CROSS-MEDIA TOPIC DETECTION: A MULTI-MODALITY FUSION FRAMEWORK

Yanyan Zhang[1], *Guorong Li*[1], *Lingyang Chu*[2], *Shuhui Wang*[2], *Weigang Zhang*[3], *Qingming Huang*[1,2]

[1]University of Chinese Academy of Sciences
[2]Key Lab of Intell. Info. Process., Inst. of Comput. Tech., Chinese Academy of Sciences
[3]School of Computer Science and Technology, Harbin Institute of Technology
{yyzhang, grli, lychu, shwang, wgzhang, qmhuang}@jdl.ac.cn

ABSTRACT

Detecting topics from Web data attracts increasing attention in recent years. Most previous works on topic detection mainly focus on the data from single medium, however, the rich and complementary information carried by multiple media can be used to effectively enhance the topic detection performance. In this paper, we propose a flexible data fusion framework to detect topics that simultaneously exist in different mediums. The framework is based on a multi-modality graph (MMG), which is obtained by fusing two single-modality graphs together: a text graph and a visual graph. Each node of MMG represents a multi-modal data and the edge weight between two nodes jointly measures their content and upload-time similarities. Since the data about the same topic often have similar content and are usually uploaded in a similar period of time, they would naturally form a dense (namely, strongly connected) subgraph in MMG. Such dense subgraph is robust to noise and can be efficiently detected by pair-wise clustering methods. The experimental results on single-medium and cross-media datasets demonstrate the flexibility and effectiveness of our method.

Index Terms— Topic detection, fusion, cross-media, multi-modality, graph

1. INTRODUCTION

With the rapid development of Internet, the huge amount of information is delivered by diversified types of mediums, such as News article, News video, Web video, Microblog, etc. Compared with the relatively limited intrinsic information capacity of a single medium, the complementary cross-media information delivered by multiple mediums is much richer. Moreover, cross-media information often has a broader acceptance group and can reflect social realities from more aspects. Therefore, it would be interesting and beneficial to jointly detect topics from the multi-modal data in different mediums. As it is shown in Fig. 1, data in different mediums generally involve multiple modalities, such as text, image, video and audio. Each data of a certain medium may contain single or multiple modalities and the data structure of the same modality varies a lot among different mediums.

Fig. 1. A toy example of cross-media topics.

A common solution to detect topics from the multi-modal data is to robustly fuse them together, which however is a difficult task in the scenario of cross-media topic detection. Firstly, the noise degree and length of text data in different mediums vary significantly, leading to the non-uniformity of text data. Secondly, the data structures of different mediums are often incomplete, since every single medium generally does not contain all the potential modalities at the same time. Thirdly, the granularity (*i.e.* time duration) of different topics varies much, which is a common problem in topic detection.

1.1. Related Work

The task of Topic Detection and Tracking (TDT) [1] has been studied for decades and many effective TDT methods [2] [3] [4] [5] have been developed to deal with News articles and News videos. Anaya-Sánchez *et al.* [3] introduce a clustering algorithm to discover and describe the topics in text collections. Pan *et al.* [4] propose a 3S-LDA model which combines the Latent Dirichlet Allocation (LDA) model with temporal and spatial clustering for topic detection. Although performing well on the professionally organized data, these methods are not suitable for the less structured and more complex data from various mediums, which is generally made by non-professional web users. Some other works [6] [7] [8] detect topics on Web video data. The bipartite graph method [6] uses the correlation between Web videos and their keywords for co-clustering, which however cannot utilize multiple fea-

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

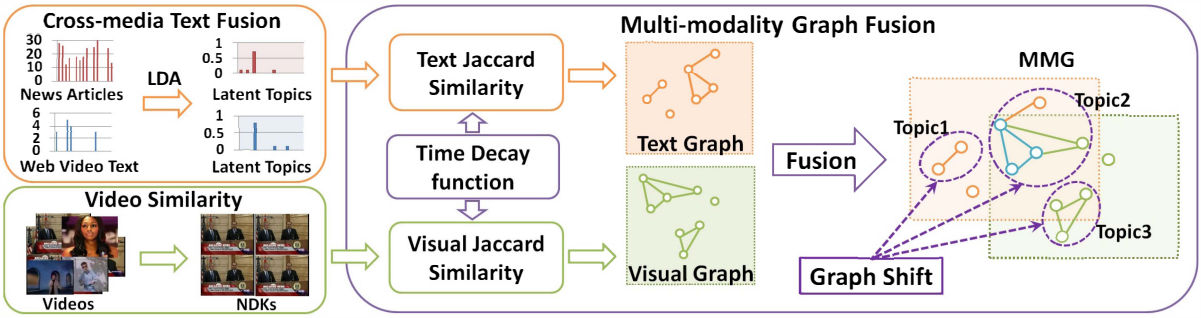

Fig. 2. The flow chart of our method. Please refer to the color pdf for a better view.

tures. The salient trajectory method proposed by Cao *et al.* [7] builds the tag and visual information into a topic evolution link graph for topic detection. The tag group method proposed by Chen *et al.* [8] fuses the dense-bursty tag groups with near-duplicate keyframes to efficiently detect Web video topics; they also use the hot queries of search engines to identify the hot topics. However, both Cao's and Chen's methods are limited by their strong reliance on video tags, which may be noisy or unavailable in some mediums. Moreover, their hard quantization strategies in dealing with the upload time would inevitably lose some accuracy. In sum, most of the above mentioned methods are not suitable to sufficiently utilize the rich cross-media information that can effectively improve the topic detection performance.

1.2. Our Method

In this paper, we propose a flexible multi-modality fusion framework, which robustly fuses multi-modal information in different mediums to simultaneously detect topics in both single and multiple mediums. Fig. 2 shows the framework of our method, and it consists of two fusion stages. The first one is a cross-media text fusion stage, which utilizes the Latent Dirichlet Allocation (LDA) model [9] to learn descriptive latent topics from the non-uniform text data in different mediums. The second one is a multi-modality graph (MMG) fusion stage. It initially utilizes the time-decay coefficient and the Jaccard similarity to build two single-modality graphs for text and video modalities. The time-decay coefficient uses the upload-time interval of two data to estimate their probability of being about the same topic. The Jaccard similarity is widely used by many applications [10] [11] to measure the set similarity; here it is used to measure the consistency of two nearest neighborhoods, which enables us to compare the inhomogeneous feature similarities of different modalities. Then, the two graphs are fused by an efficient merging process to obtain the multi-modality graph (MMG), where each graph node represents a data and the edge weight between two nodes embodies the joint similarities of their contents and upload times. Since data about the same topic always contain similar contents and are often uploaded in the same period of time, they would naturally form a dense subgraph in MMG, which can be efficiently detected by the pair-wise clustering method of graph shift [12]. In summary, the merits of our method are:

1) The LDA-based text fusion stage effectively alleviates the influence of the noisy and non-uniform text data in different mediums. Moreover, simultaneously learning the latent topics from all the text data in different mediums is able to utilize the cross-media information more sufficiently. Such latent topics generally have stronger descriptive power than other text features, since they are more related to real topics.

2) The MMG fusion stage properly fuses the inhomogeneous text and video similarities into one graph, so that the multi-modal data about the same topic would form a dense subgraph, which is highly robust to noise and could be efficiently detected. Moreover, the flexibly defined graph nodes in MMG are not required to have the same number of potential modalities, which enables us to simultaneously detect the topics from various multi-modal data of both single or multiple mediums.

3) The time-decay coefficient smoothly models the topic granularity and embeds the time information in the MMG, which effectively reduces the accuracy loss caused by the hard time quantization strategies.

2. ALGORITHM

This section first introduces the two proposed fusion stages, then illustrates how to use the graph shift method [12] to detect topics from the multi-modality graph (MMG). Data from different mediums typically involve the potential modalities of text, image, video, audio, etc. Our method mainly focuses on text (denoted by T) and video (denoted by V), which are widely used by many topic detection methods. A multi-modal data is denoted by $d_i = (d_i^T, d_i^V)$, where d_i^T and d_i^V represent the data in text and video modalities respectively. For certain single-modality mediums that do not completely contain both the text and video modalities, either d_i^T or d_i^V would be null.

2.1. Cross-media Text Fusion

Text data from different mediums usually have quite different text length and are sometimes noisy. For example, user-provided video annotations are always short, and usually contain a few key words with a relatively high degree of noise; however, the News articles often have long text length with relatively low noise degree. Although the tf-idf histogram

and the key-word group can be proper text features for certain kind of text data, they are not able to effectively handle all kinds of text data in different mediums at the same time.

We propose to extract text feature from the text data in multiple mediums by the LDA model, which is able to robustly learn the latent topics from noisy and non-uniform text data. We describe the text data d_i^T by the normalized distribution of latent topics in Eqn.1, where p_{ik} is the normalized probability of d_i^T over the k-th latent topic and C is the total number of latent topics.

$$d_i^T = [p_{i1}, ...p_{ik}..., p_{iC}] \tag{1}$$

In the scenario of cross-media topic detection, although the latent topics learned by LDA are different from real topics, their semantics are, to some extent, related to each other. Therefore, using the distribution of latent topics as text feature is able to describe the real topics of text data more accurately. The latent topics can also be learned from the text data of single medium. However, simultaneously learning them from multiple mediums is a natural text fusion process, which further enhances their descriptive power by utilizing the cross-media information.

2.2. Multi-modality Graph Fusion

As it is shown in Fig. 2, we first build two independent single-modality graphs: the text graph (denoted by G^T) and the video graph (denoted by G^V). Then, the single-modality graphs are merged into a multi-modality graph (MMG) (denoted by G), in which the topics would naturally form dense subgraphs. The key technique is to transform the incomparable text similarities and video similarities into the comparable Jaccard similarities, which makes the graph fusion possible. Another effective technique is the time-decay coefficient, which alleviates the over-split and over-merge of topics by smoothly embedding time information in the edge weights of both G^T and G^V. For convenience, we assume that the connectionless graph nodes are connected by zero-weighted edges, so all graphs can be regarded as fully connected.

Given two multi-modal data d_i and d_j, the time-decay coefficient α_{ij} smoothly measures their probability of being about the same topic by the interval between their upload times. It is defined by Eqn.2, where β is a positive scale parameter to control the rate of decay, Δ is a small fixed quantization factor and t_i,t_j are the upload time of d_i and d_j. Note that $\lfloor\cdot\rfloor$ denotes the round down operation. Apparently, when the time interval $|t_i - t_j|$ increases, the time-decay coefficient decreases exponentially, which further indicates that d_i and d_j are less likely to be about the same topic.

$$\alpha_{ij} = e^{-\beta\left(\lfloor\frac{|t_i - t_j|}{\Delta}\rfloor\right)^2} \tag{2}$$

The text graph is denoted by $G^T = (\{n_i^T\}, \{w_{ij}^T\})$, where each node n_i^T corresponds to a text data d_i^T and the edge weight w_{ij}^T is calculated by three steps. Firstly, we use dot product as similarity to find the k-nearest neighbors $N_i^T(k)$, $N_j^T(k)$ of d_i^T and d_j^T respectively. Then, the Jaccard similarity J_{ij}^T is calculated by Eqn.3. Finally, the edge weight w_{ij}^T is obtained by Eqn.4, which further incorporates the time-decay coefficient α_{ij} (see Eqn.2) with the corresponding Jaccard similarity J_{ij}^T (see Eqn.3).

$$J_{ij}^T = \frac{|N_i^T(k) \cap N_j^T(k)|}{|N_i^T(k) \cup N_j^T(k)|} \tag{3}$$

$$w_{ij}^T = \alpha_{ij} \cdot J_{ij}^T \tag{4}$$

The video graph $G^V = (\{n_i^V\}, \{w_{ij}^V\})$ is constructed in a similar way. Each node n_i^V corresponds to a video data d_i^V and the edge weight w_{ij}^V is calculated by Eqn.6, where α_{ij} is the same time-decay coefficient in Eqn.2 and J_{ij}^V is the Jaccard similarity between d_i^V and d_j^V (see Eqn.5). When calculating the k-nearest video neighbors of $N_i^V(k)$ and $N_j^V(k)$, we evaluate the similarity between videos by the number of their near duplicate keyframes (NDK). Note that the number of nodes in G^V and G^T are possibly not equal to each other, since some data are from the single-modality mediums, which do not fully involve both the text and video modalities.

$$J_{ij}^V = \frac{|N_i^V(k) \cap N_j^V(k)|}{|N_i^V(k) \cup N_j^V(k)|} \tag{5}$$

$$w_{ij}^V = \alpha_{ij} \cdot J_{ij}^V \tag{6}$$

After obtaining the text graph G^T and the video graph G^V, we fuse them into the multi-modality graph (MMG) $G = (\{n_i\}, \{w_{ij}\})$, where the node set is $\{n_i\} = \{n_i^T\} \cup \{n_i^V\}$ and the edge weight w_{ij} is obtained by Eqn.7.

$$w_{ij} = w_{ij}^T + w_{ij}^V \tag{7}$$

In this way, the single-modality nodes n_i^T and n_i^V, which correspond to the text and video modalities of the same data d_i, are fused into one MMG node n_i. Other single-modality nodes, which correspond to the data from single-modality mediums, are directly transformed to MMG nodes without fusion. All nodes in MMG are treated equally. For the fusion of edge weights w_{ij}^T and w_{ij}^V, although the dot-product similarity of text data is not directly comparable with the NDK-based similarity of video data, the corresponding Jaccard similarities are comparable, since both of them reflect the consistency of two k-nearest neighborhood. Considering that there is no prior about the relative importance of each modality, a proper solution is to treat all modalities equally by simply summing up the edge weights (see Eqn.7). Apparently, this fusion framework of MMG is flexible enough to robustly incorporate the multi-modal data in different mediums.

In short, the edge weights $\{w_{ij}\}$ of MMG jointly evaluate both the upload time similarities and the content similarities of multi-modal data $\{d_i\}$. Since the data about the same topic

are generally similar with each other in both content and upload time, the corresponding MMG nodes would be strongly connected (*i.e.* large edge weight) with each other and naturally form a dense subgraph. Such dense subgraph is a topic-sensitive pattern, which is quite robust to noise and can be efficiently detected by pair-wise clustering methods.

2.3. Topic Detection on MMG

We use the pair-wise clustering method of graph shift (GS) [12] to detect the dense subgraphs (*i.e.* topics) on MMG. The input of the GS method is the symmetric adjacency matrix A of MMG, whose element $a_{ij} \in A$ equals to the MMG edge weight w_{ij}. A subgraph of MMG is represented by a probabilistic cluster $x \in \triangle^m$, where $\triangle^m = \{x | x \in R^m, x \geq 0, |x|_1 = 1\}$ and m is the total number of graph nodes in MMG. In fact, x is a unit mapping vector, which maps a cluster of graph nodes to the standard simplex R^m. The i-th bin of x is denoted by x_i, which is the probability that the subgraph x contains the MMG node n_i. Particularly, $x_i = 0$ means that n_i is not contained by subgraph x. The GS method measures the average connection strength of subgraph x by $g(x)$ in Eqn.8 and efficiently finds all the local maximums $\{x^*\}$ of $g(x)$ (see Eqn.9). Each local maximum indicates a dense subgraph of MMG, which is defined as a topic in our method.

$$g(x) = x^T A x \tag{8}$$

$$x^* = \max_x \ g(x), \quad s.t.\ x \in \triangle^m \tag{9}$$

3. EXPERIMENT

In this section, we compare the topic detection performances of the proposed MMG method with the salient trajectory method (ST) [7] and the tag group method (TG) [8]. For MMG, the NDKs are extracted by the method proposed in [13] and the latent topics are generated by the Topic Modeling Toolbox published in [14]. All the experiments are conducted on a common PC with Core i-5 CPU and 12 GB memory.

Two multi-modality datasets are used. The "core dataset" of MCG-WEBV [15] is built with the "Most Viewed" videos (along with surrounding text) on *YouTube* from Dec 2008 to Feb 2009. It contains 3,282 Web videos and 73 manually annotated ground truth topics, whose average topic-duration is 42.2 days. Since MCG-WEBV contains only one medium (*i.e.* Web video on *Youtube*), it is not sufficient to analyze the cross-media topic detection performance of MMG. Therefore, we build a cross-media dataset YKS by crawling Web videos and News articles from *YouKu* [1] and *Sina* [2], respectively. YKS consists of 2,131 "hot" Web videos and 7,325 News articles from May 2012 to June 2012. Its ground truth contains 20 pure Web video topics, 225 pure News article topics and 73 hybrid topics which involve both the two mediums. The average topic-duration is 13.0 days.

We adopt the same performance evaluation methods as [7] and [8]. The *Precision* and *Recall* are defined by Eqn.10, where E_D is the data set of a detected topic, E_T is the data set of the ground truth topic best matched with E_D, E_C is the set of correctly detected data in E_D. After obtaining the *Precision* and *Recall* for each topic, we calculate the *F-Measure* by Eqn.11, which is a comprehensive evaluation on both *Precision* and *Recall.* Then, we evaluate the average detection performances by the same strategy as ST, which sort all the topics by their *F-Measure* and calculate the average *Precision*, *Recall* and *F-Measure* of the top-10 detected topics. We also adopt the *CP* measurement proposed in [8] to evaluate the percentage of correctly detected topics, whose *F-Measure* is bigger than 0.5. *CP* is defined by Eqn.12, where NDT is the total number of detected topics and $NCDT$ is the number of correctly detected topics.

$$Precision = \frac{|E_C|}{|E_D|} \quad Recall = \frac{|E_C|}{|E_T|} \tag{10}$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \tag{11}$$

$$CP = \frac{NCDT}{NDT} \tag{12}$$

3.1. Parameter Analysis

This section analyzes the influences of two parameters on the performance of MMG: the number of latent topics C in Eqn.1 and the scale parameter β of the time-decay coefficient (see Eqn.2). The number of nearest neighbors k (see Eqn.3 and Eqn.5) and the quantization factor Δ (see Eqn.2) do not have so much influences. Therefore they are empirically set as $k = 30$, $\Delta = 3$ for all experiments.

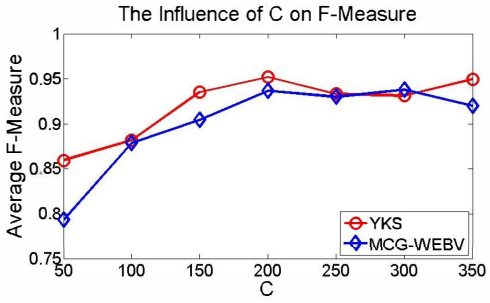

Fig. 3. The influence of C on the average *F-Measure* of top-10 detected topics in different datasets.

The number of latent topics C mainly affects the descriptive power of text feature. Fig.3 shows the influence of C on the average *F-Measure* performance of MMG. The average *F-Measure* first increases with the growth of C, which is attributed to the increasing descriptive power of latent topic

[1] http://www.youku.com/
[2] http://news.sina.com.cn/

distribution. Then, it meets an upper-bound when C becomes too large, which may be due to the bottleneck of the intrinsic descriptive power of the latent topic. We choose the optimal value $C = 200$ for both MCG-WEBV and YKS.

The scale parameter β controls the decaying rate of the time-decay coefficient α_{ij} in Eqn.2, which affects the average granularity of detected topics. A small value of β strengthens the edge weights between MMG nodes, which increases the chance of the positive nodes from large granularity topics to form a dense subgraph. However, this may also bring some noise to small granularity topics, since the edge weights between some positive and negative nodes are also strengthened. On the contrary, a big value of β weakens the MMG edge weights, which favors small granularity topics, however, may lose some positive nodes of large granularity topics. This phenomenon is shown in Fig.4, where the average granularity of top-20 detected topics monotonously decreases with the growth of β. Table. 1 shows the influence of β on the average *Precision*, *Recall* and *F-Measure* of MMG on two datasets. We can see that the optimal value for MCG-WEBV and YKS are $\beta = 0.01$ and $\beta = 0.1$ respectively. Note that the average topic granularity of MCG-WEBV is larger than YKS, so its optimal value of β is smaller.

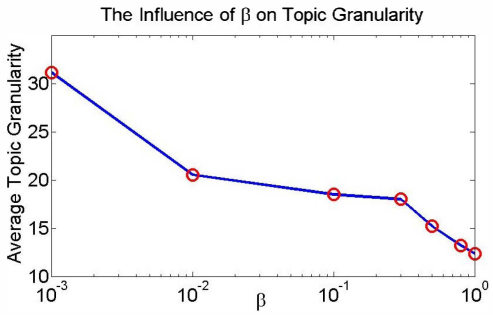

Fig. 4. The influence of β on the average granularity of top-20 detected topics on MCG-WEBV. Note that the x-axis is plotted in log scale.

Table 1. The influence of β on the average topic detection performances of top-10 detected topics. ($C = 200$).

Dataset	β	*Precision*	*Recall*	*F-Measure*
MCG-WEBV	**0.01**	**0.9366**	**0.9418**	**0.9367**
	0.1	0.8797	0.8674	0.8701
	1	0.8602	0.8526	0.8475
YKS	0.01	0.9561	0.9103	0.9305
	0.1	**0.9750**	0.9358	**0.9517**
	1	0.9124	**0.9408**	0.9199

3.2. Performance Evaluation on MCG-WEBV

We conduct our experiment on the "core dataset" of MCG-WEBV [15] to analyze the effectiveness of MMG on multi-modal data from the single-medium of Web video. For MMG, we use the optimal parameter values: $\beta = 0.01$, $C = 200$. The salient trajectory based method of ST [7] is regarded as baseline; the tag group based method of TG [8] is compared as well. Since the published MCG-WEBV dataset has been used by both ST and TG, we compare with their best reported performances. Fig. 5 shows comparative results on the average *Precision*, *Recall* and *F-Measure* of top-10 detected topics. We can see that both TG and MMG perform much better than ST. Besides, the *Precision* and *Recall* of MMG are more balanced than TG, hence it achieves a slightly better *F-Measure* performance. We further analyze the effectiveness of MMG under the measurement of CP which is proposed by TG. As it is shown in Table 2, the CP performance of MMG significantly outperforms TG; this may be attributed to the robustness and topic-sensitive property of the naturally formed dense subgraphs in MMG.

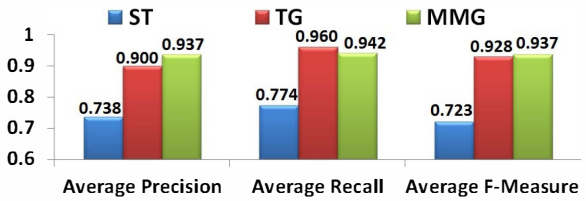

Fig. 5. The average performances evaluated on the top-10 detected topics on MCG-WEBV.

Table 2. The CP performances on MCG-WEBV.

Method	*NDT*	*NCDT*	*CP*
TG	83	31	37.35%
MMG	33	32	**96.97%**

3.3. Performance Evaluation on YKS

In this section, we analyze the performance of MMG on YKS, which consists of multi-modal data from both the mediums of News article and Web video. We compare with TG, whose source code is kindly provided by T. Chen [8]. Considering that TG is only able to use the Web video data in YKS, we fairly compare with it by running MMG on exactly the same Web video data set. Furthermore, we run MMG on the full cross-media data set of YKS to prove the advantage of cross-media topic detection. For clarity, we refer to the Web video version of our method as MMG-V. All the three methods are evaluated under the same ground truth of YKS. The results reported are the best performances of the three approaches. The optimal parameters are: for TG, $\theta = 0.2, \beta_1 = -0.25$ and $\eta = 0.43$ (θ, β_1, η are the parameters of TG [8]); for MMG-V, $\beta = 1$, $C = 200$; for MMG, $\beta = 0.1$, $C = 200$.

Fig. 6 shows the average performances of TG, MMG-V and MMG. We can see that MMG-V performs better than TG in all aspects. This proves the effectiveness of MMG-V in dealing with pure Web video data. Moreover, MMG performs better than both TG and MMG-V, which demonstrates the effectiveness of the complementary cross-media information in improving the topic detection performance.

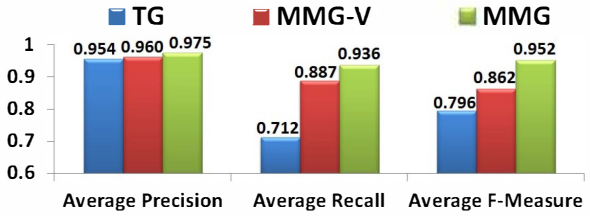

Fig. 6. The average performances evaluated on the top-10 detected topics on YKS.

Table 3 shows that the CP of all the methods on YKS are 100%. However, the number of correctly detected topics (*i.e.* $NCDT$) differs significantly, where the number of topics correctly detected by MMG is the largest. We further analyze this phenomenon in Fig. 7, where the correctly detected topics are divided into three groups according to their related modalities. Apparently, MMG-V detects more topics than TG in both the groups of "pure video topics" and "hybrid topics". This proves the effectiveness of our method in dealing with the multi-modal data from the single medium of Web video. Moreover, the MMG method not only detects more video-related topics than MMG-V, but also detects 69 pure News article topics, which shows the significant performance enhancement brought by the complementary cross-media information and the effectiveness of MMG in dealing with the incomplete multi-modal data from different mediums.

Table 3. The CP performances on YKS.

Method	NDT	NCDT	CP
TG	11	11	100%
MMG-V	23	23	100%
MMG	115	115	100%

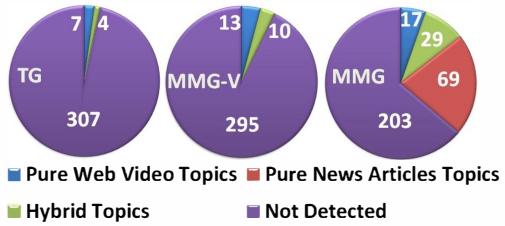

Fig. 7. Analysis on the types of detected topics on YKS.

4. CONCLUSION

We propose a multi-modality graph fusion framework to effectively detect topics from both single-medium and multi-media data sources. Such framework is highly flexible and can be efficiently extended to incorporate other modalities of data from various mediums. In our future work, we will investigate online method for cross-media topic detection.

5. REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study final report," *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, pp. 194–218, Feb. 1998.

[2] Q. Mei and C. Zhai, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," in *Proceedings of KDD 05*, 2005, pp. 198–207.

[3] H. Anaya-Sánchez, A. Pons-Porrata, and R. Berlanga-Llavori, "A document clustering algorithm for discovering and describing topics," *Pattern Recogn. Lett.*, vol. 31, no. 6, pp. 502–510, apr 2010.

[4] C. Pan and P. Mitra, "Event detection with spatial latent dirichlet allocation," in *JCDL*, 2011, pp. 349–358.

[5] H. H. Winston and S. Chang, "Topic tracking across broadcast news videos with visual duplicates and semantic concepts," in *ICIP*, 2006, pp. 141–144.

[6] L. Liu, L. Sun, Y. Rui, Y. Shi, and S. Yang, "Web video topic discovery and tracking via bipartite graph reinforcement model," in *WWW*, 2008, pp. 1009–1018.

[7] J. Cao, C. Ngo, Y. Zhang, and J. Li, "Tracking web video topics: Discovery, visualization, and monitoring," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 21, no. 12, pp. 1835–1846, 2011.

[8] T. Chen, C. Liu, and Q. Huang, "An effective multi-clue approach for web video topic detection," in *ACM Multimedia*, 2012, pp. 781–784.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[10] Y. Wang, H. Sundaram, and L. Xie, "Social event detection with interaction graph modeling," *CoRR*, vol. abs/1208.2547, 2012.

[11] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *ECCV (2)*, 2012, pp. 660–673.

[12] H. Liu and S. Yan, "Robust graph mode seeking by graph shift," in *ICML*, 2010, pp. 671–678.

[13] L. Xie, A. Natsev, J. R. Kender, M. L. Hill, and J. R. Smith, "Visual memes in social media: tracking real-world news in youtube videos," in *ACM Multimedia*, 2011, pp. 53–62.

[14] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, pp. 5228–5235, Apr. 2004.

[15] J. Cao, Y. Zhang, Y. Song, Z. Chen, X. Zhang, and J. Li, "Mcg-webv: A benchmark dataset for web video analysis," *Technical Report, ICT-MCG-09001*, May. 2009.