Hindawi
Mobile Information Systems
Volume 2021, Article ID 9947644, 11 pages
https://doi.org/10.1155/2021/9947644

Research Article

A Cross-Media Retrieval Method Based on Semisupervised Learning and Alternate Optimization

Junzheng Li,[1] **Wei Zhu**,[2] **Yanchun Yang**,[3] **and Xiyuan Zheng**[3]

[1]*Network Information Management Center, Shandong Management University, Jinan 250357, China*
[2]*Department of Ophthalmology, Jinan Central Hospital, Cheeloo College of Medicine, Shandong University, Jinan 250013, China*
[3]*School of Data and Computer Science, Shandong Women's University, Jinan 250300, China*

Correspondence should be addressed to Xiyuan Zheng; 306732399@qq.com

Received 31 March 2021; Revised 20 June 2021; Accepted 24 August 2021; Published 27 September 2021

Academic Editor: Yugen Yi

With the continuous advancement in Internet technology, we are gradually stepping into an era of big data where a large amount of multimedia data is produced every day at any given time. In order to properly utilize these data, the research on big data is also constantly evolving. Cross-media retrieval is a prime example, aiming at retrieving various forms of data, for example, text, image, audio, video, and other forms. The most difficult task for cross-media retrieval lies in the potential correlation between different modalities data and how to overcome the semantic gap. This paper proposes a cross-media retrieval method based on semisupervised learning and alternate optimization (SMDCR) to overcome the abovementioned difficulties, thereby improving the retrieval accuracy. The main advantage of this method is to make full use of the degree of correlation between the semantic information of the labeled data and unlabeled data. Simultaneously, we combine the linear regression term, correlation analysis term, and feature selection term into a joint cross-media learning framework. Furthermore, the projection matrices are trained with the alternate optimization method. Finally, experimental results on two public datasets demonstrate the effectiveness of the proposed method.

1. Introduction

1.1. Research Background. At present, with the development and popularization of the Internet and digital media, the speed of data generation is accelerating, which results in a data torrent. The data types are also expanding rapidly, from traditional text types to new ways of presenting information by text, image, audio, animation, video, 3D animation, 3D video, etc. These data types have different presentation characteristics with semantic correlation. When users search semantic features through keywords, they can retrieve multimedia objects with the same theme in different modalities (Figure 1); for example, when searching for the violin, the results will simultaneously display violin-related pictures and violin music videos and 3D graphics. This retrieval mode is called cross-modal retrieval or cross-media retrieval [1, 2], which greatly improves the scope of information retrieval and helps the information retriever understand many aspects of different types of information through visual, auditory, and other perceptual ways.

Image data and text data are the most common and frequently used means to present complex data. In order to improve the accuracy and effectiveness of retrieval, machine learning has been studied by scholars in related fields. Image recognition technology [3] is the recognition of an image based on its main features as each image has its features and uniqueness. Simple images use color depth as the main feature, while complex images are distinguished by different levels of image texture. Invalid information interferes with the effectiveness of the extracted information and should be avoided during image recognition. The data retrieval process is illustrated in Figure 2. Users can submit the key information on the retrieval platform; then, multimedia information retrieval methods are used to measure the similarity between multimedia data which are finally exhibited on the retrieval platform.

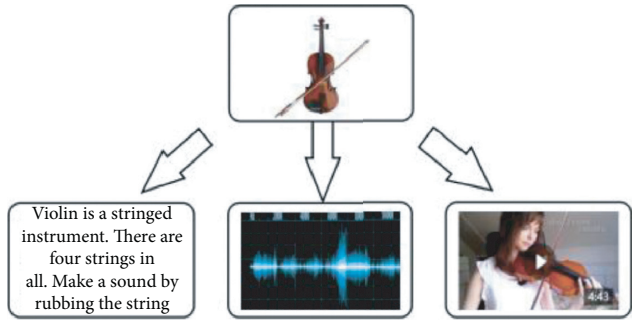


Figure 1: Demonstration of cross-media retrieval.

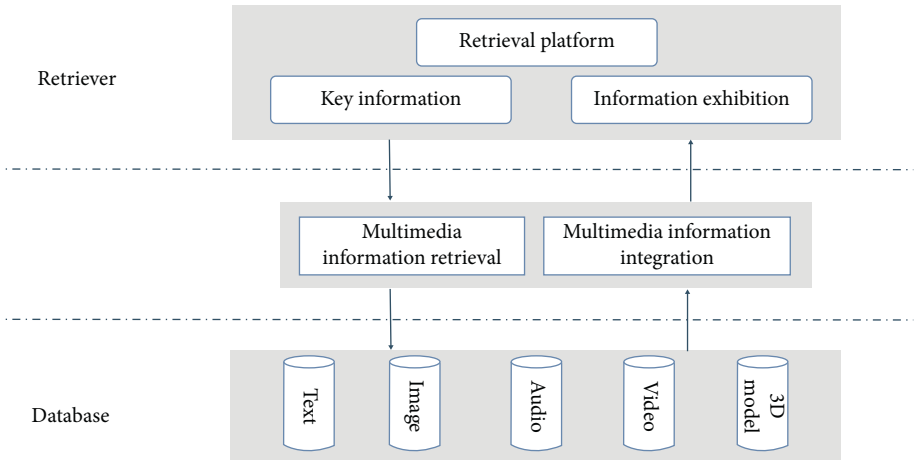


Figure 2: Data retrieval process.

Due to the popularity of the Internet and wide use of all kinds of multimedia equipment, hundreds of millions of multimedia data are produced every second with a wide range of sources, types, large volume, and other characteristics. If these data are fully utilized by technical means, the results of the research will be applied to everyday life, economic development, and even national security, which will bring unexpected gains. Research in this area will greatly improve the approach to the advancement in artificial intelligence.

1.2. Development of Cross-Media Retrieval Technology

1.2.1. Overview of Information Retrieval Models. Information retrieval refers to the process and technology of organizing information in a certain way, through which the relevant information according to the needs of the users can be located. The four most traditional categories of information retrieval models [4] are Boolean, vector space, probability, and language models. In modern Internet, the basic unit for the amount of information has risen to 10 billion, including texts, images, and videos, consequently making retrieval more difficult. If a large-scale resource set wants to be easily retrieved, the most general way is to carry out semantic annotation, and a data matrix corresponding to the semantic matrix results from a new matrix with a lot of 0's, which is generally called sparse matrix [5]. It is fundamental for information retrieval to classify information with different technologies to find information for users.

1.2.2. Image Recognition Technology. Image recognition technology has already walked into our life, such as fingerprint attendance, face unlock, and face scan payment. Due to the obvious deficiency of human recognition at present, image recognition technology is constantly being explored. The main steps of image recognition technology are divided into data information acquisition, data preprocessing, feature selection and extraction, classification decision, and classifier design [3].

The relatively new type of image recognition techniques are based on neural networks, for example, neural network image recognition models based on genetic algorithms [6] and back propagation (BP) networks [7] which are very typical. Convolutional neural network (CNN), as one of the most prevalent deep learning algorithm, has been widely applied in the field of image recognition in recent years [8, 9]. Because the image is multidimensional, dimensionality reduction becomes the most effective method to improve the recognition ability, which is further divided into linear dimensionality reduction [10] and nonlinear dimensionality reduction [11]; for example, the Principal Component Analysis (PCA) [12] and Linear Discriminant Analysis (LDA) [13] are commonly used as linear dimension

reduction methods. The deepening research on image recognition and the application of this technology will expand and aid our human society since image is one of the main sources for us to obtain information.

1.2.3. Cross-Media Search Technology. What should be done to deal with multimedia data in the big data era? Many important research aspects on cross-media research have been proposed, including cross-media understanding, cross-media retrieval, and large spatiotemporal data search. The essence of cross-media research is to mine for the relationship between different modalities data, and its research results are helpful to complete the retrieval transformation between different modalities. The distribution of multimedia data in the feature space of different modalities shows that the data features of text, image, and video are essentially different, so it is impossible to directly contrast the similarity between different modalities data in cross-media retrieval. In order to solve the above problems, the commonly used method is subspace learning [14], which learns common feature representation of different modalities data. In order to associate different data modalities, the subspace learning takes multilevel nonlinear features from different data modalities and maps them onto a common subspace of the same dimension (Figure 3). Then, it measures the similarity to establish correlations between different modalities data.

Typical methods are Canonical Correlation Analysis (CCA) [15], Semantic Matching (SM) [16], Semantic Relevance Matching (SCM) [17], T-V CCA [18], Generalized Multiview Analysis Linear Discriminant Analysis (GMLDA) [19], Generalized Multiview Analysis Marginal Fisher Analysis (GMMFA) [20], Modality-Dependent Cross-Media Retrieval (MDCR) [21], Cross-Modal Online Low-Rank Similarity function learning (CMOLRS) [22], and Semi-supervised Learning Based Semantic Cross-Media Retrieval (S3CMR) [23]. CCA is adapted to obtain the correlation matching degree of different modalities data; from the perspective of data semantics, different modalities data are matched; thus, the triple constraints are added and the matching accuracy is improved. SM method is a semantic expansion of CCA, and SCM is a simple combination of SM and CCA, while the innovation of T-V CCA is that it adds the semantic perspective and carries out high-level semantic analysis from three aspects. T goal of GMLDA is to find the optimal direction of projection, which learns a common subspace through a supervised extension of CCA. GMMFA is mainly a collection of GMLDA and CCA; MDCR combines the correlation of data pairs with the semantic, and the projection matrix learning is based on different retrieval tasks. CMOLRS approach learns a low-rank bilinear similarity measure of different modalities data; S3CMR is also a semisupervised cross-media retrieval method which makes full use of both labeled data and the unlabeled data.

1.3. Organizational Structure of This Paper. There are five sections in this paper. The organizational structure of each section is summarized as follows:

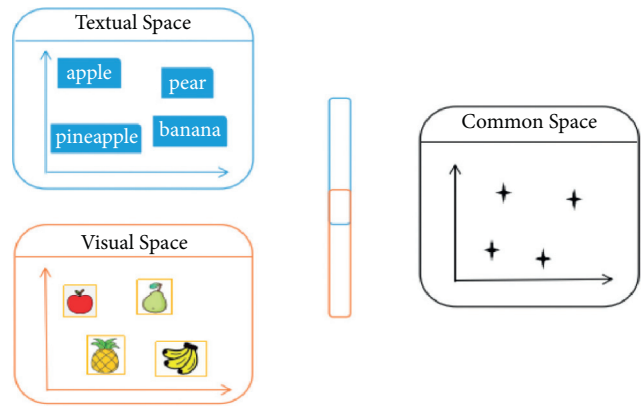


Figure 3: Subspace learning.

Section 1 mainly introduces the relevant background of our research, the development of cross-media retrieval technology, the main research direction of this paper, and the summary of the organizational structure.

Section 2 mainly introduces some classic cross-media retrieval methods, which are CM, SM, SCM, PLSR, and MDCR.

Section 3 proposes a cross-media retrieval method based on semisupervised learning and alternate optimization; it introduces the objective function, the process of algorithm derivation, and the optimization steps of the algorithm.

Experimental results and analysis on two public datasets are shown in Section 4.

Section 5 summarizes the content of alternate optimized cross-media retrieval method based on semi-supervised learning and innovation points.

2. Introduction of Cross-Media Retrieval Methods

2.1. Introduction. In the field of cross-media retrieval, researchers have proposed a variety of methods, in which the most widely used one is based on matrix projection in shared subspace to achieve unified representation of different modalities data. According to the enlightenment of related methods, this paper mainly proposes a cross-media retrieval method based on semisupervised learning and alternate optimization, which forms two pairs of projection matrices using labeled and unlabeled data samples, and calculates the sample feature distance in the alternate optimized training process, and the semantics of unlabeled sample data are represented by the trained feature distance. Wikipedia [24] and Pascal Sentence [8] datasets are used for experiments to verify the effectiveness of the proposed method. The results show that this method has a better prediction effect and improves the retrieval accuracy.

2.2. Introduction of Classic Cross-Media Retrieval Methods. For linear subspace learning, one classic method is Confirmatory Factor Analysis (CFA), which projects different

modalities data into a common space with two regularization transformation matrices, and achieves subspace projection by minimizing the matrix containing corresponding features obtained from the image and text set. The other one is the Classical Correlation Analysis (CCA) method, which is a widely used method in data correlation analysis. CCA transforms high-dimensional data into low-dimensional data through linear transformation. Then, it maximizes the correlation of different modalities data and analyzes the data using a correlation coefficient.

For nonlinear subspace learning, since CCA relies on linear methods to represent data, the idea of a pair of nonlinear transformation kernel functions [25] is used to transform data features into high-dimensional space called Kernel Canonical Correlation Analysis (KCCA) [26] when images and texts as well as other data cannot be linearly represented.

Semantic Matching (SM) [16] is originally used to measure the similarity between texts, and now it can be used to accomplish the task of cross-media retrieval. First, different modalities data are extracted from low-level features, and then the feature is semantically learned to map multimedia data objects into an isomorphic subspace, in which each dimension represents a semantic category. A multiclass logical regression is used to classify texts and images, that is, to calculate the posterior probability [27], and then matches the similarities based on semantic relevance.

Semantic relevance matching (SCM) [17] consists of CM and SM, improving the performance of both at the same time by simple combination. Firstly, the CCA method is used to map text features and image features to a subspace. Then it learns to maximize the correlation subspace and calculate the projection of each image-text pair. Next, the SM method is used to produce the same semantic subspace of the related subspace transformation matrix respectively. Retrieval is then carried out based on the distance in semantic matching.

PLSR (Partial Least Squares Regression) [28] method is a regression modeling method of multidependent variable Y for multi-independent variable X. In the process of regression, the method considers both extracting the principal components in Y and X as much as possible (PCA-Principal Component Analysis) [29] and maximizing the correlation between the extracted principal components respectively (CCA). Briefly, PLSR is the combination of three basic methods, PCA, CCA, and multivariate linear regression.

The above methods either ignore feature consistency correlation or semantic consistency correlation. Besides, they are designed to learn the same couple of projection matrices for the involved subretrieval tasks. They fail to capture the characteristic of each subretrieval task and probably deteriorate the retrieval performance.

In literature [30], Wei et al. proposed a Modality-Dependent Cross-media Retrieval (MDCR) method to learn two couples of projection matrices for two different subretrieval tasks. Although it achieves certain improvements in the accuracy of the algorithm, it still fails to explore the relationships between the projected feature representation and the semantic label for modality correlation enhancement. Different from existing methods, our work makes full use of the degree of correlation between the semantic information of labeled and unlabeled data. Simultaneously, we combine the linear regression term, correlation analysis term, and feature selection term into a joint cross-media learning framework. Furthermore, the projection matrices are trained with an alternate optimization method. Finally, experimental results on two public datasets demonstrate the effectiveness of the proposed method.

3. A Cross-Media Retrieval Method Based on Semisupervised Learning and Alternate Optimization

Modality-Dependent Cross-media Retrieval (MDCR) [30] method improves the accuracy of each single retrieval and keeps the distribution of data information consistent with semantic information. The problem is that this method ignores the effect of unlabeled data on the results when processing the optimization. This paper proposes a new cross-media retrieval method called SMDCR to better utilize semisupervised learning and alternate optimization of labeled data, unlabeled data, and semantic information. Based on MDCR method proposed by Wei, the objective function of SMDCR is as follows:

$$\min_{U,V} \lambda_1 \left\| XU^T - YV^T \right\|_F^2 + (1-\lambda_1)\left(\left\| XU^T - S \right\|_F^2 + \left\| YV^T - S \right\|_F^2 \right) + \lambda_2 \|U\|_F^2 + \lambda_3 \|V\|_F^2, \tag{1}$$

where U and V represent the projection matrices of modal X and Y, respectively, S is the semantic matrix, λ_1, λ_2, and λ_3 are balance parameters, and$_F^2$ is the L_2 norm for feature selection. The first term is used to learn the subspace and ensure correlation of image-text pairs. The second term guarantees samples to be closed to their labels after projection. The last two terms control the scale of projection matrices, which ensure the features are selected from different modalities simultaneously and also avoid overfitting.

3.1. Objective Function. The cross-media retrieval task of this paper is the mutual retrieval between text data and image data. The notations and descriptions used in this paper are as follows.

Definition $D = (X_i, Y_i)_{i=1}^n = (X_i, Y_i)_{i=1}^l \cup (X_i, Y_i)_{i=1}^u$ represents all sample data pairs, where $n = l + u$ represents the number of sample pairs, l and u represent the number of labeled samples and the number of unlabeled samples, X_i is the i-th sample of modal data X, Y_i is the i-th data sample of modal data Y, $X_{lu} = (X_l X_u)$ represents all sample data of modal data X, $Y_{lu} = (Y_l Y_u)$ represents the total sample data of modal data Y, where $X_{lu} \in \mathrm{R}^{n \times p}$ and $Y_{lu} \in \mathrm{R}^{n \times q}$, p is the dimension of image modal data, and q is the dimension of text modal data. There are c classes in dataset D; then the semantic matrix is represented with $S = (S_l, S_u) \in \mathrm{R}^{n \times c}$. By learning the image projection matrix U and the text projection matrix V, the two

different modalities data are projected into an isomorphic subspace for mutual retrieval.

The objective function of SMDCR for Image retrieval Text (I2T) is as follows:

$$f_1(U_1, V_1) = \min_{U_1, V_1} \lambda_1 \left\| X_{lu} U_1^T - Y_{lu} V_1^T \right\|_F^2 + \beta \left\| X_{lu} U_1^T - S \right\|_F^2 + \lambda_2 \left\| U_1 \right\|_F^2 + \lambda_3 \left\| V_1 \right\|_F^2. \tag{2}$$

The objective function of SMDCR for Text retrieval Image (T2I) is as follows:

$$f_2(U_2, V_2) = \min_{U_2, V_2} \lambda_1 \left\| X_{lu} U_2^T - Y_{lu} V_2^T \right\|_F^2 + \beta \left\| Y_{lu} V_2^T - S \right\|_F^2 + \lambda_2 \left\| U_2 \right\|_F^2 + \lambda_3 \left\| V_2 \right\|_F^2, \tag{3}$$

where $\|X_{lu}U_1^T - Y_{lu}V_1^T\|_F^2$ and $\|X_{lu}U_2^T - Y_{lu}V_2^T\|_F^2$ denote the feature consistency between the multimedia data in the new isomorphic space and $\|X_{lu}U_1^T - S\|_F^2$ and $\|Y_{lu}V_1^T - S\|_F^2$ represent the semantic consistency in the new isomorphic space, while the latter two terms represent the regularization term preventing overfitting. λ_1 represents the balance parameter of the feature and the projection ratio, where $\beta = 1 - \lambda_1$, λ_2 and λ_3 are the balance parameters for the feature selection in image and text modality. The role of the two expressions (2) and (3) is to learn two pairs of projection matrices U and V.

3.2. Optimization Process. The method of alternate optimization is used for training; projection matrices U, V are initialized; U is fixed and V is updated, and then V is fixed and U is updated.

For I2T, by differentiating the partial derivative of $f_1(U_1, V_1)$ for U_1, we can get

$$\frac{\partial f_1}{\partial U_1} = U_1 X_{lu}^T X_{lu} + 2\left[\lambda_3 U_1 - \lambda_1 V_1 {Y_{lu}}^T - \beta S^T X_{lu}\right] = 0,$$

$$U_1 = \frac{2\left(\lambda_1 V_1 Y_{lu}^T X_{lu} + \beta S^T X_{lu}\right)}{\left(X_{lu}^T X_{lu} + 2\lambda_2 I_1\right)}. \tag{4}$$

Similarly, by differentiating the partial derivative of $f_1(U_1, V_1)$ for V_1, we can get

$$V_1 = \frac{\left(\lambda_1 U_1 X_{lu}^T Y_{lu}\right)}{\left(\lambda_1 Y_{lu}^T Y_{lu} + \lambda_4 I_2\right)}. \tag{5}$$

For T2I, by differentiating the partial derivative of $f_2(U_2, V_2)$ for U_2, we can get

$$\frac{\partial f_2}{\partial U_2} = 2\left[\lambda_2 U_2 + \lambda_1\left(U_2 X_{lu}^T X_{lu} - V_2 {Y_{lu}}^T X_{lu}\right)\right],$$

$$U_2 = \frac{\left(\lambda_1 V_2 Y_{lu}^T X_{lu}\right)}{\left(\lambda_1 X_{lu}^T X_{lu} + \lambda_2 I_1\right)}. \tag{6}$$

By differentiating the partial derivative of $f_2(U_2, V_2)$ for V_2, we can get

$$V_2 = 2\frac{\left(\lambda_1 U_2 X_{lu}^T Y_{lu} + \beta S^T Y_{lu}\right)}{\left(Y_{lu}^T Y_{lu} + 2\lambda_3 I_2\right)}. \tag{7}$$

In the above expressions, I_1 denotes the p-dimensional unit matrix and I_2 denotes the q-dimensional unit matrix. The values of U and V are trained by the expressions above. The proposed SMDCR method uses the final optimization results for cross-media retrieval between text and images by adding semantic information of unlabeled data samples to the process of alternate optimization.

Taking the image retrieval text as an example, the optimization algorithm is shown in Algorithm 1. In the SMDCR method, the procedure of the text retrieval image algorithm is consistent with Algorithm 1, so we do not repeat it.

4. Experiments

4.1. Dataset Presentation. Wikipedia dataset selects more featured articles from Wikipedia and is the most frequently used dataset in cross-media retrieval experiments. It is also the first publicly available dataset in this field. Wikipedia dataset consists of 2,866 labeled image-text pairs, which are divided into 10 semantic classes according to semantic type. 2,173 data pairs are randomly selected as the training data and the remaining 693 pairs are used as the testing data. In this paper, two feature representation methods are used for Wikipedia dataset. One method is a representation of image data by 128-dimensional SIFT (Scale Invariant Feature Transformation) feature [31] and text data by 10-dimensional LDA (Latent Dirichlet Allocation) feature [32], which is recorded as Wikipedia-1. The other method is that image data are represented by 4096-dimensional CNN (Convolution Neural Network) feature, and text data are represented by 100-dimensional LDA feature, which is recorded as Wikipedia-2.

Pascal Sentence dataset contains 20 semantic classes, 1,000 pairs of image-text data; each semantic class contains 50 data pairs. In the experiments, 30 data pairs are randomly selected from each class as the training data, and the remaining pairs are used as the testing data. The image data are represented by 4096-dimensional CNN feature [9], while the text data are represented by 10-dimensional LDA feature.

Repeat
Repeat
value 1 = $f_1(U_1, V_1) = \min_{U_1,V_1} \lambda_1 \|X_{lu}U_1^T - Y_{lu}V_1^T\|_F^2 + \beta\|X_{lu}U_1^T - S_t\|_F^2 + \lambda_2\|U_1\|_F^2 + \lambda_3\|V_1\|_F^2$
$U_1 = 2(\lambda_1 V_1 Y_{lu}^T X_{lu} + \beta {S_t}^T X_{lu}) / (X_{lu}^T X_{lu} + 2\lambda_2 I_1)$
$V_1 = (\lambda_1 U_1 X_{lu}^T Y_{lu}) / (\lambda_1 Y_{lu}^T Y_{lu} + \lambda_3 I_2)$
value 2 = $f_1(U_1, V_1) = \min_{U_1,V_1} \lambda_1 \|X_{lu}U_1^T - Y_{lu}V_1^T\|_F^2 + \beta\|X_{lu}U_1^T - S_t\|_F^2 + \lambda_2\|U_1\|_F^2 + \lambda_3\|V_1\|_F^2$
$j = j + 1$
Until value1 − value2 < ε_1
Repeat
value 3 = $f_1(U_1, V_1) = \min_{U_1,V_1} \lambda_1 \|X_{lu}U_1^T - Y_{lu}V_1^T\|_F^2 + \beta\|X_{lu}U_1^T - S_t\|_F^2 + \lambda_2\|U_1\|_F^2 + \lambda_3\|V_1\|_F^2$
$i = i + 1$
Until value3 − value1 < ε_2
$S_u = X_u U^T - Y_u V^T$
$t = t + 1$
$S_t = (S_l, S_u)$
Until $t >$ maximum number of iterations

Algorithm 1: The whole learning algorithm for our proposed SMDCR.

4.2. Experimental Results and Analysis. This paper mainly compares SMDCR with nine methods: CCA, SM, SCM, T-V CCA [18], GMLDA [19], GMMFA [20], MDCR [21], CMOLRS [22], and S3CMR [23]. CCA is the most traditional cross-media retrieval method, which maximizes the correlation of different modalities data and analyzes the data with a correlation coefficient. SM method is a semantic expansion of CCA, and SCM is a simple combination of SM and CCA, while the innovation of T-V CCA is that it adds semantic perspective and carries out high-level semantic analysis from three aspects. The goal of GMLDA is to find the optimal direction of projection, which learns a common subspace through a supervised extension of CCA. GMMFA is mainly a collection of GMLDA and CCA; MDCR combines the correlation of data pairs and the semantic, and the projection matrix learning is based on different retrieval tasks; the CMOLRS approach learns the low-rank bilinear similarity measure of different modalities data; S3CMR is also a semisupervised cross-media retrieval method which makes full use of both labeled data and unlabeled data, but the optimization method of SM3CR is not very effective.

In our experiment, we verify our proposed method SMDCR for two retrieval tasks: Image retrieval Text (I2T) and Text retrieval Image (T2I). The main test indexes of the experimental results are Mean Average Precision (MAP) and Precision Recall (P-R) curves [33]. The parameters in equations (2) and (3) are set differently in different datasets.

On Wikipedia dataset, for I2T retrieval, the parameters are set as follows: $\lambda_1 = 0.001$, $\beta = 1 - \lambda_1$, $\lambda_2 = 0.001$, $\lambda_3 = 0.1$, $\varepsilon_1 = 0.00001$, $\varepsilon_2 = 0.00001$; for T2I retrieval, the parameters are set as follows: $\lambda_1 = 0.001$, $\beta = 1 - \lambda_1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.1$, $\varepsilon_1 = 0.00001$, $\varepsilon_2 = 0.00001$. The MAP scores of SMDCR method versus other compared methods are shown in Table 1. From Table 1, we can see that supervised learning methods (SM, SCM, T-V CCA, GMMFA, GMLDA, MDCR, and CMOLRS) outperform unsupervised learning method (CCA). Our method with semisupervised learning achieves a higher performance than supervised methods. This is because SMDCR can exploit the relationship between unlabeled data features and semantic classes. Our method adopts an effective alternate optimization strategy, and therefore SMDCR achieves higher performance than S3CMR. Furthermore, our method achieves better performance based on the P-R curves, as shown in Figure 4.

On Pascal Sentence dataset, for I2T retrieval, the parameters are set as follows: $\lambda_1 = 0.65$, $\beta = 1 - \lambda_1$, $\lambda_2 = 0.3$, $\lambda_3 = 0.1$, $\varepsilon_1 = 0.00001$, $\varepsilon_2 = 0.00001$; for T2I retrieval, the parameters are set as follows: $\lambda_1 = 0.2$, $\beta = 1 - \lambda_1$, $\lambda_2 = 0.05$, $\lambda_3 = 0.01$, $\varepsilon_1 = 0.00001$, $\varepsilon_2 = 0.00001$. The MAP scores of SMDCR method versus other compared methods are shown in Table 2. The P-R curves of I2T and T2I retrieval are shown in Figure 5. We find that SMDCR outperforms the best performance of previous papers. Among the compared methods, CCA performs the worst performance compared to other cases. This is because it is the only method that employs paired samples to learn the shared space and thereby no explicit semantics are exploited. Although MDCR learns the same couple of projections for different subretrieval tasks, it ignores paired semantic consistency. Thus, MDCR obtains a suboptimal retrieval performance compared to our method. From the experiment results, we can also see that our method achieves the best performance when semantic classes are increased.

In addition, the MAP scores for each class on Wikipedia and Pascal Sentence dataset are shown in Figures 6 and 7, respectively. Figures 4–7 further confirm the effectiveness of our method. From the above experiments, it is clear that our method with semisupervised learning and alternate optimization outperforms other compared methods.

There are four parameters λ_1, β, λ_2, λ_3 in the proposed approach. λ_1 is the weighting parameter of the feature consistency term, and λ_2 and λ_3 are the balance parameters for the feature selection in image and text modalities. In our experiments, $\beta = 1 - \lambda_1$, λ_1, λ_2, λ_3 is adjusted from 0.001, 0.01, 0.1, 1 on Wikipedia and Pascal Sentence dataset with one fixed parameter and the performance with the other two parameters is observed. The selected parameter, through experimental results, is the one that makes the performance of our method most stable.

Table 1: The MAP scores on Wikipedia dataset.

Method	MAP		
	I2T	T2I	Average
CCA	18.2	20.9	19.5
SM	22.5	22.3	22.4
SCM	27.7	22.6	25.2
GMMFA	26.4	23.1	24.7
GMLDA	27.2	23.2	25.2
TC-CCA	22.8	20.5	21.6
MDCR	27.1	22.5	24.8
CMOLRS	28.46	20.85	24.64
S3CMR	**29.3**	22.3	25.8
SMDCR	29.0	**23.5**	**26.2**

Numbers in boldface are the best.

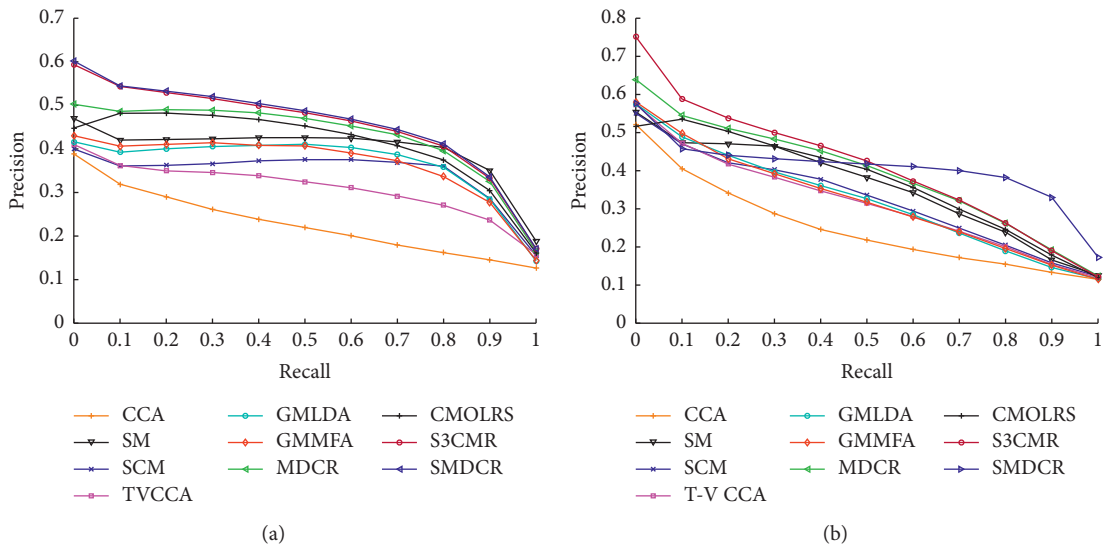


Figure 4: P-R curves of SMDCR method versus compared methods on Wikipedia-2 dataset. (a) I2T. (b) T2I.

Table 2: The MAP scores on Pascal Sentence dataset.

Method	MAP		
	I2T	T2I	Average
CCA	26.1	35.6	30.9
SM	42.6	46.7	44.6
SCM	36.9	37.5	37.2
GMMFA	45.5	44.7	45.1
GMLDA	45.6	44.8	45.2
TC-CCA	33.7	43.9	38.8
MDCR	45.5	47.1	46.3
CMOLRS	41.48	42.29	41.88
S3CMR	45.9	45.8	45.9
SMDCR	**47.5**	**48.3**	**47.9**

Numbers in boldface are the best.

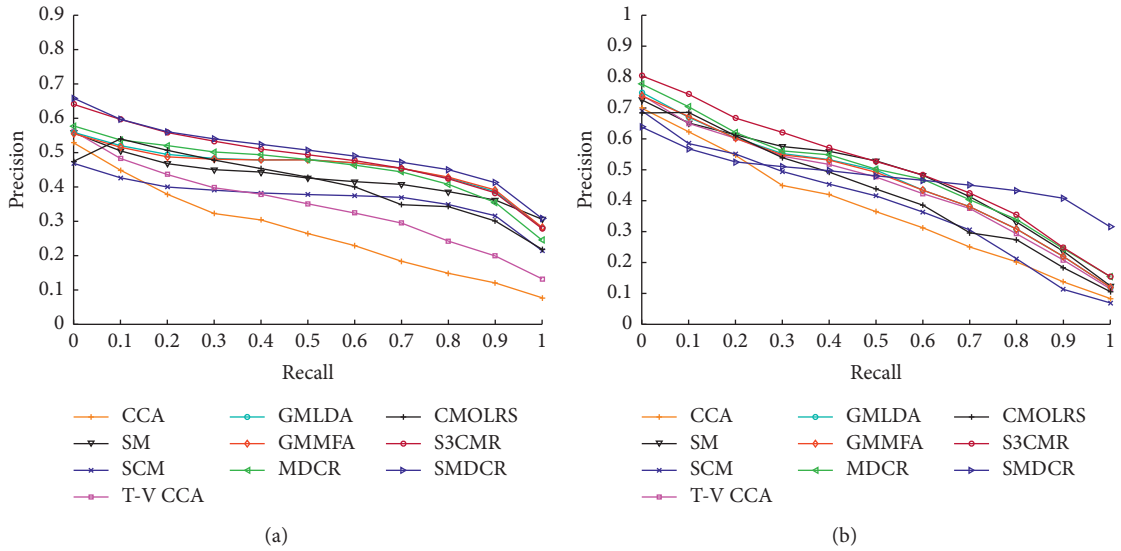


Figure 5: P-R curves of SMDCR method versus compared methods on Pascal Sentence dataset. (a) I2T. (b) T2I.

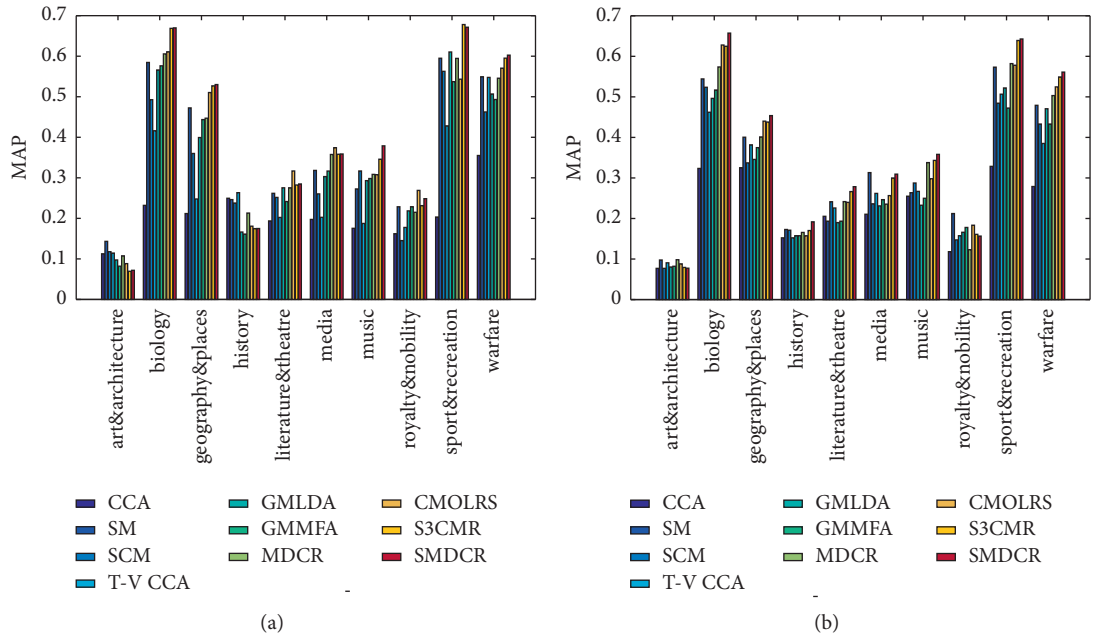
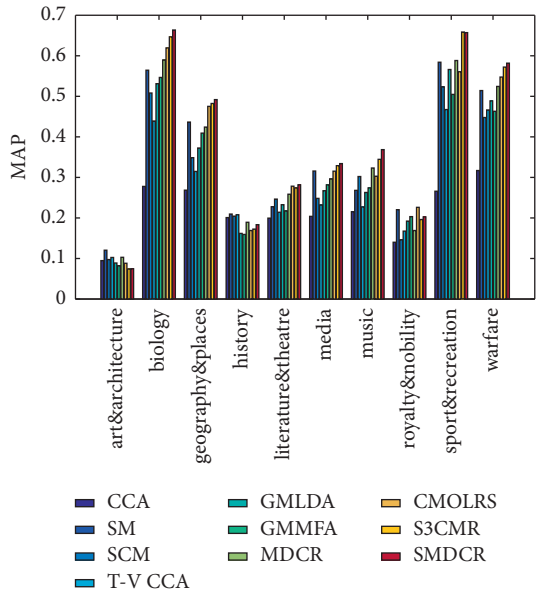
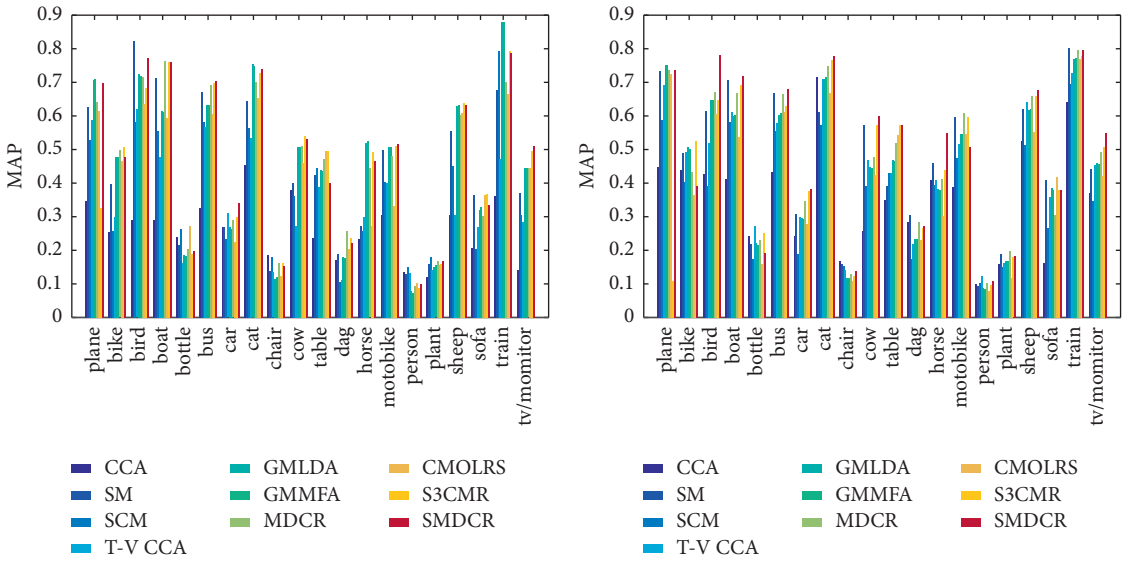


Figure 6: Continued.



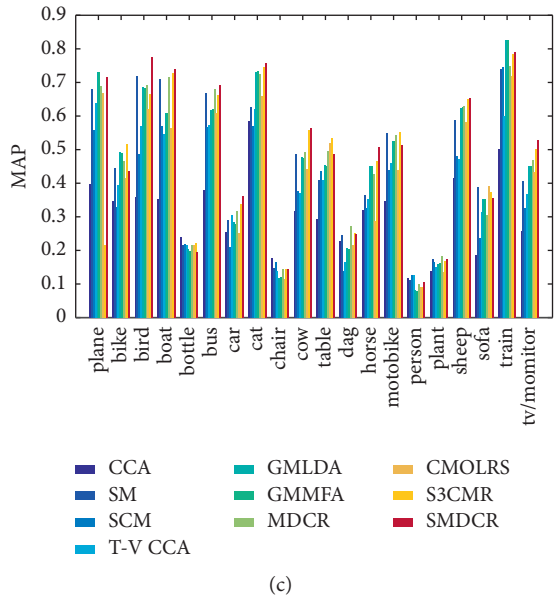
(c)

Figure 6: The MAP scores for each class on Wikipedia-2. (a) I2T on Wikipedia-2. (b) T2I on Wikipedia-2. (c) Average MAP on Wikipedia-2.



(a) (b)

Figure 7: Continued.



(c)

Figure 7: The MAP scores for each class on Pascal Sentence. (a) I2T on Pascal Sentence. (b) T2I on Pascal Sentence. (c) Average MAP on Pascal Sentence.

5. Conclusion

In recent years, with the wide spread of the concept of big data expanding to various fields, the concept of big data retrieval is gradually being established. The task of data retrieval is becoming more and more serious, and the need to solve retrieval accuracy problems is becoming more and more important. Additionally, the accuracy of the method often determines the speed and the application width. Therefore, an improved and efficient retrieval method should be proposed to serve people in daily life, work efficiency, scientific and technological development, etc. For cross-media retrieval, the biggest difficulty lies in the heterogeneity of different modalities data, which causes a semantic gap. In this paper, a cross-media retrieval method based on semisupervised learning and alternate optimization is proposed for the mutual retrieval between the text and image.

Cross-media retrieval method based on semisupervised learning and alternate optimization mainly involves semantic gap. Using a large amount of media data either labeled or unlabeled for semisupervised learning, semantic learning of labeled data to unlabeled data is applied to improve the accuracy of cross-media retrieval between texts and images. The proposed method shows a high power in cross-media retrieval tasks and brings a significant improvement in retrieval accuracy. Extensive experiments show the effectiveness of our proposed method as compared to the nine other state-of-the-art methods.

Data Availability

All data included in this study are available upon request by contact with the corresponding author.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the Talent Project of Shandong Women's University (2018GSPGJ08, 2018RC34061, and 2020RCYJ21) and the Discipline Talent Team Cultivation Program of Shandong Women's University under Grant 1904.

References

[1] M. Din, X. Zhai, and Y. Peng, "Cross-media retrieval by cluster-based correlation analysis," in *Proceedings of the 2013 IEEE International Conference on Image Processing*, Melbourne, Australia, September 2013.

[2] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: concepts, methodologies, benchmarks and challenges," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2372–2385, 2017.

[3] F. Cheng, Z. Hong, W. Fan, and H. Barry, "Image recognition technology based on deep learning," *Wireless Personal Communications*, vol. 102, pp. 1017–1933, 2018.

[4] C. Ma, W. Xia, F. Chen et al., "A content-based remote sensing image change information retrieval model," *ISPRS International Journal of Geo-Information*, vol. 6, no. 10, p. 310, 2017.

[5] H. Zhang, J. Li, Y. Huang, and Z. Liangpei, "A nonlocal weighted joint sparse representation classification method for hyperspectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2056–2065, 2017.

[6] H. Guo and Y. Zhou, "An algorithm for mining association rules based on improved genetic algorithm and its

application," in *Proceedings of the 2009 Third International Conference on Genetic and Evolutionary Computing*, pp. 117–120, Guilin, China, October 2009.

[7] Q. Deng, "A BP neural network optimisation method based on dynamical regularization," *Journal of Control and Decision*, vol. 6, no. 2, 2019.

[8] Y. Wei, Y. Zhao, C. Lu et al., "Cross-modal retrieval with cnn visual features: a new baseline," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 449–460, 2017.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research Archive*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[10] F. Lan, "The discriminate analysis and dimension reduction methods of high dimension," *Open Journal of Social Sciences*, vol. 3, no. 3, pp. 7–13, 2015.

[11] Y. Shen, P. A. Traganitis, and G. B. Giannakis, "Nonlinear dimensionality reduction on graphs," in *Proceedings of the 2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Curacao, December 2017.

[12] A. A. Pchelkin and A. N. Borisov, "On the sensitivity of the neural network implementing the principal component analysis method," *Automatic Control and Computer Sciences*, vol. 43, no. 4, pp. 195–202, 2009.

[13] C. H. Park and H. Park, "A comparison of generalized linear discriminant analysis algorithms," *Pattern Recognition*, vol. 41, no. 3, pp. 1083–1097, 2008.

[14] J. Wu, Z. Lin, and H. Zha, "Joint latent subspace learning and regression for cross-modal retrieval," in *Proceedings of the ISIGIR'17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Tokyo, Japan, August 2017.

[15] N. J. Roseveare and P. J. Schreier, "Model-order selection for analyzing correlation between two data sets using CCA with PCA preprocessing," in *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Australia, April 2015.

[16] Z. Wu, H. Zhu, G. Li et al., "An efficient wikipedia semantic matching approach to text document classification," *Information Sciences*, vol. 393, pp. 15–28, 2017.

[17] S. Ma and X. Sun, "A semantic relevance based neural network for text summarization and text simplification," *Computer Science*, 2017.

[18] J. Costa Pereira, E. Coviello, G. Doyle et al., "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 521–535, 2014.

[19] A. Sharmay, A. Kumar, H. Daume III, and D. W. Jacobs, "Generalized multiview analysis: a discriminative latent space," in *Proceedings of the Computer Vision and Pattern Recognition (CVPR), 2012*, Providence, RI, USA, June 2012.

[20] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 210–233, 2012.

[21] X. Chang, H. Shen, S. Wang, J. Liu, and X. Li, "Semisupervised feature analysis for multimedia annotation by mining label correlation," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD-14)*, Tainan, Taiwan, May 2014.

[22] Y. Wu, S. Wang, W. Zhang, and Q. Huang, "Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Hong Kong, China, July 2017.

[23] X. Zheng, W. Zhu, Z. Yu, and M. Zhang, "Semi-supervised learning based semantic cross-media retrieval," *IEEE Access*, vol. 9, pp. 75049–75057, 2021.

[24] H. Zhang, F. Wu, Y.-T. Zhuang, and J.-X. Chen, "Cross-media retrieval method based on content correlations," *Chinese Journal of Computers*, vol. 31, no. 5, pp. 820–826, 2009.

[25] W. Xia, W. Tang, and W. Xiao, "Support vector machine based on hybrid kernel function," *Journal of Chongqing University of Technology*, vol. 154, pp. 127–133, 2011.

[26] W. Wang and K. Livescu, "Large-scale approximate kernel canonical correlation analysis," *Computer Science*, vol. 4, 2016.

[27] V. Girotto and S. Pighin, "Basic understanding of posterior probability," *Frontiers in Psychology*, vol. 6, Article ID 680, 2015.

[28] I. Gialampoukidis, A. Moumtzidou, D. Liparas, T. Theodora, V. Stefanos, and I. Kompatsiaris, "Multimedia retrieval based on non-linear graph-based fusion and partial least squares regression," *Multimedia Tools and Applications*, vol. 76, no. 1, pp. 22383–22403, 2017.

[29] Z. Fan, Y. Xu, W. Zuo et al., "Modified principal component analysis: an integration of multiple similarity subspace models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1538–1552, 2017.

[30] Y. Wei, Y. Zhao, Z. Zhu et al., "Modality-dependent Cross-media Retrieval," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 4, pp. 1–13, 2015.

[31] Z. Ye and Z. He, "Fast image registration method based on Harris and sift algorithm," *Chinese Optics*, vol. 8, no. 4, pp. 574–581, 2015.

[32] A. A. Aburomman and M. B. I. Reaz, "Ensemble SVM classifiers based on PCA and LDA for IDS," in *Proceedings of the 2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)*, Putrajaya, Malaysia, November 2016.

[33] Y. Yang and J. O. Pederson, "A comparative study on feature selection in text categorization," in *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*, pp. 412–420, San Francisco, CA, USA, July 1997.