SPARSE CONSTRAINT NEAREST NEIGHBOUR SELECTION IN CROSS-MEDIA RETRIEVAL

Zechao Li, Jing Liu, Hanqing Lu

Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{zcli, jliu, luhq}@nlpr.ia.ac.cn

ABSTRACT

With the rapid increasing multimedia documents including videos, images or text, the cross-media retrieval is being focused on. Currently, most of state-of-art methods belonging to the retrieval methods are developed within the scope of the transductive learning. And as soon as the query samples are outside the database, $k$-nearest-neighbor method is always adopted. However, under such circumstances the fixed global parameter $k$ is not robust for all queries with diverse semantics. In this paper, we propose an alternative method based on the sparse representation. The query sample is considered as a sparse linear combination of all training samples, and the number of nearest neighbors is determined automatically according to the sparse coefficients to the query. Then we import the selection of nearest neighbors into a cross-media ranking model with Local Regression and Global Alignment (LRGA) to get the relevant documents to the query. We conduct extensive experiments for cross-media retrieval to demonstrate the efficiency and effectiveness of our methods.

Index Terms— Cross-Media Retrieval, Transductive Learning, Sparse Representation, $L_1$-Minimization

1. INTRODUCTION

With the development of the Internet and digital multimedia, various multimedia data including images, text and audio greatly enrich resource depository on the web. Typically, a webpage designer employs a mixture of text, images and videos, viewed as a multimedia document (MMD), to describe a specific topic more comprehensively and vividly. Such rich multimedia content enables a new type of information retrieval, i.e., the cross-media retrieval. Different from traditional single-media information retrieval like CBIR or TBIR, the research results of the cross-media retrieval may be of different modalities from the query samples. That is, the users can query whatever they want by submitting whatever they have in the cross-media retrieval.

The intrinsic problem for cross-media retrieval is to mine the semantic correlations among the heterogeneous multimedia data. There are some works focusing on the content-based cross-media retrieval [1, 2, 3, 4]. Wu et.al. [1] constructed an isomorphic subspace based on Canonical Correlation Analysis to learn the multi-modal correlations. Yang et.al. [3] proposed a two-level manifold learning method to find an optimal representation of media objects as well as MMDs for cross-media retrieval. In [4], the researchers proposed a ranking algorithm for the cross-media retrieval, namely ranking with Local Regression and Global Alignment (LRGA), which learns a robust Laplacian matrix for data ranking. For each data point, they employed a local linear regression model to predict the ranking value of its top $k$ nearest neighboring points. Furthermore, they proposed a global objective function to assign an optimal ranking value to each point.

Most of the above methods belong to the transductive learning, which means that they are test set dependent model. As a consequence, such methods should be calculated when the test data are known. If the test data are outside the database, most of existing solutions attempt to approximate the data with their $k$ nearest neighbors, in which the parameter $k$ is usually assigned to be global constant. In fact, such experiential assignment lacks in reasonability. Considering that there are various media items from diverse semantic topics or object categories, how to learn an adaptive value of $k$ to different situations is really necessary.

The sparse representation has been widely applied by the computer vision community, due to its powerful descriptive ability. Wright et.al. [5] proposed a robust face recognition method via sparse representation, which can handle occlusion and corruption well and achieve striking performance in the task. Wang et.al. [6] applied the sparse representation to large scale natural image classification.

In this paper, we utilize the sparse constraint query representation for the cross-media retrieval. We treat the query as a sparse linear combination of all the training samples, and consider the training samples, whose corresponding coefficients are more than a very little positive number, as the nearest neighbors of the query. Thus, we have no need to care the availability of query sample in database and avoid the problem of parameter selection in $k$-nearest-neighbor method. Sparse representation can be viewed as a transductive version of the $k$-nearest-neighbor method. So we can combine the sparse representation with many other ranking methods. In this paper, we import the sparse query representation into a...
cross-media ranking model with Local Regression and Global Alignment (LRGA) [4] so as to get the relevant multimedia documents to the query. For each data point, the model employed a local linear regression model to predict the ranking values of its neighboring points, which are decided by the sparse query representation in this paper. In order to assign an optimal ranking value to each data point, LRGA utilized a unified objective function to globally align local linear regression models for all the data points.

The rest of this paper is organized as follows: Section 2 introduce the basic idea of LRGA. The details of SLRGA algorithm are presented in Section 3. Section 4 shows the experimental results when the query samples are in or out of database. Conclusions are given at last.

2. BASIC IDEA OF LRGA

We will introduce the basic idea of LRGA as follows. Formally, there are \( N \) MMDs in the training set, represented as a matrix \( X = [x_1, x_2, ..., x_N] \), \( x_i \in \mathbb{R}^d \), where \( d \) is the feature dimension. We also define \( f = [f_1, f_2, ..., f_N] \in \mathbb{R}^N \), where \( f_i \) is the ranking value of \( x_i \) with respect to its relevance to the query sample, and \( y = [y_1, y_2, ..., y_N] \in \mathbb{R}^N \), where \( y_i = 1 \) if \( x_i \) is the query and 0 otherwise.

The LRGA model jointly employs the information of the query/queries and the distribution of all the data points to predict the final ranking results in cross-media retrieval. To utilize the information from query/queries, we minimize the following objective function:

\[
\min_{f \in \mathbb{R}^N} \sum_{i=1}^{N} U_{ii} (f_i - y_i)^2 = \min_{f \in \mathbb{R}^N} (f - y)^T U (f - y),
\]

(1)

where \( U \) is a diagonal matrix to assign weights to different data points in order to adapt to users’ search intention. For a data point \( x_i \) that is not the query, we set its corresponding \( U_{ii} = 1 \), because we have no prior knowledge of the point, and \( U_{ii} = \infty \) (a large constant) if \( x_i \) is the query.

From another view, we employ a locally linear regression model to predict the ranking value of each data point by its neighboring points. To assign an optimal rank value to each data point, we globally align all the local regression models through minimizing the following function as in [4]:

\[
\min_{f(i), \alpha, z} \sum_{i=1}^{N} f(i) L_i f(i) = \min_f f^T L f,
\]

(2)

where \( L_i = H - H X_i^T (X_i H X_i^T + \lambda I)^{-1} X_i H, H = I - \frac{1}{N} \sum_{k=1}^{N} 1_k \), and \( 1_k \in \mathbb{R}^{k+1} \) is a vector with all ones. Specifically, \( f(i) = [f_1, f_2, ..., f_k] \) and \( X_i = [x_i, x_{i_1}, ..., x_{i_k}] \) are the ranking values and the features of \( k \) nearest neighbors of point \( i \) plus itself respectively.

Combining Eq. 1 and Eq. 2, we obtain the final objective function to get the ranking results:

\[
\min_{f} f^T L f + (f - y)^T U (f - y).
\]

(3)

The problem is equivalent to

\[
(L + U) y = U y.
\]

(4)

LRGA is a ranking model based on the transductive learning, which aims to predict the relevance of unlabeled items which are attainable during the training stage. When the examples are outside the database, most methods resort to the \( k \)-nearest-neighbor method, which has some disadvantages as described above. To overcome these problems, we propose an alternative approach that can be viewed as a transductive version of the \( k \)-nearest-neighbor method. This approach utilizes sparse coding to determine the query’s neighbors. Although there is a parameter, it is much more robust than the case in \( k \)-nearest-neighbor method.

3. SPARSE REPRESENTATION FOR RETRIEVAL

3.1. Sparse representation

The goal of sparse coding is to represent input vectors approximately as a weighted linear combination of a set of base vectors. It is well known that the sparse representation problem is NP-hard. Recent development in the theory of sparse representation [7] reveals that if the solution sought is sparse enough, the optimization problem is equal to solving a convex optimization problem of \( L_1 \)-minimization.

In this paper, we represent the query sample as a weighted linear combination of the training samples. So we get the following optimization problem:

\[
\min_{\alpha} \left\| \alpha \right\|_1 \quad \text{s.t.} \quad x^q = X \alpha,
\]

(5)

where \( X \) is the training data as defined in Section 2, and \( \alpha \) is the vector for unknown reconstruction coefficients. This problem can be solved in polynomial time by standard linear programming methods [8]. The feature-sign search algorithm [9] provides an efficient algorithm for solving the \( L_1 \)-least squares problem.

In practice, we have to consider the noise problem because our data is downloaded from the Internet. To solve the noise problem, some researchers [5] proposed to reformulate the optimization problem as:

\[
x^q = X \alpha + z,
\]

(6)

\[
\min_{\beta} \left\| \beta \right\|_1 \quad \text{s.t.} \quad x^q = B \beta,
\]

(7)

where \( z \in \mathbb{R}^d \) is a noise term, \( B = [X, I] \in \mathbb{R}^{d \times (N+d)} \) and \( \beta = [\alpha^T, z^T]^T \). The equivalent optimization problem is:

\[
\min_{\beta} \lambda \left\| \beta \right\|_1 + \frac{1}{2} \left\| x^q - B \beta \right\|_2^2,
\]

(8)
which can be solved by the feature-sign search algorithm [9].

Then, we can found the query sample’s neighbors in the training set, whose corresponding sparse coefficients are more than a little positive value. We denote the indices of the nearest neighbors as $nn_q = \{i|\beta_i > 0, i = 1, ..., N\}$.

3.2. Combined with LRGA

Based on the sparse representation described as above, we have found nearest neighbors of a query example. We can combine sparse representation with many methods. In this section, we import the sparse representation into LRGA, namely SLRGA. The refined ranking is listed below:

Algorithm 1 The procedure of Sparse LRGA

1. **Input:** The training data $X = [x_1, x_2, ..., x_N], x_i \in \mathbb{R}^d$. A query sample $x^q \in \mathbb{R}^d$. The query information $y = [y_1, y_2, ..., y_N]$.

2. **Training:** Compute $L$ according to [4].

3. **Sparse Representation:** The sparse representation is attained by solving the optimization problem as Eq. 7. The nearest neighbors of the query is $\{x_i|i \in nn_q\}$ and the query information is $\{y_i = \beta_i|\beta_i > 0, i = 1, ..., N\}$.

4. **Getting results:** Solve the linear equation in Eq. 4 and the ranking vector is $f = (L + U)^{-1}Uy$.

5. **Output:** The ranking result in descending order.

4. EXPERIMENTS

The proposed algorithm was evaluated for cross-media retrieval. We first introduce the setup of these experiments, and then provide the detailed experimental results.

4.1. Experiment Setup

The database contains 1125 multimedia objects consisting of 500 images, 125 audios and 500 texts, which are collected from Multimedia Cyclopaedia, science and E-business Web-pages, educational films and news videos shots, etc. We divided them into 5 semantic categories and each semantic category contains 100 MMDs. These multimedia objects are divided into two non-overlapped groups. The first group contains 900 multimedia objects (including 400 images, 100 audios and 400 texts) from 400 MMDs, which are from 5 semantic categories and each semantic category contains 80 MMDs. The first group is used as the training sample, and the rest are used to test the performance.

For image objects, three types of color features (color histogram, color moment, color coherence) and three types of texture features (Tamura coarseness histogram, Tamura directionality, MSRSAR texture) are used and for audios, four types of features (RMS energy, Spectral Flux, Rolloff, Centroid) are used. We use TF/IDF feature for text objects. As to the distance, we use the Euclidean distance for images, DTW distance for audios and cosine distance for texts. If the returned result and the query sample are in the same semantic category, it is regarded as a correct result.

The search performance is evaluated using average precision (AP), which is defined as

$$AP = \frac{1}{\min(R, k)} \sum_{j=1}^{k} \frac{R_j}{j} \times I_j,$$

where $R$ and $R_j$ are the total number of true positive in the training set and the top-$j$ samples respectively; $I_j = 1$ if the $j$-th sample is correct and 0 otherwise. To aggregate the performance over multiple queries, mean average precision (MAP) is used.

4.2. Cross-Media Retrieval

Our goal is to enable users to search whatever they want by submitting whatever they have. In this section, we conduct experiments to test the cross-media retrieval performances when the queries are inside the database and outside the database.

Figure 1 compares our method with LRGA and shows the experiment results of querying audio clips by images. In the experiment, we set the parameter $\lambda = 0.01$ in sparse representation and $k = 10$ in $k$-nearest-neighbor method respectively. The mean average precision of both methods are evaluated. As shown in the Figure 1, our approach is a little better than LRGA. Because the queries are in the database, the queries’ semantics could be employed in the training stage. Anyway, it could illuminate that our method is more effective for cross-media retrieval.

Figure 2 compares the performances of Sparse LRGA and LRGA when the query examples are outside the database, i.e., the users query audio clips using examples of images that are inside the database.
In this paper, we proposed an alternative approach for the transductive learning to solve the out of sample problem in retrieval area, rather than the popular $k$-nearest-neighbor method. It adopted the sparse representation model to adaptively choose the number of neighbors for each query, thereby conquered the issue of not being robust when confronted with situations of unbalance samples in different semantic categories, which has largely affected the $k$ nearest neighbors method that just introduced a fixed global number of neighbors. Experiments have demonstrated that the mean average precision of the refined ranking model outperforms the original LRGA.

5. CONCLUSIONS

In this paper, we proposed an alternative approach for the transductive learning to solve the out of sample problem in retrieval area, rather than the popular $k$-nearest-neighbor method. It adopted the sparse representation model to adaptively choose the number of neighbors for each query, thereby conquered the issue of not being robust when confronted with situations of unbalance samples in different semantic categories, which has largely affected the $k$ nearest neighbors method that just introduced a fixed global number of neighbors. Experiments have demonstrated that the mean average precision of the refined ranking model outperforms the original LRGA.

6. ACKNOWLEDGE

This work was supported by the National Natural Science Foundation of China (Grant No. 60903146 and 90920303), and 973 ProgramProject No. 2010CB327905

7. REFERENCES


