} = -1$) consists of the texts of different semantic from the j-th image. This two-class semi-supervised learning problem can be solved by the label propagation technique[14]. We call this kind of propagation *text propagation*. That is, semantic correlation is propagated according to text similarity.

However, there are some columns containing neither positive label nor negative label, which means we do not know any correlation about these media objects and all the corresponding elements are zero. In this case, we cannot propagate the semantic correlation information. As shown in Figure 2, the elements in the right bottom region of Y cannot be obtained readily by *text propagation*. Similar to *text propagation*, *image propagation* can also be performed row by row. The propagation directions of *text propagation* and *image propagation* are orthogonal to each other. As a result, through combining *text propagation* and *image propagation*, the semantic correlation on the labeled objects can be successfully propagated to the unlabeled objects.

In practice, the larger Y_{ij} is, the more probable the ith text and the jth image have the same semantic. Obviously, semantic correlation matrix can naturally meet the requirement of cross-media retrieval. Once correlation is obtained, cross-media retrieval can be accomplished based on the semantic correlation. The propagation problem can be solved efficiently through semi-supervised learning based on k-nearest neighbors graphs. We summarized our CMCP algorithm as following:

(1) Initialize the affinity (or similarity) matrix W_T and W_I with all zeros, and update as following:

$$W_{Tij} = \frac{f_{Ti} \cdot f_{Tj}}{|f_{Ti}| \cdot |f_{Tj}|}, \quad i, j \in \{1, 2, ..., m\}, i \neq j. \tag{2}$$

Here f_{Ti}, f_{Tj} indicates the feature in text space. We adopt the cosine similarity measure here. The construction of W_I is similar to W_T.

(2) Generate the k-NN graph:

$$W_{Tij} = \begin{cases} W_{Tij}, & T_i \in kNN(T_j); \\ 0, & otherwise. \end{cases} \quad (3)$$

$kNN(T_j)$ stands for the set of k-nearest neighbors of text T_j in training set. The construction of W_I is similar to W_T.

(3) Construct the matrix $\bar{S}_T = D^{-1/2} W_T D^{-1/2}$, where D is a diagonal matrix with its (i,i)-element equal to the sum of the i-th row of W_T.The construction of $\bar{S}_I$ is similar to $\bar{S}_T$.

(4) *text propagation*: iterate $F_T(t+1) = \alpha_T \bar{S}_T F_T(t) + (1-\alpha_T)Y$ until convergence, where $F_T(t)$ denotes the text propagation result and we set $F_T(0) = Y$, α_T is a parameter in the range (0,1). Here we normalize the elements of Y column by column to make sure that the sum of each column is zero.

(5) *image propagation*: iterate $F_I(t+1) = \alpha_I F_I(t) \bar{S}_I + (1-\alpha_I) F_T^*$ until convergence, where F_T^* is the limit of the sequence $\{F_T(t)\}$, $F_I(t)$ denotes the image propagation result and we set $F_I(0) = F_T^*$, α_I is a parameter in the range (0,1).

(6) Let F_I^* denote the limit of the sequence $\{F_I(t)\}$, which represents the semantic correlation between media objects of different modalities.

According to [14], the *text propagation* converges to

$$F_T^* = (1-\alpha_T)(I - \alpha_T \bar{S}_T)^{-1} Y \quad (4)$$

Then, for the *image propagation*, we have:

$$\begin{aligned} {F_I}^T(t+1) &= \alpha_I (F_I(t)\bar{S}_I)^T + (1-\alpha_I){F_T^*}^T \\ &= \alpha_I \bar{S}_I {F_I}^T(t) + (1-\alpha_I){F_T^*}^T \end{aligned} \quad (5)$$

Hence, the *image propagation* converges to

$${F_I^*}^T = (1-\alpha_I)(I - \alpha_I \bar{S}_I)^{-1} {F_T^*}^T \quad (6)$$

Combine equation (6) and equation (4), we can obtain the following solution:

$$\begin{aligned} F^* &= F_I^* = (1-\alpha_I) F_T^* (I - \alpha_I \bar{S}_I^T)^{-1} \\ &= (1-\alpha_I)(1-\alpha_T)(I - \alpha_T \bar{S}_T)^{-1} Y (I - \alpha_I \bar{S}_I)^{-1} \end{aligned} \quad (7)$$

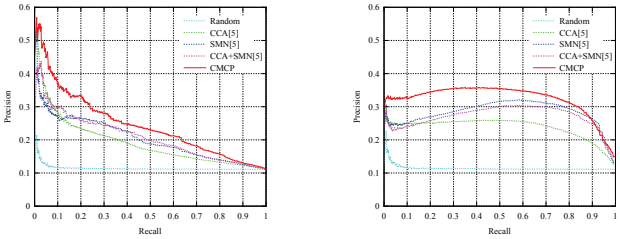
which propagate the semantic correlation throughout all of the elements in Y as shown in the right part of Figure 2.

3. EXPERIMENTS

In this section, we compare the proposed cross-modality correlation propagation algorithm with the state-of-the-art methods for cross-media retrieval.

3.1. Dataset Description

To evaluate the performance of the proposed approach, we conduct experiments on the Wikipedia dataset[9]. It is chosen from the Wikipedia's "featured articles". Both the texts and images are assigned a category label by Wikipedia. Each article is split into several sections according to its section headings. The final dataset contains a total of 2866 documents, which are text-image pairs and annotated with a label from the vocabulary of 10 semantic categories. The dataset is randomly split into a training set of 2173 documents and a test set of 693 documents.



(a) Text Query (b) Image Query

Fig. 3: Precision recall curves

We consider two cross-media retrieval tasks. The first task is to retrieve the text using an image query. Each image in the test set is used as a query and producing a ranking of all texts in the test set. The other task is to retrieve the images using a text query. Similarly, each text in the test set is used as a query and producing a ranking of all images in the test set. The precision-recall (PR) curves and mean average precision (MAP) are taken as the performance measure.

Each image is represented using a histogram of a 128-codeword SIFT codebook and each text is represented using a histogram of a 10-topic Latent Dirichlet allocation (LDA) model. All of the compared cross-media retrieval methods in the experiment section adopt the same features and training data for fair comparison purpose. We set $\alpha_I = \alpha_T = 0.6$ and $k = 30$ in Equation (7) for cross-modality correlation propagation algorithm.

3.2. Cross-media Retrieval

Table 1 shows the MAP scores of our proposed cross-modality correlation propagation approach, compared with the state-of-the-art methods[9]. In this table, four methods are compared, that is, randomly retrieving the results (Random), learning a homogeneous subspace for the modalities through canonical correlation analysis (CCA), learning a high level semantic (SMN) for each object and the combination of the above two algorithms (CCA + SMN) which firstly map the original heterogeneous features into a homogeneous subspace and then learn the high level semantic feature for each object. Both CCA and SMN are supervised learning methods which only exploit the positive correlation between media objects.

Table 1: Retrieval Performance(MAP Scores)

Experiment	Image Query	Text Query	Average
Random	0.118	0.118	0.118
CCA[9]	0.249	0.196	0.223
SMN[9]	0.225	0.223	0.224
CCA+SMN[9]	0.277	0.226	0.252
CMCP	**0.326**	**0.251**	**0.289**

It can be seen from Table 1 that the MAP scores of cross-media retrieval methods all significantly outperform those of random retrieval. The combination of CCA and SMN only shows a small improvement over SMN, while our proposed CMCP is more effective and outperforms all the previous methods. Figure 3 shows the PR curve of all of the above methods. It can be seen that CMCP attains higher precision at most levels of recall, outperforming current state-of-the-art methods.

3.3. Comparison with image retrieval system

We further compare the cross-media retrieval with the traditional unimedia image retrieval, where both query and retrieved results are images. Similar to [9], we adapt the above two kinds of cross-media retrieval task to the unimedia image retrieval task. For the first task, we complement the images in the dataset with text articles (proxy text ranking) and served as proxies for all of the images in the dataset. For the second task, we complement the image query with a text article and served as a proxy (proxy text query) to retrieve the images in the dataset.

Table 2: Content based image retrieval

Experiment	MAP Score
CMCP (Proxy Text Ranking)	**0.326**
CMCP (Proxy Text Query)	**0.251**
CCA + SMN [9] (Proxy Text Ranking)	0.277
CCA + SMN [9] (Proxy Text Query)	0.226
Image SMN [15]	0.161
Image SIFT Features [16]	0.135
Random	0.117

Table 2 shows the comparison of the cross-media retrieval approaches with a number of state-of-the-art unimedia image retrieval methods. The method of [16] represents images as distributions of SIFT features. The method of [15] projects the images to a semantic space, which is similar to the SMN cross media retrieval method [9]. The MAP score of unimedia retrieval methods has shown the difficulty of image retrieval on this Wikipedia dataset. The cross-media retrieval methods significantly improve the performance and our proposed CMCP further improve the retrieval result. Jointly modeling different modalities show a significant benefit in solving the classic image retrieval problem. This indicates that multiple modalities could be complementary to each other and boost the retrieval result.

4. CONCLUSION

In this paper, we have proposed a novel cross-modality correlation propagation approach to simultaneously deal with positive correlation and negative correlation between media objects of different modalities. Furthermore, our approach is able to propagate the correlation between heterogeneous modalities. Experiments on the wikipedia dataset show the effectiveness of CMCP, compared with state-of-the-art methods.

In the future, on one hand, we will jointly model other modalities such as audio and video; on the other hand, the correlation between categories could also be explored.

5. ACKNOWLEDGMENTS

The work described in this paper was fully supported by the National Natural Science Foundation of China under Grant Nos. 61073084 and 60873154, the Beijing Natural Science Foundation of China under Grant No. 4122035, and the National Development and Reform Commission High-tech Program of China under Grant No. [2010]3044.

6. REFERENCES

[1] G. Carneiro, A. Chan, P. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[2] J. Pan, H. Yang, C. Faloutsos, and P. Duygulu, "Automatic multimedia cross-modal correlation discovery," *ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2004.

[3] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," *Proceedings of the 26th annual international ACM SIGIR conference*, 2003.

[4] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2007.

[5] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," *ACM International Conference on Multimedia*, 2003.

[6] Y. Zhuang, Y. Yang, and F. Wu, "Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 2, pp. 221–229, February 2008.

[7] Y. Yang, Y. Zhuang, F. Wu, and Y. Pan, "Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 437–446, April 2008.

[8] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," *ACM International Conference on Multimedia*, pp. 175–184, 2008.

[9] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," *ACM international conference on Multimedia*, 2010.

[10] X. Zhai, Y. Peng, and J. Xiao, "Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval," *International Conference on MultiMedia Modeling (MMM)*, 2012.

[11] Z. Lu and M. Carreira-Perpinan, "Constrained spectral clustering through affinity propagation," *CVPR*, 2008.

[12] Z. Lu and H. Ip, "Constrained spectral clustering via exhaustive and efficient constraint propagation," *Proceedings of the European Conference on Computer Vision*, 2010.

[13] Z. Li, J. Liu, and X. Tang, "Pairwise constraint propagation by semidefinite programming for semi-supervised classification," *ICML*, pp. 576–583, 2008.

[14] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[15] N. Rasiwasia, P. Moreno, and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[16] N. Vasconcelos, "Minimum probability of error image retrieval," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2322–2336, 2004.