Google Challenge: Incremental-Learning for Web Video Categorization on Robust Semantic Feature Space

Yi-cheng Song[1,2], Yong-dong Zhang[1], Xu Zhang[1,2], Juan Cao[1], Jin-tao Li[1]
[1]Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China
[2]Graduate School of the Chinese Academy of Science, Beijing 100039, China
{songyicheng, zhyd, zhangxu, caojuan, jtli}@ict.ac.cn

ABSTRACT

With the advent of video sharing websites, the amount of videos on the internet grows rapidly. Web video categorization is an efficient methodology to organize the huge amount of data. In this paper, we propose an effective web video categorization algorithm for the large scale dataset. It includes two factors: 1) For the great diversity of web videos, we develop an effective semantic feature space called Concept Collection for Web Video Categorization (CCWV-CD) to represent web videos, which consists of concepts with small semantic gap and high distinguishing ability. Meanwhile, the online Wikipedia API is employed to diffuse the concept correlations in this space. 2) We propose an incremental support vector machine with fixed number of support vectors (n-ISVM) to fit the large scale incremental learning problem in web video categorization. Extensive experiments are conducted on the dataset of 80021 most representative videos on YouTube demonstrate that the semantic space with Wikipedia prorogation is more representative for web videos, and n-ISVM outperforms other algorithms in efficiency when performs the incremental learning.

Categories and Subject Descriptors

H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems

General Terms

Algorithms, Performance, Experimentation

Keywords

n-ISVM, Web Video Categorization, Wikipedia Propagation

1. INTRODUCTION

Facing the crazing amount of multimedia data on the web [4], retrieval by web directory is a promising solution for the web video retrieval. To our best knowledge, currently the category information on most of the video sharing websites is labeled by the user when he/she is uploading the video. Web video categories are data-driven and should flexibly adjust by the uploading interests changes. Video category varies on different video-sharing websites. So it is essential to studying the efficient automatic web video categorization algorithm. There are two important research issues: One is the robust video representation to overcome the web videos' high diversity of quality, style, and genres. The other is the classifiers with incremental learning function to meet the quickly expansion of web data.

In this paper, we construct a semantic feature space called Concept Collection for Web Video Categorization (CCWV-CD), which is consisted of concepts with small semantic gap and high distinguishing ability. Then Wikipedia is employed to diffuse the concept correlation in this space. For the universality and online characteristic of Wikipedia, even the video with the latest term "Obama" and another video tagged with "American President" can be propagated to more similar.

For the large scale problem, we propose an incremental support vector machine with fixed number of support vectors (n-ISVM), which can maintain relatively high performance comparing with the traditional support vector machine (SVM), meanwhile need less memory and computation costs. Furthermore, it can handle with training samples incrementally.

In this paper, experiments are performed on a benchmark dataset for web video analysis called MCG-WEBV [1], which gathering 80021 videos from most view and related videos of YouTube.

2. SEMANTIC FEATURE SPACE

To effectively represent the web videos, we construct a semantic space with small semantic gap and high categorization distinguishability. Then the concept correlations are diffused by Wikipedia Propagation.

Semantic Gap Measurement: Compared with the most frequent tag "video" in YouTube, obviously "cat" is easier to model and more valuable to present the video content, for the later has smaller semantic gap. For 5307 unique terms extracted from MCG-WEBV [1], we cluster the video sets including the same term in title or tag, and measure the textural and visual consistence for each set by computing the textual and visual similarities among all the videos in this set. The greater score of the set implicates its corresponding term has smaller semantic gap.

Categorization Distinguishability (CD): CD is proposed to measure the concept's distinguishability for categorization. The terms with the same distributions over all categories are less helpful to improve the video categorization. Define the Document Frequency (DF) as the total appearance of the term in the whole dataset, and Category Frequency (CF) as the number of categories where the term has appeared, then the concepts with high CD should has high DF but small CF.

Based on the above description, we construct a semantic space called CCWV-CD including top 2000 concepts with small semantic gap and great categorization distinguishability. Then, each video is represented by the classic vector space model $v[1...2000]$ based on CCWV-CD.

Wikipedia Propagation (WP): In MCG-WEBV, the average number of metadata (title and tag) for one video is only 14.5. It is too sparse in the above 2000 dimensions semantic space, and the classifiers directly trained in this feature space are less effective.

MM'09, October 19–24, 2009, Beijing, China.
ACM 978-1-60558-608-3/09/10.

In this paper, Wikipedia is employed to compute the *path frequency-inversed backward link frequency* (PFIBF) [2] between two concepts to propagate the video metadata to the 2000 dimensions CCWV-CD concepts. The WP enhances the similarity between videos with different but related words. By considering the concept correlation, the abilities of classifiers can be enhanced.

3. CATEGORIZATION ALGORITHM

The ISVM [3] can deal with the increasing training samples, but the number of support vectors increase dramatically with the number of training samples. As a result, the cost of CPU time increases with the amount of support vectors accordingly. We intend to control the computation time by limiting the number of support vectors in ISVM to a fixed value n, so our algorithm is named as n-ISVM. There are two key points for n-ISVM. The first point is to determine which support vectors should remain. The second is to decide how many support vectors should be kept.

To decide which vectors should remain, following criterion is illustrated to evaluate the effectiveness of each support vector.

$$c_i = y_i f\left(x_i\right) \tag{1}$$

In this case, training data and their labels as set $D=\left\{\left(x_i,y_i\right),i=1,2,\ldots,l\right\}$ where $x_i\in R^n,y_i=\left\{+1,-1\right\}$, the optimal separating function is $f\left(x_i\right)$.

The decrease of the number of support vectors may lead to the degradation of the performance of classifiers. Thus the value of n should be picked critically. One way to determine the value of n is decremental support vector machine. With the daily increasing training data, if a training sample is added into support vector set, a new support vectors set with m items are generated, then all the support vectors are sorted according to its c_i value. The vectors with top m-n c_i values, which mean they are the least effective, are removed from the support vector set.

In the web scale video dataset, the number of items in newly generated support vectors set m is always far larger than the limited number of items in support vectors set n. So the time cost of n-ISVM is $\mathrm{O}(n^2)$, which is efficient than ISVM $\mathrm{O}(\mathrm{m}^2)$ .So the training time will not increase with the growth of training data.

4. EXPERIMENT

We randomly spit the MCG-WEBV [1] into 3:6:1, with 3/10 videos as training data, 6/10 as testing data and 1/10 as cross validation set to adjust the C and γ for SVM.

4.1 Performance of feature spaces

In order to better understanding the utility of CCWV-CD and Wikipedia Propagation (WP) to the overall task performance, we conducted a comparison study in which we use different representation of video, including DF representation (terms within top 2000 document frequency), CCWV-CD representation without WP and CCWV-CD representation with WP. Then we retrain the classifier, the result is depicted in Table 1.

We can see that The MAP of CCWV-CD with WP is higher than DF and CCWV-CD without WP. We can conclude that the new introduced CCWV-CD improve the discriminative ability of SVM and the incorporation of concept relationship measurement from Wikipedia could closer the relationship of related videos and finally enhances the performance.

Table 1. Contribution of CCWV-CD and WP

Feature Space	MAP
DF	0.529
CCWV-CD	0.547
CCWV-CD with WP	0.552

4.2 Performance of n-ISVM

Comparing the MAP of n-ISVM(0.544) with traditional batch SVM(0.552) and ISVM(0.552), there are small degression(1.4%) in performance. However, n-ISVM can handle the training samples incrementally. As illustrated in Figure 1, we increase the number of training samples from 1000 to 10000 and record the average computation time for 15 categories on an Intel Pentium 3.20GHz and 2GB desktop. The result demonstrates that n-ISVM is an effective algorithm and more preferable while dealing with the web scale video categorization problem.

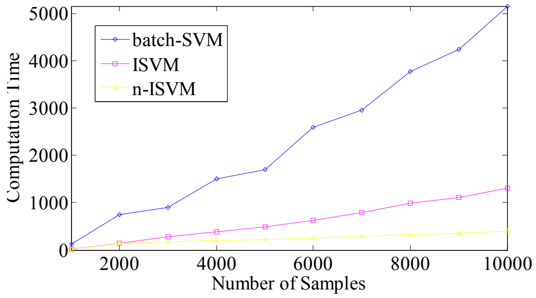

Figure 1: Training Time on MCG-WEBV data

5. ACKNOWLEDGMENTS

This work was supported by National Basic Research Program of China (973 Program, 2007CB311100), National High Technology and Research Development Program of China (863 Program, 2007AA01Z416), National Nature Science Foundation of China (60873165, 60802028).

6. REFERENCES

[1] J. Cao, Y.D. Zhang, Y.C. Song, Z.N. Chen, X. Zhang, J.T. Li. MCG-WEBV: A Benchmark Dataset for Web Video Analysis. Technical Report, MCG-ICT-CAS-09-001, May 2009.

[2] K. a. H. Nakayama, T. and Nishio, S., Wikipedia Mining - Wikipedia as a Corpus for Knowledge Extraction, in Proceedings of Annual Wikipedia Conference (Wikimania), 2008.

[3] L. Pavel, G. Christian, K. Stefan, ger, M. Klaus-Robert, and ller, Incremental Support Vector Learning: Analysis, Implementation and Applications, J. Mach. Learn. Res., vol. 7, pp. 1909-1936, 20.

[4] T.S Chua, J.H. Tang, R. C. Hong, H.J. Li, Z.P. Luo, and Y.T Zheng. NUS-WIDE: A Real-World Web Image Database from National University of Singapore, *ACM International Conference on Image and Video Retrieval*. Greece, 2009.