Towards Google Challenge: Combining Contextual and Social Information for Web Video Categorization

Xiao Wu
wuxiao@cs.cityu.edu.hk

Wan-Lei Zhao
wzhao2@cs.cityu.edu.hk

Chong-Wah Ngo
cwngo@cs.cityu.edu.hk

Department of Computer Science
City University of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong

ABSTRACT

Web video categorization is a fundamental task for web video search. In this paper, we explore the Google challenge from a new perspective by combing contextual and social information under the scenario of social web. The *semantic meaning* of text (title and tags), *video relevance* from related videos, and *user interest* induced from user videos, are integrated to robustly determine the video category. Experiments on YouTube videos demonstrate the effectiveness of the proposed solution. The performance reaches 60% improvement compared to the traditional text based classifiers.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: On-line Information Services – *Web-based services*;

General Terms

Algorithms, Design, Experimentation, Performance.

1. INTRODUCTION

Web video classification refers to the process of assigning web videos to predefined categories, which plays a vital role in many information retrieval tasks. Traditionally, video classification was undergone mainly by building classifiers on textual, audio, visual low-level features or their combination [1, 2]. However, considering the unsatisfactory performance of current high-level concept detection and the cost of feature processing, the content based classification cannot meet the expected performance and is generally expensive to implement. Fortunately, with the prosperity of social media, the social web provides rich contextual and social resources associated with videos, such as related videos, user and community information, which arouse new perspectives for web video categorization. The related videos frequently have relevant contents or similar category labels with the given video. At the same time, users share videos based on their personal interests, and therefore the uploaded videos by the same user usually have similar type. For example, videos from user "stanforduniversity" are associated with "Education", while videos from user "CBS" mainly belongs to "News and Politics".

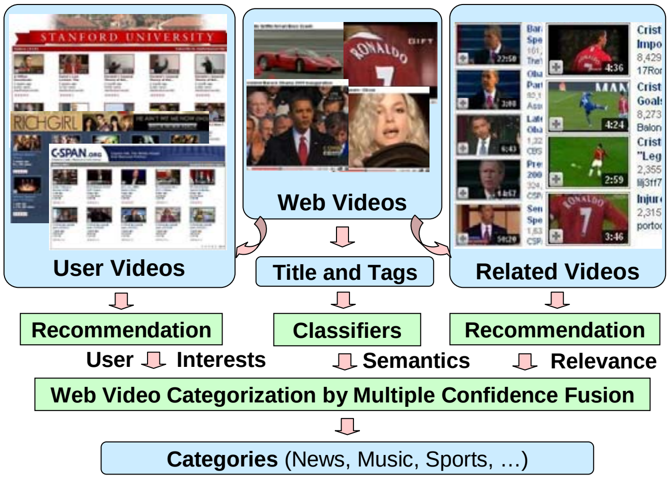

Figure 1. Framework of web video categorization

By checking the category labels of related videos and user videos, it gives constructive clues for the web video categorization. For the sake of effectiveness, efficiency, and scalability for web-scale classification, in this paper, we explore the easy-to-acquire contextual and social information to classify web videos.

2. WEB VIDEO CATEGORIZATION

Motivated by the above mentioned observation, the following contextual and social information is exploited in our web video categorization.

- **Text (title and tags)**: Text words carry semantic meaning. The title and tags of a web video are thus the most direct source for classification. However, the title and tags are noisy, inaccurate, ambiguous, incomplete, and even wrong. The discriminative power is limited for some cases.
- **Related Videos**: The related videos are usually similar or relevant videos, which is another useful hint for the web video classification. The majority of relevant videos could help to estimate the probability of the video category.
- **User Videos**: The user uploaded videos are commonly consistent with the users' personal interests. Therefore, to a large extent, the category label can be referred by the majority of videos uploaded by the same user.

The framework of the proposed web video categorization is illustrated in Figure 1. Text words, related videos, and user videos are from three different view points, that is, semantic meaning, video relevance, and user interest, respectively, which provide complementary clues for video classification. And their combination could give a more accurate and confident estimation. To make the classification robust, we fuse the confidence scores

*MM'*09, October 19–24, 2009, Beijing, China.

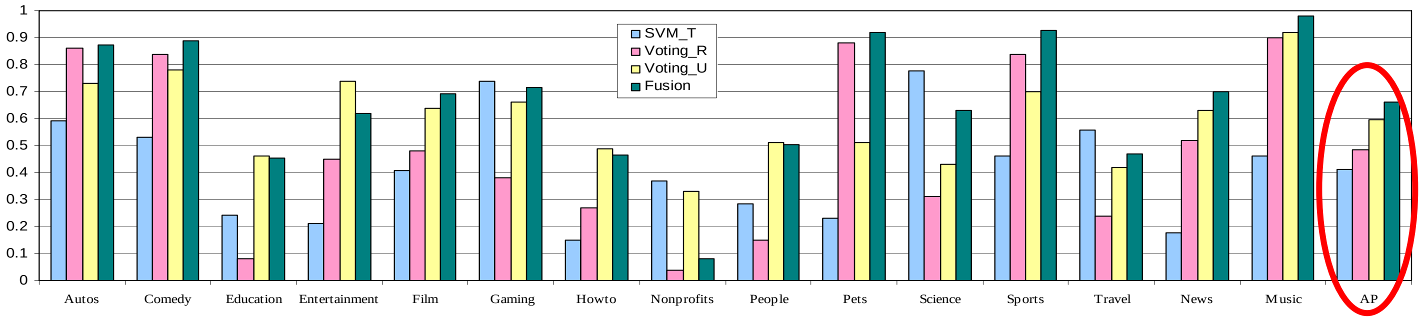

Figure 2. Performance comparison of 15 categories and overall performance

from semantic meaning (*Conf_T$_i$*), video relevance (*Conf_R$_i$*), and user interest (*Conf_U$_i$*) to give a more reasonable judgment *Score$_i$*(*V$_j$*). It is determined by the following formulation:

$$Score_i(V_j) = Conf\_T_i(V_j) + Conf\_R_i(V_j) + Conf\_U_i(V_j)$$

where *Conf_T$_i$(V$_j$)* is the probability score of video V_j belongs to the predefined category C_i, which is based on the SVM classifier on text word (title and tags), while *Conf_R$_i$(V$_j$)* and *Conf_U$_i$(V$_j$)* are confidence scores derived from video relevance and user interest according to class label distribution, which are defined as follows:

$$Conf\_R_i(V_j) = |R_{ij}| / |R_j| \qquad Conf\_U_i(V_j) = |U_{ij}| / |U_j|$$

R_j is the set of related videos for video V_j, and R_{ij} is the set of related videos having category label C_i among R_j. Similarly, U_j is the set of user videos uploaded by the same user of video V_j.

The category having the highest score determines the category label of video V_j:

$$C_i = \arg\max_{i=1}^{15} Score_i(V_j)$$

3. EXPERIMENTS

To compare the performance, we selected the “Most Viewed” videos of “This Month” (from December 2008 to March 2009) from the predefined 15 categories in YouTube as the training data, and the “Most Viewed” videos of “All Time” as the testing data. Each category has around 100 videos except “Music” category. The training and testing data sets consist of 4,610 and 2,047 web videos from YouTube, respectively. After a serial of data preprocessing (e.g. stop word removal, special character removal), there are 7,701 unique text words. In addition, we collected the related videos associated with the videos and the videos uploaded by the same users. There are totally 111,462 related videos, and 136,542 user videos for testing videos. The original video category labels are treated as the ground truth.

We use *precision* to evaluate the performance, which is defined as:

$$Precision_i = |P_i^+| / |P_i|$$

where P_i^+ is the number of correctly classified positive samples for category C_i, and P_i is the number of positive samples in ground truth. And the *average precision* (AP) is adopted to measure the overall performance for the 15 categories:

$$AP = \sum_{i=1}^{15} Precision_i$$

We compare the performance of SVM classifier (with RBF kernel) based on text feature (SVM_T), majority voting by related videos (Voting_R), majority voting by user videos (Voting_U), and the proposed fusion of these three sources (Fusion). The SVM classifiers were trained based on text features on the training set, and then predicted the testing data. The performance comparison for web video categorization is shown is Figure 2.

As it is well known that title and tags are very noisy for the web applications, the overall performance is poor (average precision is 0.412). However, to some extent, text words have sort of discriminative power. It is still a useful resource to classify the videos. For certain categories, such as “Gaming”, “Science and Technology” and “Travel and Events”, it achieves the best performance. Although the idea of majority voting from related videos and user videos is simple, they can give meaningful indication for the video categories. For most categories, they have better performance compared to text classification. The average precision for related videos and user videos are 0.483 and 0.597, respectively. And the information from user videos is more useful for offering accurate clues through users’ interests. Furthermore, semantic meaning, video relevance, and user interest, provide category confidence from different view points. Their combination achieves better performance than the individual ones, where its average precision reaches 0.662. The improvement is as high as 60% compared to the classification based on text features.

4. CONCLUSION

In this paper, we explore Google challenge from new perspectives by integrating the contextual and social information to effectively categorize web videos. A few interesting lessons have been learnt from our experiments.

- Textual title and tags are still useful features for video classification. It can achieve good performance for certain categories.
- Related videos and user videos provide constructive indication for video category.
- The information from text, related videos, and user videos complements each other. The integration of semantics, video relevance, and user interests gives significant performance improvement.
- These contextual and social features are easy to acquire, easy to use, and scalable. Therefore, the proposed solution has high potential to be applicable to the web-scale video categorization.

5. ACKNOWLEDGEMENT

The work described in this paper was supported by a grant from the RGC of HKSAR, China (CityU 119508) and a grant from City University of Hong Kong (Project No. 7002438).

6. REFERENCES

[1] D. Brezeale, and D. J. Cook. Automatic Video Classification: A Survey of the Literature. *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 38, no. 3, May 2008, pp. 416-430.

[2] L. Yang, J. Liu, X. Yang, and X.-S. Hua. Multi-Modality Web Video Categorization. *MIR’07*, pp. 265-274.