Multi-view News Video Topic Tracking Approach

Xiaoxia Huang, Zhao Lu*
Department of Computer Science and Technology, East China Normal University, 200241, Shanghai, China
xxhuang@ica.stc.sh.cn,zlu@cs.ecnu.edu.cn

Abstract. Existing researches on tracking topics of news videos require lots of labeled examples. However, video labeling is too time-consuming to generate a large number of labeled videos in real applications. In this paper, a novel approach is proposed to track news video topics through using only a few labeled samples. The three main characters of proposed approach are: (1) Multi-view learning process is used for the lack of labeled training samples existing in the topic tracking processing. (2) Principal Component Analysis is used to simplify the feature vectors and speed up processing. And (3) a collaborative training process is proposed to track the related news videos with only one labeled sample. Experiment conducted on our collected Chinese news videos show that the proposed approach can get a good performance with 82.5% precision and 96.4% recall.

Keywords: Topic tracking; Multi-view learning; Principal Component Analysis; Collaborative training

1 Introduction

News video streams are easily accessible in many forms such as news video broadcasts, blogs and podcasting because of the explosion of Internet bandwidth. It is important for users to organize news videos according to their topics. News Topic Detection and Tracking (TDT) is a research program investigating methods for automatically organizing news stories by the events that they discuss. It can be used to link evolving and historical stories according to topics.

News topic tracking is normally studied with textual features as the underlying cues [1, 2, 3]. In addition to text transcripts, news videos provide richer visual information. In news videos, there are a number of near duplicates that appear at different dates and are across various broadcast sources. Wu et al. [4] presented a system built on visual near-duplicate constraints, which is applied on top of text to improve the story clustering and mining. This work depends on manual near-duplicate labeling, which is impossible to handle with large-scale databases.

Some multi-modality fusion studies are also suggested. Hsu et al. [5] tracked four topics with visual duplicates and semantic concepts, and found that near duplicates significantly improve tracking performance. This work uses near-duplicate and textual information as two independent modalities but the potential semantic relatedness between them are not well explored.

News Videos have a complex structure and varieties of media modes. It includes images from photographer, sounds from announcer, texts from subtitle and other types of media. News topic tracking is a combination of multi-view information to reveal the

* Corresponding author
news development. Furthermore, approaches of multi-view learning [6, 7] exploit multiple redundant views to effectively learn from unlabeled data by training a set of classifiers defined in each view. Multi-view learning can get a better performance when compared to learning with only a single view, especially when the strengths of one view complement the weaknesses of the other.

Considering multi-view features of news videos and less labeled news videos in real applications, in this paper, we proposed a multi-view news video topic tracking approach (MVTT) using one labeled video. This approach starts from extracting multi-view features of news videos, takes the principal component analysis (PCA) approach to integrate the features, and a collaborative training method is used to classify and track the news videos with similar topics. Conducted on our collected Chinese news videos, the experimental results show that the MVTT approach gets a relatively better performance and its precision and recall are 82.5% and 96.4% respectively.

2 The MVTT approach

Reports of an event are usually the display of news event from multi-views, such as the announcer’s broadcast information, the visual scene information of events and specific personages’ speaking and text information. Among those views, the visual scene and broadcaster’s voice information are of the most importance. In addition, the news events develop over time and have the features of continuity and succession. In other words, the news related visual and audio data will be repeated or slightly modified in the related topic news. So we can make the hypothesis that the related visual scene data and text data have a certain degree of correlation in the news videos with the same topic.

![Fig. 1. The high-level flow of the MVTT approach](image-url)
This paper draws on the idea of multi-view method to achieve news topic tracking. The steps of the suggested approach are summarized as follows: 1) Extract the visual features of videos. Transform the RGB features into one principal component by using PCA. 2) Extract the voice text features of videos. Generate the TFIDF feature of the text features. 3) Define the confidence. Combine the two views’ classifiers by defining a joint confidence so as to lay the foundation of similarity measurement. 4) On the basis of classifiers’ combination, take the collaborative training to expand the labeled video set and track the same topic videos. The high-level flow of the MVTT approach is shown in Figure 1.

2.1 Extracting multi-view features from a news video
For the news video sample \( q \), the visual view \( x \) and the text view \( y \) are considered. Two views represent the two attribute sets of the sample \( q \), and \( c \) is used to represent the topic of video \( q \). Thus the video \( q \) can be described as \( q = (x, y, c) \), \( x \in X \) and \( y \in Y \). \( X \) and \( Y \) represent the feature sets from two views. For the easy discussion, we limit the value of \( c \) as \( c \in \{0,1\} \), here 0 presents the unrelated news videos and 1 represents the related ones. In addition, we use \( U = \{(x_i, y_i, c_i)\} \) to represent the unlabeled news videos, where \( i = 1, 2, ..., m \).
• **Visual features extraction**

To get sufficient visual features of a video, we select the frames using the equal interval sampling method, and the sampling interval is set to be 1 second. In this way, the corresponding number of frames can be selected according to the length of a video. For example, 30 frames are extracted from a 30 seconds video. For each frame, we take a K-means cluster to get the main visual information in RGB color. Through a large number of experiments, we found the 10-means is suitable for the visual feature extraction. It can avoid information dropout and reduce the complexity of information processing. Thus, the video gets \( \text{length} \times 10 \) RGB characters in the visual view, as shown in Figure 2.

Considering the RGB characters of a video, we adopt the PCA method to study the relations between RGB colors. Without loss of generality, we randomly select a video in the sample space to test in the SPSS (Statistical Product and Service Solutions) statistical analysis software [8]. That video is with a length of 124 seconds, so 124 frames are sampled and 1,240 RGB features are extracted in visual view. The transformation of PCA is defined in a way that the first principal component has the largest possible variance, and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. PCA can be done by eigenvalue decomposition of a data covariance matrix. In this way, we can obtain the eigenvalue of components shown in Table 1, where \( C_i \) means the \( i \)th component.

Based on the strategy of principal component selection, the component with eigenvalue \( \geq 1 \) will be selected as the principal component, which is marked as \( P \). From Table 1, the component1 with eigenvalue = 2.749 will be selected as the principal component of RGB features, here represented by \( P(\text{RGB}) \). In other words, the component1 can represent the features of RGB. The correlation coefficients between the component1 and RGB features are \( r(R) = 0.930 \), \( r(G) = 0.993 \), \( r(B) = 0.948 \) respectively. The high correlation coefficient shows the possibility of replacing variables of RGB features by the component1. The expression of \( P(\text{RGB}) \) is described as follows:

\[
P(\text{RGB}) = C_1 = r(R)/\sqrt{e} \times x_R + r(G)/\sqrt{e} \times x_G + r(B)/\sqrt{e} \times x_B
\]

(1)

Here \( x_R, x_G \), and \( x_B \) are the three components of RGB features and \( e \) is the eigenvalue of component1.

• **Text features extraction**

After recognizing the text of the video voice by automatic speech recognition (ASR) [9], two Chinese language preprocessing tasks are conducted: the first task is segmenting the document by Chinese word segmentation program to get the separate feature items in words [10]. The second is removing the stop words, such as modal particles, prepositions, conjunctions and punctuation that are less contributed to the text classification.

After the preprocessing, we extract the remaining words as the feature items of the video texts which can be expressed as \( (k_1, k_2, ..., k_n) \). This paper takes the commonly used Vector Space Model (VSM) to represent the text features [11]. VSM expresses a document as a space vector, that is \( W = (k_1, \omega_1; k_2, \omega_2; ...; k_i, \omega_i; ...; k_n, \omega_n) \), here \( \omega_i \) is the corresponding weight of \( k_i \). In this paper, the corresponding weight of the feature item is calculating by TF-IDF value. The formula of calculating weights generally is:

\[
\omega_i = \text{TFIDF}_i = \text{TF}_i \times \log(N/DF_i)
\]

(2)

33
In which, the expression of $TF_i$ is like this

$$TF_i = \frac{t_{fi}}{\sum_{k=0}^{D} t_{fk}}$$

(3)

where $t_{fi}$ means the frequency of feature item $k_i$ appeared in the document $d$. $N$ represents the total number of all training documents and $DF_i$ is the number of documents containing the feature item $k_i$.

### 2.2 The training process of the MVTT approach

Considering there are less labeled samples in real applications, each view is sufficient to describe the topic of a news video. The collaborative training method is used to improve the performance of the multi-view learning method with large numbers of unlabeled news videos. The collaborative training process of the suggested approach is shown in Figure 3.

As one of the most important methods in the theory of semi-supervised learning, the collaborative classification algorithm takes advantage of a small amount of labeled samples and a large number of unlabeled data for training. The collaborative training in the suggested multi-view video learning uses the labeled sample to train and gets a classifier for each view. Then, each classifier picks out a number of videos with the higher degree of confidence from the unlabeled set, views them as positive new labeled ones and puts them into the labeled training set. The training process of multi-view video learning is iteratively repeated until satisfies a stop condition.

In our collaborative training, the video with higher confidence will be selected as the related video and lower confidence will be selected as the unrelated one. First, we define the similarity between the original given sample video and a test video as follows,

$$\text{sim}_{o,i} = \alpha\left(\frac{x_0^T\cdot x_i}{||x_0||\cdot||x_i||}\right) + \beta\left(\frac{y_0^T\cdot y_i}{||y_0||\cdot||y_i||}\right)$$

(4)

Here, the $\text{sim}_{o,i}$ means the similarity between the first given labeled video $<x_0,y_0>$ and a test video $<x_i,y_i>$. $\alpha, \beta$ are weight parameters and $\alpha + \beta = 1$. We consider the $<x_0,y_0>$ as a positive instance and the confidence of $<x_i,y_i>$ is represented by $\rho_i$. The dimensions of the test videos will be modified into the same as the source one by padding 0 or truncating. Higher and lower confidence videos will be selected, thereby the number of labeled videos can be expanded. In the approach, the videos of higher similarity with the original sample will get higher confidence and highly related to the original one. In this way, we choose the higher confidence videos as positive ones which may share the same topic with the first labeled video.

The training process of the approach is as follows.

<table>
<thead>
<tr>
<th>Input:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L = {(x_0, y_0, c)}$ is the labeled sample video</td>
</tr>
<tr>
<td>$U = {(x_i, y_i, c_i)}$, in which $i = 1, 2, \ldots, m$, $c_i$ is an unknown topic of the video</td>
</tr>
</tbody>
</table>

Iterate $k$ times:

1) According to Formula 4, calculate the respective confidence $\rho_i$ between the original given sample video and the test videos $U$. |
2) According to the size of $\gamma_c$, select the higher confidence videos into the labeled positive group $P$ and lower confidence videos into the labeled negative group $M$.

3) Calculate the variances of similarities in the group $P$ and compute the difference between two variances of two iterations, represented as $\Delta_{VAR} = VAR(k) - VAR(k-1)$. If there is $\Delta_{VAR} > \varepsilon$, then the iteration ends.

Output:
Update group $L$ by combining the group $P$ and $L$, go back to (1) until $k$ times.

Output: The final updated related video group $L$ and unrelated video group $M$

$\gamma_c$: According to the different topic $c$, $\gamma_c$ represents the number of selected videos from unlabeled set. $\arg\max_{\gamma_c}(\rho_i)$ represents the positive instances, $\arg\min_{\gamma_c}(\rho_i)$ represents the negative instances.

$\varepsilon$ is set to make the algorithm converge. Here we make the assumption that once the algorithm converged, the similarities of selected videos will be stable and get the smaller differences in the variance. And $\Delta_{VAR} > \varepsilon$ is the stop condition of the iteration.

3 Experiments

In this paper, we gather Xinhua news videos as the video dataset, including 77 news videos in February 19, 2012[12]. The video length ranges from 30s to 4 minutes and each news video covers a specific topic. The types of topics range from the political, economic to culture and so on. We select one video as the labeled sample and verify the approach’s efficiency through the experiment.

We define the video group $P$ recognized by our approach as In Group and the group $N$ as Not in Group. And the video group which is judged by human beings as the same topic is represented as On topic and different topic videos are as Not on topic. The result condition is summarized in Table 2. Here, A, B, C and D are the numbers of videos in the specific condition. According to the definition of Precision, Recall, F-Measure, the evaluation method in this paper can be seen in Table 3.

In the paper, we conduct a series of experiments to check the validation of the proposed approach. The basic experiment is used to see the general performance of the MVTT approach. Second experiment is constructed to make the choice of parameters. In the third experiment, we make a comparison between multi-view and single view.
### Table 2. Result condition summarization

<table>
<thead>
<tr>
<th>In Group</th>
<th>On topic</th>
<th>Not on topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not in Group</td>
<td>B</td>
<td>D</td>
</tr>
</tbody>
</table>

### Table 3. Evaluation method

<table>
<thead>
<tr>
<th>Effective Measure</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision(P)</td>
<td>A/(A+C)</td>
</tr>
<tr>
<td>Recall(R)</td>
<td>A/(A+B)</td>
</tr>
<tr>
<td>F-Measure</td>
<td>2<em>P</em>R/(P+R)</td>
</tr>
</tbody>
</table>

### 3.1 Basic experiment

The experiment has four steps, including feature selection, vectorization, similarity calculation and training iteration. In this experiment, we suppose $\beta = 0.7$ and the reason will be discussed in the later experiment. In the iteration process, $\gamma_c$ is set as 1, here we minimize the number of selected videos so as to avoid a large number of false selected. And $\epsilon$ is set as 0.001 through a lot of experiments. We take a video talking about “Syrian” as the first labeled video $L$ and its two views features can be repressed as $<p_r(RGB), TFIDF_r>$. For the test videos, the features can be represented as $<p_l(RGB), TFIDF_l>$. And then it’s going to calculate the similarity between the sample video and the test videos using Formula 4. After the collaborative training, the higher confidence videos can be labeled and regarded as the related videos sharing the same topic with the first labeled video. There are 7 news videos are related to the topic of “Syrian” in our dataset by human judgment. Our experiment gets 6 related videos of this topic. Those 6 videos are all correct ones and there are one video missed marked. So precision, recall and F-measure of our approach in this topic are 100.0%, 85.7% and 92.3% respectively. And then we test other three topics and the performances obtained are shown in Table 4.

### Table 4. Experimental results of four topics

<table>
<thead>
<tr>
<th>Topic</th>
<th>On topic</th>
<th>In group</th>
<th>Correct</th>
<th>Missed</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syrian situation</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>1</td>
<td>100.0%</td>
<td>85.7%</td>
<td>92.3%</td>
</tr>
<tr>
<td>Xi Jinping visit USA</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>0</td>
<td>66.7%</td>
<td>100.0%</td>
<td>80.0%</td>
</tr>
<tr>
<td>Iranian nuclear issue</td>
<td>10</td>
<td>12</td>
<td>10</td>
<td>0</td>
<td>83.3%</td>
<td>100.0%</td>
<td>90.9%</td>
</tr>
<tr>
<td>Jeremy Lin in NBA</td>
<td>8</td>
<td>10</td>
<td>8</td>
<td>0</td>
<td>80.0%</td>
<td>100.0%</td>
<td>88.9%</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>82.5%</td>
<td>96.4%</td>
<td>88.0%</td>
</tr>
</tbody>
</table>

Experiment results show that the approach has a certain robustness and comprehensive 82.5% precision, 96.4% recall and 88% F-measure value. The precision rate is a little changeable because there are some topics related with each other. And the approach achieves a good recall rate which means low missed detection. So the comprehensive F-measure is comparatively good and the approach maintains a relatively stable efficiency for various topics of news videos.
3.2 Values of parameters

The experiment on the value of parameter $\alpha$, $\beta$, $\varepsilon$ is conducted below. In the experiment, the parameter $\beta$ is changed from 0.01 to 0.99 with the step length of 0.01. The Precision, Recall and F-measure value can be seen as the judgment of the pros and cons of different parameter values. Due to the limitation of length, we only express the experiment result on the F-measure, which is a comprehensive evaluation method. Experiment result is shown in Figure 4.

![Fig. 4. Different F-measure of different parameter](image)

Due to not very large amount of experimental data, the graphics come out with balance jagged. However, the trend analysis can be found when $\beta = 0.7$, the approach obtained the best performance. So in the rest of experiments, we set $\alpha = 0.3, \beta = 0.7$.

And the variance of $\varepsilon$ is selected by a large number of experiments. And the variance of similarity in the topic “Xi Jinping visit the USA” is shown in Figure 5. When the difference of variances larger than $\varepsilon$, the iteration ended and the similarities of selected videos will be stable and get the smaller differences in the variance.

3.3 Comparison between multi-view and single view

In order to maximize the use of news video features, this paper adopts a multi-view learning approach. In this experiment, we discuss the differences between multi-view and single view so as to explore the validity of multi-view method. The obtained results are shown in Table 5.

![Table 5. Comparison between multi-view and single view](image)

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual View</td>
<td>7.6%</td>
<td>15.9%</td>
<td>10.3%</td>
</tr>
<tr>
<td>Text View</td>
<td>43.5%</td>
<td>90.2%</td>
<td>58.6%</td>
</tr>
<tr>
<td>Multi-View</td>
<td>82.5%</td>
<td>96.4%</td>
<td>88.0%</td>
</tr>
</tbody>
</table>

From Table 5, the experimental results show that single visual view cannot express the semantics of the video properly with low precision and recall rate. And the single text view gets high recall rate, low precision rate and the overall F-measure value is not
good enough. So we need a multi-view approach to express the topics properly and different views can be complemented to each other. The proposed approach can achieve higher precision and recall rate and reduce the missed detection rate and false alarm rate in the video topic tracking.

4 Conclusion
In this paper, a new method for news video topic tracking is proposed. A video is a combination of multi-view features. One view is complemented to each other. First, we take the multi-view learning method to deal with the lack of manually labeled training set. Secondly, taking advantage of PCA, better efficiency can be got by projecting the original features to the principal component. Thirdly, we incorporate the collaborative training method to track the news topic from only one labeled news video. The experimental results show that the proposed approach can effectively solve the problem of the lack of the training set in the topic tracking field and obtain a good recognition performance with high robustness.

As the future work, we are planning to combine the news topic detection with topic tracking. Also in the feature representation, we will add more semantic information to consider news topic tracking in the semantic level.

Acknowledgement
This work is supported by a grant from the National High Technology Research and Development Program of China (863 Program) (No. 2011AA01A107), the grant from the Shanghai Science and Technology Foundation (No. 11511504000, No.11511502203, No. 2011XJRHT0078 and No. 10dz2253200).

Reference