Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A retrospective event detection method in news video

LING Jian[†1,2], LIAN Yi-qun[2], ZHUANG Yue-ting[1]

([1]*Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027, China*)

([2]*Department of Electronic Information, Zhejiang Institute of Media and Communication, Hangzhou 310018, China*)

[†]E-mail: lingjian@zjut.edu.cn

Received Jan. 10, 2006; revision accepted May 8, 2006

Abstract: In this work we present a probabilistic learning approach to model video news story for retrospective event detection (RED). In this approach, both content and time information on a news video is utilized to transcribe the news story into terms, which are divided into classes by their semantics. Then a probabilistic model, composed of sub-models corresponding to the semantic classes respectively, is proposed. The model's parameters are estimated by EM algorithm. Experiments showed that the proposed approach has better detection resolution.

Key words: Semantic class, EM algorithm, Retrospective news event detection, Maximum likelihood
doi:10.1631/jzus.2006.AS0193 **Document code:** A **CLC number:** TP391

INTRODUCTION

The recent development of multimedia and telecommunication led to the creation and accumulation of large amounts of news videos. To use such large-scale video data, event detection and news story classification have become more important. Motivated by such background issues, retrospective news event detection (RED) was first proposed and defined by Yang *et al*.(1998).

RED is defined as the discovery of previously unidentified events in historical news corpus of stories. Events are defined by their association with stories, so RED groups the stories in the study corpus into clusters, where each cluster represents an event, and the stories in the cluster discuss the event. Though RED has been studied for many years, it is still an open problem (Li *et al*., 2005).

The most prevailing approach for RED is Language Models first proposed by Allan *et al*.(1999) for the first story detection task of TDT. In the language modelling approach, the similarity of a story to an event is measured with the probability of its generation from the event model. Using the unigram assumption of independence of terms, the probability can be computed as the product of probabilities of generation of the terms in the story. One drawback of the unigram language model is that it treats all terms on an equal footing and ignores semantic information on the terms (Yang *et al*., 2002). Nevertheless, such information may usefully convey more information on the topic of the story (Mihalcea and Mihalcea, 2001). Besides speech in news video, the video contains other rich information on the reported event, for example the scene type such as cityscape, landscape, people and the like. Therefore, fusing multi-modal information for RED would improve its efficiency and effectiveness.

This paper proposes a novel approach for RED with the integration of multi-modal information. The remainder of this paper is organized as follows. In Section 2, we classify news video into five semantic classes to present the news video document. In Section 3, a probabilistic model of RED and the model's parameters estimating approach are introduced. Section 4 describes the experiments and results. Conclusion and future work are given in the end.

REPRESENTATION OF VIDEO NEWS STORY

An event is a specific item that happens at a specific time and place along with all necessary unavoidable consequences (Allan *et al*., 1999). Thus, there are certain elements in an event, including when it happened, where it happened, what happened, and how was it involved. These four questions give a hint of the deep understanding of the semantics of a news story in four aspects respectively when we compare two news stories. Based on such consideration, we partition the information extracted from the speech component of the news video into four semantic classes: *person*, *location*, *time*, *terms*, where *person* indicates the set of persons and organizations, *time* indicates the set of dates, *location* indicates the set of places, and *terms* indicates the remainder of the content after removing the above entities and stop words. Additionally, we classify the information extracted from the video component as another semantic class *scene*, which indicates the set of *indoor*/*outdoor*, *face*, *people*, *landscape*, and *cityscape*. For simplicity, we assume that the five kinds of information are independent.

News reports are always aroused by news events, a potentially important property to help RED is the characteristic of count-time distribution of the news report on an event. Although the counts of reports of different events are different, the characteristics of their count-time distribution are similar. Generally speaking, there are less reports at the very beginning of the event, and with the event going on, it is more attention-getting, and more reports are issued, after the count reaches its maximum point, it begins to drop. The count-time distribution is like Gaussian function, where the mean is the point when the count gets its maximum, and the variance is the event duration.

In summary, we present a video news story with five semantic classes as shown in Fig.1, with each semantic class having its own term space. For simplicity, we assume that these semantic classes are independent of each other. Let C={*person*, *location*, *scene*, *time*, *term*} be the set of semantic classes, we define the feature vector of a semantic class $F(c_i, x)$ as:

$$F(c_i, x) = \{w_1, w_2, \ \ldots, \ w_n \mid \forall_{j=1}^{n}(w_j \in x) \wedge w_j \in c_i\}, \quad (1)$$

where $c_i \in C$ is the semantic class, $n=|F(c_i, x)|$ is the feature vector dimension. That is to say, $F(c_i, x)$ is the feature of entities which belong to class C in the news story. Thus, we present a video news story as:

$$\{F(c_1, x), F(c_2, x), \ldots, F(c_5, x)\}, \quad (2)$$

where c_1, c_2, c_3, c_4, and c_5 represent semantic class: *person*, *location*, *scene*, *time*, *term* respectively.

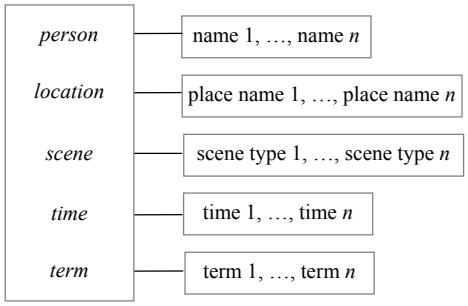

Fig.1 Representation of an video news event

RETROSPECTIVE EVENT DETECTION METHOD

Probabilistic model and parameters estimation

To differentiate entities with different semantics can improve the performance significantly (Lam *et al*., 2001). We model the above five semantic classes respectively. Among the semantic classes, *time* is much different from the others; the information contained in *time* reveals the time-count distribution of the news video documents of the event. As mentioned in the previous section, the distribution is characterized by a like-Gaussian distribution, this leads us to adopt Gaussian Mixture Model (GMM) to model *time*. The other four semantic classes are similar to the same task in text articles. The bag of words model and the Naïve Bayes (NB) classifier of this model work very well on many text classification and clustering tasks (Yang *et al*., 1999). Just like in NB, we use a mixture of unigram models to model these four semantic classes.

Thus, the whole model is the combination of the five mixture models: four unigram models and one GMM. According to the independence assumption made in the previous section, given an event e_j, the

five semantic classes of the *i*th document are conditional independent:

$$p(x_i \mid e_j) = \prod_{\lambda=1}^{5} p(F(c_\lambda, x_i) \mid e_j). \tag{3}$$

By introducing latent variable, the log-likelihood of the joint distribution $l(X; \theta)$ is:

$$\begin{aligned} l(X;\theta) &\equiv \log(p(X \mid \theta)) = \log\left[\prod_{i=1}^{M} p(x_i \mid \theta)\right] \\ &= \sum_{i=1}^{M} \log\left[\sum_{j=1}^{k} p(e_j) p(x_i \mid e_j, \theta)\right], \end{aligned} \tag{4}$$

where X represents the corpus of the news documents, M and k are number of news documents and number of events respectively (Li *et al*., 2005). We adopt Expectation Maximization (EM) algorithm to estimate the parameters of the models. EM algorithm is generally applied to maximize log-likelihood. The parameters are estimated by running E-step and M-step alternatively until a local maximum is reached. In E-step, we compute the posteriors, $p(e|x_i)$, by:

$$p(e_j \mid x_i)^{(t+1)} = \frac{p(e_j)^{(t)} p(x_i \mid e_j)^{(t)}}{p(x_i)^{(t)}}, \tag{5}$$

where the super script t indicates the tth iteration. In M-step, we update the parameters of all the sub-models respectively, in which *person*, *location*, *scene* and *term* are independent mixture of unigram models, while *time* is GMM.

The parameters of unigram models are updated by:

$$p(w_n \mid e_j)^{(t+1)} = \frac{1 + \sum_{i=1}^{M}\left[p(e_j \mid x_i)^{(t+1)} f(i,s)\right]}{N + \sum_{i=1}^{M} p(e_j \mid x_i)^{(t+1)} \times \sum_{s=1}^{N} tf(i,s)}, \tag{6}$$

where x_i is the ith document, $tf(i, n)$ is the count of entity w_n in x_i and N is the vocabulary size.

The parameters of GMM are updated by:

$$\begin{aligned} \mu_j^{(t+1)} &= \frac{\sum_{i=1}^{M}\left[p(e_j \mid x_i)^{(t+1)} time_i\right]}{\sum_{i=1}^{M} p(e_j \mid x_i)^{(t+1)}}, \\ \sigma_j^{(t+1)} &= \frac{\sum_{i=1}^{M}\left[p(e_j \mid x_i)^{(t+1)} (time_i - \mu_j^{(t+1)})^2\right]}{\sum_{i=1}^{M} p(e_j \mid x_i)^{(t+1)}}. \end{aligned} \tag{7}$$

The mixture proportions are updated by:

$$p(e_j)^{(t+1)} = \frac{\sum_{i=1}^{M} p(x_i \mid e_j)^{(t+1)}}{M}. \tag{8}$$

Because Gaussian function can be regarded as a window in time line, and the window's width can be adjusted by changing its mean and variance, GMM improves performance distinctly compared with the fixed widows or the parameter-fixed decaying functions.

Estimation of initial number of events

In the initializing stage, two initial values have to be estimated: (1) the number of events in the corpus; (2) the maximum point of count-time distribution of each event. Just like the magic number in clustering applications, an initial estimation of these two values is quite difficult. As in the analysis in Section 2, the information on the raw news video is first extracted and classified into some semantic classes. We utilize the temporal information in the semantic class *time* and the publish date of the news to determine the initial event number in the corpus and the maximum point of news number of an event. That is to say, we make a rough classification to decide the initial values needed in the farther stage.

Generally, a temporal expression needs to be evaluated with respect to the time when the document is published. We map each temporal expression to the time-axis as range with a pair of dates (*start*, *end*). Another problem is the date format; one date could be expressed in various formats, so we need to specify a uniform date format such as "yyyymmdd". Thus, for example, "3 days ago" results in pair (20041018,

20041018) if the publish day is Oct. 21, 2004.

After setting down the approximate date of each news document in the corpus, we can make a rough classification of the documents by this date. The documents with the same date are assumed to report the same event and are classified into the same group. Within a group we can check the relation between dates and count to determine the maximum point of count-time distribution of the corresponding event. Obviously, the number figured out in such way is less than the actual one. We split the group by the approach presented by Li *et al.*(2005). To select the best events number, one available measure of the fitting goodness is the log-likelihood. Given this indicator, we apply the Minimum Description Length (MDL) principle to select among values of k:

$$\begin{aligned} k &= \arg\max_k(\log(p(X;\theta)) - (m_k \log(M))/2), \\ m_k &= 3k - 1 + k(N_p - 1) + k(N_l - 1) + k(N_n - 1), \end{aligned} \quad (9)$$

where $\log(p(X;\theta))$ is expressed in Eq.(4) and m_k is the number of free parameters needed for the model. As a consequence of the principle, when models with different values of k fit the data equally well, the simplest model is selected.

EXPERIMENTS

In order to evaluate the effectiveness of the approach presented in this paper, we designed an experiment running on a corpus of news video selected from various source such as the Internet and TV news programs. The corpus collected contains 30 events annotated from 236 news stories, which span the period from June 20, 2005 to July 1, 2005.

We first segment the raw video into news story, and save each news story with one WAV file for audio and AVI file for video of the news story separately, and then extract information from audio files and video files respectively. A lot of work has been done in both automatic speech recognition (ASR) and video semantic feature extraction, and some of the approaches achieve acceptable performance. For the AVI files, we adopt the approach presented by Ling *et al.*(2005) to detect shot boundary and catalogue every shot into one of typical scene types listed in Section 2. For the WAV file, we scrub out the speech and transcript it into text by the SASDK tool, and then extract the named entities by BBN IndetiFinder tools, which can extract seven types of named entities including persons, organizations, locations, date, etc.

Experimental results use recall and precision defined in Eqs.(10) and (11) as measuring standard. To evaluate the importance of the video information we run an experiment without the scene information. The experimental results are shown in Table 1.

$$recall = \frac{correct}{correct + missed}. \quad (10)$$

$$precision = \frac{correct}{correct + false}. \quad (11)$$

Table 1 Experimental results with and without scene information

	With video information	Without video information
Event number	32	32
Detected stories	202	188
Missed stories	34	48
False stories	23	35
Recall	0.856	0.797
Precision	0.898	0.807

Table 2 shows the results of another experiment with various specified initial events numbers, the results show the affection that caused by initial number.

Table 2 Experimental results in various initial number of the events

Initial event number	Correctly detected stories	Missed stories	False stories	Recall	Precision
26	195	41	58	0.826	0.771
28	198	38	42	0.839	0.786
30*	211	25	25	0.894	0.894
32	202	34	23	0.856	0.898
34	186	50	34	0.788	0.845

*: The actual event number

The experimental results showed that the presented approach in this paper has satisfying performance and that the information extracted from video is quite helpful for event detection and that the initial event number affects very much the precision

of our approach. Table 2 shows that the estimated initial event number considerably influences the results. When the number is less than the actual one, the falsely detected number is more than the missed, while it is contrary when the number is greater than the actual one. This result accords with the result of analysis of EM algorithm. We also find another important factor that tarnishes the experimental results is the poor accuracy of the information extraction and classification approaches.

CONCLUSION

In this paper we proposed a probabilistic model for retrospective event detection in news video. It is composed three major steps. First, the information of news video is extracted and classified into five semantic classes. Then, the news video is modelled into five probabilistic sub-models corresponding to the five semantic classes, and the parameters of these models are estimated by Maximum Likelihood method. Finally, the initial number of events involved in the video news is estimated and the news video documents are grouped by events. Our experimental results over diverse video data from different sources have demonstrated that the probabilistic model is an effective approach for retrospective event detection in news video.

References

Allan, J., Jin, H., Rajman, M., Wayne, C., Gildea, D., Lavrenko, V., Hoberman, R., Captuto, D., 1999. Summer Workshop Final Report. Center of Language and Speech Processing, Johns Hopkins University.

Lam, W., Meng, H., Wong, K., Yen, J., 2001. Using contextual analysis for news event detection. *International Journal of Intelligent Systems*, **16**(4):525-546. [doi:10.1002/int.1022]

Li, Z., Wang, B., Li, W., Ma, W., 2005. A Probabilistic Model for Retrospective News Event Detection. Proceedings of SIGIR Conference on Research and Development in Information Retrieval, p.205-216.

Ling, J., Lian, Q., Zhuang, Y., 2005. A New Method for Shot Gradual Transition Detection Using SVM. Proceedings of 4th International Conference on Machine Learning and Cybernetics, Guangzhou, China, p.768-775.

Mihalcea, R., Mihalcea, S., 2001. Word Semantic for Information Retrieval: One Step Closer to the Semantic Web. Proceedings from International Conference on Tools with Artificial Intelligence, Dallas, Texas, USA, p.280-287.

Yang, Y., Pierce, T., Carbonell, J.G., 1998. A Study on Retrospective and On-Line Event Detection. Proceedings of the SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, p.28-36.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B.T., Liu, X., 1999. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems: Special Issue on Applications of Intelligent Information Retrieval*, **14**(4):32-43.

Yang, Y., Zhang, J., Carbonell, J., Jin, C., 2002. Topic-Conditioned Novelty Detection. Proceedings of the International Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, p.688-693.