IMPROVING CROSS-MODAL CORRELATION LEARNING WITH HYPERLINKS

Shuhui Wang[1], *Yiling Wu*[1,2], *Qingming Huang*[1,2]

[1]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{wangshuhui,qmhuang}@ict.ac.cn, yiling.wu@vipl.ict.ac.cn

ABSTRACT

We propose a new cross-modal correlation learning framework which boosts the performance of correlation learning models using the hyperlink information. First, we design a neighborhood selection paradigm using the hyperlink structure and content similarities to identify a set of semantically related documents for each multi-modal document in both training and testing stage. Based on the neighborhood structure, we revise two well-established content-based correlation learning models, i.e., canonical correlation analysis (CCA) and kernel canonical correlation analysis (KCCA) with a structure coding matrix. Third, we develop a correlation score aggregation technique to discover more semantically relevant cross-modal documents. To our best knowledge, this is the first to introduce hyperlink information into cross-modal correlation learning. Experimental results demonstrate that our proposed framework can significantly improve the model generality towards real-world cross-modal retrieval.

Index Terms— Correlation learning, hyperlinks, neighborhood information

1. INTRODUCTION

The aim of multimedia retrieval is to make capturing, storing, finding and using digital media an everyday occurrence in our computer environment [1]. Unlike decades ago, information is delivered by text, image, audio and video, which describes topics and real world events in a more vivid way. However, it's difficult to obtain relevant information from multiple modalities given a query from one modality. The demand for retrieving other modalities by providing a certain modality of query has become more and more urgent.

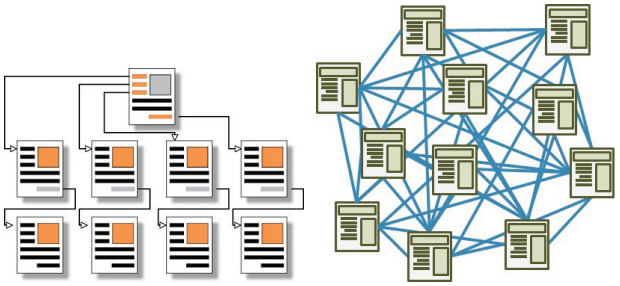
Mining correlation among different modalities is challenging due to the heterogeneity and complexity of multi-modal data. For instance, there is a wide gap between low-level pixel-based image description and the semantic interpretation. Research efforts have been devoted to constructing a measureable representation for multi-modal data [2] [3] [4]. Subspace learning is a widely accepted paradigm which seeks low-dimensional subspaces that maximize the multi-modal correlation [5] [6] [7] [8] [9]. However, existing studies address the correlation learning problem from purely content and semantic modeling perspectives. The cross-modal data that are assumed to be independently generated from certain intrinsic content distributions with ground-truth semantic labels. There are two critical issues that need to be considered when dealing with web multi-modal data.

First, due to the intrinsic *interconnection* nature of the internet, web documents are naturally inter-correlated via structure information such as *hyperlinks*. For example, in Figure 1, the web documents are organized as tree or graph structure rather than isolated from each other. Previous studies are conducted on web cross-modal datasets whose hyperlink information is discarded during the data collection and preprocessing stages. However, these datasets can not reflect the reference or co-reference relations among web documents. Such possibly incorrect assumption or prior knowledge has been inappropriately introduced during the data construction process, which at the same time misleads the study of correlation models.

Second, the hyperlinks contain rich description of the true correlation among multi-modal documents. See example from Wikipedia in Fig. 2, the outgoing hyperlinks of *Tank* provide certain level of supplementary descriptions to the original page. Therefore, the original page and the linked pages are likely to belong to the same semantic category. From the methodology perspective, existing pure content-based and semantic-based models can not well capture the semantic relations encoded in the hyperlinks, thus they can not effectively fit to the real multi-modal data.

In this paper, we develop a new cross-modal correlation learning framework which aims to boost the correlation learning models using the hyperlink information. Our contribution are three folds. First, we design a neighborhood selection paradigm using the structure and content similarities to identify a set of semantically related neighbors for each multi-modal document in both training and testing stage. Based on the neighborhood structure, we revise two well-established content-based correlation learning models, i.e., canonical cor-

This work was supported in part by National Basic Research Program of China (973 Program): 2012CB316400 and 2015CB351802, National Natural Science Foundation of China: 61025011, 61332016, 61390511 and 61303160, 863 program of China: 2014AA015202, and Basic Research Program of Shenzhen: JCYJ20140610152828686.



(a) Tree (b) Graph

Fig. 1. Typical graph structures of web pages

relation analysis (CCA) and kernel canonical correlation analysis (KCCA). From the representation perspective, we propose new neighborhood features for cross-modal retrieval. From the model perspective, we enhance the model capacity of CCA and KCCA with a structure coding matrix, which is informative in capturing the structure relation among multi-modal documents. Third, we develop a correlation score aggregation technique which expands the query based on the neighborhood structure. Consequently, more semantically relevant cross-modal documents can be effectively discovered. To our best knowledge, this is the first to introduce hyperlink information into cross-modal correlation learning. Experiments demonstrate that our proposed framework can significantly improve the model generality towards real-world cross-modal retrieval.

2. RELATED WORK

The hyperlink information has been well-studied in information retrieval research domain. By combining content information with web structure information, better models have been achieved than purely content-based algorithms in many tasks such as text classification, text clustering and document ranking. Chakrabarti [10] developed a robust relaxation labeling model for text classification by exploiting hyperlinks. Taskar [11] proposed a discriminative probabilistic model for relational data by introducing the framework of relational Markov networks to model the relational dependencies of texts. Besides the text classification task, the hyperlink information can also contribute to measure the sematic relatedness between web documents. For example, Milne et al.[12] developed a model that learns to link the semantically rich Wikipedia pages. However, the importance of hyperlink has not drawn considerable attention in the context of cross-modal correlation learning.

3. APPROACH

3.1. Overview

Given a cross-modal document corpus $\{(x_I^i, x_T^i), i = 1, ..., N\}$, where x_I^i denotes the i-th image, which is co-occurred with the i-th text x_T^i. To effective retrieve data from heterogeneous modalities, we take advantage of the hyperlink information to boost the cross-modal correlation learning on Web data. It mainly contains three key steps:

Step 1. Neighborhood selection (Section 3.2). Given a set of candidate neighbors selected by the outgoing hyperlinks, we calculate their importance score according to their discriminative ability based on the hyperlink structure and content similarity. The semantically related neighbors can be identified by selecting the most important candidates.

Step 2. Correlation learning (Section 3.3 and 3.4). Based on the neighborhood information, we revise the correlation learning with a local structure coding matrix. Consequently, each cross-modal document is projected into a d-dimensional projected representation.

Step 3. Correlation score aggregation (Section 3.5). Given a query, we retrieve the cross-modal documents by the original query and their neighbors. and output the final result in section 3.5. We will describe each part in details.

3.2. Neighbor selection

We regard the web documents as a directed graph. The i-th multi-modal document $\{(x_I^i, x_T^i)\}$ is represented by a vertex, and the hyperlink between two documents represented by a directed edge. Given a specific multi-modal document, the *outgoing hyperlinks* are pointers to other documents that provide supplementary explanation of the original document. Generally, the documents pointed by such links describe a certain topic or concept in the original document. Therefore, a set of related multi-modal documents can be retrieved by collecting the documents pointed by outgoing hyperlinks in each document, which are referred to as *candidate neighbors*.

However, not all candidate neighbors contribute equally for providing supplementary information of the original document. The reasons are two-folds. First, the contents in some candidate neighbor may be too general[1]. Second, some candidate neighbor may not be highly semantically related with the original document. For example, in Fig. 2, we have the original document of *Tank* from Wikipedia. Candidate neighbors selected by outgoing hyperlinks are marked with blue colors on the page. However, the co-occurred image only include content of *main gun*, *machine gun* and *Tracked*. Candidates like *Calibre* and *Front-line* have less semantic relation with *Tank* itself, and candidates like *Offensive* and *Defensive* are too general to provide sufficient supplementary explanation of the original document t.

Given a multi-modal document t, we aim to determine the relative importance of a candidate neighbor by considering both hyperlink and content similarity. Intuitively, a candidate neighbor should not be too general or too specific. If it is too general, the document may be neither relevant nor dis-

[1]Intuitively, when the content of a document is general, it may involve many semantic concepts. On platforms such as Wikipedia, the document with more general content will contain more outgoing hyperlinks and incoming hyperlinks.

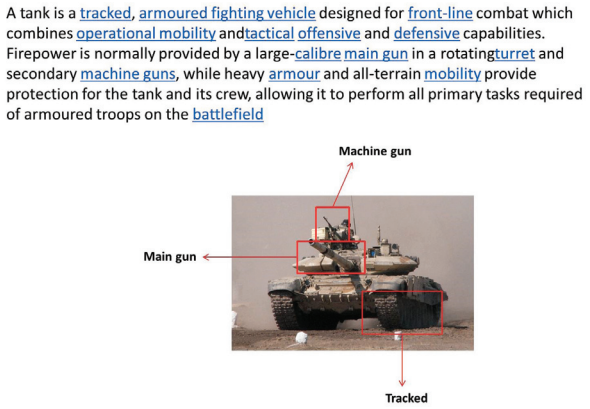


Fig. 2. A Web page *Tank* in Wikipedia (http://en.wikipedia.org/wiki/Tank)

criminative. If it is too specific, its content is not adequate to further explain the document t. We first calculate the inverse incoming link score as:

$$\alpha_p = \log(\frac{N}{IH_p + \epsilon_0}) \tag{1}$$

where IH_p represents the number of incoming hyperlinks of document p, N is the number of whole multi-modal document corpus and $\epsilon_0 > 0$ is a small positive value that avoids zeroes in the denominator. IH_p measures how likely document p will be pointed by other documents. A larger α_p indicates that p is rarely referred to by other documents, while a smaller α_p indicates that p is frequently referred to by others. Similarly, we define the inverse outgoing link score as:

$$\beta_p = \log(\frac{N}{OH_p + \epsilon_0}) \tag{2}$$

where OH_p represents the number of outgoing hyperlinks of document p. A larger β_p indicates that p rarely refer to other documents, while a smaller β_p indicates that p frequently refer to others.

Accordingly, we calculate the truncated structure importance score for each candidate neighbor p of t as:

$$w_t(p) = \begin{cases} \alpha_p \cdot \beta_p & \mu_1 \geq \alpha_p \geq \varepsilon_1, \mu_2 \geq \beta_p \geq \varepsilon_2 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

where $\varepsilon_1, \varepsilon_2$ and μ_1, μ_2 are truncating parameters that remove those candidate neighbors that are either too specific or too general. After the double-side truncation, a larger $w_t(p)$ means that document p is sufficient to provide supplementary explanation on certain concept or attribute in t by appropriate divergent description that is not included in t.

Moreover, the patterns that each document refers to other documents provide strong indication on how the topic of two documents are related. Such relation can be modeled by:

$$sl(t,p) = \frac{|C_t \cap C_p|}{|C_t \cup C_p|} \tag{4}$$

where C_t and C_p represent the document indexes that are pointed by t and p, respectively. $|\cdot|$ represent the set cardinality. The larger $sl(t,p)$ is, the more possible t and p discuss similar topics, which means that p is more likely to be the neighbor with high semantic relevance to t.

We measure the importance score of each candidate neighbor p given t as:

$$r_t{}^p = w_t(p) \cdot sl(t,p) \cdot sim(t,p) \tag{5}$$

where $sim(t,p)$ denotes the content similarity of the two multi-modal documents $\{(x_I^t, x_T^t)\}$ and $\{(x_I^p, x_T^p)\}$ as:

$$sim(t,p) = \kappa_I(x_I^t, x_I^p) + \kappa_T(x_T^t, x_T^p) \tag{6}$$

where κ_I and κ_T are domain specific similarity (kernel) definition, e.g., RBF kernel for visual modality and cosine similarity for textual modality.

With the importance score calculation and ranking, we identify the most semantically related documents with sufficient supplementary description to each document t. We denote the set of selected neighbors given t as $\mathcal{N}_t$.

3.3. Neighborhood features and similarities

Denote the training data as $\mathbf{X}_I$ and $\mathbf{X}_T$, respectively, where $X_I(t,:) = x_I^t$ and $X_T(t,:) = x_T^t$. We calculate the new feature for each document x_I^t or x_T^t based on the neighborhood structure $\mathcal{N}_t$ and the importance score $r_t(p), p \in \mathcal{N}_t$ as:

$$\begin{aligned} \bar{x}_I^t &= \lambda x_I^t + (1-\lambda) \sum_{p \in \mathcal{N}_t} \bar{r}_t^p x_I^p \\ \bar{x}_T^t &= \lambda x_T^t + (1-\lambda) \sum_{p \in \mathcal{N}_t} \bar{r}_t^p x_T^p \end{aligned} \tag{7}$$

where $\bar{r}_t^p = r_t^p / \sum_{p' \in \mathcal{N}_t} r_t^{p'}$, and $0 \leq \lambda \leq 1$. Since the neighborhood feature is in fact a linear transformation on the original representation $\mathbf{X}_I$ and $\mathbf{X}_T$, we define a neighbor structure matrix $\mathbf{R} \in \mathbb{R}^{N \times N}$ where:

$$R(t,p) = \begin{cases} \lambda, \ p = t \\ (1-\lambda)\bar{r}_t^p, \ p \in \mathcal{N}_t \\ 0, \ otherwise \end{cases} . \tag{8}$$

Therefore, the neighborhood features and neighborhood kernel (similarity) are represented as:

$$\begin{aligned} &\overline{\mathbf{X}}_I = \mathbf{R}\mathbf{X}_I, \overline{\mathbf{X}}_T = \mathbf{R}\mathbf{X}_T, \\ &\overline{\mathbf{K}}_I = \mathbf{R}\mathbf{K}_I\mathbf{R}^\top, \overline{\mathbf{K}}_T = \mathbf{R}\mathbf{K}_T\mathbf{R}^\top. \end{aligned} \tag{9}$$

where $\mathbf{R}$ represents the local structure coding matrix.

3.4. Correlation learning based on neighborhoods

We apply the neighborhood feature and similarity on CCA and KCCA, respectively.

Canonical Correlation Analysis. Given image space $\mathcal{R}^I \in \mathbb{R}^{d_I}$ and text space $\mathcal{R}^T \in \mathbb{R}^{d_T}$, the goal is to find subspace $\mathbf{U}^I \in \mathcal{R}^I$ and $\mathbf{U}^T \in \mathcal{R}^T$ that maximizes the correlation between the two modalities by learning a projection pair $\{w_I, w_T\}$ for both spaces via the following problem:

$$\begin{aligned}
&\max_{w_I, w_T} \frac{w_I^\top \overline{\mathbf{C}}_{IT} w_T}{\sqrt{w_I^\top \overline{\mathbf{C}}_{II} w_I}\sqrt{w_T^\top \overline{\mathbf{C}}_{TT} w_T}} \\
&s.t.\ ||w_I||^2 \leq 1, ||w_T||^2 \leq 1. \\
&\overline{\mathbf{C}}_{II} = \mathbf{X}_I^\top \mathbf{R}^\top \mathbf{R} \mathbf{X}_I,\ \overline{\mathbf{C}}_{TT} = \mathbf{X}_T^\top \mathbf{R}^\top \mathbf{R} \mathbf{X}_T, \\
&\overline{\mathbf{C}}_{IT} = \mathbf{X}_I^\top \mathbf{R}^\top \mathbf{R} \mathbf{X}_T.
\end{aligned} \tag{10}$$

By solving Eqn. 10 using eigen-decomposition, we obtain projection pairs $\{w_i^1, w_t^1\}$, ..., $\{w_i^{d'}, w_t^{d'}\}$, where $d' = \min(d_I, d_T)$. We choose projection pairs with d largest eigenvalues, where $d < d'$. Thus, the original data are projected into the d-dimension isometric subspace $\mathbf{U}^I$ and $\mathbf{U}^T$, respectively. The projection functions can be written as:

$$f_I(x_I^q) = \bar{x}_I^q \mathbf{W}_I,\ \ f_T(x_T^q) = \bar{x}_T^q \mathbf{W}_T, \tag{11}$$

where $\mathbf{W}_I = [w_i^1, ..., w_i^d]$ and $\mathbf{W}_T = [w_t^1, ..., w_t^d]$. After the projection, data from different modalities can be directly compared.

Kernel Correlation Analysis. The linear projection constructed by CCA can be kernelized to map data into a high-dimensional space without an explicit mapping function. According to the theory of Reproducing Kernel Hilbert Space (RKHS), Eqn. 10 can be rewritten as:

$$\max_{\alpha,\beta} \frac{\alpha^\top \overline{\mathbf{K}}_I \overline{\mathbf{K}}_T \beta}{\sqrt{\alpha^\top \overline{\mathbf{K}}_I^2 \alpha}\sqrt{\beta^\top \overline{\mathbf{K}}_T^2 \beta}}.\ s.t.\ ||\alpha||^2 \leq 1,\ ||\beta||^2 \leq 1 \tag{12}$$

Similarly, by solving Eqn. 12 with eigen-decomposition, we obtain projection pairs $\{\alpha^1, \beta^1\}$,..., $\{\alpha^d, \beta^d\}$ of the top d large eigen-values for image and text, respectively. Given any image x_I^q or text x_T^q, the projection can be written as:

$$\begin{aligned}
f_I(x_I^q) &= \left(\lambda \mathbf{K}_I(x_I^q, \cdot) + (1-\lambda) \sum_{p \in \mathcal{N}_q} \mathbf{K}_I(x_I^p, \cdot)\right) \mathbf{R}^\top \mathbf{A} \\
f_T(x_T^q) &= \left(\lambda \mathbf{K}_T(x_T^q, \cdot) + (1-\lambda) \sum_{p \in \mathcal{N}_q} \mathbf{K}_T(x_T^p, \cdot)\right) \mathbf{R}^\top \mathbf{B} \\
\mathbf{A} &= [\alpha_1, ..., \alpha_d], \mathbf{B} = [\beta_1, ..., \beta_d].
\end{aligned} \tag{13}$$

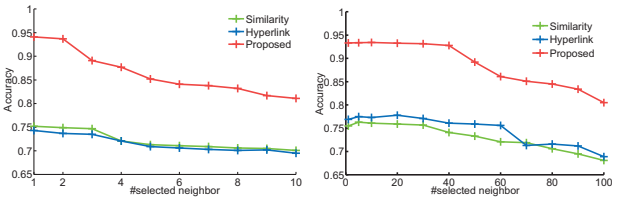
where $\mathbf{K}_I(x_I, \cdot)$ ($\mathbf{K}_T(x_T, \cdot)$) represent the similarity vectors between x_I (x_T) and the training images (texts).

3.5. Correlation score aggregation

Given an image or textual query x_I^q or x_T^q, we obtain its neighborhood $\mathcal{N}_q$ on the training data. Based on the neighborhood information, we obtain the d-dimensional projected representation $f_I(x_I^q)$ or $f_T(x_T^q)$ by Eqn. 11 or Eqn. 13. Then each document k from other modality in the retrieval database is assigned with a relevance score τ_q^k by calculating their cosine similarity to the query q.

We further aggregate the similarity scores by using some neighbors in $\mathcal{N}_q$ as the additional queries to q. Specifically, we select the top m similar neighbors $\mathcal{N}_q^m$ from $\mathcal{N}_q$, and calculate their similarities to document k. Therefore, the final similarity score of document k given query q is:

$$\overline{\tau}_q^k = \tau_q^k + \eta \sum_{q' \in \mathcal{N}_q^m} r_q{}^{q'} \tau_{q'}^k \Big/ \sum_{q' \in \mathcal{N}_q^m} r_q{}^{q'} \tag{14}$$



(a) WIKI-FA. (b) WIKI-CMR.

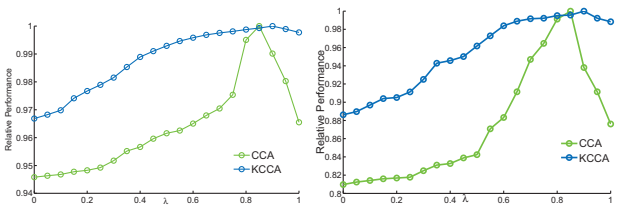
Fig. 3. Accuracy curves of neighbor searching.

4. EXPERIMENT

We conduct experiment on two datasets. WIKI-FA dataset [7] is built on Wikipedia featured articles with 2866 text-image pair annotated with labels from 10 semantic classes. The training set includes 2173 document pairs, and the test set includes 693 document pairs. We use a simple crawling strategy to collect the web structure information that is not included in the original dataset [7]. Specifically, the outgoing hyperlinks of each document in WIKI-FA are collected, including hyperlinks pointing to documents in WIKI-FA and hyperlinks pointing to other documents in Wikipedia. The hyperlinks pointing to Web documents outside Wikipedia are removed. WIKI-CMR [13] contains 8567 text-image document pairs with both content information and Web structure information. We represent each text with 2000-dimension feature using Latent Semantic Analysis on TF-IDF representation, and represent each image using a 10754-dimensional sparse-coding-based representation [14] on dense SIFT descriptors with vocabulary size 512. Similar to WIKI-FA, each document pair is labeled with 10 semantic classes. It is partitioned into training set including 6567 documents, and test set including 2000 documents. Unless specified, other parameters are set to guarantee the optimal performance for both our approach and the compared approaches.

4.1. Neighbor searching accuracy

The quality of neighborhood selection determines the performance of correlation learning and score aggregation. To evaluate the quality of the selected neighborhood with the proposed scheme (Section 3.2), we measure the precision of the top ranked candidate neighbors by checking if their labels are identical to the query document. Given top m candidate neighbors, we compare three neighborhood searching strategies, and record the label accuracy in Fig. 3. We see that by combining hyperlink structure and content information, the accuracies of neighbor searching have been significantly improved on both datasets, compared to only using hyperlink or content similarity. On WIKI-FA, the highest accuracy (94.1%) is achieved when $m = 1$ using our strategy. On WIKI-CMR, the highest accuracy (93.4%) is achieved when $m = 10$ for our strategy. We observe obvious drops on our accuracy curves when increasing the selected number of neighbors. Therefore, to ensure both precision and recall, we set the number of selected neighbors as $m = 2$ for WIKI-FA and



(a) WIKI-FA. (b) WIKI-CMR.

Fig. 4. Relative MAP curves on different λ.

$m = 40$ for WIKI-CMR in the subsequent experiments.

4.2. Effectiveness of the neighborhood features

The parameter λ (Eqn. 7) determines the relative importance of the original feature and neighborhood feature. We first conduct sensitivity test on both datasets by randomly choosing 20% from the training partitions as the validation sets, and then we evaluate the performance of the proposed neighborhood feature on existing cross-modal retrieval paradigms.

We evaluate the influence of λ by setting $\eta = 0$ (no correlation score aggregation). The results are measured by the average of Mean Average Precision (MAP) of text-to-image and image-to-text retrieval. For a better illustration, we show the relative performance in Fig. 4, which is the MAP divided by the maximum performance under all the settings of λ. When $\lambda = 1$, it is equivalent to using the original feature without neighborhood information. When $\lambda = 0$, it is equivalent to replace the original feature with neighborhood feature. From Fig. 4, we see that the neighborhood feature provides more discriminative information. The best performances have been achieved when $\lambda = 0.85$ for CCA and $\lambda = 0.9$ for KCCA on both WIKI-FA and WIKI-CMR. Even when we remove the original feature ($\lambda = 0$), the performance drops on both datasets are not very significant. We also see that KCCA is more tolerant to the setting of λ than CCA, as kernel functions in KCCA provide local smoothing in the RKHS, which tends to be more robust to representation variation.

Furthermore, we evaluate the effectiveness of the proposed neighborhood feature based on the three cross-modal retrieval paradigms proposed by [7], namely, correlation matching (CM), semantic matching (SM), and semantic correlation matching (SCM). In CM, we directly report the results of CCA and KCCA. In SM, we build multi-class linear and kernel logistic regression models to map cross-modal data into the semantic space. In SCM, we perform correlation learning by CCA and KCCA, respectively, and then we construct multi-class linear logistic regression model on the d-dimensional representation. The average results of MAP are recorded in Table 1 and 2. We see that by appropriately incorporating the neighborhood feature, better performances have been achieved on both datasets. The results show that the neighborhood feature is effective in incorporating more semantic information in cross-modal correlation learning.

Table 1. Effectiveness of neighborhood feature on WIKI-FA

Method	CM	SM	SCM
CCA+origin ($\lambda = 1$) [7]	0.223	0.224	0.252
CCA+neighborhood ($\lambda = 0$)	0.201	0.198	0.224
CCA+best ($\lambda = 0.85$)	**0.238**	**0.237**	**0.283**
KCCA+orgin ($\lambda = 1$)	0.579	0.578	0.613
KCCA+neighborhood ($\lambda = 0$)	0.531	0.524	0.588
KCCA+best ($\lambda = 0.9$)	**0.594**	**0.591**	**0.637**

Table 2. Effectiveness of neighborhood feature on WIKI-CMR

Method	CM	SM	SCM
CCA+origin ($\lambda = 1$)	0.292	0.274	0.337
CCA+neighborhood ($\lambda = 0$)	0.281	0.247	0.295
CCA+best ($\lambda = 0.85$)	**0.318**	**0.298**	**0.364**
KCCA+orgin ($\lambda = 1$)	0.591	0.582	0.635
KCCA+neighborhood ($\lambda = 0$)	0.567	0.545	0.594
KCCA+best ($\lambda = 0.9$)	**0.617**	**0.596**	**0.658**

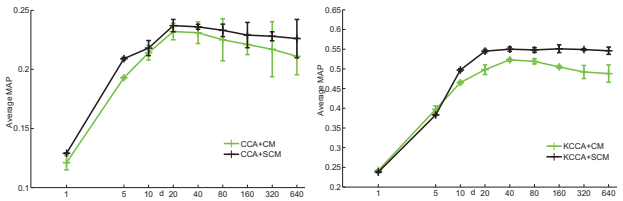
4.3. Correlation score aggregation

We conduct evaluation on the effectiveness of correlation score aggregation. There are two parameters that influence the performance. m denotes the number of aggregated neighbor queries, and η (Eqn. 14) determines the relative importance of the neighborhood queries. We conduct experiment on the training/validation sets by setting λ with the optimal values, and find that $\eta = 0.15$ is the optimal setting on both WIKI-FA and WIKI-CMR.

Based on the near-optimal setting of η, we conduct experiment on how correlation score aggregation with semantically related neighbors influences the retrieval performance. The results of average MAP are shown in Table 3 and 4. We see that by aggregating certain number of neighbors, the retrieval performance of CCA and KCCA has been enhanced on both datasets. The reason can be explained by the fact that correlation score aggregation is likely to enhance the recall of the top ranked results. However, if too many neighbors are treated as the expanded queries, the precision will be decreased. Therefore, an appropriate setting of the number of expanded queries should be carefully selected.

4.4. Sensitivity on d

We discuss how the reduced dimensions d influence the performance of the proposed model. According to the analysis



(a) CCA (b) KCCA

Fig. 5. Sensitivity on d on WIKI-CMR dataset

Table 4. Correlation score aggregation on WIKI-CMR

#neighbor	0	1	2	4	6	8	10	16	24
CCA+CM	0.318	0.323	**0.329**	0.325	0.323	0.321	0.319	0.316	0.314
CCA+SM	0.298	**0.303**	0.297	0.295	0.293	0.291	0.288	0.284	0.281
CCA+SCM	0.364	0.369	0.373	**0.377**	0.374	0.372	0.369	0.367	0.364
KCCA+CM	0.617	0.626	0.633	**0.636**	0.631	0.627	0.624	0.618	0.612
KCCA+SM	0.596	0.605	0.612	**0.619**	0.616	0.612	0.609	0.607	0.593
KCCA+SCM	0.658	0.663	0.668	**0.673**	0.669	0.666	0.661	0.659	0.652

Table 3. Correlation score aggregation on WIKI-FA

neighbors	0	1	2
CCA+CM	0.238	0.241	0.247
CCA+SM	0.237	0.243	0.248
CCA+SCM	0.283	0.293	**0.308**
KCCA+CM	0.594	0.606	0.613
KCCA+SM	0.591	0.605	0.615
KCCA+SCM	0.637	0.650	**0.662**

in [7], the algorithms are not sensitive on d, since the textual feature dimension of WIKI-FA is only 10. However, on WIKI-CMR, the numbers of feature dimensions of image and text are 10754 and 2000, respectively. Therefore, we conduct sensitivity evaluation on WIKI-CMR with respect to d based on 5-folds cross validation training data. Specifically, we evaluate the performance on the training/validation set with $d = [1, 5, 10, 20, 40, 80, 160, 320, 640]$, and report the average performance and standard deviation in Fig. 5. The best performances have been achieved when d is around 20, which means that by encoding the neighborhood information with correlation model, a more compact representation can be established on data with heterogeneous modalities. Specifically, SCM strategy are less sensitive than CM to d, since SCM projects multi-modal data into the 10-dimensional semantic space, and the parameters of the mapping function are regularized by square-norm penalty.

5. CONCLUSION

We propose a novel framework which combines both content and hyperlink information for cross modal retrieval. Extensive experiments prove that the hyperlink information is useful in selecting semantically related cross-modal documents for calculating more discriminative features, and improving the retrieval performance by structure coding and query expansion. In future work, we will study how to take full advantage of the Web structure and other social attributes to better discover the inter-relation among cross-modal documents. We will evaluate how hyperlinks can improve other correlation learning models, and focus on developing joint models using content, semantic and structure information.

6. REFERENCES

[1] L. A. Rowe and R. Jain, "Acm sigmm retreat report on future directions in multimedia research," *ACM TOMCCAP*, vol. 1, pp. 3–13, 2004.

[2] J. S Hare, P. AS. Sinclair, P. H. Lewis, K. Martinez, P. GB. Enser, and C.J. Sandom, "Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches," 2006.

[3] C. Wang, L. Zhang, and H.-J. Zhang, "Learning to reduce the semantic gap in web image retrieval and annotation," in *ACM SIGIR*, 2008, pp. 355–362.

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

[5] D. 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, 2004.

[6] D. R. Hardoon and J. Shawe-Taylor, "Sparse canonical correlation analysis," *Machine Learning*, vol. 83, no. 3, pp. 331–353, 2011.

[7] 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," in *ACM Multimedia*. ACM, 2010, pp. 251–260.

[8] A. Kimura, H. Kameoka, M. Sugiyama, T. Nakano, E. Maeda, H. Sakano, and K. Ishiguro, "Semicca: Efficient semi-supervised learning of canonical correlations," in *ICPR*, 2010, pp. 2933–2936.

[9] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *CVPR*. IEEE, 2010, pp. 3594–3601.

[10] S.Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in *SIGMOD*, 1998, pp. 307–318.

[11] B. Taskar, P. Abbeel, and D. Koller, "Discriminative probabilistic models for relational data," in *UAI*, 2002, pp. 485–492.

[12] D. N. Milne and I. H. Witten, "Learning to link with wikipedia," in *CIKM*, 2008, pp. 509–518.

[13] W. Xiong, S. Wang, C. Zhang, and Q. Huang, "Wiki-cmr: A web cross modality dataset for studying and evaluation of cross modality retrieval models," in *ICME'13*, 2013, pp. 1–6.

[14] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*, 2009, pp. 1794–1801.