Online Web-Video Topic Detection and Tracking with Semi-supervised Learning

Guorong Li[1], Weigang Zhang[3], Junbiao Pang[4],
Qingming Huang[1,2], and Shuqiang Jiang[2]

[1] Graduate University, Chinese Academy of Sciences (CAS), Beijing, China
[2] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS, Beijing, China
[3] School of Computer Science and Technology, Harbin Inst. of Tech., China
[4] Beijing Municipal Key Lab. of Multimedia and Intelligent Software Tec., Beijing Univ. of Tech, China
{grli,wgzhang,jbpang,qmhuang,sqjiang}@jdl.ac.cn

Abstract. With the rapid growth of web data, a large amount of web videos are available online. However, how to organize them for facilitating users' experience and government supervision remains a problem yet to be seriously investigated. Topic detection and tracking, which has been a hot research topic for decades, could cluster web videos into different topics according to their semantic content. However, how to online discover topic and track them from web videos and images has not been fully discussed. In this paper, we formulate topic detection and tracking as an online tracking, detection and learning problem. First, by learning from historical data including labeled data and plenty of unlabeled data using semi-supervised multi-class multi-feature method, we obtain a topic tracker which could also discover novel topics from the new stream data. Second, when new data arrives, an online updating method is developed to make topic tracker adapt to the evolution of the stream data. We conduct experiments on public dataset to evaluate the performance of the proposed method and the results demonstrate its effectiveness for topic detection and tracking.

Keywords: Topic Detection and Tracking, web video, Multi-feature fusion, semi-supervised learning.

1 Introduction

With the development of network technology, many forms of social media such as video, twitter, blogs spring up. The openness, immediacy and reality of the Internet information made it an important indication of the trend of the society. More and more people use them to share their experiences or to find out what was happening outside and the web data is increasing at a unprecedented rate. According to the report, only in 2011 we had 1.8 trillion GB data and the data volume will grow with the speed of 50%. It is estimated that in 2020 global data will reach 35.2ZB, 90 percent of which are images or videos.

B. Huet et al. (Eds.): PCM 2013, LNCS 8294, pp. 750–759, 2013.

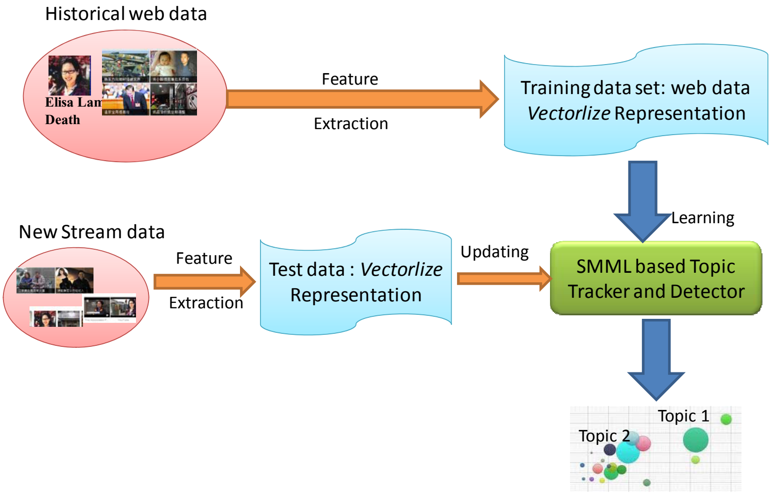

Fig. 1. The framework for topic tracking and detection from web data

Facing so large volume video or image data, it is inconvenient for users to get the main contents and find out what they are interested with. Therefore how to organize those video and image data is an important but very difficult problem. Topic detection and tracking [1](TDT) aims to discover topical related data and thus could classify web data by its semantic content. This brings at least three advantages:
(1) For users, it is easy to quickly understand what happens from the Internet and find out the materials related to the topic they are concerned with.
(2) For administrators, it is able to organize different modal data according to their contents. Furthermore, they can recommend topics which are related with that users are taking part in.
(3) For governments, they can discover hot or sensitive topics such as “watch” quickly and understand the peoples’ necessities. Moreover, for some topics which are rumors but becoming more and more attractive, such as “the salt is polluted because after the earthquake shook Japan on March 11, 2011”, it is possible for governments to make timely response and prevent its further spread.

During the past years, as text is the main form of the media data in the past, researchers focus on TDT from a collection of documents and have achieved great success [2] [3–5]. However, directly apply those methods to other types of media data (e.g. video, microblog) often generates poor results because of their different characteristics. For example web video contains short text tags and rich visual information. Intuitively, different modal features could complement with each other, but how to fuse them is still an open problem. In [6], a new Co-clustering method is developed to detect emerging topics from social streams and could leverage three sources of data including news article, micro-blog and picture. Recently, some works such as [7–10] are developed to mine topics from web videos through fusing multi-modal features, but most of them are based on clustering method and perform topic detection in an off-line fashion (mainly rely on textual features). However, the number of topics could not be determined in advance and new topics will be revealed along the time, so perhaps online learning feature fusion for topic detection would be a better choice.

As web videos or images usually arrives in streams over time, their content contains strong correlations. Historical data is very useful to better understand and predict the underlying topics in the future data. Therefore, instead of analyzing them in an off-line manner, we propose to learn topic detector from historical data and process and classify the stream of video data as soon as they are available. In this way, it is possible to track the topics of interest, predict their evolutions and discover the emerging trends in time. First, provided historical data including labeled and unlabeled samples, we learn a classifier using semi-supervised multi-feature multi-class learning method[11]. This method can fuse different modality features without destroying the local structure of individual feature. Second, when new data arrives, identify whether they are new topics or which topic they are belong to. Then updating classifier to adapt to the evolution of the streaming data. This approach has the added advantage that it could build models for the interested topics and track them.

2 The Proposed Algorithm

In this section, we will describe our algorithm for topic detection and tracking in detail. Here we need to introduce some notations.

Table 1. Notations of variables

X	the labeled sample set at time t
Y	the corresponding labels of X
x_i	the i^{th} sample
x_i^g	the g^{th} feature representation for sample x_i
X^g	the g^{th} feature representation for labeled sample set
N_i^g	the k nearest neighbors of sample x_i according to the g^{th} feature
U	unlabeled sample set at time
h_i^g	the local classifier learned from N_i^g
b^g	the global classifier learned from X^g

2.1 Semi-supervised Multi-feature Multi-class Learning Problem

Provided the historical data including labeled samples $\mathcal{X}$ and unlabeled samples $\mathcal{U}$, we want to learn a multi-topic classifier which could be used to track known topics and detect new topics. To exploit the structure information implied in the training data, for each sample x_i we assume there exists local classifiers h_i^g, which can classify N_i^g into c classes. Let f_i^g denote the expected predicting result of x_i using according to the g^{th} feature. Then the loss function of local classifiers derived from g^{th} feature is defined as following:

$$L_h^g = \sum_{i=1}^{n} \sum_{x_j^g \in N_i^g} (loss(h_i^g(x_j^g), f_j^g)) + \mu_1 \parallel h_i^g \parallel \tag{1}$$

where $\parallel h_i^g \parallel$ denotes the model complexity of h_i^g and λ is a parameter. Then through minimizing the loss function Equ. 2, the v features are combined together.

$$L_h = \sum_{g=1}^{v} L_h^g \tag{2}$$

To classify samples outside the training set, we assume there exist global classifies b_g whose expected predicted results are denoted by F. Similarly, the loss function of the global classifier is defined as:

$$L_b = \sum_{g=1}^{v} \sum_{i=1}^{n} loss(b_g(x_i), F_i) + \mu_1 \parallel b_g \parallel \tag{3}$$

Finally the consistency loss function among f, F and Y is simply formulated by Equ. (4).

$$L_c = \sum_{g=1}^{v} \parallel F - f_g \parallel_2^2 + \mu_3 \sum_{i=1}^{n_L} \parallel F_i - y_i \parallel_2^2 \tag{4}$$

where μ_3 is a parameter. Then by summing over the above three loss functions, we obtain the final objective function:

$$L(F, f_g, h_g^i, b_g) = L_h + \alpha L_b + \beta L_c \tag{5}$$

where α and β are parameters. There are many selections for the form of local and global classifiers, such as multiclass boosting[12], SVM. We select the most common construction: linear classifiers for its generalization[13]. Let $h_i^g(x_i^g) = (w_i^g)^T x_i^g + c_i^g$, $b_g(x_i^g) = (W^g)^T x_i^g + C^g$ and the complexity of the linear classifier is defined as the norm of its coefficient. The objective loss function can be rewritten as:

$$\begin{aligned} L(w_i^g, c_i^g, W^g, C^g) = & \sum_{g=1}^{v} \sum_{i=1}^{n} \sum_{x_j^g \in N_i^g} \parallel (w_i^g)^T x_j^g + c_i^g - f_j^g \parallel^2 \\ & + \alpha \sum_{g=1}^{v} \sum_{i=1}^{n} \parallel (W^g)^T x_i^g + C^g - F_i \parallel^2 + \gamma_1 \sum_{g=1}^{v} \parallel F - f^g \parallel^2 \\ & + \beta \sum_{i=1}^{n_L} \parallel F_i - Y_i \parallel^2 + \gamma_2 \sum_{i=1}^{n} \sum_{g=1}^{v} \parallel w_i^g \parallel^2 + \gamma_3 \sum_{g=1}^{v} \parallel W^g \parallel^2 \end{aligned} \tag{6}$$

where the γ_1, γ_2 and γ_3 are parameters. To obtain the minimum value of L, similar to [11], we set the derivative L with respect to variables c_i^g, w_i^g, C^g, W^g to be zero, we have

$$\begin{cases} c_i^g = \frac{\sum_{x_j^g \in N_i^g} f_i^g - (w_i^g)^T x_j^g}{k+1} \\ w_i^g = \left(\sum_{x_j^g \in N_i^g} x_j^g (x_j^g - \bar{N_i^g})^T + \gamma_2 I\right)^{-1} \sum_{x_j^g \in N_i^g} x_j^g (f_j^g - \bar{f_i^g})^T \\ C^g = \frac{\sum_{i=1}^n (F_i - (W^g)^T x_i^g)}{n} \\ W^g = \left[\sum_{i=1}^n x_i^g (x_i^g - \bar{X^g})^T + \gamma_3 I\right]^{-1} \sum_{i=1}^n x_i^g (F_i - \bar{F})^T \end{cases} \tag{7}$$

where $\bar{N_i^g} = \frac{\sum_{x_j^g \in N_i^g} x_j^g}{k+1}$, $\bar{f_i^g} = \frac{\sum_{x_j^g \in N_i^g} f_j^g}{k+1}$, $\bar{F} = \frac{\sum_{i=1}^n F_i}{n}$ and $\bar{X^g} = \frac{\sum_{i=1}^n x_i^g}{n}$. To simplify the equation, we introduce some variables $D_i^g = \left[\sum_{x_j^g \in N_i^g} x_j^g (x_j^g - \bar{N_i^g})^T + \gamma_1 I\right]^{-1}$, $B^g = \left[\sum_{i=1}^n x_i^g (x_i^g - \bar{X_g})^T + \gamma_3 I\right]^{-1}$, $M_i^g(j) = \begin{cases} 1, x_j^g \in N_i^g \\ 0, otherwise \end{cases}$, $S_i^g(j,j) = \begin{cases} 1, x_j^g \in N_i^g \\ 0, otherwise \end{cases}$ then w_i^g and W^g become

$$\begin{cases} c_i^g = \frac{(f^g - (w_i^g)^T X^g) M_i^g}{k+1} \\ w_i^g = D_i^g X^g S_i^g \left[I - 1_n \frac{(M_i^g)^T}{k+1}\right] (f^g)^T \\ C^g = \frac{(F - (W^g)^T X^g) 1_n}{n} \\ W^g = B^g X^g (I - \frac{1_n (1_n)^T}{n}) F^T \end{cases} \tag{8}$$

where $f^g = [f_1^g, f_2^g, \cdots, f_n^g]$, $X^g = [x_1^g, x_2^g, \cdots, x_n^g]$. Similarly, through setting the derivative L with respect to variables f_i^g, F_i to be zero and substituting c_i^g, w_i^g into the derivation, we have

$$f^g \left\{ \sum_{x_i^g \in N_j^g} \left[(\varOmega_j^g)^T (X_i^g - \frac{X^g M_i^g}{k+1}) + \frac{M_i^g}{k+1} - P_i \right] + \gamma_1 P_i \right\} \\ = -\gamma_1 F_i \tag{9}$$

where $\varOmega_j^g = D_j^g X^g S_j^g (I - \frac{M_j^g (1_n)^T}{k+1})$. Let $T_i^g = \sum_{x_i^g \in N_j^g} \left((\varOmega_j^g)^T (X_i^g - X^g M_i^g) + M_i^g - P_i\right)$, $T^g = \{T_1^g, T_2^g, \cdots, T_n^g\}$, easily, we derive

$$f^g = -\gamma_1 F (T^g + \gamma_2 I)^{-1}. \tag{10}$$

Algorithm 1. Semi-supervised Multi-Feature Multi-class learning

```
Input: training examples: $X, Y$ and stream web data $x_t, x_{t+1}, \cdots$, parameters
       $\alpha$, $\gamma_1$, $\gamma_2$, $\gamma_3$
Output: paratmers of the learned global and local classifiers:
       $\{< W^1, C^1 >, < W^2, C^2 >, \cdots, < W^v, C^v >\}$
       $\{< w_1^1, c_1^1 >, < w_2^1, c_2^1 >, \cdots, < w_n^1, c_n^1 >, < w_1^2, c_1^2 >, \cdots, < w_n^2, c_2^n >$
       $, \cdots, < w_1^v, c_1^v >, \cdots, < w_n^v, c_n^v >\}$
for $g = 1; g \leq v; g++$ do
    for $i = 1; i \leq n; i++$ do
        Compute $M_i^g$, $S_i^g$, $N_i^g$;
        Compute inverse of the $D_i^g$;
        Calculate $\Omega_i^g$;
    end
    Calculate $B^g$;
    for $i = 1; i \leq n;i++$ do
        Calculate $T_i^g$;
    end
end
Compute $F$;
for $g = 1; g \leq v; g++$ do
    Compute $f^g$, $W^g$ and $C^g$;
    for $i = 1; i \leq n; i++$ do
        Calculate $w_i^g$, $c_i^g$;
    end
end
```

Next substituting C^g, W^g and f^g into $\frac{\partial L}{\partial F_i}$, we obtain the optimal solution for F is:

$$\begin{aligned} F = & \, diag(I(1 \leq n_L), \cdots, I(n \leq n_L))Y \\ & \left\{ \left[\alpha \sum_{g=1}^{v} (B^g X^g (I - \frac{1_n (1_n)^T}{n}))^T (X^g - \frac{X^g 1_n (1_n)^T}{n}) \right] \right. \\ & + \frac{1_n (1_n)^T - I}{n} + \left[\gamma_1 \sum_{g=1}^{v} I - \gamma_1 (T^g + \gamma_3 I)^{-1} \right] \\ & \left. + diag(I(1 \leq n_L), \cdots, I(n \leq n_L)) \right\}^{-1} \end{aligned} \tag{11}$$

After obtaining the solution, we could compute f^g, W^g, w_i^g, C^g and c_i^g sequtially according to Equ. (10),Equ. (8). Alg. 1 provides the details of semi-supervised multi-feature multi-class learning procedure.

2.2 Online Topic Tracking and Detection

Provided a new data $x_t = \{x_t^1, x_t^2, \cdots, x_t^v\}$, we first use the learned global classifier to predict its label

Algorithm 2. Online updating method for Topic Tracker

```
Input: test examples: $x_{n+1}, \cdots, x_{n+p}$, $\{< M^g_i, D^g_i, \omega^g_i, T^g_i >\}_{i=1,g=1}^{i=n,g=v}$, $< iB^g >$,
       $F$
Output: parameters of the learned global and local classifiers:
        $\{< W^1, C^1 >, < W^2, C^2 >, \cdots, < W^v, C^v >\}$
        $\{< w^1_1, c^1_1 >, < w^1_2, c^1_2 >, \cdots, < w^1_n, c^1_n >, < w^2_1, c^2_1 >, \cdots, < w^2_n, c^n_2 >$
        $, \cdots, < w^v_1, c^v_1 >, \cdots, < w^v_n, c^v_n >\}$
for $g = 1; g \leq v; g++$ do
    Select $k$ neighbors from $\tilde{N}^g_{n+1}$ to compose $N^g_{n+1}$; Calculate $N^{\bar{g}}_{n+1}$ and set
    $M^g_{n+1}$;
    $D^g_{n+1} = \left[N^g_{n+1} * (N^g_{n+1})^T - (k+1)N^{\bar{g}}_{n+1}(N^{\bar{g}}_{n+1})^T + \gamma_1 I\right]^{-1}$;
    $\Omega^g_{n+1} = D^g_{n+1} X^g S^g_{n+1}(I - \frac{M^g_{n+1}(1_{n+1})^T}{k+1})$;
    for $i = 1; i \leq n; i++$ do
        if $M^g_{n+1}(i) == 1$ then
            $T^g_i += (\Omega^g_{n+1})^T (x^g_i - \bar{N}^g_i) + M^g_i - P_i$;
        end
    end
    $T^g_{n+1} = 0$;
    $\bar{X}_g = \frac{n * \bar{X}_g + x^g_n}{n+1}$;
    $B_g = \left[X_g * (X_g)^T + x^g_n * (x^g_n)^T - (n+1)\bar{X}_g * (\bar{X}_g)^T + \gamma_3 I\right]^{-1}$;
end
Compute $F$;
for $g = 1; g \leq v; g++$ do
    Compute $f_g$, $W^g$ and $C^g$;
    for $i = 1; i \leq n; i++$ do
        Calculate $w^g_i$, $c^g_i$;
    end
end
```

$$\tilde{F}_t = \frac{1}{v}\sum_{g=1}^{v}(W^g)^T x^g_t + C^g \tag{12}$$

$$c_t = max_c\{F^c_t\}. \tag{13}$$

Then we want to verify whether x_t belong to the c^{th} topic. Generally, if x_t belong to the c^{th}_t topic, its feature vector should be similar to those training samples coming from c^{th}_t topic. Therefore, the local classifiers $h_c(x_t) = \{h^g_i : y_i(c_t) = 1\}$ are likely to classify x_t correctly. We define credible local classifiers set of x_t $CLC(x_t) = \{h^g_i : y_i(c_t) = 1\}$. If the ratio of the credible local classifiers which classifies x_t into the c^{th} topic is larger than the threshold $thresh$, $y(x_t) = e_{c^{th}}$ (e_c denote unit vector where the c^{th} element is one); otherwise, x_t is assumed to come from a novel topic and $y(x_t) = e_{M+1}$ and the set of known topic classes is updated accordingly, $Y = \begin{pmatrix} Y \\ 0 \end{pmatrix}, y(x_t)\Big)$.

2.3 Online Updating

To be adaptive to the evolution of the web data, when new data x_{n+1} arrives the topic tracker and detector should be updated. However, it is time consuming if we re-minimize the loss function considering into the new sample. Therefore, we select an approximate method to online update the variables of the local classifier and global classifier. The k-nearest neighbors of $x_i, i = 1, 2, \cdots, n$ are considered to remain unchanged, while that of x_{n+1} are selected from historical data. Then M^g_{n+1}, N^g_{n+1} and $N^{\bar{g}}_{n+1}$ could be easily obtained. The detail procedures for online updating are shown in Algorithm. 2.

3 Experiments

Table 2. F-measure of the best 10 topics which are tracked successfully by SMMLTDT/MMLTDT

best topic	1	2	3	4	5	6	7	8	9	10	average
SMMLTDT	0.9637	0.9616	0.9549	0.9509	0.9235	0.9117	0.8833	0.8718	0.8458	0.8276	0.9095
MMLTDT	0.8758	0.8754	0.8709	0.8678	0.8500	0.8353	0.8287	0.8197	0.7845	0.7822	0.8390

The experiments are mainly performed on public dataset MCG-WEBV[14], which includes 248,887 videos downloaded from YouTub. The CoreData are 14,473 videos, which are the “Most Viewed” videos of “This Month” from Dec. 2008 to Nov. 2009. Different types of visual feature such as Color Histogram, Edge Histogram, Texture Concurrence, and textual features including title, tag, description etc. are provided. There are 73 topics annotated as ground truth. Each key frame of a video is considered as an example. The label of video is obtained by voting of the images belong to it. Moreover, to verify whether the unlabeled data is useful, we test the proposed method without using unlabeled data (MMLTDT). $k = 20$, $\alpha = \beta = \gamma_1 = \gamma_2 = \gamma_3 = 1.0$, $thresh = 0.5$ in the following experiment.

3.1 Online Topic Tracking

We select 10 hottest topics from the annotated 73 topics, the web data from the other topics are considered as from 11^{th} class. The data are divided into two parts according to their upload time: training data and testing data. Test data are used as unlabeled samples. Moreover, to test whether unlabeled data could be useful, we set $\mathcal{U} = NULL$ and SMMLTDD degrades into supervised multi-class multi-feature learning (referred as MMLTDT). Precision, Recall and F-measure defined in [9] are used for quantitative evaluations. Table. 2 shows the top 10 F-measure of the topics that are successfully tracked by MMLTDT and SMMLTDT respectively. Figure. 2 provides comparisons of SMMLTDT, MMLTDT and [9] in terms of F-measure. We can see that MMLTDT could integrate multi-modal features to build effective classifier for target topics and mining information implied in the unlabeled data SMMLTDT could further improve the performance.

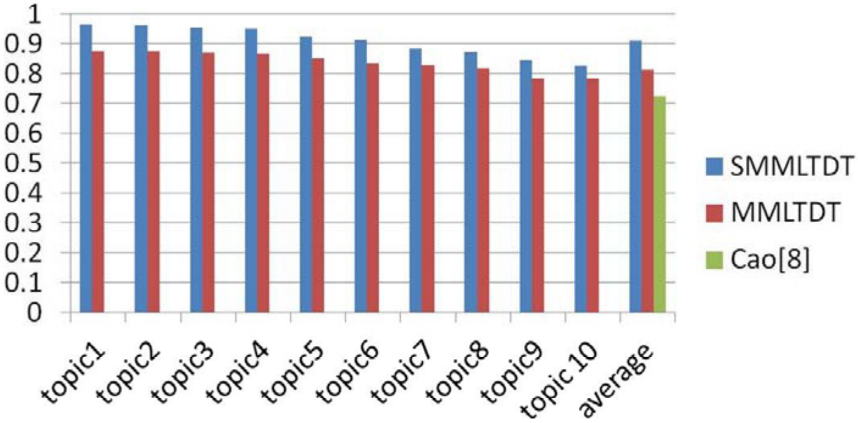

Fig. 2. The F-measure of 10 topics respectively and average F-measure of MMLTDT, SMMLTDT and Cao[9]

3.2 Online Topic Discovery

The data of the earliest one month is used as historical data. For each topic, we randomly choose part of samples belong to it as unlabeled samples. If the F-measure of a discovered topic is larger than 0.5, we considered this topic as successfully detected and tracked (referred as Topic Successfully Detected and Tracked TSDT). The number of TSDT is 30.

4 Conclusions

To analyze the massive web video streaming data, we propose an online topic tracking and detection method, which fuses different types of features through multi-class multi-feature learning algorithm. To be adaptive to the dynamic of streaming data, we develop an online updating method to update our topic tracker and detector. New topics could be incremental revealed as learning and updating go on. Experiments on public dataset show that the proposed method can not only track the interested topics but also can discover new topics and build efficient classifiers for them.

Acknowledgments. This work was supported by China Postdoctoral Science Foundation: 2012M520436, in part by National Basic Research Program of China (973 Program): 2012CB316400, National Natural Science Foundation of China: 61025011 and 61202322.

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report (1998)
2. Chen, K., Luesukprasert, L., Chou, S.: Hot topic extraction based on timeline analysis and multi-dimensional sentence modeling. IEEE Transactions on Knowledge Data Engeering 19(8), 1016–1025 (2007)
3. Sun, A.X., Hu, M.: Query-guided event detection from news and blog streams. IEEE Transactions on Systems, Man and Cybernetics 41(5), 834–839 (2011)
4. Zhai, Y., Shah, M.: Tracking news stories across different sources. In: Proceedings of the 20th ACM International Conference on Multimedia, MM 2005, pp. 2–10. ACM (2005)
5. Kasiviswanathan, S.P., Melville, P., Banerjee, A., Sindhwani, V.: Emerging topic detection using dictionary learning. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, pp. 745–754. ACM (2011)
6. Bao, B.K., Min, W., Sang, J., Xu, C.: Multimedia news digger on emerging topics from social streams. In: Proceedings of the 20th ACM International Conference on Multimedia, MM 2012, pp. 1357–1358. ACM, New York (2012)
7. Shao, J., Ma, S., Lu, W., Zhuang, Y.: A unified framework for web video topic discovery and visualization. Pattern Recognition Letters 33(4), 410–419 (2012)
8. Hong, R., Tang, J., Tan, H., Ngo, C., Yan, S., Chua, T.: Beyond search: event driven summarization for web videos. ACM Transactions on Multimedia Computing, Communications and Applications 33(4), 410–419 (2011)
9. Cao, J., Ngo, C.W., Zhang, Y.D., Li, J.T.: Tracking web video topics: Discovery, visualization, and monitoring. IEEE Transactions on Circuits and Systems for Video Technology 21(12), 1835–1846 (2011)
10. Chen, T., Liu, C., Huang, Q.: An effective multi-clue fusion approach for web video topic detection. In: Proceedings of the 20th ACM International Conference on Multimedia, MM 2012, pp. 781–784. ACM, New York (2012)
11. Yang, Y., Song, J., Huang, Z., Ma, Z., Sebe, N., Hauptmann, A.: Multi-feature fusion via hierarchical regression for multimedia analysis. IEEE Transactions on Multimedia 15(3), 572–581 (2013)
12. Freund, Y., Schapire, R.: A decision theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
13. Hastie, T., Simard, P.: Models and metrics for handwritten character recognition. Statistical Science 13(1), 54–65 (1998)
14. Cao, J., Zhang, Y., Song, Y., Chen, Z., Zhang, X., Li, J.: Mcg-webv: A benchmark dataset for web video analysis (2009)