Cross-Media Retrieval Based on Canonical Correlation Analysis and Decision Tree

Hongye Duan, Meng Wang, Peiqing Zou
Tus College of Digit, Guangxi University of Science and Technology, LiuZhou 545006, China
mwang007@gxust.edu.cn
fiora1027@163.com

Abstract. The retrieval of approximate data from massive multimedia data is the key of computer science research. With the explosive growth of data scale, data retrieval needs to face the overlapping massive multimodal data called “dimension disaster”, and the traditional retrieval methods appeared to be inefficient. The cross-media retrieval with canonical correlation analysis (CCA) is a kind of way to map different media features to the largest correlation isomorphism subspace and then compare the similarity between cross-media data in this subspace. However CCA is a linear model and it is difficult to map the lower data to the higher semantic. This paper proposed a cross-media retrieval method based on CCA and Decision Tree to solve the problem. CCA was used to depict the correlation between image and text feature, Decision Tree was used to approach to feedback and repeatedly adjust the correlation. Experiments on the Wikipedia text-image datasets verified that the Tree-CM model can mine the complex correlation between cross-media data and has better performance than other state-of-the-art models.

1. Introduction
With the rapid development of computer technology and smartphones, massive multi-dimensional data is stored and collected. Information explosion followed, which creates difficulties in data mining. Among them, the analysis and processing of multimedia data has become the keystone and difficulty in deep learning, artificial intelligence, computer vision and other fields. The traditional information retrieval method is difficult to satisfy people's demand for mining multimedia information from massive data. For example, when a user enters the keyword "horse" as a query, not just want to get the text description of horse, also hope to get multimedia information associated with it, such as pictures, sounds and videos and so on. CMR (Cross-Media Retrieval, CMR)[1] is intended to address the interrelated problems between the different forms of data, to achieve mutual retrieve between different multimedia data.

However, for the similar high-level semantics of multimedia data, their low-level features are quite different. How to mine the correlation between text, language, image, video and other multimedia data in order to cross the semantic gap between multimedia data, that is the importance of cross-media retrieval research, and the "semantic gap" also restricts the development of cross-media retrieval technology[2].

For low-level features of multimedia information analysis, early CMR done a lot of research in this area, so that they ignored the relationship between the low-level features of multimedia information and the existence of high-level semantics. This paper started from the relationship...
between the low-level features and the high-level semantic, proposed a cross-media information retrieval method based on canonical correlation analysis and decision tree algorithm. This method firstly mapped multimedia data to the same subspace through canonical correlation analysis, established the mapping relationship between image data and text data to realize the preliminary cross-media retrieval function, and further corrected the error of canonical correlation mapping through decision tree to improve the retrieval accuracy. The finally results in the Wikipedia text image data set showed that, the proposed method based on canonical correlation analysis and decision tree has a better improvement in the accuracy of retrieval compared with the traditional cross-media retrieval methods.

2. Related Work

Currently, researchers have proposed a variety of methods to solve the problem of feature heterogeneity of multimedia data, most studies have focused on learning method based on common space. Such methods projected the heterogeneous description of multimedia into a subspace of semantic sharing, thus mining the semantic complementarity between multimedia, mining the semantic features shared between multimedia descriptions, that is, semantic consistency.

Hong Zhang et al.[3] analyzed the correlation between the feature matrix of multimedia data in the process of subspace mapping, proved and found the maximum correlation between multimedia data objects by using canonical correlation analysis algorithm. Liang Dong et al.[4] in order to solve the limitation of underlying semantics, they put forward a clustering method based on the idea of logarithm retrieval, and made a cross-media retrieval method based on correlation analysis in time basis. Tao Jiang et al.[5] proposed a cross media retrieval based on the similarity of the model and used fuzzy transformation technology to measure the correlation between visual features and text features and information similarity. At present, there are still many scholars committed to applying CCA algorithm to cross media retrieval with a wider range and more complex data mode. Nikhil rasiwasia et al.[6] found that CCA algorithm can effectively solve the problem of semantic gap between text data and image data, and based on CCA modeling, it was found that this model is more effective in feature space with higher level of abstraction. Based on CCA, Jile Zhou et al.[7] proposed a latent semantic sparse hashing method, which used sparse coding and matrix decomposition to perform similarity retrieval across media, hash coding is used to capture the dominant structure of the image, learning latent concepts from texts through matrix decomposition.

The above cross-media retrieval methods based on CCA attempted to integrate the shallow semantics with the deep semantics, but the retrieval accuracy is still not good. This paper analyzed the mapping between low-level features and high-level semantic to further mine the deep relations among multimedia data. A method based on canonical correlation analysis and decision tree was proposed and the modified learning method also applied to improve the retrieval accuracy.

3. Cross—Media Retrieval Model

In this section we proposed a method named Tree-CM which based on canonical correlation analysis algorithm and decision tree to deal with the cross-media retrieval task. Firstly, we put the processed image feature values and txt feature values in the same isomorphism subspace to map different media features to the largest correlation. Secondly, we used Decision Tree approach to feedback, repeatedly adjust the correlation. Finally, cross-media retrieval can be realized in the isomorphism subspace, on both image-to-text and text-to-image query tasks.

This method is described as follows:

1. Extracting feature values from relevant text-image pairs. The representation of text is derived from a latent Dirichlet allocation (LDA)[8] model, text documents are represented by their topic assignment probability distributions. Image representation is based on the popular scale invariant feature transformation (SIFT)[9] and the extracted features are clustered by K-means to form a 128 dimensional matrix.

2. Mapping feature vectors of text and image to the same isomorphism subspace by CCA.
3. Dividing the image-text pairs dataset into training set and test set and then a decision tree classifier was trained.

4. After the mapping spaces for images data and text data are established respectively. Comparing the similarity between cross-media data in the the largest correlation isomorphism subspace, and then using queried multi-media data to correct the query result by Decision tree model in order to get the final results.

3.1. Correlation Matching

The given mapping relation between two different data space: \( \text{Map} : X \rightarrow Y \) was reversible. It means that the nearest query \( \text{Map}(q) \) can be found in \( X \) when a query task \( q \in X \) was given. Similarly, the nearest query \( \text{Map}^{-1}(q) \) can be found in \( Y \) when a query task \( q \in Y \) was given. After mapping relationship \( \text{Map} \) was established, we mapped the processed image features and text features to the corresponding subspace \( s \) and \( t \). The features of image and text can be corresponded, the overall representation relation can be reformulated as:

\[
\text{Map} : s \rightarrow t
\]  

The image and text space were mapped into two reversible isomorphic subspaces \( s \) and \( t \), \( \text{Map} : s \rightarrow t \). The nearest image data \( Y \) will be found in the subspace \( t \) and the nearest text data \( X \) will be found in the subspace \( s \) as well.

3.2. Subspace-based Related Query

The mapping relationship on \( s \) and \( t \) is reversible, which helps to form a compact and efficient expression. Vector \( \rho_i \) was the spatial coordinate mapped from image space \( X \) to maximum subspace \( s \), and vector \( \rho_T \) was the spatial coordinate mapped from image space \( Y \) to maximum subspace \( t \) in this expression. Given a query text \( T \) and its projection \( \rho_T \), \( R \) best matched images will be found by minimizing the formula \( \text{dis}(\rho_i, \rho_T) \).

\[
\text{Dis}(I, T) = \text{dis}(\rho_i, \rho_T)
\]  

Similarly, given a query image \( I \) and its projection \( \rho_I \), \( R \) best matched texts will also be found by minimizing the formula \( \text{dis}(\rho_I, \rho_T) \).

3.3. Tree-CM model

This paper proposed a method based on the CCA and Decision Tree cross modal multimedia retrieval method named Tree-CM(Decision Tree-Correlation Matching). Using CCA methods to depict the correlation between image and text feature and Decision Tree approach to feedback, repeatedly adjust the correlation.

The Tree-CM method acted on the maximal correlation subspace of canonical correlation rather than the initial image and text feature space. Therefore, the distribution of subspace has a direct impact on the efficiency of the feedback method. Since the subspace is based on data correlation mapping, which has a clustering effect on images and texts. Therefore, we may think that data with similar semantics and the same modality will be distributed in more concentrated areas, which provided a good basis for our decision tree classification.

In order to achieve a better correction, we make an adjustment on formula (3):

\[
\text{Dis}(I, T) = \text{dis}(\rho_i, \rho_T) + c(\rho_i, \rho_T)
\]
\( c(\rho_t, \rho_I) \) is a correction factor, the distance that the error mapping by the training set needs to be modified after feedback. The initial value of this distance is 0. If the mapping relationship between text data and image data is error, the value will correct to positive value and increases in ascending order of distance.

Set \( I^q \) as the image to be queried, and use the trained decision tree model to judge the returned text to get \( N \) negative cases, the correction process is as follows:

1. \( \forall n_i \in N, \) get adjacented \( k \) neighbor of \( n_i \) in text dataset \( Z = \{z_1, \ldots, z_j, \ldots, z_k\} \), and arranged in ascending order of distance;
2. Let \( c(I^q, n_i) = \alpha(\alpha > 0) \), modify the corresponding correction factor \( s \) value of each element in the set \( Z \) in an isometric manner in turn:
   \[
   c(I^q, n_i) = \alpha - j \times \frac{\alpha}{k}
   \]  
   (4)
3. The texts that similar to the query example \( I^q \) is recalculated according to formula (3) and returned as a new query result.

The Tree-CM method process is shown as follows:

1. Paired image features and text features were input into the decision tree as training sets, and marking the correctness of mapping results, 1 if the mapping is correct, otherwise mark 0.
2. Putting the test set data into the decision tree model, and also output the mapping label.
3. After getting the label, the mapping distance of the original image and text can be modified by combining the results of label feedback and formula (2), and the modified distance \( D_{Is} \) corresponds to formula (3). The adjusted distance is used to modify the mapping relationship between image and text in the test set. After repeated for many times, the nearest \( R \) retrieval results are obtained in the new distance.

**Fig.1 Framework of the proposed model**

4. **Experiments**

Experiments are conducted on English Wikipedia datasets. This dataset consists of 2866 image-text pairs labeled with 10 semantic classes in total. For each image-text pair, we extract 10-dimensional LDA textual features to represent the text view and 128-dimensional SIFT visual features to represent the image view. In this section, experiments were carried out to verify the validity of the method. The dataset was partitioned into two parts, one for training (2173 pairs) and the other for testing (693 pairs).

In the experiments, we use Euclidean Distance to measure the similarity between heterogeneous data. And we used the mean average precision (MAP) [10] of all classes, MAP of each class and to evaluate the model. Given a set of queries, MAP refers to the average precision (AP) of all queries. We compared our proposed Tree-CM with the following approaches based on Canonical Correlation
Analysis including m-CCA, CCA, DCCA, KCCA and MixPCCA, on both image-to-text and text-to-image tasks. The MAP scores of all the methods on the dataset are shown in Table 1.

| Method      | im2txt | txt2im | Average |
|-------------|--------|--------|---------|
| m-CCA       | 0.2108 | 0.1857 | 0.1983  |
| CCA         | 0.1872 | 0.216  | 0.2016  |
| DCCA        | 0.2724 | 0.2632 | 0.2678  |
| KCCA        | 0.2873 | 0.2791 | 0.2832  |
| MixPCCA     | 0.2954 | 0.3132 | 0.3043  |
| Tree-CM     | 0.2912 | 0.3316 | 0.3114  |

It showed that the MixPCCA method achieved better performances than other compared methods in terms of image-to-text task, the MAP of Tree-CM model is second only to this method. It is also clear that Tree-CM achieved the best performance on the text-to-image query task. In addition, Tree-CM also outperformed the compared methods in terms of the average MAP. Accordingly, the Tree-CM method proposed in this paper has better retrieval accuracy for both two kinds of retrieval.

In addition, we conducted extensive experiments to compare Tree-CM model with other models which are well cited work in this field including CCA, PLS, TCM, GMMFA, GMLDA. The MAP scores of the methods on the dataset are shown in Table 2.

| Method     | im2txt | txt2im | Average |
|------------|--------|--------|---------|
| CCA        | 0.1872 | 0.216  | 0.2016  |
| PLS        | 0.1633 | 0.2402 | 0.2032  |
| SCM        | 0.2930 | 0.2320 | 0.2660  |
| GMMFA      | 0.2885 | 0.3159 | 0.3022  |
| GMLDA      | 0.2964 | 0.3155 | 0.3060  |
| Tree-CM    | 0.2912 | 0.3316 | 0.3114  |

We can clearly draw the same conclusion that Tree-CM method has a highest accuracy in text-to-image query task, but presented a moderate level in text-to-image query task. In conclusion, Tree-CM method reached the highest precision in all compared methods. That also reflected the superiority of it.

Besides, we analyzed the different retrieval accuracy of different cross-media retrieval methods in 10 classes as shown in Figure 3. In the way of image-to-text tasks, there showed a big difference between different classes, among which the MAP in biology, geography and sport categories was significantly better than that in others, while art class has the lowest retrieval accuracy and lower than other classes.
The same conclusion in text-to-image tasks can be drawn from Fig.2. Combining Fig.2, it can be clearly seen that for different classes, the retrieval accuracy is not the same, among which biology, geography, and sport have higher retrieval accuracy in various methods, but art has the lowest. This may be due to the fact that the high-level semantics of art class is more complex than others, and it is also more difficult to extract feature, and difficult to carry out effective learning and recognition as well. It also reflected the importance of the high-level semantics among multimedia in the process of crossmedia retrieval.

5. Conclusion

In this paper, we proposed a novel cross-modal retrieval model named as Tree-CM that took both irregular graph-structured textual representations and regular vector-structured visual representations into consideration to jointly learn coupled feature and common latent semantic space. LDA algorithm was used to extract text feature and SIFT was used to extract image features.

Mapping different media features to the largest correlation isomorphism subspace through the canonical correlation analysis, comparing the similarity between cross-media data in the subspace. To deal with the problem of information imbalance between different modalities, Decision Tree approach to feedback repeatedly adjust the correlation. Extensive experiments on Wikipedia text image dataset verified that our method considerably outperformed among the state-of-the-art models.

Acknowledgments

This work was financially supported by the Liuzhou Science and technology project (2018DH10503).

References

[1] Hong Zhang, Fei Wu, Yueting Zuan. Research on cross-media correlation reasoning and retrieval [J]. Journal of Computer Research and Development, 2008, 45(5): 869-76.

[2] Ping Cai, Zhiqiang Wang, Xianghua Fu. Research on Semantic-based Cross-media Information Retrieval Technology [J]. Microelectronics Computer, 2010, 27(3): 102-105.

[3] Lei Zhang, Yao Zhao, Zhenfeng Zhu. Research progress of cross-media semantic sharing subspace learning [J]. Chinese Journal of Computers, 2017, 40(6): 1394-421.

[4] Dong Liang, Honghao Yang; Changqiao Xu. A cross-media retrieval method based on correlation analysis and logarithmic retrieval clustering [J]. China Sciencepaper, 2018, 13(14): 1590-5.

[5] TAN A H. Discovering Image-Text Associations for Cross-Media Web Information Fusion; proceedings of the European Conference on Principle and Practice of Knowledge Discovery in Databases, F, 2006 [C].

[6] VASCONCELOS N. A New Approach to Cross-Modal Multimedia Retrieval; proceedings of the Proceedings of the 18th International Conference on Multimedea 2010, Firenze, Italy, October 25-29, 2010, F, 2010 [C].

[7] ZHOU J, DING G, GUO Y. Latent semantic sparse hashing for cross-modal similarity search [J]. Research and development in information retrieval, 2014.

[8] LAFFERTY J. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3(993-1022).

[9] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, November 2004, Volume 60 Issue 2, pp. 91-110

[10] YUE Y, FINLEY T, RADLINSKI F, et al. A support vector method for optimizing average precision [J]. Research and development in information retrieval, 2007.