Cross-Modal Information Retrieval – A Case Study on Chinese Wikipedia

Yonghui Cong[1], Zengchang Qin[1,⋆], Jing Yu[1], and Tao Wan[2]

[1] Intelligent Computing and Machine Learning Lab
School of ASEE, Beihang University, Beijing, China
zcqin@buaa.edu.cn
[2] Department of Biomedical Engineering, Rutgers University, USA
{yonghuicong,jing.emy.yu}@gmail.com, tao.wan.wan@rutgers.edu

Abstract. Probability models have been used in cross-modal multimedia information retrieval recently by building conjunctive models bridging the text and image components. Previous studies have shown that cross-modal information retrieval system using the topic correlation model (TCM) outperforms state-of-the-art models in English corpus. In this paper, we will focus on the Chinese language, which is different from western languages composed by alphabets. Words and characters will be chosen as the basic structural units of Chinese, respectively. We also set up a test database, named Ch-Wikipedia, in which documents with paired image and text are extracted from Chinese website of Wikipedia. We investigate the problems of retrieving texts (ranked by semantic closeness) given an image query, and vice versa. The capabilities of the TCM model is verified by experiments across the Ch-Wikipedia dataset.

Keywords: Cross-modal information retrieval, topic correlation model (TCM), word-based topics, character-based topics, Ch-Wikipedia.

1 Introduction

The amount of multimedia information on the Internet is growing by an explosive rate in recent years. Much attention has been attracted to build more efficient search engines for multi-modal information including music, videos, texts, images and so on. As we know that, most of popular information retrieval systems we use at present such as Google and Baidu[1] are still uni-modal. Relations between different modalities are not well modeled. Captions or category labels are used to build information manually, which are both time-consuming and laborious. Many techniques have been proposed aimed at bridging information in different modalities [4,6,8,9,11,14]. It has been showed that multi-modal retrieval systems have made significant progress compared to uni-modal approaches [11,12,14].

Previous research in [8] maps the information in different modalities into a higher dimensional semantic space where the similarity is measured. In recent

⋆ Corresponding author.

[1] Websites: www.google.com, www.baidu.com

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 15–26, 2012.

studies, a new probabilistic model was proposed in [14] to investigate mid-level feature correlation between texts and images. In the new model, probabilistic correlations between the mid-level "topics" are considered. Given a query image (text), a SVM classifier can be applied to compute the probability distribution over categories.

Most of the models proposed for cross-modal information retrieval are focused on English corpus and the techniques used in studying English can be easily extended to other alphabetic languages. For example, an English text can be regarded as a collection of words, which are the basic structural units in the majority of western language. However, the basic semantic units in Chinese language are not necessarily to be the Chinese words [13,17]. In this paper, we extend the TCM model to study Chinese language by employing two language models for semantic modeling. Zhao *et al.* [17] first show that the computational evidence that character-based topic models outperform the word-based topic models. The experimental results for both two models will be given in the following sections.

This paper is structured as follows: we introduce topic representation for multimedia contents in Section 2. In Section 3, topic correlation model for cross-modal information retrieval is described in details. We also create a database extracted from the Chinese Wikipedia. In Section 4, a series of experiments are conducted based on this database and the results are given and analyzed in details. The conclusions and the future work are given in the end.

2 Topic Representation

It is a significant issue to represent multimedia information by appropriate features. Low-level features have many limitations such as colors, textures for images and keywords, captions for texts. When considering the multi-modal documents such as Wiki articles, each article contains the semantically similar contents in different modalities (e.g., images and texts). However, the semantic relations cannot be well captured by low level features. Mid-level features such as visual words in the bag-of-features model and latent topics in the topic models [5] can be used to model semantic correlations between contents in different modalities. For example, given a corpus of Wiki articles with paired text and image, we may find that the texts with words like 'sky, blue, sunny' may somehow occurs more often with images containing blue colors. These latent semantic relations can be modeled by correlation between the topics of words and topics of image features (visual words in the bag-of features model). In this way, the semantic gap can be reduced by using mid-level features to model different content modalities.

2.1 Bag-of-Features Model

In this paper, scale invariant feature transform (SIFT) features are used to model the image components in a document. There have been many other low-level image features such as HOG (histograms of oriented gradients), LBP (local binary

pattern). The reason for choosing SIFT is for its effectiveness and stability. SIFT features are invariant to rotation, scaling, translation and small distortions [15]. It has been empirically proven to be one of the most robust among the local invariant feature descriptors with respect to different geometrical changes [10]. The bag-of-features (BoF) model has been getting popular recently. It has two key concepts: local features and codebook. The essential aspect of local feature concept is to extract global image descriptors and represent images as a collection of local properties calculated from a set of small sub-images called patches. Codebook is a way that an image can be represented by a set of local features. The idea is to cluster the feature descriptors of all patches based on a given cluster number and each cluster represents a "visual word" that will be used to form the codebook. After obtaining the codebook, each image can be represented by the BoF histograms of the visual vocabulary of the codebook. The similarity of images can be measured by comparing between the BoF histograms. The Bag-of-features model has been well studied as one of the most effective approaches in image classification [5,15].

2.2 Word-Based Topics and Character-Based Topics

In text modeling, the bag-of-words (BoW) assumption is also well used. For example, by using topic models such as latent Dirichelet allocation (LDA) [1], a text is represented by a mixture of latent topics, and each topic is represented by a probability distribution over vocabulary. Such word topics can be used to model text components in a document [2] or used in other natural language processing applications such as Q&A [7]. Most research on topic models only concern English language. Different from western languages such as English, the morphology of Chinese language is more complex. Characters, instead of words, are the basic structure unit for Chinese language. This has been both discussed in Chinese linguistics [13] and testified using computational model [17,18]. In this research, the word-based and character-based topic models are tested, respectively. Experiments are conducted to show the differences between segmenting Chinese in words and characters. The process of modeling for the text and image components in a Wikipedia document is schematically showed in Fig. 1.

The representation of both images and texts here are not using the low-level features directly. We construct the mid-level representations for modeling the content in a two-level hierarchical structure to make them more robust and abstract. In this paper, we may use the term "topics of features" to represent the visual words in order to highlight the similarity between BoF and the topic model. Because in this research we are interested in the correlations between these topics of different modalities.

3 Topic Correlation Model

Large and heterogeneous collections of images are usually accompanied with noisy texts. The cross-modal retrieval systems are developed for users to be

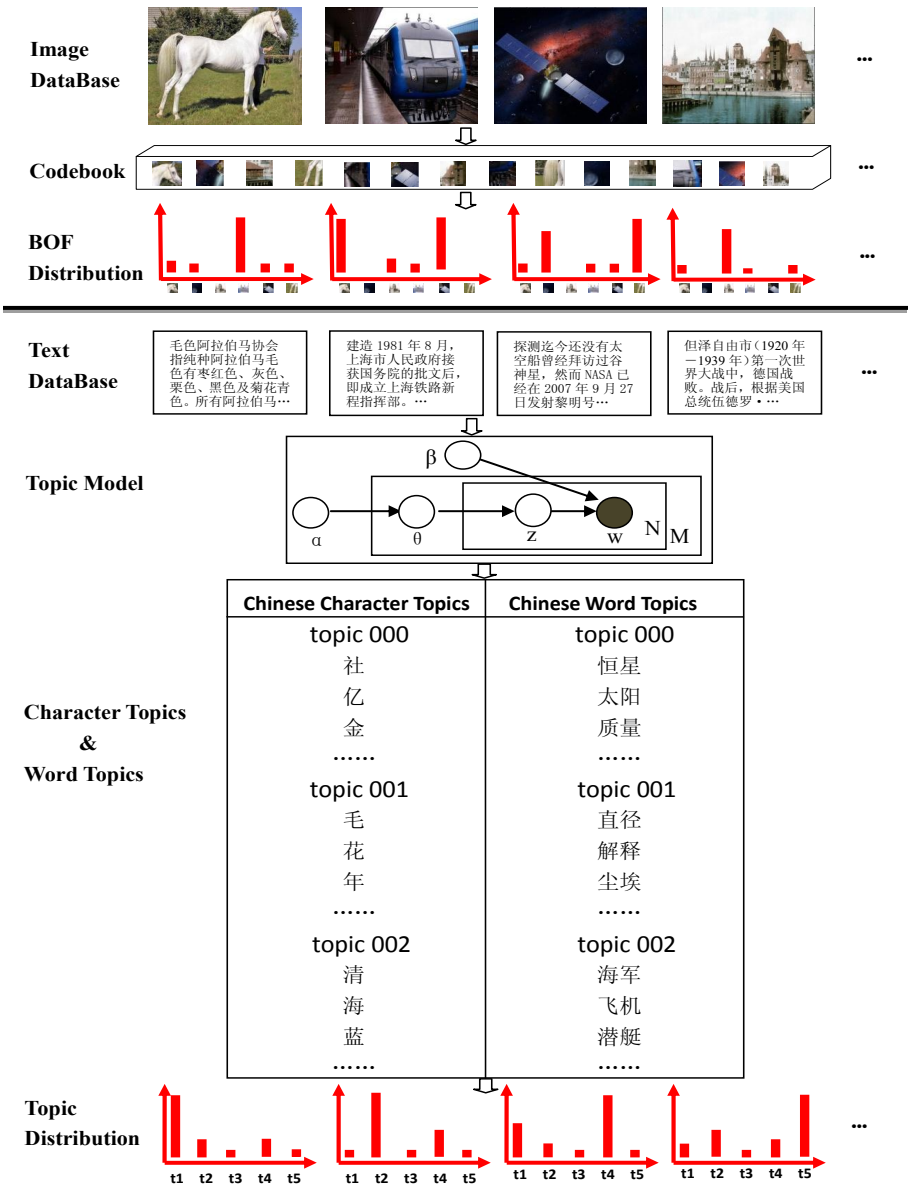

Fig. 1. Process of modeling for the text and image components for Chinese Wikipedia articles. Two modalities (image and text) are separated by a straight line. The upper part shows the procedure of extracting SIFT features from images and build distribution of BOF. The lower part shows how the topic distribution over words is computed using LDA. Chinese texts are modeled by LDAs based on words and characters, respectively. The topic list has showed that text component can be represented by a mixture of latent topics and each topic is a distribution over the vocabulary (either words or characters).

able to browse and search these collections more efficiently. Rasiwasia *et al.* [8] has demonstrated the benefits of joint model for text and image components by mapping both into one high dimensional semantic space. In recent years, the statistical correlation methods have attracted much attention. Probability models for matching words and pictures has been used to segment images with associated text [3].

The underlying relation between topics may reflect the correlation between image and text components. In [14], images are represented by distributions on topics of features and texts are represented by distributions on topics of words. In the topic correlation model, naive probabilistic correlations between image and text features are considered. Given a collection of documents is defined by $\mathbf{D} = [D_1, D_2, ..., D_K]$. We assume that there is only one-to-one mapping between image and text is acquiescent for research purpose, e.g, $D_k = [I_k, TX_k]$. where $D_k \in \mathbf{D}$, I_k and TX_k is the related image and related text in D_k respectively. Given a query I_q (or TX_q), how to find a document $D_j \in \mathbf{D}$ that has the most semantically related texts (images).

We define that $\mathbf{V} = [V_1, V_2, ..., V_M]$ is the set of visual words in the codebook where M is the codebook size. $\mathbf{T} = [T_1, T_2, ..., T_N]$ is the set of topics and N is a predefined number of topics. For a visual word V_i and a topic T_j, the underlying probabilistic relation can be computed on the training document $\mathbf{D}$,

$$P(V_i|T_j) = \sum_{k=1}^{K} P(V_i|I_k)P(I_k|TX_k)P(TX_k|T_j) \tag{1}$$

where $P(V_i|I_k)$ is the BoF distribution over V_i of the image I_k. Since the image I_k and the text TX_k appear in the same document D_k, then

$$P(I_k|TX_k) = P(TX_k|I_k) = 1$$

For the third term $P(TX_k|T_j)$, according to the Bayes theorem, we can obtain

$$P(TX_k|T_j) = \frac{P(T_j|TX_k)P(TX_k)}{\sum_{k=1}^{K} P(T_j|TX_k)P(TX_k)} \tag{2}$$

where $P(TX_k)$ is the prior probability of the text component in document D_k. Without any information on each document, we use the uniform distribution as the prior according to the principle of maximum entropy. Formally,

$$P(TX_k) = P(I_k) = \frac{1}{K} \tag{3}$$

Similarly, the likelihood of topic T_j given a visual word V_i is evaluated by

$$P(T_j|V_i) = \sum_k P(T_j|TX_k)P(TX_k|I_k)P(I_k|V_i) \tag{4}$$

where $P(T_j|TX_k)$ is the topic distribution over T_j of the text TX_k. According to the Bayes theorem,

$$P(I_k|V_i) = \frac{P(V_i|I_k)P(I_k)}{\sum_{k=1}^{K} P(V_i|I_k)P(I_k)} \tag{5}$$

When the category information is available, this information can be applied to find better matching patterns between image and text components. In this framework, SVM is used as such a classifier [14]. Given a query text (image), text classifier (image classifier) is applied to compute its probability distribution over categories. C_i denotes the ith category given a set of categories $\mathbf{C} = [C_1, C_2, ..., C_n]$. Then the probability of the image I_k given a query text TX_q is computed by summing up the conditional probabilities across all the categories. Formally,

$$P(I_k|TX_q) = \sum_i P(I_k|C_i)P(C_i|TX_q) \tag{6}$$

Based on the Bayes theorem, we can obtain

$$P(I_k|C_i) = \frac{P(C_i|I_k)P(I_k)}{\sum_k P(C_i|I_k)P(I_k)} \tag{7}$$

Similarly, given an image query, the probability of a text component is computed by

$$P(TX_k|I_q) = \sum_i P(TX_k|C_i)P(C_i|I_q) \tag{8}$$

where

$$P(TX_k|C_i) = \frac{P(C_i|TX_k)P(TX_k)}{\sum_k P(C_i|TX_k)P(TX_k)} \tag{9}$$

The values of $P(C_i|I_k)$ and $P(C_i|TX_k)$ can be obtained by the predictions from trained SVM classifiers. And these two values are not necessarily the same as the classifiers are trained individually based on the contents of different modalities.

Table 1. Summary of the Ch-Wikipedia Dataset, the articles are extracted from Chinese Wikipedia website

Category	Training	Test	Total
Culture	285	71	356
Biology & Medicine	327	82	409
Natural Science	279	70	349
Geography	374	93	467
History	424	106	530
Traffic	156	39	195
Warfare & Military	206	52	258
Scholar & Occupational Figures	145	36	181
Political & Military Figures	286	72	358

4 Experimental Studies

In this section, we conduct a series of experiments to demonstrate the effectiveness of the TCM on a new database in Chinese. In these experiments, 20% of the documents were randomly chosen as the test set and the remaining as training.

4.1 Database and Preprocessing

Since there is no well accepted image-text paired Chinese corpus for cross-modal retrieval research. We create a database by our own, which is named Ch-Wikipedia Dataset[2]. It consists 3103 documents of paired texts and images from 9 categories listed in Table 1. The documents in this corpus are crawled from a collection of articles in Chinese Wikipedia, which is one of the biggest Internet information websites in Chinese language. There are 20 classes in original corpus that cover literature, media, sports, politics and other topics. Each article is split into some parts by section heading. The texts containing less than 100 Chinese characters will be ignored in our research. The first image associated with a text is chosen as its related image and other texts without images are removed. Topics of some classes are similar, such as "humanities" and "culture & society", which can be classified into one bigger category "culture". On the other hand, some classes with less than 150 documents will be abandoned as well if they can not be merged to a larger category. We list the final categories in Table 1 after merging similar ones manually.

Table 2. Accuracy rate of SVM classifiers with different kernels

Database	*Linear*	*Polynomial*	*RBF*	*Sigmoid*
Training images	0.356	0.203	0.316	0.303
Test images	0.309	0.198	0.308	310
Word-based training texts	0.642	0.548	0.663	0.627
Word-based test texts	0.654	0.385	0.665	0.668
Character-based training texts	0.820	0.639	0.811	0.776
Character-based test texts	0.712	0.533	0.721	0.704
Average	**0.582**	0.418	0.576	0.564

A score function is introduced to evaluate the likelihood of an image I_k given a text query TX_q by $S(I_k) = P(I_k|TX_q)$ and used to get a ranked list of returned data. We test the probabilistic correlation model on Ch-wikipedia dataset for the following tasks: (1) obtaining a ranked list of texts from the training database given a query image, (2) obtaining a ranked list of images from the training database given a query text. The mean average precision (MAP) is applied to measure the performance[3]. The word-based and character-based topic models are trained separately [17]. When we do the Chinese character-based topic modeling, we remove the rare characters that appear less than 3 times across the whole corpus and the terms appearing in over 50% of the documents which we consider as stop words [17]. When modeling for word-based topics, we first get a "stopwords list" from network[4], and ignore them for building the wordlist.

[2] `http://icmll.buaa.edu.cn/zh_wikipedia`

[3] `http://en.wikipedia.org/wiki/Information_retrieval`

[4] The stopword list of Chinese language can be downloaded from: `http://www.byywee.com/page/M0/S639/639550.html`

Table 3. The MAP value with different topic numbers for TCM. The best performance is highlighted.

Topic	**Topic correlation model**		
Number	Word-based	Character-based	Average
10	0.247	0.262	0.255
50	0.266	0.275	0.271
100	0.263	0.292	0.278
200	0.261	0.30	0.281
300	0.259	0.304	0.282
500	0.261	0.306	**0.284**
1000	0.264	0.301	0.282

Words appear less than 3 times through all documents are removed as well. After the above preprocessing we get 21240 unique Chinese words and 3419 unique Chinese characters. We set the topic number to 100 for the LDA and the same visual words number for the BoF model. It's showed in Table 2 that the SVM classifiers with linear kernel have the highest accuracy rate in average. So that in the following experiments, we use the SVM with linear kernel for category prediction.

4.2 Experimental Results

We vary the topic number and the empirical evaluations are shown in Fig. 2. The exact accuracy values are shown in Table 3. When the number of topics was set to 500 for TCM, we get the best performance. Thus, appropriate numbers for topics will boost the accuracy. The MAP for the TCM based on Chinese words and characters are showed in Fig. 3. Agreed with previous research in [16,17], the character-based topic model has a better performance than word-based model for any given topic number. Along with the increase of topic number from 10 to 1000, the MAP of the the word-based model has a tendency of increasing with tiny fluctuation. As for character-based model, it keeps increasing and the MAP is up to 31%, improved by 5% in average compared with the word-based. One possible reason is that the size of word vocabulary is much larger than the size of character vocabulary.

Fig. 4 and 5 show two examples of cross-modal retrieval. For Fig. 4, the given image on the top left is a picture of train and railway. The top three outputs of the TCM are texts about airport, railway system, and airport hotel which are all semantically close to travel and train. For Fig. 5, the system is given a text describe nepenthe (the corresponding image is also shown at below). The top three outputs are all nepenthe images just in different kinds.

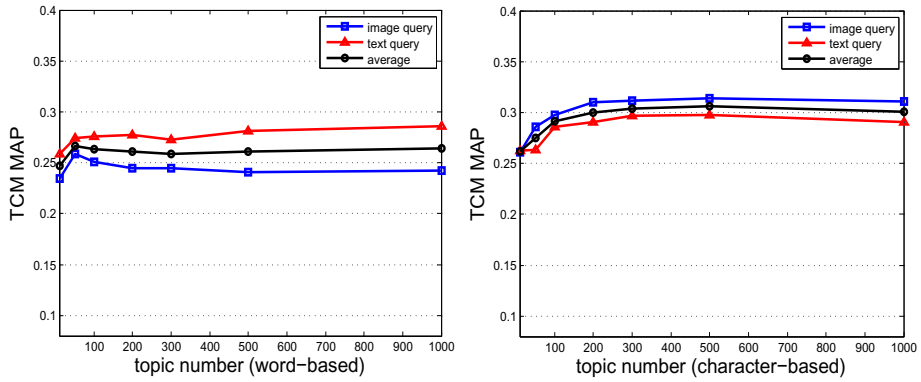

Fig. 2. The performance with increasing topic numbers. It is obvious that the model trained by character-based topic model achieved better performance than word-based one. The average retrieval accuracy for the character-based TCM model is about 31%, improved by 5% compared to the word-based model.

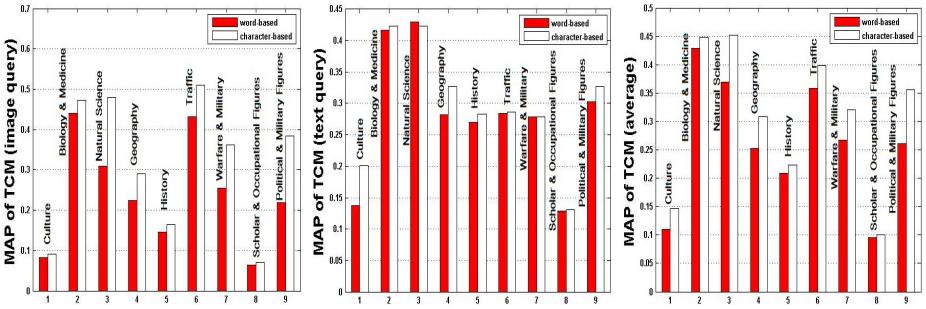

Fig. 3. Comparisons of retrieval results in each category in MAP. From the left to the right: (1) results of image retrieval; (2) results of text retrieval; (3) average results of retrieval.

Since there is no previous research about cross-modal information retrieval in Chinese language, so we cannot compare our results directly to other retrieval systems. However, pervious research [14] has shown that the TCM outperforms the other state-of-the-art cross-modal retrieval models on English Wikipedia corpus [9]. The average accuracy is up to 27%.

However, what should not be ignored is that the larger the topic number, the more time is spent to train topic distribution. The computational time is doubled when the given number of topics are doubled. But the performance can only be improved slightly. Thus, considering the time complexity, we may not like to set the number of topics too large and we can still obtain fair performance.

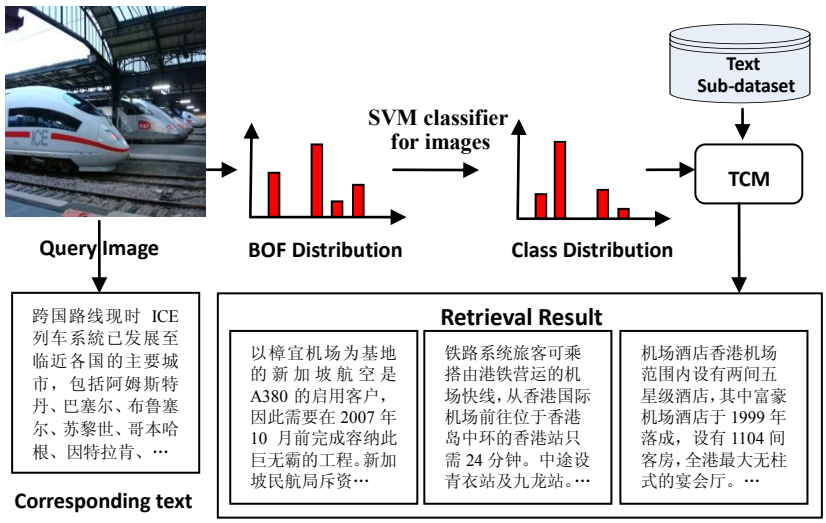

Fig. 4. An example of image query: given an image, a list of semantically related Chinese texts are returned using the topic correlation model

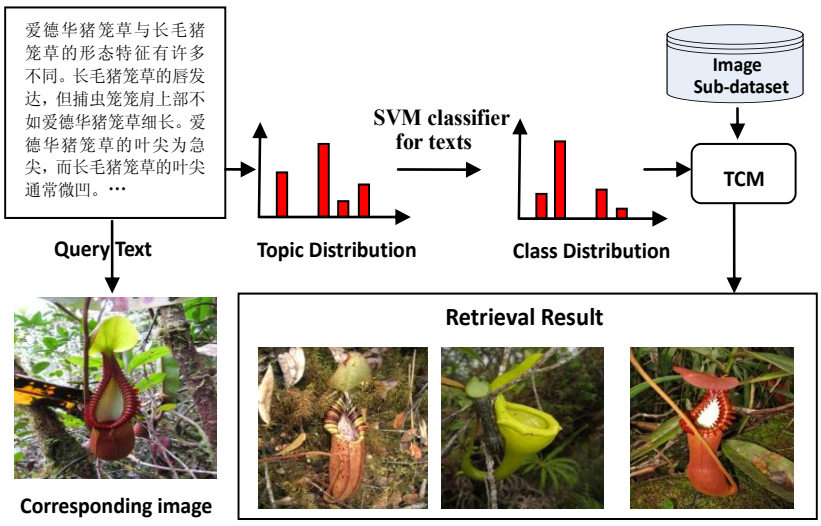

Fig. 5. An example of text query: give the description of nepenthe in Chinese, a list of images related to nepenthe are returned using the topic correlation model

5 Conclusions and Future Work

In this paper, we use a probabilistic model to study the correlations between mid-level features in different modalities. Topic correlation model is used for cross-modal information retrieval on a Chinese database Ch-Wikipedia. Comprehensive experimental results are presented to show the effectiveness. We also use the word-based and character-based topic model for text modeling. Empirical results agree with previous research that the character-based model outperforms the word-based model. We have achieved a good performance for cross-modal information retrieval in Chinese language.

In this research, the topic number is fixed for both image and text components. It is not necessarily true. The optimal topic numbers could be depend on image and text properties in the training set. How to independently find appropriate number of topics for image and text may be an interesting research problem. We can also explore more in Chinese language modeling to boost the accuracy of the TCM. For example, we can consider the word-character relations to build more accurate language model [18]. Other future work can be focused on weakening the noises when we build the correlation between different modalities.

Acknowledgments. This work is partially funded by the NCET Program of MOE, the SRF for ROCS, the Fundamental Research Funds for the Central Universities and Graduate Innovative Practice Fund of BUAA.

References

1. Blei, D.M., Ng, A., Jordan, M.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
2. Blei, D.M., Lafferty, J.D.: Topic models. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series (2009)
3. Barnard, K., Duygulu, P., Forsyth, D., Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research 3, 1107–1135 (2003)
4. Jeon, J., Lavreko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: ACM SIGIR Conf. Research and Development in Information Retrieval, New York, pp. 119–126 (2003)
5. Jiang, Y., Ngo, C., Yang, J.: Towards optimal Bag-of-features for object categorization and semantic video retrieval. In: CIVR, pp. 494–501 (2007)
6. Metzler, D., Manmatha, R.: An inference network approach to image retrieval. In: Image and Video Retrieval, pp. 42–50 (2005)
7. Qin, Z., Thint, M., Huang, Z.: Ranking Answers by Hierarchical Topic Models. In: Chien, B.-C., Hong, T.-P., Chen, S.-M., Ali, M. (eds.) IEA/AIE 2009. LNCS, vol. 5579, pp. 103–112. Springer, Heidelberg (2009)
8. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: ACM Multimedia (MM), pp. 251–260 (2010)
9. Rasiwasia, N., Moreno, P., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Transactions on Multimedia 9(5), 923–938 (2007)

10. Schmid, C., Mikolajczyk, K.: A performance evaluation of local descriptors. In: ICPR, vol. 2, pp. 257–263 (2003)
11. Snoek, C., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. Multimedia Tools and Applications 25(1), 5–35 (2005)
12. Westerveld, T.: Probabilistic multimedia retrieval. ACM 25, 438 (2002)
13. Xu, T.Q.: Fundamental structural principles of Chinese semantic syntax in terms of Chinese Characters. Applied Linguistics 1, 3–13 (2001) (in Chinese)
14. Yu, J., Cong, Y., Qin, Z., Wan, T.: Cross-modal topic correlations for multimedia retrieval. To appear in ICPR (2012)
15. Yuan, X., Yu, J., Qin, Z., Wan, T.: A SIFT-LBP image retrieval model based on bag-of features. In: International Conference on Image Processing (ICIP), pp. 1061–1064 (2011)
16. Zhang, Y., Qin, Z.: A topic model of observing Chinese characters. In: Proceedings of the International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 2, pp. 7–10 (2010)
17. Zhao, Q., Qin, Z., Wan, T.: What Is the Basic Semantic Unit of Chinese Language? A Computational Approach Based on Topic Models. In: Kanazawa, M., Kornai, A., Kracht, M., Seki, H. (eds.) MOL 12. LNCS, vol. 6878, pp. 143–157. Springer, Heidelberg (2011)
18. Zhao, Q., Qin, Z., Wan, T.: Topic Modeling of Chinese Language Using Character-Word Relations. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 139–147. Springer, Heidelberg (2011)