A Graph Model for Cross-modal Retrieval

Shixun Wang • Peng Pan • Yansheng Lu

Abstract With the rapid growth of multimedia document on the web, cross-modal retrieval has become an important issue. The modality of a query is different from that of the retrieved results in cross-modal retrieval. In this paper, we propose a novel graph model, which not only combines content and semantics similarities through two Markov chains, but also utilizes the interaction between different modalities to attain the whole semantics information of a multimedia document. Content similarity focuses on the original features within each modality, while semantics similarity focuses on the semantic vectors in a common space. Both of them are very significant. Random forests method is used to map the original features into a semantic space. The ranked list for a query is achieved by highlighting an optimal path across the corresponding chain. Experiments on the Wikipedia dataset show that the performance of our model significantly outperforms those of existing approaches for cross-modal retrieval.

Keywords Cross-modal retrieval • Graph model • Content similarity • Semantics similarity • Interaction

1 Introduction

In recent years, there has been a massive explosion of multimedia content on the web. Many researchers have devoted themselves to the retrieval methods of unimodal content data, where a query and the retrieved results are of the same modality. However, many different modalities data such as texts, images and music often co-exist in a multimedia document to better express the same semantic information. Such documents include newspaper articles, personal blogs and E-commerce web-pages. Actually, a user may require texts or other modalities with a query image. In this case, the unimodal retrieval methods cannot measure the content similarity among media data of different modalities, so the cross-modal retrieval becomes increasingly important.

S. Wang • P. Pan (✉) • Y. Lu
School of Computer Science & Technology,
Huazhong University of Science and Technology, Wuhan 430074, China
e-mail: wsxun@hust.edu.cn, {panpeng, lys}@mail.hust.edu.cn

Different media data supply complementary information that strongly helps people to comprehend the multimedia document. Therefore, modeling the interaction between different modalities seems to be significant. The data space of unimodal object can be divided into content space and semantic space. The latter is a probability simplex, in which each dimension represents a predefined concept, and data point is a vector of posterior concept probabilities. The methods that only consider single space have limitations, and a combination of two spaces may be beneficial. To the best of our knowledge, such work has not been found for cross-modal retrieval in previous papers.

The beginning of a cross-modal system is the automatic annotation of media [3, 9], but the annotations are restricted to describe visible or auditory object. The key problem in cross-modal retrieval is how to measure the similarity among different media modalities, the existing methods usually focus on a common space in which the classical measure can be directly applied. The common spaces include correlative subspace [6, 7, 8], semantic space [8] and hash space [12].

Rasiwasia et al [8] apply CCA to learn the subspace that maximizes the correlation between image and text. Lmura et al [7] use GCCA to simultaneously focus on the correlation among image, sound and location information. Li et al [6] introduce CFA to seek transformations that best represent the association between two different modalities. But these methods do not consider semantic information.

In [8], the semantic vectors of media objects can be got by multi-class logistic regression. But this method does not consider the content similarity. Zhen et al [12] propose a model for learning hash functions from different modalities data automatically. Zhai et al [11] propose a cross-media correlation propagation approach to deal with positive and negative correlations between different modalities data. The above two methods depend on concept label similarity rather than semantic similarity. Xie et al [10] use a semantic generation model to describe the semantic correlation of different modalities, but they do not simultaneously combine the two different information sources.

There are multiple independent Markov chains of latent variables in the factorial Hidden Markov model (FHMM) [5]. The distribution of observed variable at a given time step is conditional on the states of all of the corresponding latent variables at that same time step. Our model is not treated as the FHMM because it has some variations against the standard model, as well as its algorithm used to highlight the most likely path is different from the standard algorithm.

In this paper, the goals are to retrieve texts in response to a query image and vice-versa. The proposed model, which takes two Markov chains, combines content similarity and semantics similarity together in a graph. The intuition is that the full semantic information of multimedia document can be generated by the interaction between two modalities. Every state represents a media object in each chain. Each edge connecting two states is weighted by the corresponding content similarity, while each edge between a hidden state and semantic space is indicated by the semantic vector of the state. The retrieval process simultaneously depends on both of the two modalities.

2 The Graph Model

In this section, we design a graph model for cross-modal retrieval. Although the fundamental thoughts of our model can be applied to other modalities, the discussions are limited to multimedia documents containing images and texts.

The multimedia dataset is denoted as $\Delta = \{D_1, \ldots, D_{|\Delta|}\}$, in which each document consists of an image and a text, namely $D_i = (\mathbf{I}_i, \mathbf{T}_i)$. Images and texts are described as low-level feature vectors $\mathbf{I}_i \in \Re^I$ and $\mathbf{T}_i \in \Re^T$, respectively. Consider a vocabulary L consisting of $|L|$ unique labels, each label $l_i \in L$ is a semantic concept such as "Geography". The goal of cross-modal retrieval is to, given an image (text) query $\mathbf{I}_q \in \Re^I$ ($\mathbf{T}_q \in \Re^T$) in the test set, search for the closest match in the text (image) space $\Re^T$ ($\Re^I$) of the test set.

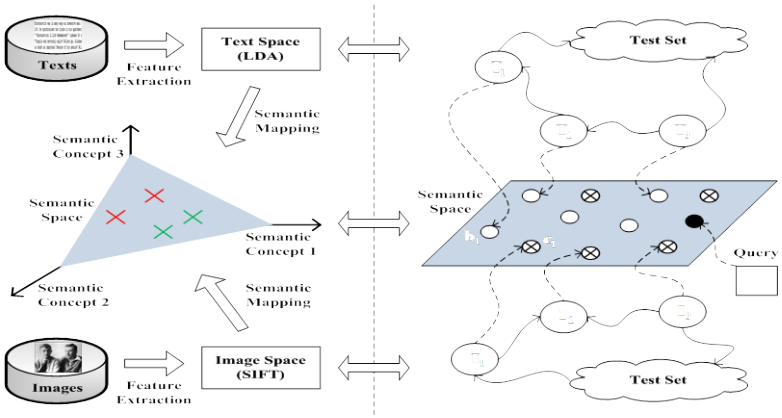

Fig. 1 Framework of the CCSSI model. There are two Markov chains of hidden states for texts and images. The observation at a time step depends on both text and image at that time step

In the retrieval process, our proposed model not only combines content and semantics similarities together but also considers the interaction between images and texts. This model is named as CCSSI for brevity. A general framework of the CCSSI model is shown in Fig. 1, it can be seen that the framework mainly consists of the following two parts.

- Feature representation: The original features derived from images and texts, and the semantic features based on the former.
- Retrieval scheme: Utilizing two Markov chains, we rank the retrieved media data. The objects in each chain can provide content and semantics information.

When a user retrieves something he wants, the goal is to observe consistently the fulfillment of his need during the time of accessing the test set. Therefore, the semantic vector of a query can be treated as an observation which does not change for all time steps. Each text in the test set is treated as a state in one hidden state

space, and the content similarity matrix is represented as the transition probabilities matrix where each element denotes the weight between two states. The semantic vector of each text is described as emission probability which characterizes the mapping mechanism from an original feature into the semantic space. Likewise, the other hidden space for images can be processed. After such a process, an object in one chain can proceed to another object in this chain according to transition probability, but not to any object in the other chain. All of objects in both chains can emit semantic vector at any time step. Images can provide complementary information if texts are retrieved, and vice-versa. The retrieval procedure simultaneously relies on the transition and emission probabilities in the two chains. As a result, the retrieved object at each time step implies the interaction between texts and images, and the neighboring states in retrieval path carry similar information of content and semantics.

Some symbols in the CCSSI model are listed as follows.

- K is the number of time steps, and k is a certain time step whose range is from 1 to K; N is the size of the test set, and $N \geq K$.
- $T(k)$ and $I(k)$ are the indexes of text and image at kth time step, respectively.
- $\mathbf{Q}$, $\mathbf{b}_i$ and $\mathbf{c}_i$ are the semantic vectors of the query, $\mathbf{T}_i$ and $\mathbf{I}_i$, respectively.
- a_{ij} is the probability of moving from $\mathbf{T}_i$ to $\mathbf{T}_j$ in a single step, and e_{ij} is the probability of moving from $\mathbf{I}_i$ to $\mathbf{I}_j$ in a single step.
- $\mathbf{A} = \{a_{ij} \mid i, j = 1, \ldots, N\}$ and $\mathbf{E} = \{e_{ij} \mid i, j = 1, \ldots, N\}$ are the state transition probability matrices of $\mathbf{T}$ and $\mathbf{I}$, respectively.
- $\mathbf{B} = \{\mathbf{b}_i \mid i = 1, \ldots, N\}$ and $\mathbf{C} = \{\mathbf{c}_i \mid i = 1, \ldots, N\}$ are the emission probability matrices of $\mathbf{T}$ and $\mathbf{I}$, respectively.
- $S_i(k)$ is the semantics similarity between $\mathbf{Q}$ and $\mathbf{b}_i$ at kth time step, and $P_i(k)$ is the semantics similarity between $\mathbf{Q}$ and $\mathbf{c}_i$ at kth time step.
- Sim_c and Sim_s are similarity measures in the original feature space and in the semantic space, respectively.

Self-transition is set to zero because that it is useless in the retrieval process. For keeping the probabilistic attribute of the transition, the elements of $\mathbf{A}$ and $\mathbf{E}$ must be, row by row, normalized to one. Therefore, the transition probability matrix is usually not symmetric. The CCSSI model can be specified by the compact notation $\lambda = (\mathbf{A}, \mathbf{E}, \mathbf{B}, \mathbf{C})$.

Random forests [2] can bag an ensemble of decision trees for classification, where the types of randomness contain bagging and random feature selection. To classify a new object, each tree in the forest gives a unit vote and the forest chooses the classification having the most votes. The forest can output the posterior probabilities of concepts:

$$P(l_i \mid \mathbf{x}) = V_i / V \tag{1}$$

where $\mathbf{x}$ is the new object, V the number of decision trees, and V_i the number of trees which vote for the concept l_i.

3 Cross-modal Retrieval Algorithm

In cross-modal retrieval, the modality of a query is different from that of the retrieved results. Based on the CCSSI model, the algorithm of cross-modal retrieval is summarized as follows.

1. Initialize the state transition probability matrix **A** with all zeros, and update its elements:

$$a_{ij} = Sim_c(\mathbf{T}_i, \mathbf{T}_j) \Big/ \sum_{j=1}^{N} Sim_c(\mathbf{T}_i, \mathbf{T}_j), \quad i \neq j \tag{2}$$

   The construction of **E** is similar to **A**.
2. Calculate the emission probability matrices **B** and **C** for texts and images, respectively.
3. If the query object is an image $\mathbf{I}_q$, then we

   a Calculate the semantic similarities $S_i(k)$ and $P_i(k)$:

$$\begin{aligned} S_i(k) &= S_i(k+1) = Sim_s(\mathbf{Q}, \mathbf{b}_i), \quad k \in \{1, \ldots, K\} \\ P_i(k) &= \begin{cases} 1, & k = 1, i = q; \\ 0, & k = 1, i \neq q; \\ Sim_s(\mathbf{Q}, \mathbf{c}_i), & k \geq 2. \end{cases} \end{aligned} \tag{3}$$

   b Initialize the path, and calculate:

$$\left(T(1), I(1)\right) = \arg\max_{1 \leq i \leq N} \max_{1 \leq j \leq N} \left(S_i(1) \cdot a_{qi} \cdot P_j(1)\right) \tag{4}$$

   c For k=2, …, K, let (u, v) = ($T(k-1)$, $I(k-1)$), and grow the paths $T(k)$ and $I(k)$ for each step as follow:

$$\left(T(k), I(k)\right) = \arg\max_{1 \leq i \leq N} \max_{1 \leq j \leq N} \left(S_i(k) \cdot a_{ui} \cdot P_j(k) \cdot e_{vj}\right) \tag{5}$$

   where $i \notin \{T(1), \ldots, T(k-1)\}$ and $j \notin \{I(1), \ldots, I(k-1)\}$.

   d Output the path $\{T(1), \ldots, T(K)\}$.
4. If the query object is a text $\mathbf{T}_q$, then we

   a Calculate the semantic similarities $S_i(k)$ and $P_i(k)$:

$$\begin{aligned} S_i(k) &= \begin{cases} 1, & k = 1, i = q; \\ 0, & k = 1, i \neq q; \\ Sim_s(\mathbf{Q}, \mathbf{b}_i), & k \geq 2. \end{cases} \\ P_i(k) &= P_i(k+1) = Sim_s(\mathbf{Q}, \mathbf{c}_i), \quad k \in \{1, \ldots, K\} \end{aligned} \tag{6}$$

   b Initialize the path, and calculate:

$$\left(I(1),T(1)\right)=\arg\max_{1\le i\le N}\max_{1\le j\le N}\left(P_i(1)\cdot e_{qi}\cdot S_j(1)\right) \tag{7}$$

c For k=2, …, K, let (u, v) = ($I(k-1)$, $T(k-1)$), and grow the paths $I(k)$ and $T(k)$ for each step as follow:

$$\left(I(k),T(k)\right)=\arg\max_{1\le i\le N}\max_{1\le j\le N}\left(P_i(k)\cdot e_{ui}\cdot S_j(k)\cdot a_{vj}\right) \tag{8}$$

where $i \notin \{I(1), \ldots, I(k-1)\}$ and $j \notin \{T(1), \ldots, T(k-1)\}$.

d Output the path $\{I(1), \ldots, I(K)\}$.

At each time step, the algorithm calculates a probabilistic product that relies on two different modalities. The product is just the interaction for generating the whole semantics information. The matrices $\mathbf{A}$, $\mathbf{E}$, $\mathbf{B}$ and $\mathbf{C}$ can be computed beforehand to decrease the computation cost. As can be seen, the time and space complexities are $O(KN^2)$ and $O(KN)$, respectively. So the algorithm can be performed more efficiently by making the value of K smaller.

4 Experiments

In this section, some experimental evaluations of our model are described, which are compared with those of the existing methods for cross-modal retrieval.

To evaluate the performance of the proposed approach, we carry out some experiments on the Wikipedia dataset [8], which is chosen from the Wikipedia's "featured articles". Each article is split into several sections according to its section headings, and the images are assigned to the respective sections according to image position in the article. The final multimedia corpus contains a total of 2866 documents, which are image-text pairs and annotated with a label from the vocabulary of 10 semantic concepts. The dataset is randomly split into a training set of 2173 documents and a test set of 693 documents.

The low-level original features of images and texts are represented by bag-of-words (BOW) model [4] and topic model [1], respectively. Each image is represented using a histogram of a 128-codeword SIFT codebook, and each text is represented using a histogram of 10-topic LDA text model. The precision-recall (PR) curves and mean average precision (MAP) are taken as performance measures. The content similarities of images and texts are measured by histogram intersection and inner product, respectively. The semantic similarities are measured by normalized correlation. We set V= 500 in (1).

The MAP scores of our model are compared with those of the existing methods in Table 1. It can be seen that our method significantly outperforms the compared methods. For instance, CCSSI improves about 10% over CMCP and 26% over SCM, achieving an average MAP score of 0.317. Fig. 2 shows the PR curves of cross-modal retrieval with CM, SM, SCM, and CCSSI. Note that CCSSI gets higher precision at most levels of recall, outperforming the previous methods.

Table 1 Retrieval performance (MAP scores)

Experiment	Image Query	Text Query	Average
CCSSI	0.376	0.257	0.317
CMCP [11]	0.326	0.251	0.289
SCM [8]	0.275	0.226	0.251
SM [8]	0.271	0.212	0.242
CM [8]	0.242	0.198	0.220

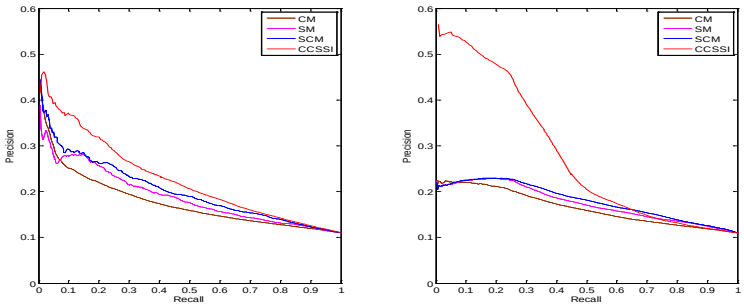

Fig. 2 Precision recall curves for text query on the left and image query on the right

An example of text query is presented in Fig. 3, where the top four retrieval results are shown under CCSSI and SCM. The text query is shown along with the ground truth image. As can be seen, CCSSI can return the images which have the same semantic class ("Geography") as the query text. In particular, the results also have similar visible object, namely building.

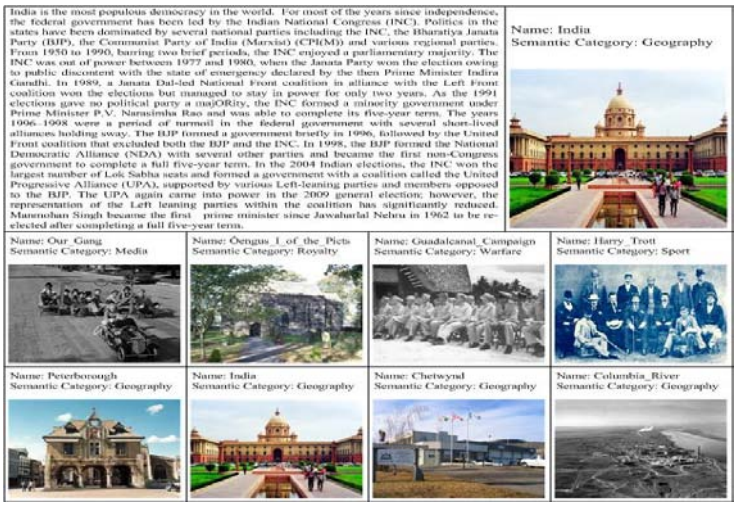

Fig. 3 An example of text query. The query text and ground truth image are shown on the top row; the top four images retrieved by SCM and CCSSI are shown on the second and bottom row, respectively.

5 Conclusions

In this paper, we have presented a graph model for cross-modal retrieval. The CCSSI model combines content and semantics similarities through two Markov chains. Content similarity focuses on the internal structure of each modality, while semantics similarity reflects the semantic correlation between different modalities. Moreover, our model also uses the interaction between different modalities to get the whole semantics information of a multimedia document. The experimental results on the Wikipedia dataset show that the performance of our model significantly outperforms those of previous approaches for cross-modal retrieval.

In the future, we will apply CCSSI to other modalities such as text and audio. We would also like to use some methods such as stacked generalization to better explore the correlation between classes.

References

1. Blei, D., Ng, A., Jordan, M. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022
2. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32
3. Carneiro, G., Chan, A., Moreno, P., Vasconcelos, N. (2007). Supervised learning of semantic classes for image annotation and retrieval. IEEE Trans. Pattern. Anal. Mach. Intell, 29(3), 394-410
4. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C. (2004). Visual categorization with bags of keypoints. In Workshop on Statistical Learning in Computer Vision, ECCV
5. Ghahramani, Z., Jordan, M. (1997). Factorial hidden Markov models. Machine Learning, 29(2-3), 245-273
6. Li, D., Dimitrova, N., Li, M., Sethi, I. K. (2003). Multimedia content processing through cross-modal association. In: Proceedings of ACM International Conference on Multimedia
7. Lmura, J., Fujisawa, T., Harada, T., Kuniyoshi, Y. (2011). Efficient multi-modal retrieval in conceptual space. In: Proceedings of ACM International Conference on Multimedia
8. Rasiwasia, N., Pereira, J. C., Coviello, E., Doyle, G., Lanckriet, G., Levy, R., Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In: Proceedings of ACM International Conference on Multimedia
9. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. IEEE Transactions on Audio, Speech, and Language Processing, 16(2), 467-476
10. Xie, L., Pan, P., Lu, Y. (2013). A semantic model for cross-modal and multi-modal retrieval. In: Proceedings of ACM International Conference on Multimedia Retrieval
11. Zhai, X., Peng, Y., Xiao, J. (2012). Cross-modality correlation propagation for cross-media retrieval. In: Proceedings of ICASSP
12. Zhen. Y., Yeung, D. (2012). A probabilistic model for multimodal hash function learning. In: Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining