Unsupervised Web Topic Detection Using A Ranked Clustering-Like Pattern Across Similarity Cascades

Junbiao Pang, Fei Jia, Chunjie Zhang, Weigang Zhang, Qingming Huang, *Senior Member, IEEE*, and Baocai Yin

***Abstract*—Despite the massive growth of social media on the Internet, the process of organizing, understanding, and monitoring user generated content (UGC) has become one of the most pressing problems in today's society. Discovering topics on the web from a huge volume of UGC is one of the promising approaches to achieve this goal. Compared with classical topic detection and tracking in news articles, identifying topics on the web is by no means easy due to the noisy, sparse, and less-constrained data on the Internet. In this paper, we investigate methods from the perspective of similarity diffusion, and propose a clustering-like pattern across similarity cascades (SCs). SCs are a series of subgraphs generated by truncating a similarity graph with a set of thresholds, and then maximal cliques are used to capture topics. Finally, a topic-restricted similarity diffusion process is proposed to efficiently identify real topics from a large number of candidates. Experiments demonstrate that our approach outperforms the state-of-the-art methods on three public data sets.**

***Index Terms*—Maximal clique, Poisson deconvolution, similarity cascade (SC), unsupervised ranking, web topic detection.**

I. Introduction

WITH the rapid development of web technology, social media websites have become convenient platforms for people to assess the world and exchange their opinions. However, the unprecedented explosion in the volume of webpages from social media has made it difficult for web users to quickly access hot topics [34] and for web administrators to systematically monitor web activities [10], [17]. Driven by such practical requirements, for instance, Google, YouTube and Sina have supplied a similar service called "Hot Topic Trend" to help people grasp what are recently interested contents. For this reason, there is an increasing need for techniques to organize data into a meaningful and effective manner.

The task of Topic Detection and Tracking (TDT) [3] is one such effort to automatically organize news articles into topics. Nevertheless, TDT mainly focuses on discovering topics from professionally edited news articles [2] which are totally different from User-Generated Content (UGC) on social media. The content of social media is more unconstrained and less predictable than that of new articles. Due to the fact that the textual and visual information from social media tend to be short, sparse and noisy [47], traditional approaches based on long and structured text are not competent for this problem. Moreover, it is reasonable for TDT to assign each news article into a topic [2]. Because each professionally edited article focuses on a certain topic. In contrast, social media usually contains many low-valued contents which never evolve into any topic.

Therefore, it is a natural idea to detect topics from social media with facing noisy, sparse and less-constrained data. To adhere to the definition of TDT [2], *web topic detection* in this paper is defined as the task of discovering of a tiny fraction of webpages strongly connected by a seminal event from a large amount of social media. An *event* is something that is coincidentally concerned by most of web users. Note that web topic detection is totally different from topic models [5], [6] that are a suite of algorithms that model each text corpora as a mixture of hidden themes. Recognizing that social media is heterogenous data, such as hyperlink, time stamp, textual and visual information, many literatures on web topic detection consider web topics as clusterings from multi-modal data. Among the most popular approaches is the similarity graph method which first fuses multiple cues into a graph and applies graph-based clustering algorithms [30], [46].

The key parameter in topic-as-clustering is the number of clusterings. Intuitively, if the number of topics is large, the discovered topics tend to be broken into multiple fragmental ones which are highly sensitive to noisy features. On the other side,

Manuscript received August 18, 2014; revised December 02, 2014; accepted April 13, 2015. Date of publication April 21, 2015; date of current version May 13, 2015. This work was supported in part by National Basic Research Program of China (973 Program) under Grant 2012CB316400, in part by Natural Science Foundation of China under Grant 61332016, Grant 61202234, Grant 61202322, Grant 61303154, Grant 61390510, and Grant 61472387, by the Beijing Natural Science Foundation under Grant 4132010 and Grant KZ201310005006, and by the Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PHR). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jing-Ming Guo.

J. Pang and B. Yin are with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China (e-mail: junbiao_pang@bjut.edu.cn; ybc@bjut.edu.cn).

F. Jia and C. Zhang are with the School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jiafei@jdl.ac.cn; cjzhang@jdl.ac.cn).

W. Zhang is with the School of Computer Science and Technology, Harbin Institute of Technology at Weihai, Weihai 264209, China (e-mail: wgzhang@jdl.ac.cn).

Q. Huang is with the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Metropolitan Transportation, Beijing University of Technology, Beijing 100124, China, and is also with the University of Chinese Academy of Sciences, Chinese Academy of Sciences (CAS), Beijing 100049, China, and the Institute of Computing Technology, CAS, Beijing 100190, China (e-mail: qmhuang@jdl.ac.cn).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes the proof of Proposition 1 and appended experimental results. This material is 45.5 KB in size.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2425143

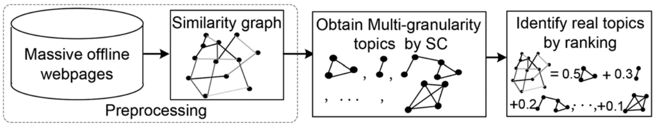


Fig. 1. System framework of our approach.

if the number of topics is small, a detection system would have a relatively low recall, posing one challenge to clustering algorithms. That is, clustering algorithms [15] should handle a large number of irrelevant and noisy data. In fact, practical applications will not always know the real number of topics. Because people have to face the polysemous phenomenon caused by short and less-constrained data during organizing topics–different ways to understand the semantic meanings of a webpage. For instance, without enough contextual information, it is reasonable to accept that a webpage about "car insurance" can be categorized into the "traffic accident" topic or not.

The key notion of our solution is that unsupervised ranking on multi-granularity topics can avoid the problem of determining the number of topics. Multi-granularity topics have two potential benefits: naturally modeling the difference among person in organizing topics and partially reducing the damage of miss-coding of descriptors caused by spare and noisy data. However, multi-granularity approach usually generates a large number of overcomplete topics with mutual exclusion (e.g. a webpage cannot simultaneously belong to two topics), overlapping (e.g., a topic about "car accident" may or not be the "traffic accident" and vice versa), and subsumption (e.g. all topics about "car accident" are "traffic problem"). If topic patterns perfectly capture all real topics, retrieval real topics from enormous topic candidates by interestingness is an intuitive approach. However, ranking by interestingness is a non-trivial task, especially when supervised information is expensive and difficult to be obtained.

In this paper, we seek both an *universal* scheme to generate multi-granularity topics and an *unsupervised* ranking method, based on two motivations. First, although an enormous volume of literatures have been devoted to design topic patterns, there is little attention about how to generate multi-granularity topics with off-the-shelf algorithms. Second, we want to avoid disadvantages of classical ranking technologies such as learning to ranking [22]: the requirement of training information. In summary, instead of following topic-as-clustering approach, we factorize web topic detection into two stages: generating multi-granularity topics and identifying real topics by unsupervised ranking.

Fig. 1 shows the overview of proposed framework. The input are offline webpages associated with metadata possibly including user-generated tags, titles and visual information. After these webpages are converted into a similarity graph, we approach multi-granularity topics by considering the generation of topics as Similarity Cascade (SC). SC assumes that seed topics germinate at high similarity layers, and then these seeds gradually grow up by absorbing more webpages at low similarity layers. When different webpages at different layers are absorbed according to predefined topic patterns, multi-granularity topics are generated. During this process, Maximum Clique (MC) is investigated to capture topics. Rather than designing a special clustering pattern [23], MC only requires that the similarities among all samples in a topic are larger than a threshold, in the hope of alleviating the problem of improper feature coding caused by noisy data.

Further in determining which topic candidate tends to be real one, interestingness of topics is estimated by two factors: the sizes of topics and the weights of topics to change a data set. During computing these weights, Similarity Diffusion Process (SDP) is proposed to decompose the similarity between two webpages into several topics. Therefore, a webpage is approximately organized into different semantic meanings, and the polysemous problem is naturally handled.

The proposed method is simple, and yet exceptionally powerful. By adopting MC as a clustering-like pattern to capture topics and by introducing SDP to rank topics, we develop a web topic detection method that exceeds the state-of-the-art approaches. In the experiments, we compare our method with four state-of-the-art approaches [10], [17], [11], [46] and two alternatives on three public data sets. Our main contributions are summarized as follows.

- By considering the polysemous phenomenon, SC is proposed to generate multi-granularity topics for the similarity graph method [30]. To the best of our knowledge, this is the first to address the relation between multi-granularity topics and the polysemous semantic meanings of topics.
- SDP is formulated as deconvolution process to decompose a similarity into the relative weights, converting identification of topics into an unsupervised ranking problem.
- To the best of our knowledge, we first propose accuracy v.s. False Positives Per Topic (FPPT) to evaluate performances at the topic-wise level, establishing a new benchmark for web topic detection.

The rest of this paper is organized as follows: Section II reviews the related work. We describe the details of our approach in Section III. Experimental results are presented in Section IV and the paper is concluded in Section V.

II. Related Work

In the past decades, many approaches have been proposed to find topics from news [3], [7], [4], blogs [37], web videos [10], social images [29], etc. Therefore, we will not be confined to web topic detection, and mainly survey the computational approaches to discover topics in different tasks.

A. The Unsupervised Approaches

In the unsupervised approaches, it is reasonable to assume that elements in a topic have higher similarities between each other. Based on this assumption, existing approaches often cluster social media into topics. Therefore, the main difference among these approaches is how to define the objective functions of clusterings.

The most popular way to define a clustering is to calculate average intra-similarity within a topic. For example, Yang *et al.* [45] proposed a classical framework in which the group-average clustering was used to discover topics. In [42], an agglomerative clustering method based on average pair-wise similarities was proposed to group news into topics. Zhang *et al.* [46] proposed that the dense similarity in a subgraph is used to grasp topics by Graph Shift (GS) [21]. Cao *et al.* [10] first generated events on video tags by k-means, and then linked these ones into topics based on textual-visual similarity. The intra-similarity approach tends to discover a small number of real topics (i.e., recall can be relatively low), because simple intra-similarity often fails to handle the sparse and noisy data which widely occur in social media.

Instead of calculating the intra-similarity [10], [17], some approaches adopt ad-hoc clustering schemes. In [44], Nonnegative Matrix Factorization (NMF) showed more accurate performance than that of the spectral methods in document clustering. He *et al.* [16] proposed periodic, aperiodic features and the characteristics of word trajectory, for event detection in news. Recently, topic models have been proposed to infer hidden themes for document analysis, including Latent Dirichlet Allocation (LDA) [5], Hierarchical Dirichlet Processes (HDP) [38], probabilistic Latent Semantic Analysis (pLSA) [18] and various variations. These topic models generally work well on long and structured documents [15]. However, these models tend to fail on short and noisy text from social media since they are heavily dependent on word co-occurrence. Therefore, text from social media is often "cleansed" by NLP methods [15] or is enhanced by auxiliary information such as hyperlink [23]. As a result, these approaches are not directly extensible for topic detection from social media.

B. The Side Information-Based Approaches

To handle these problem in the unsupervised approaches, incorporating the possible side information into clusterings could probably guide clustering process. The existing approaches can be divided into two groups: the outer information group and the inner information group.

The approaches in the outer information group aim to utilize possible cues from the *other* information channels beyond itself. For instance, [37] proposed to use queries recorded in searching engines to filter out false positive topics. Similarly, [11] detected topics in a user-oriented manner and proposed a query-guided topic detection method. In [26], by leveraging the external sources such as online news and blogs, news videos are clustered into a hierarchical structure. Often the most severe drawback of these approaches is that the quality of side information should be close to that of the supervised one. This hampers the extension of this approach into web topic detection.

Instead of using the outer information, the inner information group aims to exploit the possible complementary modalities from data itself. One modality is often considered as the mutual side information of the others. In this approach, there are two important threads of research. One extends clustering algorithms into the multi-modality data [6], [31], and the other is the similarity graph method [30], a work based on multi-modalities fusion.

In the former case, discovery of topics involves extending single-modality based models into multi-modal data. For instance, [28] proposed a 3S-LDA model which combines LDA with temporal and spatial clusterings for news. Multi-modal LDA [31] was proposed to group image with tags into topics. In similarity graph method, multi-modal information is fused into edges of a similarity graph. For instance, In [43], Wu *et al.* used weighted similarity between Nearly-Duplicated Keyframe (NDK) and text based on speech transcripts for news videos. Compared with the extension of topic modelings [5] into multi-modal data, similarity graph is computationally simple and noise-resistant [30], and is easily extendable for other graph-based algorithms [1].

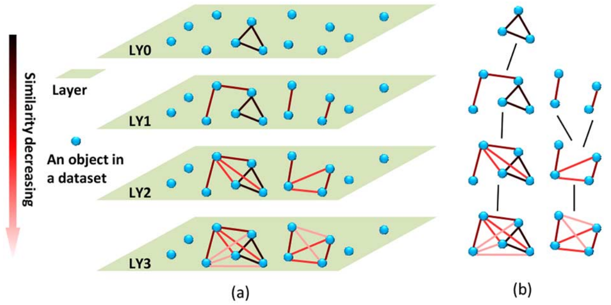


Fig. 2. Illustration of similarity cascade. (a) The similarity cascade. (b) The multi-granularity topics.

III. Generate and Identify Web Topics

This paper aims at proposing an approach which is expected to generalize well in different modalities or multi-modalities. Therefore, from the viewpoint of similarity, we use text information as a case study to unveil the nature of web topics.

A. Preprocessing by Converting Data Into Similarity Graph

Given a data set, we build a graph to represent similarities among data. The nodes v_i represent samples, and edges e_{ij} between two nodes v_i and v_j denote their similarity. Any similarity mapping can be used to convert the similarity between two samples into a graph $G = (V, E)$, where $V = \{v_i\}, E = \{e(i,j)\}$. Given any two samples represented as feature vectors h_i ($h_i \in \mathbb{R}^K$) and $h_j (h_j \in \mathbb{R}^K)$, instead of utilizing cosine distance in classical text analysis [25], this paper adopts Normalized Histogram Intersection (NHI) to measure the similarity

$$e(i,j) = \begin{cases} 0, & i = j \\ \frac{\sum_k \min(h_i(k), h_j(k))}{\sum_k \max(h_i(k), h_j(k))}, & i \neq j \end{cases} \tag{1}$$

where $h_i(k)$ and $h_j(k)$ are the k-th bin of the histogram h_i and h_j, respectively. Because NHI has the noise-resistance ability, which was experimentally verified in computer vision [14] and was also successfully used in tag matching [11]. Moreover, our unreported results discover that NHI has achieved better results that both Euclidean and cosine distances.

As a preprocessing stage, similarity graph is a general framework [1], [30], [10], where multi-modal fusion can be handled with off-the-shelf methods [41], [1], [46]. Note that handling noise and spare data in terms of enriching text feature space [19] is not an optimal choice for preprocessing, as it is not directly extendible for other modalities. In contrast, similarity graph is expected to generalize well, as similarity can be computed by any descriptor, such as, Fisher Vector (FV) [33] for images and enriched text features for short text [19].

B. Multi-Granularity and Similarity Cascade

The polysemous phenomenon in organizing ambiguous data tends to generate multi-granularity topics, if not enough contextual information is provided. From the viewpoint of similarity, Fig. 2(a) illustrates the generation of multi-granularity topics: the semantic meanings of a topic gradually shift by absorbing

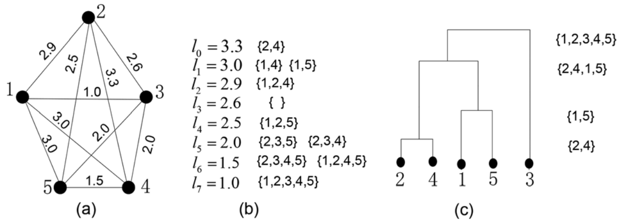


Fig. 3. Comparison between HAC and SC on a toy data. (a) A toy data where the weights of edges represent similarities. (b) The thresholds of SC and the corresponding MCs at each layer. (c) Dendrogram by HAC and the corresponding clusterings.

different webpages at lower similarity layers, or new seed topics generate at lower ones. This recursively happens in or across different similarity layers. We call this process as SC.

This process drives topics into multi-granularity ones over a series of similarity layers, clearly explaining the polysemous phenomenon by overlapping and subsumption relationship among topics. When a topic C propagates over cascade, it leaves a trace as multi-granularity topics, in the form of a set of tuple $(C, V_+)_l$ which means that topic C absorbs a set of nodes V_+ above layer l. If we denote the fact that cascade initially starts from some seminal topic C at the layer l_0 as $(\oslash, V_+)_{l_0}$, two-granularity topics across the t-th layer are represented as

$$(C^{l_t}, V_+^{l_t})_{l_t} \rightarrow (C^{l_{t+1}}, V_+^{l_{t+1}})_{l_{t+1}}, \quad t = 0, \dots \tag{2}$$

where $C^{l_{t+1}} = (C^{l_t}, V_+^{l_t})_{l_t}$ is the topic at the $t+1$-th layer where similarity among two nodes is above l_{t+1}. For instance, Fig. 2(b) illustrates that a seed topic propagates from LY0 to LY3, and two seed topics generated at LY1 evaluated into the same topic at LY3.

In practice, SC can be simulated by quantizing similarities into different layers, i.e., $\mathcal{L} = \{l_0, l_1, \dots, l_T\}$ where T is the number of thresholds. At the layer l_t, any off-the-shelf method can be used to generate topics $\mathcal{C}^t$ which finally form a candidate pool $\mathcal{C}$ (see Alg. 1).

We would like to mention Hierarchical Agglomerative Clustering (HAC) [25], which is quite similar to our SC at the first glance. Both of them seek for discovery of patterns in a hierarchical manner. However, they are different in motive and technique: 1) HAC is originally designed for hierarchically cluster data and for analyzing the structure of a data set; SC is not for a hierarchical clustering, but for handling the difference among people to organize topics; 2) the clusterings of HAC depend on the selection of a distance between a sample and a clustering, while SC is a general framework where any off-the-shelf method can be used to generate topics. In Fig. 3, HAC with maximal distance and SC with MC are used to produce topic candidates, respectively. As illustrated in Fig. 3, the simple distance and hierarchical clustering make HAC produce a smaller number of topics than SC with MC. Therefore, SC has better ability to model the polysemous phenomenon.

C. Maximum Clique as Clustering-Like Pattern

Rather than using dense clustering pattern [46], MC, a connection-based pattern, is proposed to capture topics. The reasons of introducing the connection-based pattern are three folds: 1) it is unnecessary to design a perfect pattern, as topics from social media tends to have multiple patterns; 2) connectivity in MC can be considered as a relaxed clustering constraint, handling multiple topic patterns; 3) connectivity in MC can model weak correlations among data, since descriptors are not sufficient to represent the semantic meanings. Therefore, by measuring connectivity, MC is a clustering-like pattern to grasp diverse topics.

Algorithm 1: SC Generates candidates across cascade

```
Input A graph $G$, thresholds $\mathcal{L} = \{l_1, l_2, \cdots, l_T\}$
Output Topic candidates $\mathcal{C}$
Initialize $\mathcal{C} = \oslash$;
for each $l \in \{l_0, l_1, \cdots, l_T\}$ do
    $SG^l(i,j) = \begin{cases} e(i,j), & \text{if } e(i,j) \geq l \\ 0, & \text{if } e(i,j) < l \end{cases}$
    $\mathcal{C}^t$ = all candidates output by any algorithm on the $SG^l$;
    $\mathcal{C} = \mathcal{C} \cup \mathcal{C}^t$;
end
```

Bron-Kerbosch algorithm [9] is a classical method to find MC, but it is time-consuming. Although a fast algorithm [39] has been proposed, it is still too slow for our application scenario, since we try to find all MCs from all similarity layers.

The Accelerated Algorithm (AA): The multi-granularity topics at different layers have the same "seed" topic. Based on this observation, many candidates that do not contain these seed topics can be safely filtered out.

Definition 1: [The similarity between a node and a topic] Given a node v_i and a topic candidate C, the similarity between a node and a topic is defined as $sim(v_i, C) = \min_{\forall v_j \in C} sim(v_i, v_j)$, where v_j is a node in the topic C.

Based on Definition 1, the AA algorithm is proposed to generate seed-related MCs as follows:

1. Divide the layers $\mathcal{L}$ into the subset $\mathcal{L}^H$ and the subset $\mathcal{L}^L$ by the parameter τ, i.e., $\mathcal{L}^H = \{l_t | l_t > \tau \cdot \max(\mathcal{L})\}$, $\mathcal{L}^L = \mathcal{L} \backslash \mathcal{L}^H$.
2. Identify seed topics $C^h (C^h \in \mathcal{C}^H)$ by Alg. 3.3 from these similarity layers $\mathcal{L}^H$.
3. Given a seed topic C^h from the set $\mathcal{C}^H$, find the node set $I = \arg \underset{v \in V}{\text{sim}}(v, C^h) \geq l_t^l (l_t^l \in \mathcal{L}^L)$ for each layer in $\mathcal{L}^L$, and then extract the subgraph SG built by the node set I.
4. Run Alg. 1 on the subgraph SG at the layers $\mathcal{L}^L$ to find new MCs set $\mathcal{C}^L$.

Fig. 4 explains this procedure on a toy example. A "seed" topic {2,3,4} is first identified in Fig. 4(a), and then Fig. 4(b) finds the subgraph SG with the corresponding nodes $I = \{1, 2, 3, 4, 5, 6\}$. Fig. 4(c) expands the seed topic {2,3,4} into two MCs, {1,2,3,4} and {2,3,4,5,6} from subgraph SG. Moreover, MC {2,7} at the low layer has been safely filtered out.

Theoretical Justification: In this subsection, we theoretically justify that all candidates output by the AA algorithm are also MCs.

Theorem 1: $\forall$ *clique* C^l $(C^l \in \mathcal{C}^L)$ identified by AA algorithm at the layer *l from the subgraph SG, whose corresponding*

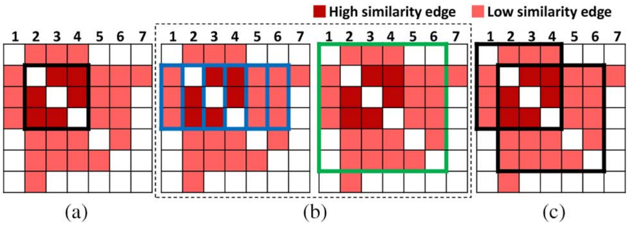


Fig. 4. Illustration of AA algorithm on a 7-node graph with two similarity layers (best viewed in color). (a) A seed MC (in black box). (b) Node set I (in blue box), and extracted subgraph SG (in green box). (c) All MCs from SG (in black box).

"seed" clique is $C^h (C^h \in \mathcal{C}^H)$, we have: 1) $C^h \subseteq C^l$, and 2) C^l *is also a maximum clique of graph G at the layer* l.

Proof: We can prove all above claims by applying the principles of reduction to absurdity. Due to the limited length of this paper, we supply the detailed proof in the supplementary material. ■

D. Identifying Topics by Deconvolution

Topic-Restricted SDP: The premise underlying SDP is that a similarity between two webpages should be decomposed into a set of non-negative ones, which really reflects the similarity according to the context within each topic. The k-th topic ($k = 1, \ldots, K$) can be formally represented as

$$C_k = b_k^T \circ b_k \tag{3}$$

where the indicator vector b_k, $b_k \in \Delta^N$ ($\Delta = \{0, 1\}$), and the operation $\circ$ means that the diagonal of matrix $b_k^T b_k$ is set to zero. The i-th bin of the b_k is denoted as b_{ki}, where $b_{ki} = 1$ or 0 means that topic C_k whether contains the i-th node or not.

Ideally, if topics C_k ($k = 1, \ldots, K$) have the intersected edge $e(i,j)$, we wish to model the relative weights as SDP

$$e(i,j) = \sum_{k=1}^{K} \mu_k(i,j) C_k(i,j) + noise \tag{4}$$

where $\mu_k(i,j) (\mu_k(i,j) \geq 0)$ are the edge-wise relative weights, and $noise$ is a noise term determined by different applications. The edge-wise weights $\mu_k(i,j)$ represent the possibility of a webpage to be organized into the k-th topic. However, SDP (4) assumes that edge-wise weights in a topic are independent, without considering correlation with other webpages in a topic.

To identify real topics by ranking, we restrict that all edges in a topic share the same edge-wise weight by the topic-restricted SDP

$$e(i,j) = \text{Possion}\left(\sum_{k=1}^{K} \mu_k C_k(i,j)\right) \tag{5}$$

where $\mu_k (\mu_k \geq 0)$ are the topic-wise weights. Compared with $\mu_k(i,j)$ in (4), the shared weights μ_k in (5) are more reasonable to rank topics: 1) webpages in a topic are presumed to describe the samething; 2) the shared weights tend to avoid overfitting problem during parameter estimation.

Ranking by Deconvolution: Estimating the weights μ_k in (5) is a special case of deconvolution problem [32], [24]. The solution of (5) is based on the maximum likelihood estimation

$$P(G|\mu_k, C_k) = \prod_{(i,j)\in G} \frac{e(i,j)^{G(i,j)} e^{-e(i,j)}}{G(i,j)!}. \tag{6}$$

Equation (6) assumes that the likelihood probability is conditionally independent for each edge (i,j) and similarities at intersected edges are exchangeable among topics.

Reference [35] has proven that (6) is a concave function by showing the second derivatives of (6) are negative semidefinite. In order to optimize (6), we follow Theorem of [40] that the sufficient conditions for μ_k to be a maximizer of (6) are Kuhn-Tucker conditions, where all μ_k satisfy

$$\mu_k \frac{\partial}{\partial \mu_k} \ln \left(\prod_{(i,j)\in G} \frac{e(i,j)^{G(i,j)} e^{-e(i,j)}}{G(i,j)!} \right) = 0 \tag{7}$$

and

$$\frac{\partial}{\partial \mu_k} \ln \left(\prod_{(i,j)\in G} \frac{e(i,j)^{G(i,j)} e^{-e(i,j)}}{G(i,j)!} \right) \leq 0, \qquad \text{if } \mu_k = 0. \tag{8}$$

To obtain the iterative solution for μ_k, we use the first condition in (7)[1] for μ_k

$$\mu_k \sum_{(i,j)\in G} \frac{\partial}{\partial \mu_k} (G(i,j)\ln(e(i,j)) - e(i,j) - \ln(G(i,j)!)) = 0.$$

After adding the iteration index t, we get

$$\mu_k^{t+1} = \mu_k^t \sum_{(i,j)\in G} \frac{G(i,j) C_k(i,j)}{T_k \sum_{m=1}^{K} \mu_m^t C_m(i,j)} \tag{9}$$

where $T_k = \sum_{(i,j)\in C_k} C_k(i,j), k = 1 \ldots, K$. To understand (9), we can consider that $G' = \sum_{m=1}^{K} \mu_m C_m(i,j)$ is the prediction of a similarity diffused graph according to the current estimation of the relative weight μ_t^t. Thus, $\frac{G(i,j)}{G'(i,j)}$ can be considered as residual errors by the edge-wise division between the similarity graph G and the predicted one G'. The correlation integrates residual errors according to the connected edges in topics $C_k(i,j)$. The division operation by T_k computes the relative weights μ_k^{t+1} according the number of the nonzero edges in C_k. Therefore, the updating rule (9) essentially computes a non-diffused graph G^∞ that would generate the similarity graph G. Typically, the algorithm starts with $\mu_k^0 = 1$. The iterations are stopped when $\|\mu^{t+1} - \mu^t\| < \varepsilon$, where ε is a predefined tolerance to terminate this process.

Note that although (9) requires to compute the edge-wise sum in a topic C_k, sparse C_k favors to estimate μ_k by the nonzero edges. Therefore, the computational cost at each iteration is

$$\mathcal{O}(N_{e_k} + K \cdot N_{v_k}^2) \tag{10}$$

where N_{e_k} is number of the nonzero edges in topics C_k, and N_{v_k} is number of the nodes in topics C_k.

[1] The second condition in (8) is used to prove the convergence of the iteration algorithm. For further details, we refer readers to [35].

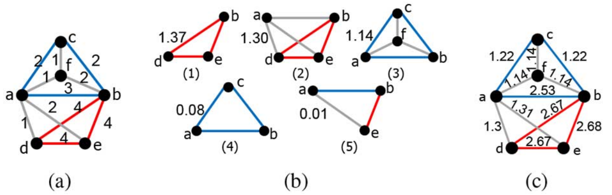


Fig. 5. Decomposition of a similarity graph by (9). (a) Original similarity graph. (b) Topic-wise weights. (c) Reconstructed similarity graph. The numbers mean similarities between two nodes (best viewed in color).

After the weights μ_k are estimated, the interestingness of a topic is computed as

$$i_k = \mu_k \cdot |C_k| \tag{11}$$

where $|C_k|$ is the number of objects in topic C_k. Note that our method during the evaluation adopts the Non-Maximal Suppression (NMS) [13] to handle the problem that which one is selected as the real topic if several topics intersect with each others.

Example: Reconstruction and Ranking. A simple example, illustrated in Fig. 5, decomposes a similarity graph in Fig. 5(a) based on (9). The reconstruction, results from topic-wise weights in Fig. 5(b), is shown in Fig. 5(c). The larger topic-wise weight a topic has, the important this topic is to reconstruct a similarity graph.

IV. Experiment and Discussion

A. Benchmark Data Sets

1. *MCG-WEBV*[10]. This popular data set is first proposed to detect web video topics from UGC. MCG-WEBV is downloaded from the "Most viewed" videos of "This month" on *YouTube* and covers 15 *YouTube* categories. For this data set, the surrounding text of each video is considered as a set of words. Bag-of-Words (BoW) are used to encode features.
2. *YKS*[46]. It is a cross-media and cross-platform data set crawled on *YouKu* and *Sina* respectively. We only use text cues on YKS in the following experiments. During the pre-processing, YKS is tokenized by *NLTK* package, and then TF-IDF is used to measure the importance of each word. Finally, BoWs are used to code these TF-IDF into descriptors.
3. *Social Event Detection 2012 (SED2012)*[29]. It consists of 167,332 photos captured between the beginning of 2009 and the end of 2011 with metadata including tags, geotags, time-stapes, etc. Although SED2012 is originally designed to perceive social events happened in the real world, we here consider social events as special web topics if the geographic information is ignored. We evaluate our method on Challenge 2 in SED2012:

Challenge 2 of SED2012, "find all soccer events taking place in Hamburg (Germany) and Madrid (Spain) in the test collection", is used to verify the effectiveness of our method.

The statistics of data sets are summarized in Table I. Dictionaries of these sets contain multi-language words, as well as user-defined abbreviations. Obviously, the text from social media is shorter and noisier than news articles [47].

B. Evaluation

For MCG-WEBV and YKS: First, we follow top-10 F_1 to evaluate performance on MCG-WEBV [10] and YKS [46]. Every detected topic D_t is matched with the ground truth, and then the highest top $10 F_1$ scores are averaged to measure the performance

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

where $Precision = \frac{|D_t \cap G_t|}{|D_t|}$ is the precision, $Recall = \frac{|D_t \cap G_t|}{|G_t|}$ is the recall, D_t is a detected topic, G_t is a ground truth topic, and $|\cdot|$ denotes the number of webpages in a topic. Top-10 F_1 just considers the best matched 10 topics without measuring false positives. Therefore, this paper proposes to replace top-10 F_1 by top-10 F_1 v.s. number of generated topics, if we only need to measure top-n best results of a system. For top-10 F_1 v.s. number of generated topics, if two methods have the same top-10 F_1 score, the one with smaller number of topics has better performance.

Second, we propose a topic-wise evaluation method, accuracy v.s. FPPT: if a topic is correctly detected, how many false positives are caused by a detection system. More concretely, the accuracy is defined as

$$\text{Accuracy} = \frac{\#\text{Successful}}{\#\text{Groundtruth}}. \tag{13}$$

A topic candidate D_t is recognized as a successful detection, if Normalized Intersected Ratio (NIR) $r = \frac{|D_t \cap G_t|}{|D_t \cup G_t|}$ is larger than a threshold. Because grouping data into topics has to face coherence problem: "even with relatively well written text, one can learn topics that are a mix of concepts or hard to understand" [27]. If a detected topic contains an element from the groundtruth one, we still obtain a F_1 score. However, human barely understand the meaning of this detected topic. The threshold of NIR is used to remove the incoherent topics. The NIR with 0.5 threshold is also widely used to indicate a successful detection in computer vision [13]. In this work, we also follow this choice. In fact, if we want the detected topics to be more coherent, this threshold should be assigned a higher value, e.g., 0.7. During evaluation, a higher accuracy indicates a better result, if two systems have the same FPPT value.

For SED2012: SED2012 proposes to measure the performance by Normalized Mutual Information (NMI) [36], [29]. NMI in [36] is estimated as follows:

$$NMI = \frac{\sum_{i=1}^{\hat{m}} \sum_{j=1}^{\hat{m}} \hat{m}_{ij} \log\left(\frac{m \hat{m}_{ij}}{\hat{m}_i (\hat{m}^*)_j}\right)}{\sqrt{\left(\sum_{i=1}^{\hat{m}} \hat{m}_i \log\frac{\hat{m}_i}{m}\right)\left(\sum_{j=1}^{\hat{m}} (\hat{m}^*)_j \log\frac{(\hat{m}^*)_j}{m}\right)}}, \tag{14}$$

where $\hat{m}_i$ is the number of data contained in the i-th obtained topic, $(\hat{m}^*)_j$ is the number of data in the j-th ground truth, and $\hat{m}_{ij}$ denotes the number of data that are in the intersection between the i-th obtained topic and the j-th ground truth. NMI receives values in the range [0,1] where a higher value indicating a better agreement with the ground truth.

TABLE I
SUMMARY OF WEB TOPIC DATA SETS USED IN THE EXPERIMENTS

Dataset	#Topic	#Webpage	#Webpage in all topics	Dictionary size	Average #word/page	min (#word in a webpage)	max (#word in a webpage)	Comments (the cues used in our experiments are indicated in bold.)
MCG-WEBV	73	3,660	832	9,212	35	1	140	Videos and their surrounding **titles, tags and descriptions** on *Youtube* from Dec 2008 to Feb 2009
YKS	298	8,660	990	80,294	228	3	2,865	**News articles on *Sina*** and **titles, tags and descriptions** web videos on *YouKu* from May 2012 to June 2012
Challage2 SED2012	79	167,332	1,684	1,973	8	1	62	**titles, description, tags** and geographical information.

C. Baselines and Alternative Approaches

1. *Discriminative Probabilistic Models (DPM)* [17]. The first baseline comes from the unsupervised temporal discriminative probabilistic model for news streams. We first resort to its offline version to embed documents into the discriminative feature space, and then the soft partition, vMF mixture model [4], is used to generate topics. DPM has reported better performance than LDA [5] in terms of discovering topics on several testbeds for TDT. We implement DMP to give fair comparisons on MCG-WEBV, YKS, and SED2012.
2. *Event-Clustering Based Method (ECBM)* [10]. Different from our scheme, the work [10] first clusters the tags in each time units, and then uses both NDK and the tag events are linked into topics. Note that this approach involves many engineering details and hyper-parameters. Therefore, we directly copy the reported results to give fair comparisons on MCG-WEBV. While we also implement this method by ourself and report the best tuned results on YKS.
3. *Multi-Modality Based Method (MMBM)*[46]. The method [46] uses multi-modal information, i.e., NDK of videos and text information, to build the similarity graph [30], and utilizes GS [23] on this graph. Different from our method, this work assumes that the elements in topic pattern should be closely correlated. In addition, the visual and text information is also used in this approach [46]. We compare our work to [46] on MCG-WEBV, YKS and SED2012 data sets.
4. *Side-Information Based Method (SIBM)*[11]. This method extracts hot searched queries from search engines, and refines the topics with an ad-hoc approach. This baseline demonstrates that our approaches can achieve superior results without any outer side information on both MCG-WEBV and YKS.
5. *Ranking by Intra-similarity.* As an alternative approach, we do not decompose the similarity into topic-wise weights, but directly compute Normalized intra-similarity either By the number of Nodes (NBN) or By the number of nonzero Edges (NBE)

$$\mathrm{NBN}_k = \frac{\sum_{(i,j)\in C_k} e(i,j)}{|C_k|_N}, \quad \mathrm{NBE}_k = \frac{\sum_{(i,j)\in C_k} e(i,j)}{|C_k|_E}$$

where $|\cdot|_N$ computes the number of nodes, and $|\cdot|_E$ computes the number of nonzero edges. These alternatives demonstrate the effectiveness of the topic-wise weights in identifying real topics.

TABLE II
TOP-10 F_1 (#TOPIC) SCORES OF DIFFERENT SCHEMES TO REPRESENT TOPICS ON MCG-WEBV

Granularity	LDA [5]	NMF [20]	GS [23]	MC
Single	0.457 (50) 0.606 (70) 0.723(100)	0.609 (50) 0.688 (70) 0.740(100)	0.852(179)	0.873 (4,462)
Multiple	0.942(19,094)	0.945(13,331)	0.921(320)	**0.953**(13,649)

D. Analysis of Our Approach

Analysis of Maximum Cliques: In this experiment, GS [21] is considered as one of baselines, recently achieving the state-of-the-art performance on MCG-WEBV [46]. LDA [5] and NMF [20] are used as the other baselines, treating each topic as a multinomial distribution of words.

First, the performances of different topic patterns are compared in single-granularity scenario. Noticing that the groundtruth number of topics in MCG-WEBV is 73 (see Table I), for LDA and NMF, we generate a series of number of topics ranging from 50 to 100. The number of GS [23] automatically generates 179 topics while the number of MC is determined by running BK algorithm on the quantized similarity graph. As illustrated in Table II, top-10 F_1 scores of both LDA and NMF achieve better performances when the number of topics is larger than 73. Because the assumption that every webpage should belong to a topic in both NMF and LDA does not hold in social media, due to many noise and irrelevant data. Moreover, MC achieves the best result among GS, LDA and NMF at the cost of generating enormous number of topics.

Second, we consider the performances of topic patterns in multi-granularity scenario. To obtain multi-granularity topics, the numbers of topics in both NMF and LDA are assigned from 50 to 1000 with the step size 20. While both GS and MC are generated across SC with a series of thresholds $\mathcal{L} = [0.05 : 0.03 : 0.3, 0.3 : 0.1 : 1]$. As shown in Table II, MC consistently achieves best performance among LDA, MMF and GS.

The Effectiveness of SC: We use a SC with 11 layers, and identify topics both in and across these layers. As shown in Fig. 6, MC always achieves better accuracies (13) than GS [23], [46] at each similarity layer. In particular, after the 0.7 level, MC outperforms GS at least 20% accuracy. It means that the effectiveness of SC is closely related to what kind of topic pattern is used. Moreover, if multi-granularity topics are generated by SC, the performances of both GS and MC are all consistently improved. For instance, compared with accuracies at the 0.3 level,

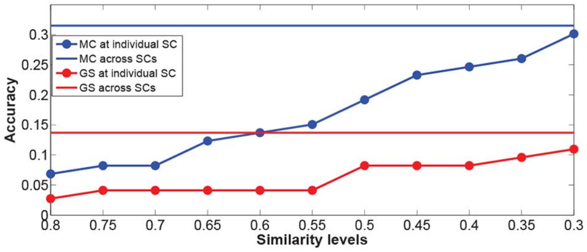


Fig. 6. Comparisons of generating topics by GS [46] and MC across SC.

TABLE III
COMPARISONS BETWEEN TWO ALGORITHMS IN GENERATING TOPIC CANDIDATES

	BK [9]	AA
Running time	11 days	0.8 days
#Topic candidates	623,488	**13,646**
Top-10F_1 after ranking	**0.969**	0.953

TABLE IV
EFFECTIVENESS OF RANKING FOR IDENTIFYING REAL TOPICS

Algorithms	Top-10 *Recall*	Top-10 *Precision*	Top-10 F_1
No ranking	0.580±0.021	0.79±0.018	0.654±0.013
Our Approach	**0.948**	**0.967**	**0.953**

the multi-granularity MC and GS increase 4.55% accuracy rate and 25.00% one, respectively.

The Performance of the AA Algorithm: In this subsection, we compare the performances of the AA algorithm with the BK one. For the AA algorithm, we divide $\mathcal{L}$ into two subsets by setting $\tau = 0.03$.

Table III shows clearly that the AA algorithm accelerate the time of identifying meaningful candidates. Moreover, compared with the BK, the top-10 F_1 of the AA algorithm slightly decreases from 0.969 to 0.953. These results indicate that the AA algorithm may filter out some meaningful candidates at the middle similarity layers.

The Effectiveness of Ranking by Deconvolution: We select 73 highest weighted topics from 13,646 ranked MC topics. For the unranked scenario, we randomly select 73 candidates as detected topics. This process is done 5 times, and the mean and the standard deviation are calculated.

As shown in Table IV, top-10 F_1 score has been improved about 45% after the unsupervised ranking. Unsupervised ranking successfully retrievals more meaningful topics, justifying the correctness of our hypothesis: the similarity diffusion follows Possion noise.

E. Qualitative Comparisons With Other State-Of-The-Art Methods

In this subsection, we compare the proposed approach on other benchmark data sets. To make comparisons as meaningful as possible, we use the same experimental settings proposed by each data set.

Web-Video Topic Detection in MCG-WEBV: Table V shows the comparisons among SIBM [11], ECBM [10], MMBM [46], DPM [17] and our method. The results of SIBM, ECBM and MMBM are directly copied from their papers, while the results of DPM are implemented and tested on MCG-WEBV by ourself. Note that MMBM [46] recently reported the state-of-the-art results on this data set.

TABLE V
PERFORMANCES OF VARIOUS APPROACHES ON MCG-WEBV

Methods	Top-10 *Recall*	Top-10 *Precision*	Top-10 F_1
SIBM [11]	0.900	0.960	0.928
ECBM [10]	0.738	0.774	0.723
MMBM [46]	0.937	0.942	0.937
DPM [17]	0.988	0.809	0.887
Our approach	**0.948**	**0.967**	**0.953**

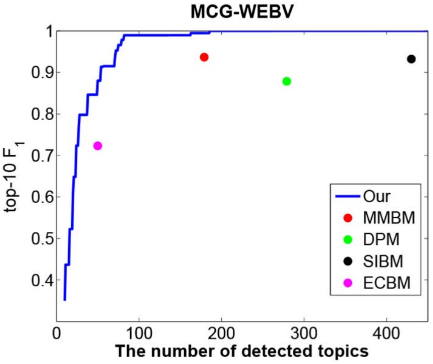


Fig. 7. Comparisons between the state-of-the-art approaches and our method by top-10 F_1 vs. number of generated topics on MCG-WEBV (best viewed in color).

From Table V, we can see that our approach demonstrates the best overall performance when only short text cue is used. ECBM achieves the worst performance than other methods. The main explanation is that ECBM [10] totally depends on clustering on tags and then utilizes the visual and temporal consistency to link clusterings into topics. However, a few noise in tags would greatly change the results due to the sparsity of tags per video, making top-10 *Recall* remarkably low. Compared with SIBM [11], we can see that top-10 *Precision* is very close to our approach. As expected, the well selected key words from queries naturally filter out many false positives. On the contrary, top-10 *Recall* of SIBM [11] is lower than both MMBM [46] and our approach. The explanation is that these key words from searching engines tend to have no correlation with these topics generated from social media. Consequently, SIBM [11] largely depends on the quality of side information. Among all these approaches, MMBM [46] is difficult to achieve a high top-10 *Recall* since the dense topic pattern is adopted [21]. In contrast, our approach starts with connected-based topic pattern and then tries to find all possible candidates. Therefore, top-10 *Recall* of our approach is much higher than that of all the other approaches.

Top-10 F_1 v.s. number of generated topics is illustrated in Fig. 7. Top-10 F_1 of our method increases quickly along with number of generated topics. It means that our ranking method can effectively retrieve top-10 topics at the cost of a small number of false positives. For instance, to achieve approximate 0.9 top-10 F_1 score, MMBM [46], SIBM [11] and DPM [17] generated 179, 430, and 275 topics respectively on

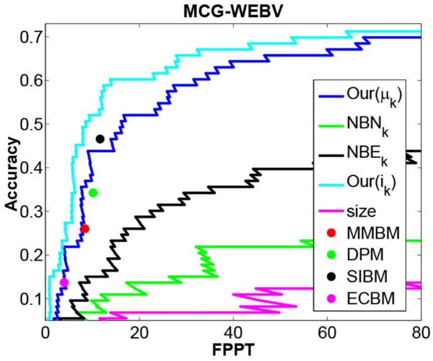


Fig. 8. Comparison among the state-of-the-art methods, alternatives and our methods by accuracy vs. FPPT on MCG-WEBV (best viewed in color).

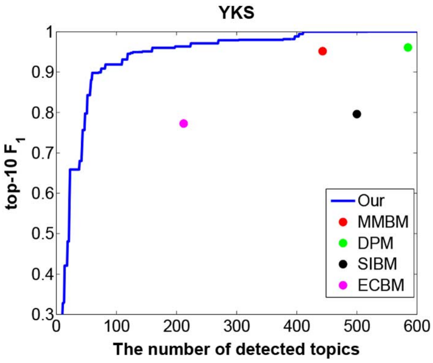


Fig. 9. Comparisons between the state-of-the-art approaches and our method by top-10 F_1 vs. number of generated topics on YKS (best viewed in color).

MCG-WEBV, while our method only requires 70 topic candidates. Consequently, top-10 F_1 v.s. number of generated topics not only compares top-10 matched topics, but also measures the number of false positives.

To further evaluate the topic-wise performance, the accuracy v.s. FPPT, is computed. As shown in Fig. 8, our approach (“Our(i_k)”) (11) is consistently better than MMBM [46], DPM [17], SIBM [11] and ECBM [10]. In addition, there are two interesting observations in Fig. 8, as follows.

1) . Ranking by topic-wise weights (“Our(μ_k)”) outperforms the alternatives (“NBE$_k$“ and “NBN$_k$“). It verifies that the weights of reconstructing a data set is a nature choice to estimate interestingness of a topic.
2) The combination of both size of a topic and its topic-wise weight (“Our(i_k)”) outperforms the topic-wise weights (“Our(μ_k)”), although ranking by the size of a topic (“Size”) has the worst performance. It means that the interestingness of a topic has close relation with size of topics, and we should combine it with the topic-wise weights during ranking.

Web Topic Detection in YKS: YKS, a cross-platform data set, requires to grasp more diverse types of topics than MCG-WEBV. Fig. 9 shows that our method consistently outperforms MMBM [46], DPM [17], SIBM [11] and ECBM [10], if the same number of topics is generated. For instance, top-10 F_1 of our method is 0.96 while ECBM [10] is 0.78, when 200 topics are used for both methods. Compared with the results on MCG-WEBV, there is an additional observation on YKS. Fig. 9 shows that both MMBM [46] and DPM [17] require to generate much more number of topics than that of our approach. For instance, MMBM [46] and DPM [17] have to generate 435 and 590 topics in order to archive 0.95 top-10 F_1. In contrast, our approach only generates 110 topic candidates. This illustrates the generalization ability of our approach across different data sets.

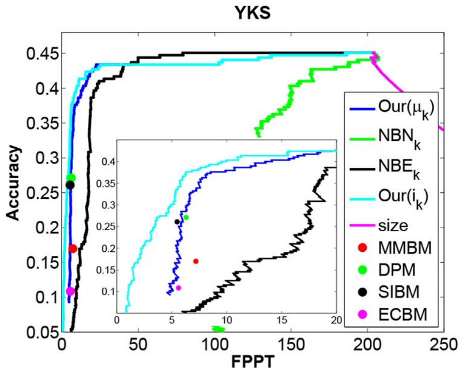


Fig. 10. Comparison with other methods by accuracy vs. FPPT on YKS. Note that the overlapped subfig zooms up the part of this figure where FPPT is from 0 to 20 (best viewed in color).

Fig. 10 further shows that accuracy v.s. FPPT curves on YKS data set. Our approach consistently outperforms these state-of-the-art methods. By ranking interestingness of topics, our approach (“Our(i_k)”) also consistently achieves better results than these alternatives (“NBN$_k$“ and “NBE$_k$“) in the overlapped sugfig of Fig. 10.

Interestingly, if we compare Figs. 8 with 10, sizes of topics have a positive effect on MCG-WEBV but play a negative role on YKS. That is, accuracy on MCG-WEBV is increased by ranking sizes of topics in Fig. 8, while Fig. 10 shows that accuracy on YKS is decreased by ranking sizes of topics. On the other hand, combination of both sizes of topics and topic-wise weights consistently achieves the best performance on both data sets.

Challenge 2 in SED2012: For Challenge 2, SED2012 is mixed with a lot of non-topic and non-soccer topic samples. To reduce the search space, we follow the approach in [8] to identify multiple soccer stadiums for each given cities, and extract all the dates and times of soccer matches from the beginning of 2009 to the end of 2011 from *playerhistory.com*. Then, we select the positive samples and the negative ones which are not taken in given times. Finally, a linear Support Vector Machine (SVM) is trained to reduce the number of samples to 6,448.

Note that the provided text metadata are partially cleaned by removing the stop words, html tags and camera related words,i.e., “Nikon”, “35 mm”. Finally, TF-IDF is computed from these text metadata including title, description and tags. In the following experiments, both DPM [17] and our method only use text cue; while MMBM [46] uses both text and visual cues. Table VI shows that our approach consistently outperforms MMBM [46] and DPM [17].

TABLE VI
NMI (#Topic) of Different Approaches on SED2012

Algorithms	MMBM [46](6)	DPM [17](160)	Our approach(120)
NMI	0.472	0.399	**0.518**

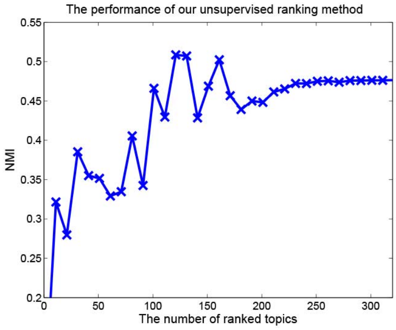


Fig. 11. NMI vs. the number of ranked topics for our approach on SED2012.

Fig. 11 further illustrates the effectiveness of our unsupervised ranking on SED2012. As illustrated in Fig. 11, using all topic candidates (328 topics in our approach) does not guarantee the optimal performance since the best result is achieved at the top 120 topics. It means that ranking by deconvolution can efficiently identify these real topics. Moreover, NMI curve before the top 160 topics oscillates intensively, but gradually becomes smooth after the top 200 topics. This phenomenon may be caused by two reasons: 1) if a candidate is wrongly ranked, NMS tends to suppress the real ones which overlap with the wrongly ranked topic, and 2) beyond sizes of topics and topic-wise weights, there should have other factors to influence interestingness of topics.

V. Conclusions

In this paper, we have described a web topic detection method for social media based on multi-granularity topics and ranking interestingness of topics, leading to results matching or surpassing the state-of-the-art methods on several data sets. Moreover, a new benchmark is established to measure topic-wise performance for this task. The promising results of this paper motive a further examination of combination of multi-granularity topics and estimating interestingness of topics. First, more effective topic patterns, like random methods, may scale up well to large-scale problems over MC used here. Furthermore, the heterogeneous cues should be embedded into a graph, as our approach only depends on the similarity between two samples. Moreover, more potential factors should be investigated into SPD such as entropy rate [12].

Acknowledgment

The authors would like to thank the associate editor and the anonymous reviewers whose comments improved this paper greatly.

References

[1] M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimers, “Sensing trending topics in Twitter,” *IEEE Trans. Multimedia*, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.

[2] J. Allan, *Topic Detection and Tracking: Event Based Information Organization*. Norwell, MA, USA: Kluwer, 2000.

[3] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, “Topic detection and tracking pilot study final report,” in *Proc. DARPA Broadcast News Transcription Understand. Workshop*, Feb. 1998, pp. 194–218.

[4] A. Banerjee and S. Basuy, “Topic models over text streams: A study of batch and online unsupervised learning,” in *Proc. SIAM Int. Conf. Data Mining*, 2007, pp. 431–436.

[5] D. Blei, M. David, A. Ng, M. Jordan, and J. Lafferty, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[6] D. Blei and J. Lafferty, “A correlated topic model of science,” *Ann. Appl. Sci.*, vol. 3, no. 3, pp. 17–35, 2007.

[7] T. Brants, F. Chen, and A. Farahat, “A system for new event detection,” in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2003, pp. 330–337.

[8] M. Brenner and E. Lzquierdo, “Social event detection and retrieval in collaborative photo collections,” in *Proc. Int. Conf. Multimedia Retrieval*, 2012, pp. 1–8.

[9] C. Bron and J. Kerbosch, “Algorithm 457: Finding all cliques of an undirected graph,” *Commun. ACM*, vol. 16, no. 9, pp. 575–576, 1973.

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

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

[12] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.

[13] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” *Int. J. Comput. Vis.*, vol. 88, no. 3, pp. 303–338, 2010.

[14] T. Gevers and H. Stokman, “Robust histogram construction from color invariants for object recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 1, pp. 113–121, Jan. 2004.

[15] B. Han, P. Cook, and T. Baldwin, “Automatically constructing a normalisation for microblogs,” in *Proc. Joint Conf. Empirical Methods Natural Lang. Process. Comput. Natural Lang. Learn.*, 2012, pp. 421–432.

[16] Q. He, K. Chang, and E. Lim, “Analyzing feature trajectories for event detection,” in *Proc. ACM SIGIR Res. Develop. Inf. Retrieval*, 2007, pp. 207–214.

[17] Q. He, K. Chang, E. Lim, and A. Banerjee, “Keep it simple with time: A re-examination of probabilitic topic detection models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1795–1808, Oct. 2010.

[18] T. Hofmann, “Probabistical latent semantic indexing,” in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 1999, pp. 50–57.

[19] X. Hu, N. Sun, C. Zhang, and T. Chua, “Exploiting internal and external semantics for the clustering of short texts using world knowldege,” in *Proc. ACM Conf. Inf. Knowl. Manage.*, 2009, pp. 919–928.

[20] D. Lee and H. Seung, “Learning the parts of objects using non-negative matrix factorizations,” *Nature*, vol. 401, pp. 788–791, 1999.

[21] H. Liu and S. Yan, “Robust graph mode seeking by graph shift,” in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 671–678.

[22] T.-Y. Liu, “Learning to rank for information retrieval,” *Found. Trends Inf. Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, “Topic-link lda: Joint models of topic and author community,” in *Proc.Int. Conf. Mach. Learn.*, 2009, pp. 338–349.

[24] L. Lucy, “An iterative technique for the rectification of observed distributions,” *Astron. J.*, vol. 79, no. 6, pp. 745–754, 1974.

[25] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval.*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[26] S. Neo, Y. Ran, H. Goh, Y. Zheng, T. Chua, and J. Li, “The use of topic evaluation to help users browse and find answer in new video corpus,” in *Proc. Int. Conf. ACM Multimedia*, 2007, pp. 198–207.

[27] D. Newman, E. Bonilla, and W. Buntine, “Improving topic coherence with regularized topic models,” *Neural Inf. Process. Syst.*, pp. 496–504, 2011.

[28] C.-C. Pan and P. Mitra, “Event detection with spatial latent Dirichlet allocation,” in *Proc. IEEE Joint Conf. Digit. Libraries*, Jun. 2011, pp. 349–358.

[29] S. Papadopoulos, E. Schinas, V. Mezaris, R. Troncy, and I. Kompatsaris, “Social event detection at mediaeval 2012: Challenges, dataset and evaluation,” in *Proc. MediaEval 2012 Workshop*, Pisa, Italy, Oct. 2012.

[30] S. Papadopoulous, C. Zigkolis, Y. Kompatsiaris, and A. Vakali, “Cluster-based landmark and event detection on tagged photo collections,” *IEEE Multimedia Mag.*, vol. 18, no. 1, pp. 52–63, Jan. 2011.

[31] D. Putthividhy, H. Attias, and S. Magarajan, “Topic regression multi-modal latent Dirichlet allocation for image annotation,” *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3408–3415, Jun. 2010.

[32] W. H. Richardson, “Bayesian-based iterative method of image restoration,” *J. Optical Soc. Amer. (1917-1983)*, vol. 62, pp. 55–59, Jan. 1972.

[33] J. Sánchez, T. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the fisher vector: Theory and practice,” *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[34] D. Shahaf and C. Guestrin, “Connecting the dots between news articles,” in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 623–632.

[35] L. Shepp and Y. Vardi, “Maximum likelihood reconstruction for emission tomography,” *IEEE Trans. Med. Imag.*, vol. 1, no. 2, pp. 113–122, Oct. 1982.

[36] A. Strehl and J. Ghosh, “Cluster ensembles–a knowledge reuse framework for combining multiple partitions,” *J. Mach. Learn.Res.*, vol. 3, no. 1, pp. 583–617, 2003.

[37] A. Sun and M. Hu, “Query-guided event detection from news and blog streams,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 5, pp. 834–839, Sep. 2011.

[38] Y. Teh, M. Jordan, M. Beal, and D. Blei, “Hierarchical Dirichlet processes,” *J. Amer. Statist. Assoc.*, vol. 101, pp. 1566–1581, 2006.

[39] E. Tomita, A. Tanaka, and H. Takahashi, “The worst-case time complexity for generating all maximal cliques and computational experiments,” *Theoretical Comput. Sci.*, vol. 363, no. 1, pp. 28–42, 2006.

[40] Y. Vardi, *Nonlinear Programming.*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1996.

[41] M. Varma and B. Babu, “More generality in efficient multiple kernel learning,” in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 1–8.

[42] C. Wang, M. Zhang, S. Ma, and L. Ru, “Automatic online news construction in web environment,” in *Proc. Int. Conf. WWW*, 2008, pp. 457–466.

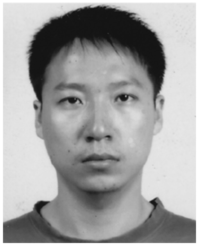
[43] X. Wu, G. Hauptmann, and C. Ngo, “Novelty detection for crosslingual news story with visual duplicates and speech transcripts,” in *Proc. Int. Conf. ACM Multimedia*, 2007, pp. 168–177.

[44] W. Xu, X. Liu, and Y. Gong, “Document clustering based on non-negative matrix factorization,” in *Proc. ACM SIGIR Res. Develop. Inf. Retrieval*, 2003, pp. 267–273.

[45] Y. Yang, T. Pierce, and J. Carbonell, “A study of retrospective and on-line event detection,” in *Proc. ACM SIGIR Res. Develop. Inf. Retrieval*, 1998, pp. 28–36.

[46] Y. Zhang, G. Li, L. Chu, S. Wang, W. Zhang, and Q. Huang, “Cross-media topic detection: A multi-modality fusion framework,” in *Proc. Int. Conf. Multimedia Expro*, 2013, pp. 1–6.

[47] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, “Comparing Twitter and traditional media using topic models,” in *Proc. Eur. Conf. Adv. Inf. Retrieval*, 2011, pp. 338–349.



Junbiao Pang received the B.S. and M.S. degree in computational fluid dynamics and computer science from the Harbin Institute of Technology, Harbin, China, in 2002 and 2004, respectively, and the Ph.D. from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011.

He was an Assistant Professor of computer science and technology with the Beijing University of Technology (BJUT), Beijing, China, from 2011 to 2013. He is currently a Faculty Member with the College of Metropolitan Transportation, BJUT, Beijing, China. He has authored or coauthored approximately 20 academic papers in publications such as the IEEE Transactions on Image Processing, ECCV, ICCV, and *ACM Multimedia*. His research interests include multimedia and machine learning.

Fei Jia received the B.E. degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2012, and is currently working towards the M.S. degree in computer science and technology at the University of the Chinese Academy of Sciences, Beijing, China.

His research interests include machine learning, image content analysis, and information retrieval.

Chunjie Zhang received the B.E. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2006, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

He is currently a Faculty Member with the University of the Chinese Academy of Sciences, Beijing, China. His research interests include machine learning, image content analysis, and object categorization.

Weigang Zhang received the B.S. and M.S. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2003 and 2005, and he is currently working toward the Ph.D. degree at the School of Computer Science and Technology, HIT, Harbin, China.

He is an Associate Professor with the School of Computer Science and Technology, HIT at Weihai, Weihai, China. His research interests include multimedia computing and computer vision.

Qingming Huang (A’04–M’04–SM’08) received the B.S. degree in computer science and the Ph.D. degree in computer engineering from Harbin Institute of Technology, Harbin, China, in 1988 and 1994, respectively.

He is currently a Professor with the University of the Chinese Academy of Sciences (CAS), Beijing, China, and an Adjunct Professor with the Institute of Computing Technology, CAS, China and with Beijing University of Technology, Beijing, China. He has authored or coauthored more than 200 academic papers in prestigious international journals including the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Image Processing, and top-level conferences such as ACM Multimedia, ICCV, CVPR, ECCV, and VLDB. His research interests include multimedia computing, image processing, computer vision, pattern recognition, and machine learning.

Dr. Huang is the Associate Editor of *Acta Automatica Sinica* and the reviewer of various international journals including the IEEE Transactions on Multimedia, the IEEE Transactions on Circuits and Systems for Video Technology, and the IEEE Transactions on Image Processing. He has served as Program Chair, Track Chair, Area Chair, and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, and PSIVT.

Baocai Yin received the M.S. and Ph.D. degree in computational mathematics from Dalian University of Technology, Dalian, China, in 1988 and 1993, respectively.

He is currently a Professor with the Beijing University of Technology (BJUT), Beijing, China. He is also the Director of the Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing, China, and the Dean of College of Metropolitan Transportation, BJUT, Beijing, China. He has authored or coauthored more than 200 academic papers in prestigious international journals including the IEEE Transactions on Multimedia and the IEEE Transactions on Circuits and Systems for Video Technology, and top-level conferences such as INFOCOM and ACM SIGGRAPH. His research areas include multimedia, image processing, computer vision, and pattern recognition.

Dr. Yin is currently the Editorial Member for the Journal of Information and Computational Science (USA).