Method of subspace cross-modal retrieval based on hypergraph ranking

Lingyun Huang*
School of Management, Jinan University, Guangzhou, Guangdong, China

ABSTRACT

Based on the research of subspace cross-modal retrieval method based on high-level semantic correlation, a subspace cross-modal retrieval method based on hypergraph ranking is proposed. The method is mainly divided into two parts: subspace mapping and hypergraph sorting. In the subspace mapping part, the high-order semantic correlation and sparse structural constraints of multi-modal data are considered. The hypergraph sorting part uses hypergraphs to describe multiple targets. The characteristics of multiple correlations of multimodal data are comprehensively considered, and the similarity between multimodal data is comprehensively considered, and the semantic correlation between multimodal data is further mined. Combining hypergraph sorting and subspace learning to further mine semantic associations between cross-modal data can further improve cross-modal retrieval accuracy.

Keywords: Cross-modal retrieval, subspace, higher-order semantics, hypergraph sorting

1. INTRODUCTION

In the context of the Internet era, information resources are gradually showing the characteristics of modal diversification and complex structure. When humans are exposed to new information, they perceive information through the co-stimulation and complementation of organs such as “eyes” (vision) and “ears” (hearing)[1]. People can disseminate data in multiple forms (i.e. modalities) anytime, anywhere[2]. As shown in Figure 1, the data modalities represented as tigers can be presented in different modalities such as text, images, audio, and video, and the information between multiple modalities is complementary or cross-correlated[3-8]. Compared with the traditional retrieval method between single modalities, data retrieval between multiple modalities, that is, retrieval between cross-modalities, is more in line with people’s needs and has important research and application value.

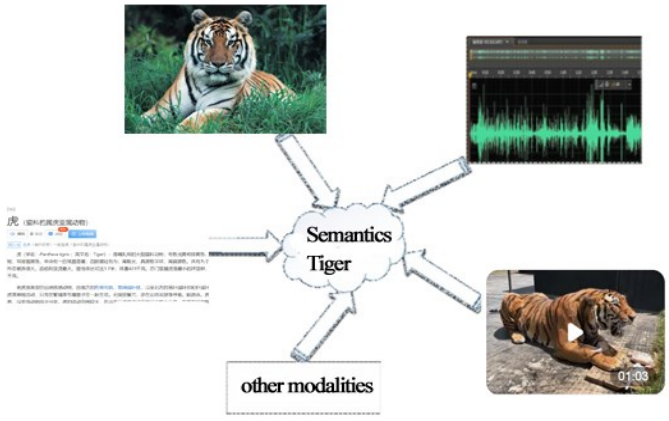

Figure 1. The semantic correlation of multimodal.

Cross-modal retrieval integrates data forms of text, image, audio, video and other modalities, and cross-retrieval between various data modalities is a major research hotspot in information retrieval at present. As shown in Figure 2, it is an example of cross-modal retrieval: input the text “forest” to return a picture of a forest, or enter a picture of a forest to

* huanglinwan@163.com

Third International Conference on Computer Science and Communication Technology (ICCSCT 2022)
edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12506, 1250648

return a text description of the forest, or enter a paragraph Voice description of the forest, return related text description, etc. The form of cross-modal retrieval is that the input is one modal information, and the return is the data description of another one or more modalities. Since the data and models of different modalities usually show heterogeneity. For example, the feature expressions of image and text data are fundamentally different, and it is difficult to directly measure the similarity between them, so the main problem faced by cross-modal retrieval is different the data of the modality has heterogeneity in the underlying features, that is, the semantic gap. Therefore, the most important part in the cross-modal retrieval process is to break the semantic gap between different modal data, so that the features between different modalities can be related through the semantic information contained in them, and can be directly measured, thereby breaking the Heterogeneity of underlying features. In order to achieve this goal, the key step in current cross-modal retrieval is usually to convert features of different dimensions in different modalities into features of the same dimension through various algorithms, and return the cross-modal retrieval results through calculation.

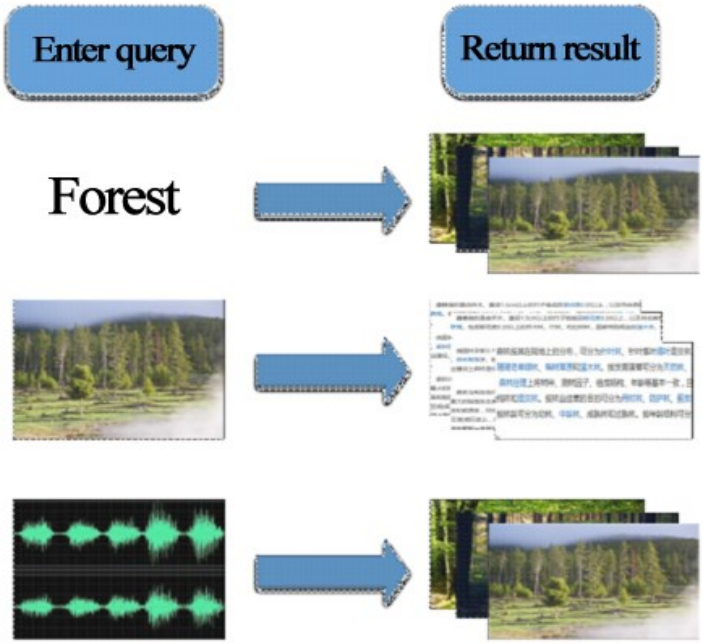

Figure 2. The example of cross modal retrieval.

At present, cross-modal retrieval still has the following problems. (1) For images with rich perspectives and structural features, there are problems such as insufficient mining and low matching degree during feature extraction and matching. (2) None of the existing models fails to fully mine cross-modal data features and semantic correlations is not effective in the process of data isomorphism, and the retrieval accuracy is low.

Cross-modal retrieval methods such as CCA, JGRHML, SCM, and JFSSL are all based on the construction of common subspaces. Such methods are not sufficient in mining the semantic associations between multimodal data, and ignore the structural features of multimodal data. There is still some room for improvement in mapping common subspaces[9-14].

The subspace cross-modal retrieval method based on higher-order semantic correlation includes two main parts: subspace mapping and cross-modal retrieval metrics[15-19]. The model framework is shown in Figure 3. On this basis, we propose research on subspace cross-modal retrieval method based on hypergraph ranking. The model consists of two parts, the subspace mapping model and the hypergraph ranking model. The overall block diagram is shown in Figure 4. The first half is the subspace mapping model. The semantic correlation matrix is constructed by using the semantic annotation information of images and texts, and the feature selection is carried out in combination with the structural sparsity requirements of cross-modal data; the second half is the hypergraph sorting. The model utilizes the first half of the subspace mapping results and the semantic relationship between cross-modal data to mine data correlations within and between modalities, construct a hypergraph sorting model, and use sorting for cross-modal retrieval.

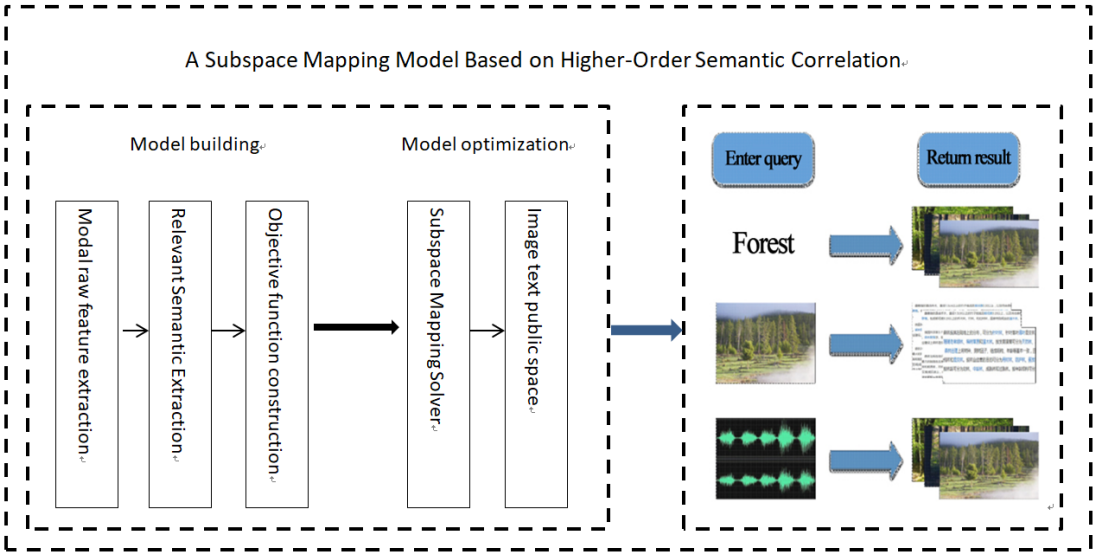

Figure 3. The framework of subspace cross modal retrieval method based on high-order semantic correlation research content flow chart.

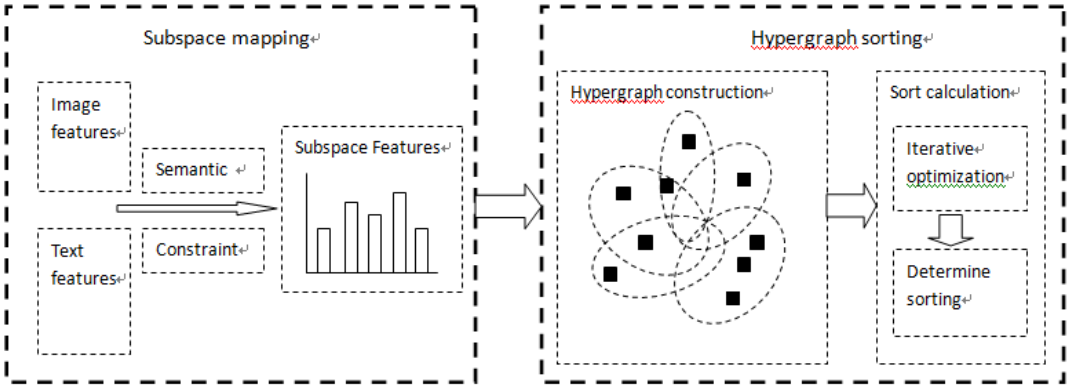

Figure 4. The system diagram of subspace method for cross modal retrieval based on hypergraph ranking.

2. CONSTRUCTION AND SOLUTION OF SUBSPACE MAPPING MODEL BASED ON HYPERGRAPH SORTING

Let $X_I \notin R^{N\times d_1}$ and $X_T \notin R^{N\times d_2}$ denote the image and text datasets, respectively, and $L \notin R^{N\times C}$ the label matrix of the data, where N represents the number of samples, d_1 and d_2 represent the data dimensions of the image and text, respectively, and C represents the label type. The goal of subspace mapping is to obtain the projection matrices P_I and P_T of images and texts, and map the original data with large differences in dimensions to data of the same dimension. The subspace mapping adopts the model: $\min_{P_I,P_T} \sum_{a=I,T} \left\| X_a^T P_a - S \right\| + \lambda \sum_{a=I,T} \left\| P_a \right\|_{21}$, which is divided into two parts: constraints based on semantics and constraints based on data structure. Among them, S represents the semantic subspace of the mapping, which is represented by high-order semantic correlation: $S_{ij} \begin{cases} 1, & I_i \text{ and } T_j \text{ belong to the same class of L} \\ 0, & others \end{cases}$, and the subspace projection matrices P_I, P_T and data matrices of the image and text are obtained respectively: $Y_I = X_I^T P_I$, $Y_T = X_T^T P_T$.

A hypergraph model is built using formula $G=(V,E,\omega)$, where V the vertex set is, E is the hyperedge set, and ω represents the weight vector of the hyperedge. The correlation matrix $H \notin R^{V\times E}$ is usually used to represent the correlation between hypergraphs. The value of H is usually determined according to the requirements for correlation. The hypergraph correlation matrix between images and text can be expressed as: $H(v_i,e_j)=\begin{cases}1 & if\ v_i\in e_j\\ 0 & otherwise\end{cases}$.

Based on the correlation matrix H, the degree $v\in V$ of the vertex $d(v)$ and the degree $e\in E$ of the hyperedge $\delta(e)$ can be obtained, where: $d(v)=\sum_{e\in E}\omega(e)h(v,e)$, $\delta(e)=\sum_{v\in V}h(v,e)$, in addition, let D_v and D_e be the diagonal matrices of the vertex degree and the hyperedge degree of the hypergraph model, respectively, define W represents the diagonal matrix of the hyperedge weight matrix, and the optimization function of the ranking score can be expressed as:

$$\Omega(f)=\frac{1}{2}\sum_{e\in E}\sum_{u,v\in V}\frac{\omega(e)H(u,e)H(v,e)}{\delta(e)}\times\left(\frac{f(u)}{\sqrt{d(u)}}-\frac{f(v)}{\sqrt{d(v)}}\right)^2+\mu\sum_{u\in V}(f(u)-q(u))^2 \tag{1}$$

where μ is the equilibrium constant, q is the original ranking score, and f is the vector of final ranking scores. The definition of $\Theta=D_v^{-1/2}HWD_e^{-1/2}H^TD_v^{-1/2}$ can be expressed as:

$$\begin{aligned}\Omega(f)&=\sum_{e\in E}\sum_{u,v\in V}\frac{\omega(e)H(u,e)H(v,e)}{\delta(e)}\left(\frac{f^2(u)}{d(u)}-\frac{f(u)f(v)}{\sqrt{d(u)d(v)}}\right)+\mu(f-q)^T(f-q)\\ &\sum_{u\in V}f2(u)\sum_{e\in E}\frac{\omega(e)H(u,e)}{d(u)}\sum_{v\in V}\frac{H(v,e)}{\delta(e)}\\ &\sum_{e\in E}\sum_{v\in e}\frac{f(u)H(u,e)\omega(e)H(v,e)f(v)}{\sqrt{d(u)d(v)}\delta(e)}+\mu(f-q)^T(f-1)\\ &=f^T(I-\Theta)f+\mu(f-q)^T(f-q)\\ &s.t.\ \sum_{v\in V}\frac{H(v,e)}{\delta(e)}=1\quad\sum_{e\in E}\frac{\omega(e)H(u,e)}{d(u)}=1\end{aligned} \tag{2}$$

where I represents the identity matrix and $\Delta=I-\Theta$ represents the hypergraph Laplacian matrix. Differentiating f in the above formula, the calculation formula of the ranking score can be obtained: $f=\left(\frac{\mu}{\mu+1}\right)\left(I-\frac{1}{\mu+1}\Theta\right)^{-1}q$, the similarity matrix in the text mode and the image mode can be expressed as the following formula:

$$S_{TT}(i,j)=\begin{cases}\exp\left(-\frac{\|v_i-v_j\|^2}{\delta^2}\right) & if\ i\neq j\\ 0 & else\end{cases} \tag{3}$$

The purpose of constructing intra-modal and inter-modal similarity matrices is to build a joint hypergraph model, which can build a joint similarity matrix of text modalities: $S_T=\theta*S_{II}+(1-\theta)*S_{TI}$, which is used to balance the proportion of text and image modalities in constructing a joint similarity matrix. Likewise, the joint similarity matrix: $S_I=\theta*S_{TT}+(1-\theta)*S_{IT}$.

The method of k nearest neighbours is used to construct and solve hyperedges: when performing cross-retrieval of images and texts, k nearest neighbours are selected to construct hyperedges according to different query modalities. Then the weight value of each hyperedge can be obtained: $\omega(e_i)=\sum_{vj\in ei}S(i,j)$, then given a query mode T or I, the corresponding query score f can be obtained. The higher the f, the more similar the data are.

Experimental dataset selection and parameter setting during the experiment, two retrieval tasks, image retrieval text and text retrieval image, were also set. In the Wiki data set, 2173 sample pairs were selected as the training set and 693 sample pairs were used as the test set; 50% of the sample pairs in the NUS-WID data set were respectively selected as the training set and test set; 4000 samples were selected in the XMedia data set pair as the training set, and the remaining 1000 sample pairs as the test set. Through multiple experiments, the parameter settings are: k is 5, μ is set to 0.9, and θ is set to 0.2 when the experimental results are the best.

3. EXPERIMENTAL RESULTS AND COMPARATIVE ANALYSIS

We compare several cross-modal retrieval methods to verify the effectiveness of the methods, and the precision-recall curves on three public datasets: Wiki, NUS-WIDE, and XMedia.

From the experimental results shown in Tables 1-3 and Figures 5-7, it can be seen that the MAP value obtained by using the Wiki data set is 0.1971, 0.1694, and 0.1170 higher than that of the CCA and other comparison methods, respectively increased by 0.1971, 0.1694, 0.1170, 0.1500, 0.0057; the MAP values obtained with the NUS-WIDE dataset were increased by 0.3023, 0.1878, 0.1700, 0.1133, and 0.0223, respectively; the MAP values obtained with the XMedia dataset were increased by 0.8721, 0.5820, 0.3662, 0.1989, and 0.0139, respectively. Through the comparative analysis of the experimental results, it can be found that compared with the CCA, SCM, JGRHML, JFSSL, and HOSC methods, they have achieved better retrieval results and improved the cross-modal retrieval accuracy. Comparing the differences in the experimental results of several methods, breaking the semantic barrier in the process of cross-modal retrieval is the basis for improving retrieval accuracy. Combining semantic information to narrow the distance of multimodal data is the key to improve the accuracy of cross-modal retrieval. The subspace mapping model is constructed by using the semantic annotation information of the multimodal data, and then the hypergraph model is constructed by combining the similarity between the multimodal data and the modalities, and the correlation between the multimodal data is further mined, so that the Semantically closely related data is easier to retrieve, improving cross-modal retrieval accuracy.

Table 1. Comparison of MAP values of different methods on the WIKI dataset.

	Image retrieval text	Text retrieve images	Retrieve the average
CCA	0.2549	0.1846	0.2198
JGRHML	0.2830	0.2119	0.2475
SCM	0.3501	0.2496	0.2999
JFSSL	0.3063	0.2275	0.2669
HOSC	0.4184	0.4039	0.4112
SCMRHR	0.4231	0.4106	0.4169

Table 2. Comparison of MAP values of different methods on the WIKI dataset.

	Image retrieval text	Text retrieve images	Retrieve the average
CCA	0.2178	0.1824	0.2001
JGRHML	0.3425	0.2866	0.3146
SCM	0.3746	0.2902	0.3324
JFSSL	0.4035	0.3747	0.3891
HOSC	0.4975	0.4628	0.4801
SCMRHR	0.5214	0.4833	0.5024

Table 3. Comparison of MAP values of different methods on the WIKI dataset.

	Image retrieval text	Text retrieve images	Retrieve the average
CCA	0.1220	0.1207	0.1214
JGRHML	0.4601	0.3629	0.4115
SCM	0.6335	0.6210	0.6273
JFSSL	0.8126	0.7765	0.7946
HOSC	0.9839	0.9752	0.9796
SCMRHR	0.9976	0.9894	0.9935

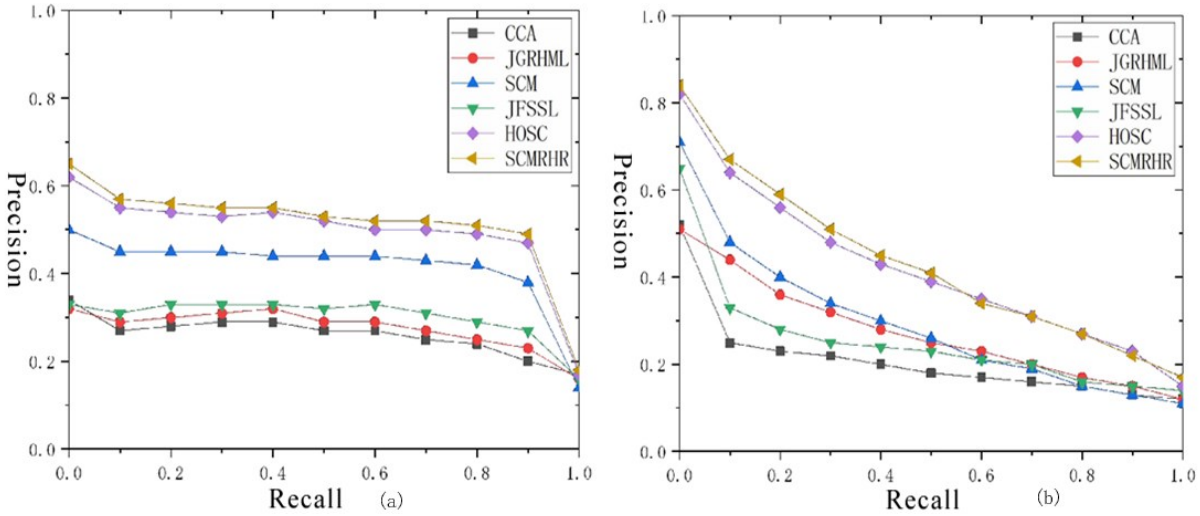

Figure 5. The Precision-recall curve on Wiki dataset: (a) Image to text; (b) Text to image.

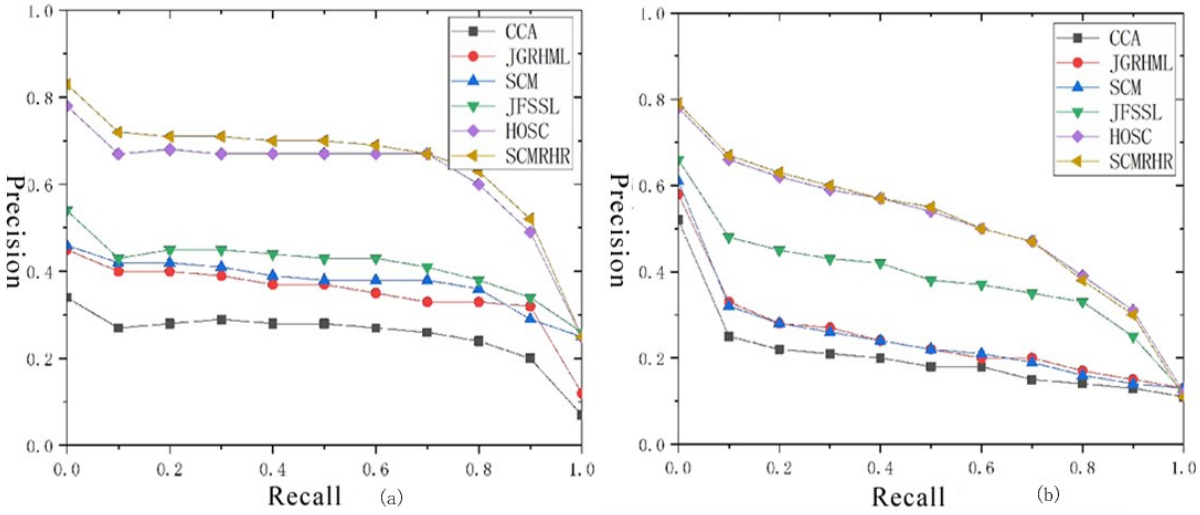

Figure 6. The precision-recall curve on NUS-WIDE dataset: (a) Image to text; (b) Text to image.

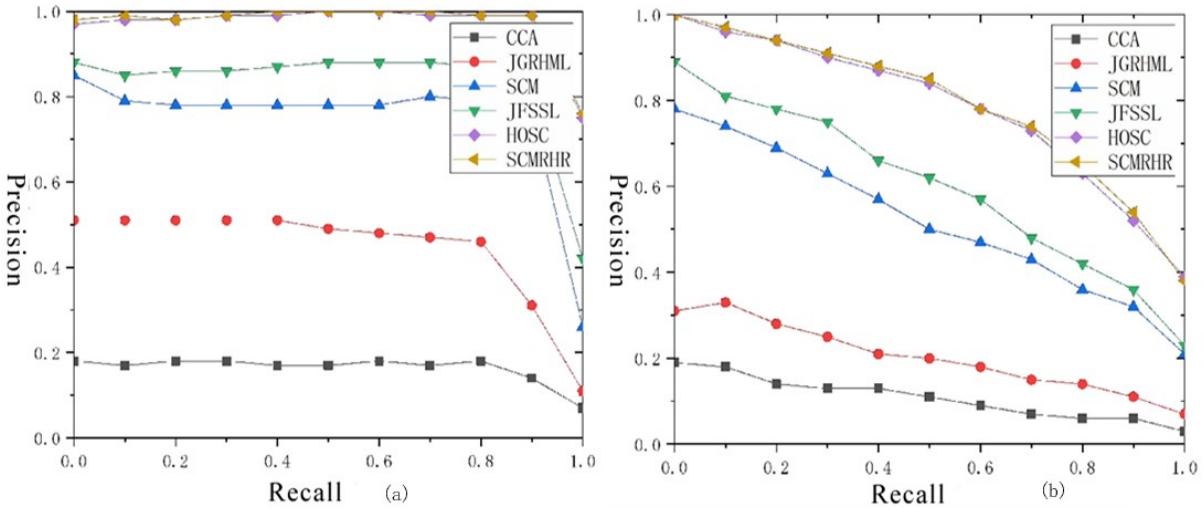

Figure 7. The Precision-recall curve on XMedia dataset: (a) Image to text; (b) Text to image.

4. CONCLUSION

Research on subspace cross-modal retrieval method based on hypergraph sorting: This method is based on the research on subspace cross-modal retrieval method based on high-order semantic correlation (HOSC), and uses the image mapped by high-order semantic subspace. A hypergraph ranking model is constructed with textual feature data and semantic annotation information of multimodal data. This method takes the image and text feature data mapped by semantic subspace as the vertices of the hypergraph, takes the similarity relationship between the data as the hyperedge, uses the method of neighbours to determine and solve the hyperedge, and then obtains the hypergraph ranking score. This method makes the original multimodal data with the same semantics closer, and further improves the cross-modal retrieval accuracy. However, due to the addition of a hypergraph sorting model, this method also increases a certain computational cost, and this method cannot achieve better retrieval results for unlabeled sample data.

In general, on the basis of considering the semantic correlation of multi-modal data and processing the correlation between multi-modal data and between modalities, the hypergraph can be used to describe the multi-sample relationship between multiple samples. The characteristics of the correlation relationship are used to build a hypergraph model, which makes the mapped multimodal data with related semantics closer. Although the cross-modal retrieval accuracy has been improved to a certain extent, it can only be applied to labelled datasets, and there is still a lot of room for improvement in semantic mining and heterogeneous feature isomorphism.

REFERENCES

[1] Lu, C., Li, F. and Chen, Q., “Image retrieval method based on improved hash algorithm,” Electronic Science and Technology, (5), 28-32(2020).
[2] Li, Y., Miao, Z. and Wang, J., “Deep binary constraint Hashing for fast image retrieval,” Electronics Letters, 54(1), 25-27(2018).
[3] Tang, X., Wang, Y. and Zou, F., “A fast large-scale image retrieval method based on multi-hash algorithm,” Computer Engineering and Science, (7), 1316-1321(2016).
[4] Li, S., Tao, Z. Q., Li, K. and Fu, Y., “Visual to text: Survey of imagand video captioning,” IEEE Transactions on Emerging Topics in Computational Intelligence, 3(4), 297-312(2019).
[5] Lin, Z. J., Ding, G. G. and Hu, M. Q., “Semantics-preserving hashing for cross-view retrieval,” IEEE Computer Society, 3864-3872(2015).
[6] Ding, G. G., Guo, Y. C. and Zhou, J. L., “Collective matrix factorization Hashing for multimodal data,” IEEE Computer Society, 2083-2090(2014).
[7] Zhang, D. Q. and Li, W. J., “Large-scale supervised multimodal hashing with semantic correlation maximization,” Proc. of the AAAI Conf. on Artificial Intelligence, 2177-2183(2014).

[8] Wang, D., Gao, X. B. and Wang, X. M., “Semantic topic multi-modal Hashing for cross-media retrieval,” Proc. of the Inter. Joint Conf. on Artificial Intelligence, 3890-3896(2015).
[9] Liu, F. and Zhang, H., “Discriminant cross-modal hash retrieval algorithm based on multi-level semantics,” Computer Applications, 41(8), 2187-2192(2021).
[10] Zhang, Q., Tian, X. and Yang, F., “Fusion retrieval model based on mathematical text and expression transformation,” Computer Engineering, 45(3), 175-181+187(2019).
[11] Miao, F., Jia, H. D. and Xiong, Y. N., “Approximate neighbor selection method for mobile users based on service similarity,” Computer Engineering, 44(5), 168-173+179(2018).
[12] Peng, Y., Zhai, X. and Zhao, Y., “Semi-supervised cross-media feature learning with unified patch graph regularization,” IEEE Transactions on Circuits and Systems for Video Technology, 26(3), 583-596(2016).
[13] Zhuo, Y. K., Qi, J. W. and Peng, Y. X., “Cross-media deep fine-grained associative learning method,” Journal of Software, 30(4), 884-895(2019).
[14] Deng, C., Tang, X. and Yan, Y., “Discriminative dictionary learning with common label alignment for cross-modal retrieval,” IEEE Transactions on Multimedia, 18(2), 208-218(2016).
[15] Zhang, L., Ma, B. and Li, G., “Cross-modal retrieval using multi-ordered discriminative structured subspace learning,” IEEE Transactions on Multimedia, 19(6), 1220-1233(2017).
[16] Zhang, L., Ma, B. and Li, G., “Generalized Semi-supervised and structured subspace learning for cross-modal retrieval,” IEEE Transactions on Multimedia, 20(1), 128-141(2017).
[17] He, R., Tan, T. and Wang, L., “Regularized correntropy for robust feature selection,” IEEE Conf. on Computer Vision and Pattern Recognition, 1633(2012).
[18] Yang, Y., Shen, H. and Ma, Z., “1-norm regularized discriminative feature selection for unsupervised learning,” Proc. of the 22nd Inter. Joint Conf. on Artificial Intelligence (AAAI), 1538(2011).
[19] Nikolova, M. and Ng, M. K., “Analysis of half-quadratic minimization methods for signal and image recovery,” SIAM Journal on Scientific Computing, 27(3), 937-966(2005).