Boosting Web Video Categorization with Digraph Propagation

Sansan Hong and Feng Wang*

Abstract—**Web video categorization plays an important role in information retrieval. Most existing approaches employ text, visual, or semantic features to determine the category of a given video. Meanwhile, besides these features extracted from the videos, the webpages on video sharing sites contain abundant external information such as user comments and links between videos, which are good hints for video categorization. In this paper, we propose to employ the hyperlinks between different videos with digraph propagation to boost the performance of traditional video categorization. Our approach is based on the observation that the related videos usually belong to similar categories, and thus a video can vote for the categories of other videos linked to it. Given a set of videos, a digraph is first constructed with the videos as the vertexes and the hyperlinks between all related videos as the edges. Initial weights are assigned to the vertexes with the output category probabilities of the traditional text-based classifiers. We then update the weight of each video by the votes from the linked videos. This is carried out by iteratively propagating the category probabilities from each video to its neighbors. Experimental results show that our approach can significantly improve the performance of the traditional video categorization approaches.**

Index Terms—**Digraph propagation, related video, web video categorization**

I. Introduction

WITH the rapid development of multimedia and network technologies, in the past few years, the number of web video increased explosively. In addition, due to the popularity of video recording equipment, it becomes much easier for people to share homemade videos. The massive amount of video data poses more challenges to the management of the web video corpus.

Video categorization plays an important role in managing and retrieving large-scale video corpus by automatically labeling given videos with a set of predefined categories [7]. Traditionally, researchers focus on extracting various features such as text, visual and audio information for video categorization [1] [2]. In recent years, multi-modality approaches by combining the above three features to compute video similarity are proposed [3] [4] [5]. However, there are some problems with existing approaches for large-scale web video categorization. On one hand, the extraction of semantic features is difficult and has very high computational complexity. Thus, it is not efficient and impractical for massive web video categorization. In addition, web videos contain professional videos as well as homemade videos of poor quality. The uneven quality results in more noise for visual and audio features, and affects the performance of categorization. On the other hand, although video categorization based on the associated textual information is simple, the textual data (such as title, tags, abstract and so on) heavily relies on what the uploaders provide together with the video [6]. Therefore, in most cases, the textual data is incomplete and usually cannot accurately represent the content of the video, which will eventually affect the performance of web video categorization.

Based on the observations of web video sharing sites, almost all video pages provide abundant information associated with videos such as title, tags, abstract, and links to other related videos as illustrated in Fig.1. The videos on video sharing sites are usually dependent on each other, and have mutual relations represented by hyperlinks between related videos. Thus, the videos and the links between them compose a graph to represent the relationships among all videos. Generally, the related videos in a video page belong to the same category as the current video, which are good hints for determining the genre of a given video.

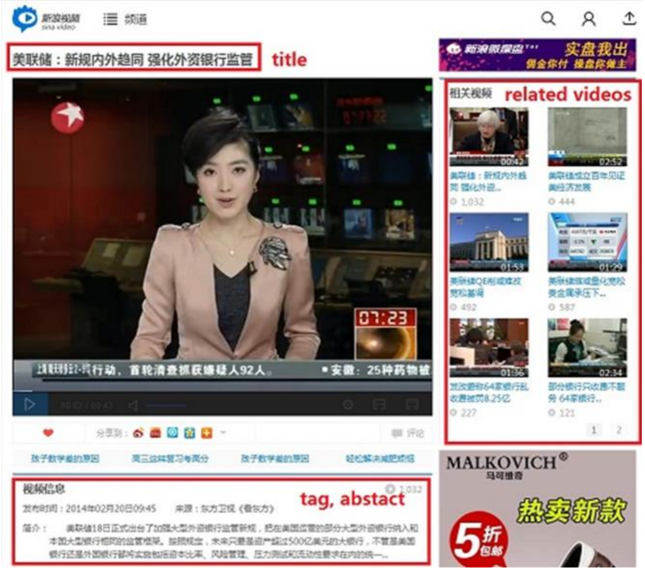

Fig. 1. External information associated with videos on sina video website.

Based on the above observation, in this paper, we propose a novel approach to boost web video categorization by employing the hyperlinks between videos. The basic idea of our approach is as follows. In a video page, the hyperlink from video A to video B can be thought as a vote of category support of video A to video B. In other words, the category of

Manuscript received July 8, 2014; revised August 5, 2014.

The work described in this paper was sponsored by the National Natural Science Foundation of China (No. 61103127, No. 61375016), SRF for ROCS, SEM, Shanghai Pujiang Program (No. 12PJ1402700), and the Fundamental Research Funds for the Central Universities.

Sansan Hong is with the Department of Computer Science and Technology, East China Normal University, Shanghai, P.R.C 200241 China(e-mail: sanshong2013@gmail.com).

Feng Wang is with the Department of Computer Science and Technology, East China Normal University, Shanghai, P.R.C 200241 China(e-mail: fwang@cs.ecnu.edu.cn).

* Corresponding Author.

B can be refined by the category of A. First, we construct a digraph by taking all videos as the vertexes and the hyperlinks between them as the edges. Second, SVM classifier is trained to get the probability scores of each test video for each category, and thus the probability scores can be taken as the initial weights of the vertexes of the graph. Third, the weight of each vertex is iteratively updated by the votes from its linked videos. Finally, the weights are used to determine the category of each video.

The remaining of this paper is organized as follows. Section 2 overviews the related work. Section 3 describes our proposed digraph-based approach. Our experimental results are presented and discussed in Section 4. Finally, Section 5 concludes this paper.

II. Related Work

This section reviews the related work on video categorization. First, we introduce the new research trends of video categorization. Second, we discuss some related work of web video categorization by employing social information.

A. *Studies on Video Categorization*

Most existing works employ textual, audio, visual features or their combination [7] for classifier training in video categorization. In [2], X. Yuan et al. propose a relatively scalable video type and subtype hierarchical structure. Ten computable spatio-temporal features are extracted to classify videos by using a hierarchical Support Vector Machines. In [8], audio soundtrack, temporal structure, and color content are exploited to classify video genres. The improvement of this approach is verified respectively from classification perspective, retrieval perspective, and relevance feedback perspective.

In recent years, due to the unsatisfactory performance of video categorization by using a single video feature, some researchers attempt to combine various features such as text, audio, and visual information to improve the performance of video categorization. In [3], Yang et al. propose an approach including two modalities: semantic modality and text modality. The semantic modality includes three features, i.e. concept histogram, visual word vector model, and visual word Latent Semantic Analysis. The text modality includes video title, abstract and labels. The experimental results indicate that the proposed approach can improve the performance of video categorization on three classifiers including Support Vector Machine, Gaussian Mixture Model and Manifold Ranking. Xu et al. [5] extract visual features, Automatic Speech Recognition transcripts, and text data to train a one-vs-all SVM classifier for video categorization.

Darin Brezeale et al. present a survey on automatic video classification [9]. They comprehensively analyze and compare the literatures of video categorization including text-based approaches, audio-based approaches, visual-based approaches, and the combination of the above three modalities.

B. *Web Video Categorization with Social Information*

Generally，video pages of social video sites include additional social information such as the playlist, user interest, user comments, and links to other related videos. These social information can be utilized for video categorization.

In [6], user comments are extracted and a text-based classifier is employed to improve the performance of video categorization. By combining text-based classifier and content-based classifier, the proposed approach outperforms the single feature based classifiers. In [10], Yew et al. address web video categorization from a social construction perspective. The interactions such as user comments on videos, the information about user sharing videos, and user chatting information are valuable to determine the video genre. A Na ïve Bayes classifier is employed to predict the video categories.

In [7], Xiao Wu et al. boost web video categorization by combining the model-based and data-driven approaches. Text features from the title and tags are used to train the model-based classifiers. The related videos and user videos are utilized to meet the shortfall of text features. Moreover, the data-driven approach complements experimental results from the perspective of video relevance and user interest [7]. However, they simply assign a confidence score to each video by counting the total number of its related videos. The mutual links between different videos are not taken into account.

III. Web Video Categorization with Digraph Propagation

In this section, we present our approach for web video categorization by employing the hyperlinks between related videos with digraph propagation. Figure 2 illustrates the framework of our proposed approach. First, SVM classifiers are trained with the textual metadata (video title, tags, and abstract) extracted from video pages. Second, for each video page, we extract all links to other related videos on the video sharing site. A digraph is constructed with the videos as the vertexes and the links between videos as the edges. The category scores obtained by the SVM classifiers are used as the initial weights for the vertexes. Finally, we iteratively update the weights by digraph propagation to reflect the category votes from the related videos.

A. *SVM Classifier with Textual Features*

SVM classifiers are first trained by employing the textual information associated with the web videos. In each video

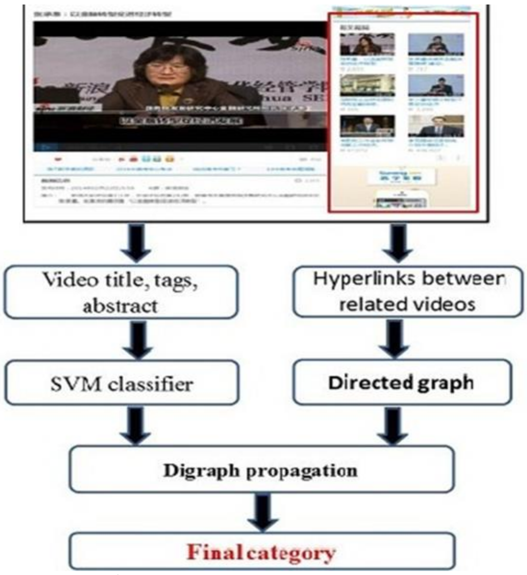

Fig. 2. Framework of web video categorization with digraph propagation

page, we extract the video title, abstract, and tags for SVM training.

By the text preprocessing including word segmentation, stop words removal, and feature selection, a word dictionary is constructed. We use VSM (vector space model) to represent the video documents. TF-IDF is employed to compute the weight of each word. In our implementation, LIBSVM is adopted for SVM training and prediction. LIBSVM is a popular open source machine learning library, developed at the National Taiwan University and both written in C++ though with a C API. LIBSVM implements the SMO algorithm for kernelized support vector machines (SVMs), supporting classification and regression [12]. By the SVM classifiers, we get the probability scores of each video for each category. Generally, the texts extracted from video pages are short and noisy, which will limit the performance of video categorization. Thus, it is necessary to integrate other information in the classification process.

B. Construction of Digraph

Besides the textual information, we also extract the links to other related videos on each video page. These links are valuable for video categorization since they present the relationships between different videos from certain perspectives. In most cases, the related videos are usually of similar topics, and belong to the same category. This is a good hint for determining the video categories. In this section, we employ the hyperlinks between different videos to boost the performance of text-based classifier. For this purpose, in our approach, a graph is first constructed to represent the relationships between different videos based on the links.

For a given video set, we construct a graph $G = (V, E)$ where $V = \{v_1, v_2, \cdots, v_M\}$ is the vertex set and each vertex represents a video. Given a video $v_i (i = 1,2, \cdots, M)$, for each category label c_j, we get the probability score $p(v_i, c_j)$ that v_i belongs to c_j by the text-based SVM classifier trained in Section III.A. We then assign $p(v_i, c_j), j = 1,2, \cdots, N$ (N is the number of categories) to the vertex v_j as the initial weight vector.

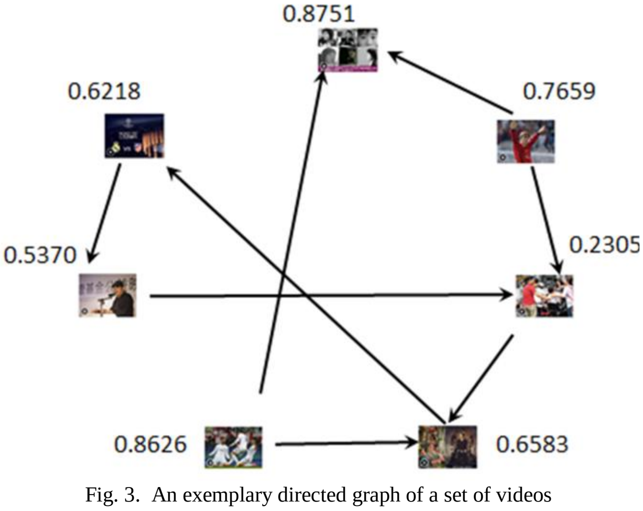

Fig. 3. An exemplary directed graph of a set of videos

For each video, we extract all incoming links from other related videos. The edge set $E = \{e_1, e_2, \cdots, e_K\}$ in the graph is used to present the links between different videos. Each edge e_k represents a link from video v_a to v_b. Figure 3 shows an example digraph constructed from a set of 7 videos where the number associated with each video is the probability score that the video belongs to a given category.

C. Boosting Video Categorization with Digraph Propagation

Graph propagation is widely used in information retrieval. In [13], multi-graph propagation is employed to improve the ranking of the search results. Another typical example of graph propagation is Google's PageRank algorithm. The web graph is constructed with all web pages as the nodes and the hyperlinks between web pages as the edges [11]. The importance of a webpage is weighted by the number of pages pointing to it.

In this paper, we employ hyperlinks between web videos to boost the performance of video categorization with graph propagation. There are abundant links between the videos on video sharing sites, which can be used to compensate the aforementioned shortcomings of the text-based classifiers. Given two videos v_a and v_b, a link from v_a to v_b is a good hint to determine the categories of two videos since they most likely belong to the similar categories. Thus, a link from v_a to v_b can be thought of as a category vote of v_a to v_b. In this way, the classification result of a video could be refined according to the categories of its related videos.

Based on the diagraph constructed in Section III.B, in this section, we present our approach to boost web video categorization with digraph propagation. In our approach, we iteratively update the weights of the vertexes by propagating the weight of each vertex to other linked vertexes. Our algorithm proceeds as follows.

1) Initialize the weights $p(v_i, c_j), j = 1,2, \cdots, N$ for each vertex v_i using the output probabilities of the text based SVM classifier as described in Section III.B.

2) Update the weights of each vertex by

$$p'(v_i, c_j) = \frac{1}{|\mathcal{L}(v_i)| + 1} \cdot (p(v_i, c_j) + \sum_{v_l \in \mathcal{L}(v_i)} p(v_l, c_j))$$

where $\mathcal{L}(v_i)$ is the set of videos that link to v_i.

3) If $|p'(v_i, c_j) - p(v_i, c_j)| < thres$ for $i = 1,2, \cdots, M$ and $j = 1,2, \cdots, N$, goto step 4; otherwise, set $p(v_i, c_j) = p'(v_i, c_j)$ and goto step 2. In our experiment, $thres$ is empirically set to 0.001.

4) For each video v_i, its category is determined by

$$c(v_i) = \underset{1 \le j \le N}{\arg\max}\ p(v_i, c_j)$$

In step 2, the weights of each video are updated by the category votes from other videos linked to it. Similarly, each video propagates its weights to the other related videos in the graph.

Based on the exemplary digraph in Figure 3, we illustrate the digraph propagation algorithm in Figures 4 and 5. By applying the above algorithm to update the weights of all vertexes, after the first iteration, we can get the new state of the digraph as shown in Figure 4. The weight of each vertex is updated by the influence from the related videos. By iteratively executing steps 2 and 3, the digraph comes to the

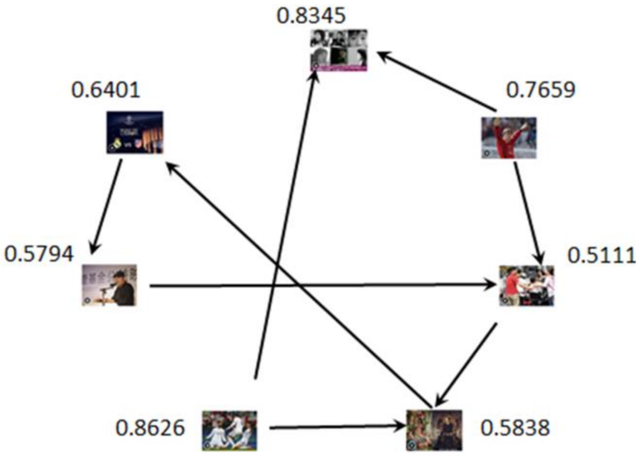

Fig. 4. The state of the directed graph after the first iteration

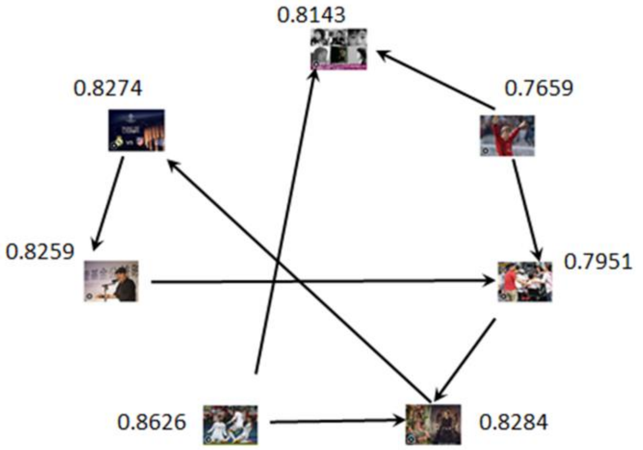

Fig. 5. The final state of the directed graph after 25 iterations

convergence after a total of 25 iterations. The final state of the digraph is shown in Figure 5.

By comparing Figures 3 and 5, we can observe that the initial weight of a vertex in Figure 3 is quite low (0.2305). The textual information associated with the video is too simple and contains much noise. After the digraph propagation, the weight is updated to 0.7951. This is because that the video is linked to two vertexes whose weights are relatively large, which means that the genres of these videos might be similar. Thus, the diagraph propagation can compensate the shortfalls of text based classifier by taking into account the mutual relation of related videos.

IV. Experiments

A. Dataset

To evaluate the performance of our proposed approach, we crawl video pages from sina video sharing sites [14]. In total we download 10500 videos from 7 predefined categories for classifier training, and 9793 videos for testing. The statistics of the training and testing sets is shown in Table I.

For text-based categorization, we extract the text information from the video pages. After the text preprocessing including word segmentation, stop words removal, and feature selection, 12449 words are selected as the feature dictionary. VSM model with TF-IDF is employed to represent the video content. SVMs are then trained on the training set. In our implementation, LIBMSVM [12] is adopted. The output scores of the SVMs are used as the initial weights of the vertexes in the digraph.

TABLE I
Statistics of The Data Set in Our Experiments

Category	**Training set**	**Test set**
Auto	1500	1400
Baby	1500	1400
Finance	1500	1399
Music	1500	1400
Sport	1500	1395
Technology	1500	1399
Entertainment	1500	1400
Total	**10500**	**9793**

For the construction of the digraph of the videos, we extract all hyperlinks between the related videos in our test set. In total 61630 hyperlinks are extracted. Finally, a digraph is constructed with 9793 videos as the vertexes and 61630 links as the edges.

B. Performance Evaluation

We adopt precision, recall and *F*-measure as the metrics to evaluate the performance of our proposed approach. F-measure is defined as follows

$$F = \frac{2 \times p \times r}{p + r}$$

where *p* is precision and *r* is recall. We compare the performance of our approach with the traditional text-based video categorization by SVM. Table II presents the performances of the two approaches.

As shown in Table II, the average precision, recall and *F*-score of the traditional text-based method are 0.7176, 0.6191 and 0.6177 respectively. Text data does contain some useful information for video categorization. However, for some categories such as *Music* and *Technology*, the recall values are only 0.2014 and 0.3967, and *F*-scores are also very low (0.3144 and 0.5391 respectively). For these two categories, the text data of the videos contains too much noise, and the categories of these videos cannot be effectively identified by using only the text data.

Compared with the text-based approach, our approach significantly improves the performance of video categorization. Specifically, the precision, recall, and *F*-score are improved by 14.41%, 24.05%, and 23.44% on average respectively. For the two categories *Finance* and *Sport*, *F*-scores are improved by 30.40% and 34.58% respectively. This is mainly because that there are rich hyperlinks between related videos in these two categories. Moreover, for the other two categories *Music* and *Technology* which have poor performances in the traditional text-based video categorization, our approach improves the *F*-scores by 45.55% and 34.65% respectively. For *Auto* and *Baby*, the performances of text based classifiers are already quite good, and thus the improvement of our approach is not very significant.

For speed efficiency, our graph propagation converges after a total of 135 iterations. The whole process takes approximately 2 hours on a desktop PC. This is acceptable considering the scale of the dataset and the significant performance improvement.

V. Conclusion

In this paper, we have presented our approach to boost the

TABLE II
PERFORMANCE COMPARISON BETWEEN OUR APPROACH AND TEXT BASED APPROACH

Category	Accuracy								
	Text-based approach			Our proposed approach					
	precision	recall	F_1	precision	Improve	recall	Improve	F_1	Improve
Auto	0.8413	0.8214	0.8312	0.8806	4.67%	0.9057	10.26%	0.8930	7.44%
Baby	0.9718	0.7629	0.8547	0.9789	0.73%	0.865	13.83%	0.9185	7.46%
Finance	0.6289	0.7148	0.6691	0.8129	**29.26%**	0.9414	**31.70%**	0.8725	**30.40%**
Music	0.7157	0.2014	0.3144	0.8265	15.48%	0.3164	57.10%	0.4576	45.55%
Sport	0.6808	0.5749	0.6234	0.8589	**26.16%**	0.8201	**42.65%**	0.8390	**34.58%**
Technology	0.8409	0.3967	0.5391	0.9096	8.17%	0.6040	52.27%	0.7259	34.65%
Entertainment	0.3441	0.8614	0.4917	0.4792	39.26%	0.9236	7.22%	0.6310	28.33%
Average	0.7176	0.6191	0.6177	0.8210	**14.41%**	0.7680	**24.05%**	0.7625	**23.44%**

performance of video categorization. Besides the textual information widely used in traditional approaches, we extract the hyperlinks from the video pages and employ a digraph to represent the relationships between different videos. By propagating the category scores of each video to its related videos, we refine the results of the text based SVM classifiers. Our experiments show that the hyperlinks are good hints for determining the video categories and our approach can significantly improve the classification results. For future work, other social information on web video sharing sites such as user videos and user comments could be integrated into web video categorization for further improvement.

REFERENCES

[1] Truong B T, Dorai C, "Automatic genre identification for content-based video categorization," in *Proceedings of the 15th Int. Conf. on Patten Recognition*, 2000, 4: 230-233.

[2] Yuan X, Lai W, Mei T, et al., "Automatic video genre categorization using hierarchical SVM," *IEEE Int. Conf. on Image Processing*, 2006: 2905-2908.

[3] Yang L, Liu J, Yang X, et al., "Multi-modality web video categorization," *ACM Int. Workshop on Multimedia Information Retrieval*, 2007: 265-274.

[4] Huang H, Lu Y, Zhang F, et al., "A Multi-modal Clustering Method for Web Videos," in *Trustworthy Computing and Services*, Springer Berlin Heidelberg, 2013: 163-169.

[5] Xu P, Shi Y, and Larson M A, "TUD at MediaEval 2012 genre tagging task: Multi-modality video categorization with one-vs-all classifiers," in *MediaEval*, 2012.

[6] Filippova K and Hall K B, "Improved video categorization from text metadata and user comments," in *Proceedings of the 34th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 835-842, 2011.

[7] Wu X, Ngo C W, Zhu Y M, et al., "Boosting web video categorization with contextual information from social web," in *World Wide Web*, 2012, 15(2): 197-212.

[8] Ionescu B E, Seyerlehner K, Mironică I, et al., "An audio-visual approach to web video categorization," *Multimedia Tools and Applications*, 2012.

[9] Brezeale Dand Cook D J, "Automatic video classification: A survey of the literature," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2008, 38(3): 416-430.

[10] Yew J, Shamma D A, and Churchill E F, "Knowing funny: genre perception and categorization in social video sharing," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011: 297-306.

[11] PageRank from Wikipedia, http://en.wikipedia.org/wiki/PageRank.

[12] Chang C C and Lin C J, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011, 2(3): 27.

[13] Liu J, Lai W, Hua X S, et al., "Video search re-ranking via multi-graph propagation," in *Proceedings of the 15th ACM Int. Conf. on Multimedia*, pp. 208-217, 2007.

[14] Sina Video, http://video.sina.com.cn/.